Learning from noisy binary labels: a tale of two approaches

Aditya Krishna Menon

National ICT Australia and The Australian National University

NICTA

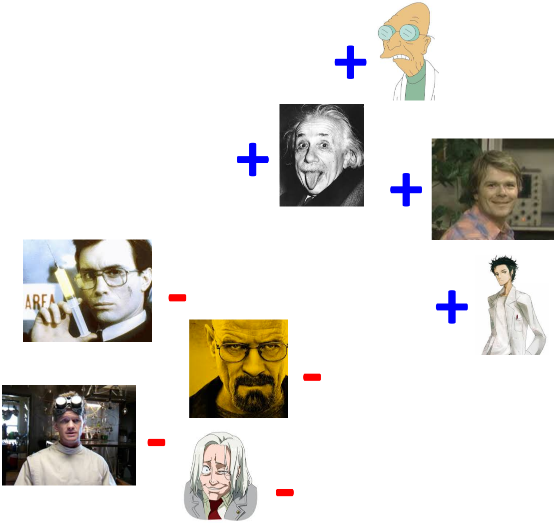
Australian National University

Learning from noisy binary labels: a tale of two approaches

Aditya Krishna Menon

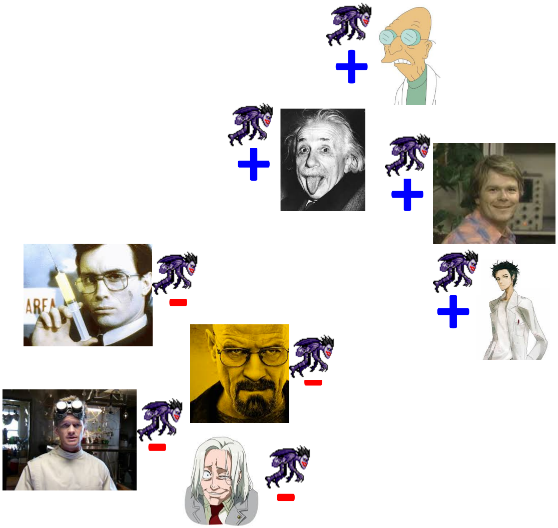
Data61 and The Australian National University

Learning from binary labels

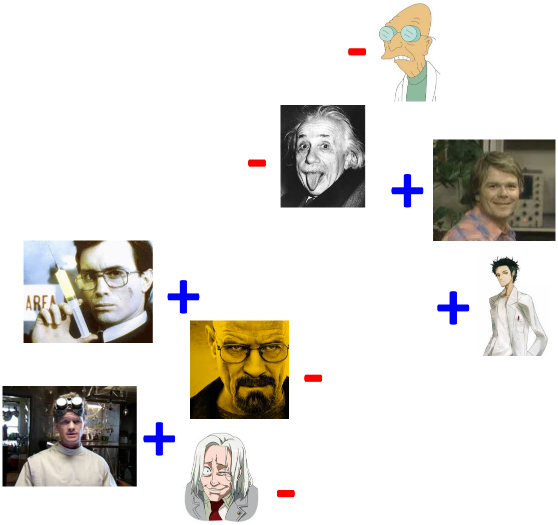


Learning from binary labels

Learning from noisy binary labels



Learning from noisy binary labels



Learning from noisy labels: applications

Learning from noisy annotators

✔ ✔ ✔ ✖

✔ ✔ ✖ ✖

✔ ✖ ✖ ✖

Learning from noisy labels: applications

Learning from noisy annotators

Positive and unlabelled learning

This talk

Can we learn a good classifier from noisy samples?

This talk

Can we learn a good classifier from noisy samples?

Yes, by either:

This talk

Can we learn a good classifier from noisy samples?

Yes, by either:

- choosing a suitably robust loss function
  - e.g. going beyond square, hinge, or logistic loss

This talk

Can we learn a good classifier from noisy samples?

Yes, by either:

- choosing a suitably robust loss function
  - e.g. going beyond square, hinge, or logistic loss
- choosing a suitably rich function or scorer class
  - e.g. going beyond linear models

Roadmap

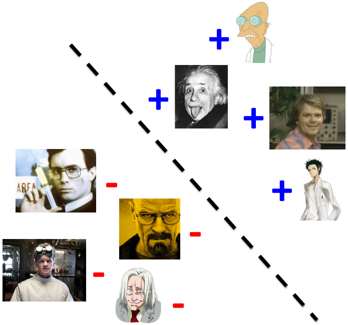
Our aim is to fill in the entries of this table

	Noise			
	Symmetric	Class-conditional	Instance	Instance and label
Loss ℓ	?	?	?	?
Scorer $\mathcal{S}$	?	?	?	?

Learning from clean binary labels

Learning with binary labels: from the trenches

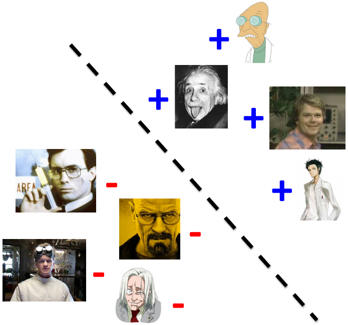
SVMs: find a large margin separator for $\{(x_i, y_i)\}_{i=1}^n$



Learning with binary labels: from the trenches

SVMs: find a large margin separator for $\{(x_i, y_i)\}_{i=1}^n$

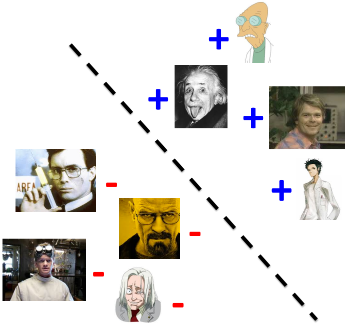
$$\min_w \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^n \max(0, 1 - y_i \cdot \langle w, x_i \rangle)$$



Learning with binary labels: from the trenches

SVMs: find a large margin separator for $\{(x_i, y_i)\}_{i=1}^n$

$$\min_w \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^n \max(0, 1 - y_i \cdot \langle w, x_i \rangle)$$



Slightly increased formalism required

Learning with binary labels: from the towers

Fix an instance space $\mathcal{X}$ (e.g. $\mathbb{R}^n$)

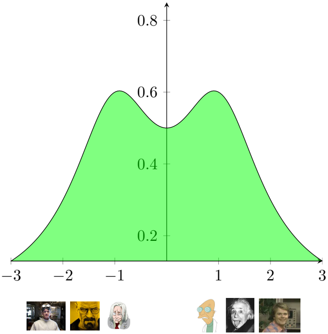
Let D be a distribution over $\mathcal{X} \times \{\pm 1\}$

Learning with binary labels: from the towers

Fix an instance space $\mathcal{X}$ (e.g. $\mathbb{R}^n$)

Let D be a distribution over $\mathcal{X} \times \{\pm 1\}$

- marginal probability over all instances

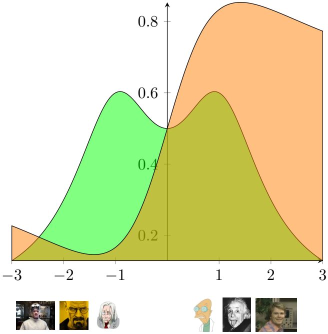


Learning with binary labels: from the towers

Fix an instance space $\mathcal{X}$ (e.g. $\mathbb{R}^n$)

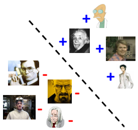
Let D be a distribution over $\mathcal{X} \times \{\pm 1\}$

- marginal probability over all instances
- class-probability for all instances



Scorers, losses, risks

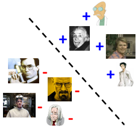
A scorer is any $s\colon \mathcal{X} \to \mathbb{R}$, and scorer class any $\mathcal{S} \subseteq \mathbb{R}^{\mathcal{X}}$



- e.g. linear scorer $s\colon x \mapsto \langle w, x \rangle$

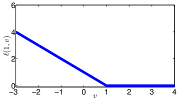
Scorers, losses, risks

A scorer is any $s\colon \mathcal{X} \to \mathbb{R}$, and scorer class any $\mathcal{S} \subseteq \mathbb{R}^{\mathcal{X}}$



- e.g. linear scorer $s\colon x \mapsto \langle w, x \rangle$

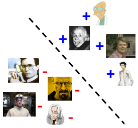
A loss is any $\ell\colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$



- e.g. hinge loss $\ell\colon (y, v) \mapsto \max(0, 1 - yv)$

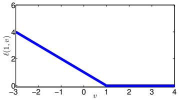
Scorers, losses, risks

A scorer is any $s\colon \mathcal{X} \to \mathbb{R}$, and scorer class any $\mathcal{S} \subseteq \mathbb{R}^{\mathcal{X}}$



- e.g. linear scorer $s\colon x \mapsto \langle w, x \rangle$

A loss is any $\ell\colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$



- e.g. hinge loss $\ell\colon (y, v) \mapsto \max(0, 1 - yv)$

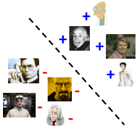
The risk of scorer s wrt loss ℓ and distribution D is

$$\mathbb{L}(s; D, \ell) = \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [\ell(\mathsf{Y}, s(\mathsf{X}))]$$

- average loss on a random sample

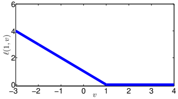
Scorers, losses, risks

A scorer is any $s \colon \mathcal{X} \to \mathbb{R}$, and scorer class any $\mathcal{S} \subseteq \mathbb{R}^{\mathcal{X}}$



- e.g. linear scorer $s \colon x \mapsto \langle w, x \rangle$

A loss is any $\ell \colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$



- e.g. hinge loss $\ell \colon (y, v) \mapsto \max(0, 1 - yv)$

The risk of scorer s wrt loss ℓ and distribution D is

$$\mathbb{L}(s; D, \ell) = \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [\ell(\mathsf{Y}, s(\mathsf{X}))]$$

- average loss on a random sample

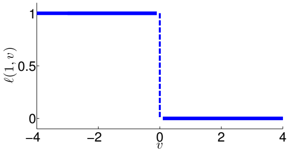
The empirical risk wrt finite sample $\mathsf{S} \sim D^n$ is

$$\mathbb{L}(s; \mathsf{S}, \ell) = \frac{1}{|\mathsf{S}|} \sum_{(x,y) \in \mathsf{S}} \ell(y, s(x)).$$

Binary classification

Binary classification concerns the 0-1 loss

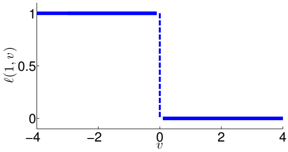
$$\ell^{01}(y,v) = [\![yv < 0]\!] + \frac{1}{2} \cdot [\![v = 0]\!]$$



Binary classification

Binary classification concerns the 0-1 loss

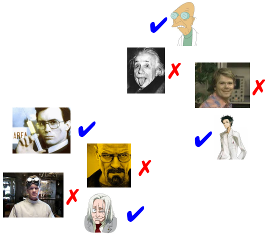
$$\ell^{01}(y,v) = [\![yv < 0]\!] + \frac{1}{2} \cdot [\![v = 0]\!]$$



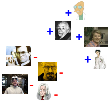
Corresponding misclassification risk is

$$\mathbb{L}(s;D,\ell) = \mathbb{P}_{(\mathsf{X},\mathsf{Y})\sim D}(\mathsf{Y} \neq \mathrm{sign}(s(\mathsf{X})))$$

- probability of misclassifying instance



Our view of learning



Samples

(iid from D^n)

S

Scorer class

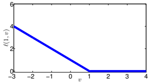
(e.g. linear)

$\mathcal{S}$

Loss

(e.g. hinge)

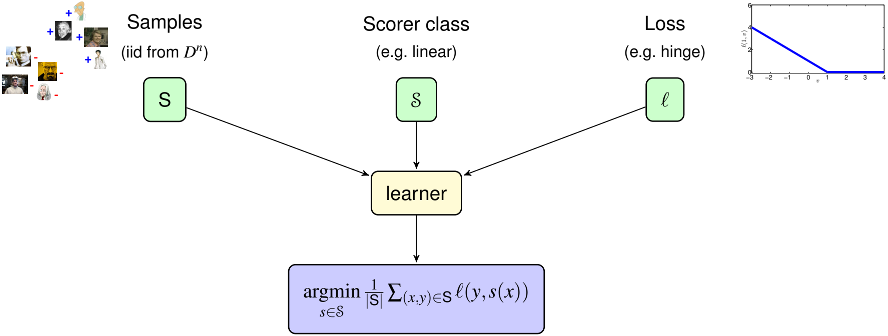
ℓ



learner

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \frac{1}{|\mathsf{S}|} \sum_{(x,y) \in \mathsf{S}} \ell(y, s(x))$$

Our view of learning

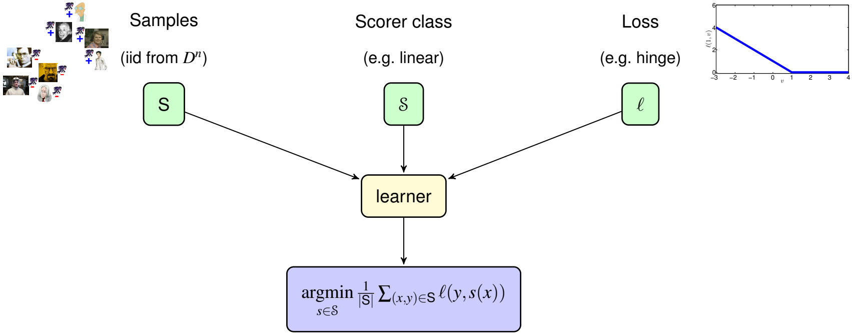


e.g. soft-margin SVM uses:

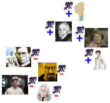
- bounded-norm linear scorers $\mathcal{S} = \{x \mapsto \langle w, x \rangle \mid \|w\|_2 \leq W\}$
- hinge loss $\ell(y, v) = \max(0, 1 - yv)$

Learning from noisy binary labels

Our view of learning



Our view of noisy learning



Samples

(iid from $\bar{D}^n$)

$\bar{\mathsf{S}}$

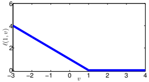
Scorer class

(e.g. linear)

$\mathcal{S}$

Loss

(e.g. hinge)



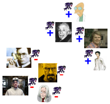
ℓ

learner

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \frac{1}{|\bar{\mathsf{S}}|} \sum_{(x,y) \in \bar{\mathsf{S}}} \ell(y, s(x))$$

Samples from some $\bar{D} \neq D$, where labels flipped with certain probability

Our view of noisy learning



Samples

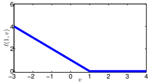
(iid from $\bar{D}^n$)

Scorer class

(e.g. linear)

Loss

(e.g. hinge)



$\bar{\mathsf{S}}$ → learner ← $\mathcal{S}$

ℓ → learner

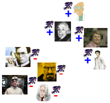
learner →

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \frac{1}{|\bar{\mathsf{S}}|} \sum_{(x,y) \in \bar{\mathsf{S}}} \ell(y, s(x))$$

Samples from some $\bar{D} \neq D$, where labels flipped with certain probability

Noisy labels might affect us in three ways:

Our view of noisy learning



Samples

(iid from $\bar{D}^n$)

$\bar{\mathsf{S}}$

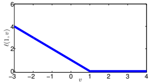
Scorer class

(e.g. linear)

$\mathcal{S}$

Loss

(e.g. hinge)



ℓ

learner

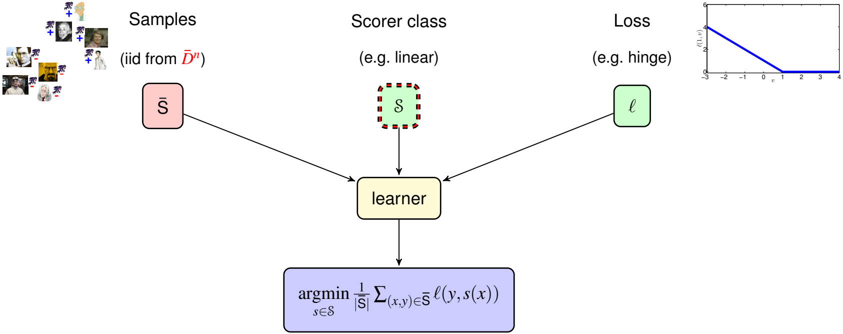
$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \frac{1}{|\bar{\mathsf{S}}|} \sum_{(x,y) \in \bar{\mathsf{S}}} \ell(y, s(x))$$

Samples from some $\bar{D} \neq D$, where labels flipped with certain probability

Noisy labels might affect us in three ways:

- insufficient samples?

Our view of noisy learning

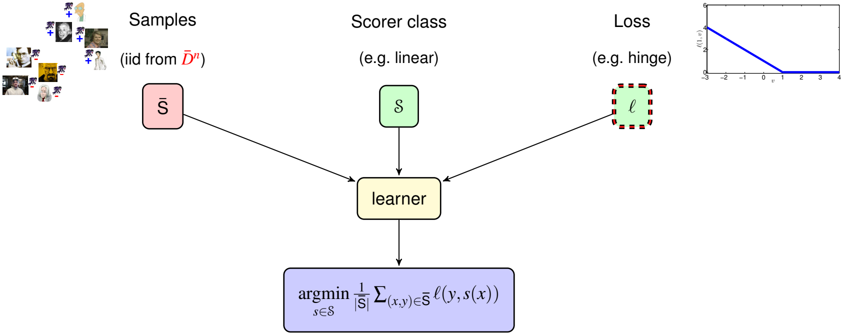


Samples from some $\bar{D} \neq D$, where labels flipped with certain probability

Noisy labels might affect us in three ways:

- insufficient samples?
- insufficiently rich model?

Our view of noisy learning

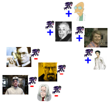


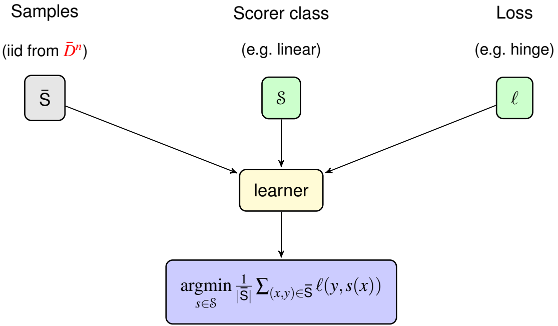
Samples from some $\bar{D} \neq D$, where labels flipped with certain probability

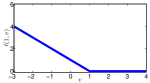
Noisy labels might affect us in three ways:

- insufficient samples?
- insufficiently rich model?
- insufficiently “robust” loss?

Our view of noisy learning







Samples from some $\bar{D} \neq D$, where labels flipped with certain probability

Noisy labels might affect us in three ways:

- insufficient samples?
- insufficiently rich model?
- insufficiently "robust" loss?

Noise-robustness

We would like our learner $(\ell, \mathcal{S})$ to be noise-robust

Noise-robustness

We would like our learner $(\ell, \mathcal{S})$ to be noise-robust

A (stringent) formalism:

Risk minimiser doesn't change under noise

- e.g. optimal classifier remains so

Noise-robustness

We would like our learner $(\ell, \mathcal{S})$ to be noise-robust

A (stringent) formalism:

Risk minimiser doesn't change under noise

- e.g. optimal classifier remains so

Ideal		**Reality**
$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}(s; D, \ell)$	$\stackrel{?}{=}$	$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}(s; \bar{D}, \ell)$

Roadmap

We have basically two ways to ensure robustness:

- pick a "good" loss ℓ
- pick a "good" scoring class $\mathcal{S}$

Recommended choice based on type of label noise...

Roadmap

We have basically two ways to ensure robustness:

- pick a “good” loss ℓ
- pick a “good” scoring class $\mathcal{S}$

Recommended choice based on type of label noise...

	Noise			
	Symmetric	Class-conditional	Instance	Instance and label
Loss ℓ	?	?	?	?
Scorer $\mathcal{S}$	?	?	?	?

Roadmap

We have basically two ways to ensure robustness:

- pick a "good" loss ℓ
- pick a "good" scoring class $\mathcal{S}$

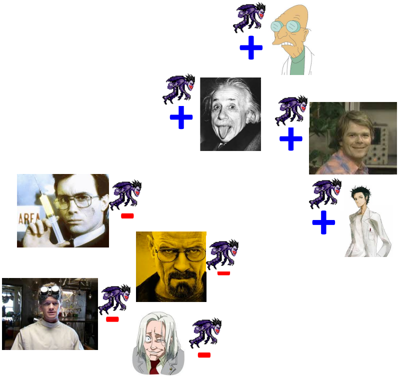
Recommended choice based on type of label noise...

	Noise			
	Symmetric	Class-conditional	Instance	Instance and label
Loss ℓ	**?**	?	?	?
Scorer $\mathcal{S}$	**?**	?	?	?

Noise-robustness via loss design

Warm up: symmetric label noise

Labels flipped with constant, instant-independent probability ρ



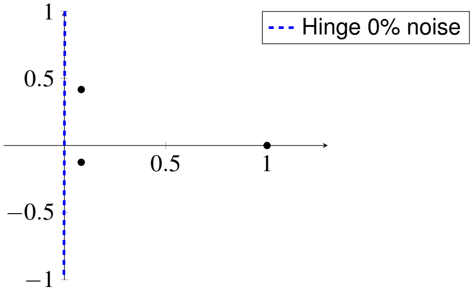
Seems innocuous enough...

Cool down: a disheartening result

Convex potentials ℓ and linear scorers $\mathcal{S}$ brittle to any such noise!

(Long and Servedio, 2010) gave constructive proof

- separable D concentrated on three points
- convex potential minimiser on $\bar{D}$ yields random guessing!

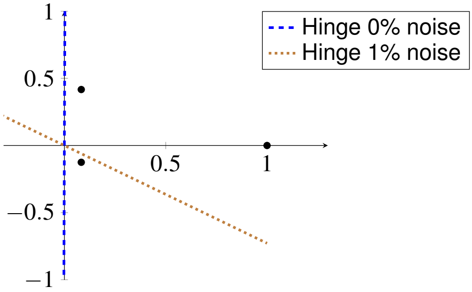


Cool down: a disheartening result

Convex potentials ℓ and linear scorers $\mathcal{S}$ brittle to any such noise!

(Long and Servedio, 2010) gave constructive proof

- separable D concentrated on three points
- convex potential minimiser on $\bar{D}$ yields random guessing!



Not all ℓ are equaℓ

Previous example relies on ℓ being convex potential

- doesn't preclude other losses being robust

Not all ℓ are equaℓ

Previous example relies on ℓ being convex potential

- doesn't preclude other losses being robust

Some hope: can show that, for any $\mathcal{S}$,

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}(s; D, \ell^{01}) = \underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}(s; \bar{D}, \ell^{01})$$

- in our earlier parlance, 0-1 loss is robust to symmetric noise

Not all ℓ are equaℓ

Previous example relies on ℓ being convex potential

- doesn't preclude other losses being robust

Some hope: can show that, for any $\mathcal{S}$,

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}(s; D, \ell^{01}) = \underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}(s; \bar{D}, \ell^{01})$$

- in our earlier parlance, 0-1 loss is robust to symmetric noise

For what other ℓ do we find, for any $\mathcal{S}$,

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}(s; D, \ell) \stackrel{?}{=} \underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}(s; \bar{D}, \ell)$$

Noise-corrected losses

(Natarajan et al., 2013) proved a useful fact:

Average loss on noisy data = average noisy loss on clean data

Noise-corrected losses

(Natarajan et al., 2013) proved a useful fact:

Average loss on noisy data = average noisy loss on clean data

Lemma

For any D*, loss* ℓ*, and* $\rho \in [0, 1/2)$*,* $\bar{D} = \mathrm{SLN}(D, \rho)$ *has*

$$\mathbb{L}(s; \bar{D}, \ell) = \mathbb{L}(s; D, \bar{\ell})$$

for noise-corrected loss

$$\bar{\ell}(y, v) = \frac{(1 - \rho) \cdot \ell(y, v) - \rho \cdot \ell(-y, v)}{1 - 2 \cdot \rho}.$$

Here, $\mathrm{SLN}(D, \rho)$ means D corrupted with symmetric noise

Noise-corrected losses: intuition

Noise-corrected loss is simply

$$\begin{bmatrix} \bar{\ell}_1(v) \\ \bar{\ell}_{-1}(v) \end{bmatrix} = \begin{bmatrix} 1-\rho & \rho \\ \rho & 1-\rho \end{bmatrix}^{-1} \begin{bmatrix} \ell_1(v) \\ \ell_{-1}(v) \end{bmatrix}$$

using shorthand $\ell_y(v) = \ell(y, v)$

Inverting noise-transition matrix to get unbiased estimate of ℓ

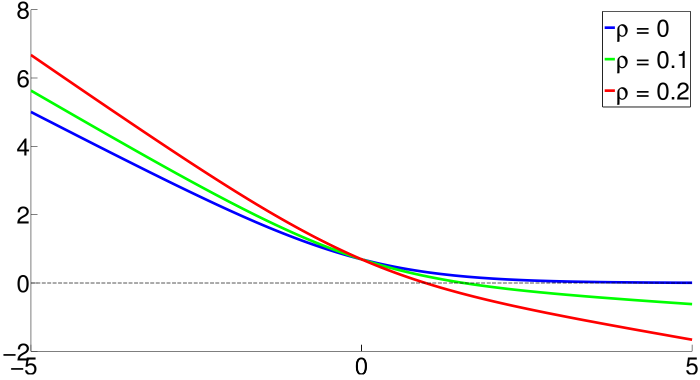
$\bar{\ell}$ (necessarily) depends on the unknown noise rate ρ

- if these can be estimated, very powerful!
- estimation possible under assumptions (for another day...)

Noise-corrected losses: example

For logistic loss, the noise-corrected losses are convex

- negatively unbounded for $\rho > 0$
- this will crop up later...



Back to risk mismatch

(Long and Servedio, 2010) example relies on

$$\operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{L}(s; D, \ell) \neq \operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{L}(s; \bar{D}, \ell)$$

for arbitrary $\mathcal{S}$

Back to risk mismatch

(Long and Servedio, 2010) example relies on

$$\operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{L}(s; D, \ell) \neq \operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{L}(s; D, \bar{\ell})$$

for arbitrary $\mathcal{S}$

Back to risk mismatch

(Long and Servedio, 2010) example relies on

$$\operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{L}(s; D, \ell) \neq \operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{L}(s; D, \bar{\ell})$$

for arbitrary $\mathcal{S}$

At least, not in general...

- and we can now compare ℓ and $\bar{\ell}$ on equal footing!

Eigen-losses

Since

$$\begin{bmatrix} \bar{\ell}_1(v) \\ \bar{\ell}_{-1}(v) \end{bmatrix} = \begin{bmatrix} 1-\rho & \rho \\ \rho & 1-\rho \end{bmatrix}^{-1} \begin{bmatrix} \ell_1(v) \\ \ell_{-1}(v) \end{bmatrix},$$

why not consider eigen-losses of this transform?

Eigen-losses

Since

$$\begin{bmatrix} \bar{\ell}_1(v) \\ \bar{\ell}_{-1}(v) \end{bmatrix} = \begin{bmatrix} 1-\rho & \rho \\ \rho & 1-\rho \end{bmatrix}^{-1} \begin{bmatrix} \ell_1(v) \\ \ell_{-1}(v) \end{bmatrix},$$

why not consider eigen-losses of this transform?

i.e. a loss ℓ for which

$$\begin{bmatrix} \bar{\ell}_1(v) \\ \bar{\ell}_{-1}(v) \end{bmatrix} = \lambda \cdot \begin{bmatrix} \ell_1(v) \\ \ell_{-1}(v) \end{bmatrix} + \mu$$

Eigen-losses

Since

$$\begin{bmatrix} \bar{\ell}_1(v) \\ \bar{\ell}_{-1}(v) \end{bmatrix} = \begin{bmatrix} 1-\rho & \rho \\ \rho & 1-\rho \end{bmatrix}^{-1} \begin{bmatrix} \ell_1(v) \\ \ell_{-1}(v) \end{bmatrix},$$

why not consider eigen-losses of this transform?

i.e. a loss ℓ for which

$$\begin{bmatrix} \bar{\ell}_1(v) \\ \bar{\ell}_{-1}(v) \end{bmatrix} = \lambda \cdot \begin{bmatrix} \ell_1(v) \\ \ell_{-1}(v) \end{bmatrix} + \mu$$

Such an ℓ would clearly have symmetric noise-robustness:

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \, \mathbb{L}(s; D, \ell) = \underset{s \in \mathcal{S}}{\operatorname{argmin}} \, \mathbb{L}(s; \bar{D}, \ell)$$

for any choice of $\mathcal{S}$

Convex eigen-losses?

Eigen-losses include any ℓ satisfying (c.f. (Ghosh et al., 2015))

$$\ell_1(v) + \ell_{-1}(v) = C$$

so that

$$\begin{bmatrix} \bar{\ell}_1(v) \\ \bar{\ell}_{-1}(v) \end{bmatrix} = \frac{1}{1 - 2 \cdot \rho} \cdot \begin{bmatrix} \ell_1(v) \\ \ell_{-1}(v) \end{bmatrix} - \frac{\rho}{1 - 2 \cdot \rho} \cdot C,$$

Convex eigen-losses?

Eigen-losses include any ℓ satisfying (c.f. (Ghosh et al., 2015))

$$\ell_1(v) + \ell_{-1}(v) = C$$

so that

$$\begin{bmatrix} \bar{\ell}_1(v) \\ \bar{\ell}_{-1}(v) \end{bmatrix} = \frac{1}{1 - 2 \cdot \rho} \cdot \begin{bmatrix} \ell_1(v) \\ \ell_{-1}(v) \end{bmatrix} - \frac{\rho}{1 - 2 \cdot \rho} \cdot C,$$

For ℓ nonnegative, this would imply ℓ must be non-convex

- e.g. 0-1 loss, ramp loss

Convex eigen-losses?

Eigen-losses include any ℓ satisfying (c.f. (Ghosh et al., 2015))

$$\ell_1(v) + \ell_{-1}(v) = C$$

so that

$$\begin{bmatrix} \bar{\ell}_1(v) \\ \bar{\ell}_{-1}(v) \end{bmatrix} = \frac{1}{1 - 2 \cdot \rho} \cdot \begin{bmatrix} \ell_1(v) \\ \ell_{-1}(v) \end{bmatrix} - \frac{\rho}{1 - 2 \cdot \rho} \cdot C,$$

For ℓ nonnegative, this would imply ℓ must be non-convex

- e.g. 0-1 loss, ramp loss

What if we remove the nonnegativity assumption?

- noise-corrected losses $\bar{\ell}$ frequently unbounded below

The unhinged loss

Removing nonnegativity, we can get a convex loss:

$$\begin{bmatrix} \ell_1(v) \\ \ell_{-1}(v) \end{bmatrix} = \begin{bmatrix} 1 - v \\ 1 + v \end{bmatrix}$$

The unhinged loss

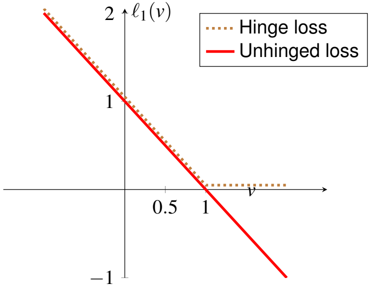
Removing nonnegativity, we can get a convex loss:

$$\begin{bmatrix} \ell_1(v) \\ \ell_{-1}(v) \end{bmatrix} = \begin{bmatrix} 1 - v \\ 1 + v \end{bmatrix}$$

We have unearthed a simple, noise-robust loss: the linear loss

$$\ell(y, v) = 1 - yv$$

- hinge loss without clamping at zero
- hence, also called the **"unhinged"** loss



Minimising the unhinged loss

Suppose we use regularised linear scorers $\mathcal{S} = \{x \mapsto \langle w, x \rangle\}$

- regularisation ensures boundedness of scores

An easy calculation reveals

$$\underset{w \in \mathcal{S}}{\operatorname{argmin}} \frac{\lambda}{2} \|w\|_2^2 + \underset{(\mathsf{X},\mathsf{Y}) \sim D}{\mathbb{E}} [-\mathsf{Y} \cdot \langle w, \mathsf{X} \rangle]$$

Minimising the unhinged loss

Suppose we use regularised linear scorers $\mathcal{S} = \{x \mapsto \langle w, x \rangle\}$

- regularisation ensures boundedness of scores

An easy calculation reveals

$$\underset{w \in \mathcal{S}}{\operatorname{argmin}} \frac{\lambda}{2} \|w\|_2^2 + \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [-\mathsf{Y} \cdot \langle w, \mathsf{X} \rangle] = \frac{1}{\lambda} \cdot \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [\mathsf{Y} \cdot \mathsf{X}]$$

Minimising the unhinged loss

Suppose we use regularised linear scorers $\mathcal{S} = \{x \mapsto \langle w, x \rangle\}$

- regularisation ensures boundedness of scores

An easy calculation reveals

$$\begin{aligned}
\underset{w \in \mathcal{S}}{\operatorname{argmin}} \frac{\lambda}{2} \|w\|_2^2 + \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [-\mathsf{Y} \cdot \langle w, \mathsf{X} \rangle] &= \frac{1}{\lambda} \cdot \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [\mathsf{Y} \cdot \mathsf{X}] \\
&= \frac{1}{\lambda} \cdot \left(\pi \cdot \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=1} [\mathsf{X}] - (1 - \pi) \cdot \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=-1} [\mathsf{X}] \right)
\end{aligned}$$

Minimising the unhinged loss

Suppose we use regularised linear scorers $\mathcal{S} = \{x \mapsto \langle w, x \rangle\}$

- regularisation ensures boundedness of scores

An easy calculation reveals

$$\underset{w \in \mathcal{S}}{\operatorname{argmin}} \frac{\lambda}{2} \|w\|_2^2 + \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [-\mathsf{Y} \cdot \langle w, \mathsf{X} \rangle] = \frac{1}{\lambda} \cdot \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [\mathsf{Y} \cdot \mathsf{X}]$$
$$= \frac{1}{\lambda} \cdot \left(\pi \cdot \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=1} [\mathsf{X}] - (1 - \pi) \cdot \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=-1} [\mathsf{X}] \right)$$

for $\pi = \mathbb{P}(\mathsf{Y} = 1)$

Minimising the unhinged loss

Suppose we use regularised linear scorers $\mathcal{S} = \{x \mapsto \langle w, x \rangle\}$

- regularisation ensures boundedness of scores

An easy calculation reveals

$$\underset{w \in \mathcal{S}}{\operatorname{argmin}} \frac{\lambda}{2} \|w\|_2^2 + \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [-\mathsf{Y} \cdot \langle w, \mathsf{X} \rangle] = \frac{1}{\lambda} \cdot \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [\mathsf{Y} \cdot \mathsf{X}]$$
$$= \frac{1}{\lambda} \cdot \left(\pi \cdot \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=1} [\mathsf{X}] - (1 - \pi) \cdot \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=-1} [\mathsf{X}] \right)$$

for $\pi = \mathbb{P}(\mathsf{Y} = 1)$

Minimiser is a weighted nearest centroid classifier

- this simple classifier is robust to symmetric label noise

Relation to square loss

Recall for square loss, $\ell(y, v) = (1 - yv)^2$, optimal linear scorer is

$$w^* = \left(\mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[\mathsf{X}\mathsf{X}^T \right] \right)^{-1} \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} \left[\mathsf{Y} \cdot \mathsf{X} \right]$$

Relation to square loss

Recall for square loss, $\ell(y, v) = (1 - yv)^2$, optimal linear scorer is

$$w^* = \left(\mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[\mathsf{X}\mathsf{X}^T \right] \right)^{-1} \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [\mathsf{Y} \cdot \mathsf{X}]$$

Unhinged solution is equivalent on whitened data

- note matrix inverse unaffected by noise
- simple proof that **square loss is also robust** (Manwani et al., 2014)

Relation to hinge loss

If $\|x\|_2 \leq X$, $\|w\|_2 \leq \frac{1}{X}$, by Cauchy-Schwartz

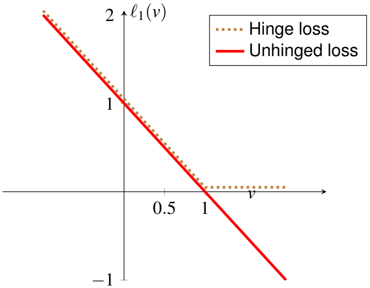
$$|\langle w, x \rangle| \leq 1$$

Relation to hinge loss

If $\|x\|_2 \leq X$, $\|w\|_2 \leq \frac{1}{X}$, by Cauchy-Schwartz

$$|\langle w, x \rangle| \leq 1$$

i.e. we **don't hit the "hinge" component**!

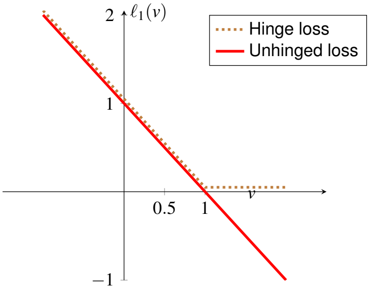


Relation to hinge loss

If $\|x\|_2 \leq X$, $\|w\|_2 \leq \frac{1}{X}$, by Cauchy-Schwartz

$$|\langle w, x \rangle| \leq 1$$

i.e. we **don't hit the "hinge" component**!

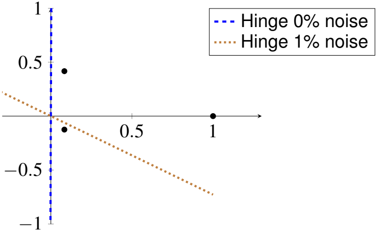


Thus, for large λ, unhinged $\equiv$ hinge loss

- unhinged minimisation $\equiv$ highly regularised SVM minimisation
- strong ℓ_2 regularisation $\implies$ symmetric noise robustness

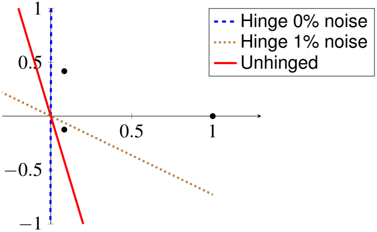
Experimental illustration

Distributional minimiser on (Long and Servedio, 2010) coherent:



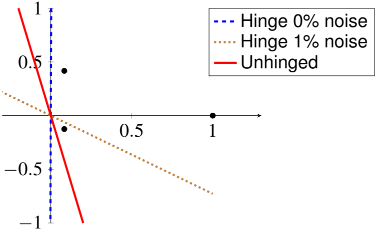
Experimental illustration

Distributional minimiser on (Long and Servedio, 2010) coherent:



Experimental illustration

Distributional minimiser on (Long and Servedio, 2010) coherent:



Empirical minimiser on sample of 800 instances also coherent:

	Hinge	**Unhinged**
$\rho = 0$	0.00 ± 0.00	0.00 ± 0.00
$\rho = 0.1$	0.15 ± 0.27	0.00 ± 0.00
$\rho = 0.2$	0.21 ± 0.30	0.00 ± 0.00
$\rho = 0.3$	0.38 ± 0.37	0.00 ± 0.00
$\rho = 0.4$	0.42 ± 0.36	0.00 ± 0.00
$\rho = 0.49$	0.47 ± 0.38	0.34 ± 0.48

Roadmap

To ensure robustness, either

- pick a "good" loss ℓ
- pick a "good" scoring class $\mathcal{S}$

	Noise			
	Symmetric	Class-conditional	Instance	Instance and label
Loss ℓ	Unhinged	?	?	?
Scorer $\mathcal{S}$	Arbitrary	?	?	?

Roadmap

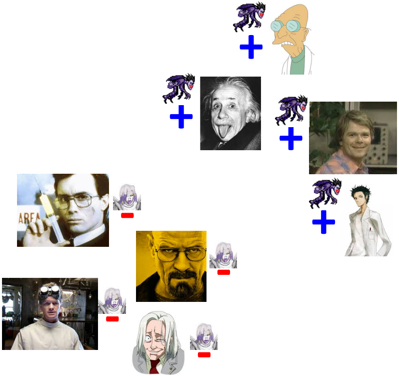
To ensure robustness, either

- pick a "good" loss ℓ
- pick a "good" scoring class $\mathcal{S}$

	Noise			
	Symmetric	Class-conditional	Instance	Instance and label
Loss ℓ	Unhinged	**?**	?	?
Scorer $\mathcal{S}$	Arbitrary	**?**	?	?

Class-conditional noise

Labels flipped with class-dependent probabilities ρ_+, ρ_-



Seems not overly different from symmetric case...

Another disheartening result

Unhinged loss is no longer robust to class-conditional noise:

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \, \mathbb{L}(s; D, \ell) \neq \underset{s \in \mathcal{S}}{\operatorname{argmin}} \, \mathbb{L}(s; \bar{D}, \ell)$$

for a generic function class $\mathcal{S} \subseteq \mathbb{R}^{\mathcal{X}}$

Another disheartening result

Unhinged loss is no longer robust to class-conditional noise:

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}(s; D, \ell) \neq \underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}(s; \bar{D}, \ell)$$

for a generic function class $\mathcal{S} \subseteq \mathbb{R}^{\mathcal{X}}$

Why? Under class-conditional noise, we have

$$\begin{bmatrix} \bar{\ell}_1(v) \\ \bar{\ell}_{-1}(v) \end{bmatrix} = \begin{bmatrix} 1 - \rho_+ & \rho_- \\ \rho_+ & 1 - \rho_- \end{bmatrix}^{-1} \begin{bmatrix} \ell_1(v) \\ \ell_{-1}(v) \end{bmatrix},$$

as per (Natarajan et al., 2013)

Another disheartening result

Unhinged loss is no longer robust to class-conditional noise:

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \, \mathbb{L}(s; D, \ell) \neq \underset{s \in \mathcal{S}}{\operatorname{argmin}} \, \mathbb{L}(s; \bar{D}, \ell)$$

for a generic function class $\mathcal{S} \subseteq \mathbb{R}^{\mathcal{X}}$

Why? Under class-conditional noise, we have

$$\begin{bmatrix} \bar{\ell}_1(v) \\ \bar{\ell}_{-1}(v) \end{bmatrix} = \begin{bmatrix} 1 - \rho_+ & \rho_- \\ \rho_+ & 1 - \rho_- \end{bmatrix}^{-1} \begin{bmatrix} \ell_1(v) \\ \ell_{-1}(v) \end{bmatrix},$$

as per (Natarajan et al., 2013)

Transition matrix no longer has noise-independent eigenvector!

Back to basics

Recall that for symmetric noise,

$$\operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{L}(s; D, \ell^{01}) = \operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{L}(s; \bar{D}, \ell^{01})$$

Is this still true in class-conditional case?

Back to basics

Recall that for symmetric noise,

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \, \mathbb{L}(s; D, \ell^{01}) = \underset{s \in \mathcal{S}}{\operatorname{argmin}} \, \mathbb{L}(s; \bar{D}, \ell^{01})$$

Is this still true in class-conditional case?

No! In fact,

$$\mathbb{L}(s; \bar{D}, \ell^{01}) = a \cdot \mathbb{L}(s; D, \ell^{(c)}) + b$$

for certain a, b, c and cost-sensitive loss $\ell^{(c)}$

- cost ratio c for false positives vs false negatives

Back to basics

Recall that for symmetric noise,

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}(s; D, \ell^{01}) = \underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}(s; \bar{D}, \ell^{01})$$

Is this still true in class-conditional case?

No! In fact,

$$\mathbb{L}(s; \bar{D}, \ell^{01}) = a \cdot \mathbb{L}(s; D, \ell^{(c)}) + b$$

for certain a, b, c and cost-sensitive loss $\ell^{(c)}$

- cost ratio c for false positives vs false negatives

Perhaps cost-sensitive losses fare better?

Loss balancing

Suppose we consider the risk for balanced 0-1 loss

$$\mathbb{L}(s; D, \ell^{\mathsf{bal}}) = \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} \left[w(\mathsf{Y}) \cdot \ell^{01}(\mathsf{Y}, s(\mathsf{X})) \right].$$

for $w(1) = \pi^{-1}, w(-1) = (1 - \pi)^{-1}$

Loss balancing

Suppose we consider the risk for balanced 0-1 loss

$$\mathbb{L}(s; D, \ell^{\text{bal}}) = \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} \left[w(\mathsf{Y}) \cdot \ell^{01}(\mathsf{Y}, s(\mathsf{X})) \right].$$

for $w(1) = \pi^{-1}, w(-1) = (1 - \pi)^{-1}$

Equally, this is the balanced error rate

$$\mathbb{L}(s; D, \ell^{\text{bal}}) = \mathbb{P}_{\mathsf{X}|\mathsf{Y}=+1}(\mathsf{Y} \neq \text{sign}(s(\mathsf{X}))) + \mathbb{P}_{\mathsf{X}|\mathsf{Y}=-1}(\mathsf{Y} \neq \text{sign}(s(\mathsf{X})))$$

- costs balance false positive and negative errors
- useful when classes are imbalanced

Balancing for class-conditional robustness

Balanced 0-1 loss **is** preserved under class-conditional noise

Lemma

For any D *and* $s \colon \mathcal{X} \to \mathbb{R}$, $\bar{D} = \mathsf{CCN}(D, \rho_+, \rho_-)$ *has*

$$\mathbb{L}(s; \bar{D}, \ell^{bal}) = a \cdot \mathbb{L}(s; D, \ell^{bal}) + b$$

for noise-dependent constants $a > 0, b > 0$.

Here, $\mathsf{CCN}(D, \rho_+, \rho_-)$ means D corrupted with class-conditional noise

Balancing for class-conditional robustness

Balanced 0-1 loss **is** preserved under class-conditional noise

Lemma

For any D *and* $s\colon \mathcal{X} \to \mathbb{R}$, $\bar{D} = \mathsf{CCN}(D, \rho_+, \rho_-)$ *has*

$$\mathbb{L}(s; \bar{D}, \ell^{bal}) = a \cdot \mathbb{L}(s; D, \ell^{bal}) + b$$

for noise-dependent constants $a > 0, b > 0$.

Here, $\mathsf{CCN}(D, \rho_+, \rho_-)$ means D corrupted with class-conditional noise

For any $\mathcal{S}$, minimisers are thus preserved:

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}}\, \mathbb{L}(s; D, \ell^{\mathsf{bal}}) = \underset{s \in \mathcal{S}}{\operatorname{argmin}}\, \mathbb{L}(s; \bar{D}, \ell^{\mathsf{bal}}).$$

Balancing and eigenvectors

Consider false negative and positive rates

$$\begin{aligned}\mathrm{FNR}(s;D) &= \mathbb{P}_{\mathsf{X}|\mathsf{Y}=+1}(\mathsf{Y} \neq \mathrm{sign}(s(\mathsf{X})))\\ \mathrm{FPR}(s;D) &= \mathbb{P}_{\mathsf{X}|\mathsf{Y}=-1}(\mathsf{Y} \neq \mathrm{sign}(s(\mathsf{X}))).\end{aligned}$$

Balancing and eigenvectors

Consider false negative and positive rates

$$\mathrm{FNR}(s;D) = \mathbb{P}_{\mathsf{X}|\mathsf{Y}=+1}(\mathsf{Y} \neq \mathrm{sign}(s(\mathsf{X})))$$
$$\mathrm{FPR}(s;D) = \mathbb{P}_{\mathsf{X}|\mathsf{Y}=-1}(\mathsf{Y} \neq \mathrm{sign}(s(\mathsf{X}))).$$

For noise-dependent α, β (c.f. Scott et al., 2013)

$$\begin{bmatrix}\mathrm{FPR}(s;\bar{D}) & \mathrm{FNR}(s;\bar{D})\end{bmatrix} = \begin{bmatrix}\mathrm{FPR}(s;D) & \mathrm{FNR}(s;D)\end{bmatrix} \begin{bmatrix} 1-\beta & -\alpha \\ -\beta & 1-\alpha \end{bmatrix} + \begin{bmatrix}\beta & \alpha\end{bmatrix}.$$

Balancing and eigenvectors

Consider false negative and positive rates

$$\mathrm{FNR}(s; D) = \mathbb{P}_{\mathsf{X}|\mathsf{Y}=+1}(\mathsf{Y} \neq \mathrm{sign}(s(\mathsf{X})))$$
$$\mathrm{FPR}(s; D) = \mathbb{P}_{\mathsf{X}|\mathsf{Y}=-1}(\mathsf{Y} \neq \mathrm{sign}(s(\mathsf{X}))).$$

For noise-dependent α, β (c.f. Scott et al., 2013)

$$\begin{bmatrix}\mathrm{FPR}(s; \bar{D}) & \mathrm{FNR}(s; \bar{D})\end{bmatrix} = \begin{bmatrix}\mathrm{FPR}(s; D) & \mathrm{FNR}(s; D)\end{bmatrix} \begin{bmatrix} 1-\beta & -\alpha \\ -\beta & 1-\alpha \end{bmatrix} + \begin{bmatrix}\beta & \alpha\end{bmatrix}.$$

This transition matrix has eigenvector $[1; 1]$

- hence balancing unaffected by noise!

Balancing for class-conditional robustness

By similarly balancing the unhinged loss, we find

$$\mathbb{L}_{\mathsf{bal}}(s; D, \ell) = a \cdot \mathbb{L}_{\mathsf{bal}}(s; \bar{D}, \ell) + b$$

for noise-dependent constants $a > 0, b > 0$, and thus

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}}\, \mathbb{L}_{\mathsf{bal}}(s; D, \ell) = \underset{s \in \mathcal{S}}{\operatorname{argmin}}\, \mathbb{L}_{\mathsf{bal}}(s; \bar{D}, \ell)$$

for any $\mathcal{S} \subseteq \mathbb{R}^{\mathcal{X}}$

- equally holds for any $\ell_1(v) + \ell_{-1}(v) = C$

Balancing for class-conditional robustness

By similarly balancing the unhinged loss, we find

$$\mathbb{L}_{\mathsf{bal}}(s; D, \ell) = a \cdot \mathbb{L}_{\mathsf{bal}}(s; \bar{D}, \ell) + b$$

for noise-dependent constants $a > 0, b > 0$, and thus

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}_{\mathsf{bal}}(s; D, \ell) = \underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{L}_{\mathsf{bal}}(s; \bar{D}, \ell)$$

for any $\mathcal{S} \subseteq \mathbb{R}^{\mathcal{X}}$

- equally holds for any $\ell_1(v) + \ell_{-1}(v) = C$

Balanced unhinged loss is robust to class-conditional noise

- corresponds to (unweighted) nearest-centroid classifier

Comment: what does it all mean?

Robustness of (weighted) mean classifier not surprising

Loss viewpoint more generally useful

- connection to ℓ_2 regularisation
- role of balancing

Mean operator useful for further analysis

- preservation implies approximate robustness (Patrini et al., 2016)

Roadmap

To ensure robustness, either

- pick a “good” loss ℓ
- pick a “good” scoring class $\mathcal{S}$

	Noise			
	Symmetric	Class-conditional	Instance	Instance and label
Loss ℓ	Unhinged	Weighted unhinged	?	?
Scorer $\mathcal{S}$	Arbitrary	Arbitrary	?	?

Roadmap

To ensure robustness, either

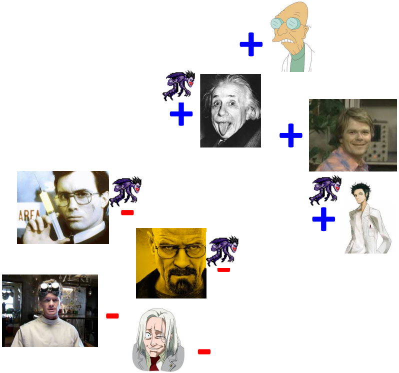
- pick a "good" loss ℓ
- pick a "good" scoring class $\mathcal{S}$

	Noise			
	Symmetric	Class-conditional	Instance	Instance and label
Loss ℓ	Unhinged	Weighted unhinged	**?**	?
Scorer $\mathcal{S}$	Arbitrary	Arbitrary	**?**	?

Noise-robustness via scorer design

Instance-dependent noise

Labels flipped with instance-dependent probability



Appears vastly more challenging...

One last disheartening result

Instance-dependent noise (unsurprisingly) breaks unhinged loss:

$$\operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{L}(s; D, \ell) \neq \operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{L}(s; \bar{D}, \ell)$$

for a generic function class $\mathcal{S} \subseteq \mathbb{R}^{\mathcal{X}}$

Why? Noise-transition is instance-dependent...

Crossroads

To ensure robustness, either

- pick a "good" loss ℓ
- pick a "good" scoring class $\mathcal{S}$

Crossroads

To ensure robustness, either

- pick a "good" loss ℓ and seek bounds
- pick a "good" scoring class $\mathcal{S}$

Crossroads

To ensure robustness, either

- pick a "good" loss ℓ and seek bounds
- pick a "good" scoring class $\mathcal{S}$

We'll follow the latter route

- progress is possible for former (Ghosh et al., 2015, van Rooyen et al., 2016)

In fact, we take $\mathcal{S}$ out of the picture altogether

- Bayes-optimal analysis of robustness

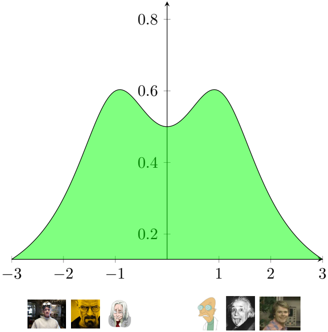
Distributions for learning with binary labels

For distribution D over $\mathcal{X} \times \{\pm 1\}$, we have

Distributions for learning with binary labels

For distribution D over $\mathcal{X} \times \{\pm 1\}$, we have

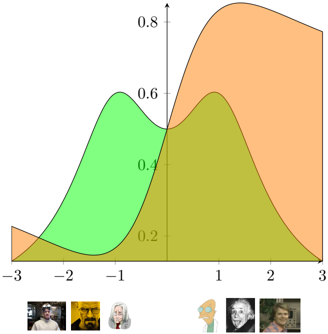
Marginal $\quad M(x) = \mathbb{P}(\mathsf{X} = x)$



Distributions for learning with binary labels

For distribution D over $\mathcal{X} \times \{\pm 1\}$, we have

Marginal	$M(x) = \mathbb{P}(\mathsf{X} = x)$
Class-probability function	$\eta(x) = \mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x)$



Bayes-optimal scorers

The theoretical best scorer for a given loss is any

$$s^* \in \underset{s \in \mathbb{R}^{\mathcal{X}}}{\operatorname{Argmin}} \mathbb{L}(s; D, \ell),$$

known as a Bayes-optimal scorer

Bayes-optimal scorers

The theoretical best scorer for a given loss is any

$$s^* \in \operatorname*{Argmin}_{s \in \mathbb{R}^{\mathcal{X}}} \mathbb{L}(s; D, \ell),$$

known as a Bayes-optimal scorer

For binary classification, any Bayes-optimal scorer has

$$\operatorname{sign}(s^*(x)) = \operatorname{sign}(2\eta(x) - 1)$$

- sign says whether, on average, instance is positive or not

A basic lemma

Lemma

For any $D = (M, \eta)$ *and* $\rho : \mathcal{X} \to [0, 1/2)$, $\bar{D} = \mathrm{IDN}(D, \rho)$ *has*

$$(\forall x \in \mathcal{X})\ \bar{\eta}(x) - \frac{1}{2} = (1 - 2 \cdot \rho(x)) \cdot \left(\eta(x) - \frac{1}{2} \right).$$

Here, $\mathrm{IDN}(D, \rho)$ means D corrupted with instance-dependent noise

A basic lemma

Lemma

For any $D = (M, \eta)$ *and* $\rho : \mathcal{X} \to [0, 1/2)$, $\bar{D} = \mathrm{IDN}(D, \rho)$ *has*

$$(\forall x \in \mathcal{X})\, \bar{\eta}(x) - \frac{1}{2} = (1 - 2 \cdot \rho(x)) \cdot \left(\eta(x) - \frac{1}{2} \right).$$

Here, $\mathrm{IDN}(D, \rho)$ means D corrupted with instance-dependent noise

The Bayes-optimal classifier is unchanged under noise:

$$\underset{s \in \{\pm 1\}^{\mathcal{X}}}{\operatorname{argmin}}\, \mathbb{L}(s; D, \ell) = \underset{s \in \{\pm 1\}^{\mathcal{X}}}{\operatorname{argmin}}\, \mathbb{L}(s; \bar{D}, \ell).$$

Crucially, this relies on using a powerful scorer class

spare me the details!

A basic lemma: proof

Proof.

By marginalising out the true label, we find

$$\bar{\eta}(x) = \mathbb{P}(\bar{\mathsf{Y}} = 1 \mid \mathsf{X} = x) = (1 - 2 \cdot \rho(x)) \cdot \eta(x) + \rho(x).$$

We have

□

A basic lemma: proof

Proof.

By marginalising out the true label, we find

$$\bar{\eta}(x) = \mathbb{P}(\bar{\mathsf{Y}} = 1 \mid \mathsf{X} = x) = (1 - 2 \cdot \rho(x)) \cdot \eta(x) + \rho(x).$$

We have

$$\bar{\eta}(x) - \frac{1}{2} = (1 - 2 \cdot \rho(x)) \cdot \eta(x) + \rho(x) - \frac{1}{2}$$

□

A basic lemma: proof

Proof.

By marginalising out the true label, we find

$$\bar{\eta}(x) = \mathbb{P}(\bar{\mathsf{Y}} = 1 \mid \mathsf{X} = x) = (1 - 2 \cdot \rho(x)) \cdot \eta(x) + \rho(x).$$

We have

$$\begin{aligned}
\bar{\eta}(x) - \frac{1}{2} &= (1 - 2 \cdot \rho(x)) \cdot \eta(x) + \rho(x) - \frac{1}{2} \\
&= \eta(x) - \frac{1}{2} + \rho(x) \cdot (1 - 2 \cdot \eta(x))
\end{aligned}$$

□

A basic lemma: proof

Proof.

By marginalising out the true label, we find

$$\bar{\eta}(x) = \mathbb{P}(\bar{\mathsf{Y}} = 1 \mid \mathsf{X} = x) = (1 - 2 \cdot \rho(x)) \cdot \eta(x) + \rho(x).$$

We have

$$\begin{aligned}
\bar{\eta}(x) - \frac{1}{2} &= (1 - 2 \cdot \rho(x)) \cdot \eta(x) + \rho(x) - \frac{1}{2} \\
&= \eta(x) - \frac{1}{2} + \rho(x) \cdot (1 - 2 \cdot \eta(x)) \\
&= \eta(x) - \frac{1}{2} + 2 \cdot \rho(x) \cdot \left(\frac{1}{2} - \eta(x) \right)
\end{aligned}$$

$\square$

A basic lemma: proof

Proof.

By marginalising out the true label, we find

$$\bar{\eta}(x) = \mathbb{P}(\bar{\mathsf{Y}} = 1 \mid \mathsf{X} = x) = (1 - 2 \cdot \rho(x)) \cdot \eta(x) + \rho(x).$$

We have

$$\begin{aligned}
\bar{\eta}(x) - \frac{1}{2} &= (1 - 2 \cdot \rho(x)) \cdot \eta(x) + \rho(x) - \frac{1}{2} \\
&= \eta(x) - \frac{1}{2} + \rho(x) \cdot (1 - 2 \cdot \eta(x)) \\
&= \eta(x) - \frac{1}{2} + 2 \cdot \rho(x) \cdot \left(\frac{1}{2} - \eta(x)\right) \\
&= (1 - 2 \cdot \rho(x)) \cdot \left(\eta(x) - \frac{1}{2}\right).
\end{aligned}$$

□

Regret of sub-optimal solutions

Optimal solutions align, but what about sub-optimal solutions?

Regret of sub-optimal solutions

Optimal solutions align, but what about sub-optimal solutions?

Assess quality of generic scorer $s\colon \mathcal{X} \to \mathbb{R}$ using regret:

$$\mathrm{regret}(s; D, \ell) = \mathbb{L}(s; D, \ell) - \min_{s^* \in \mathbb{R}^{\mathcal{X}}} \mathbb{L}(s^*; D, \ell)$$

- excess risk over best (Bayes-optimal) scorer
- calibrated losses ℓ have surrogate regret bounds:

$$\mathrm{regret}(s; D, \ell^{01}) \leq \varphi_\ell(\mathrm{regret}(s; D, \ell)).$$

Regret of sub-optimal solutions

Optimal solutions align, but what about sub-optimal solutions?

Assess quality of generic scorer $s \colon \mathcal{X} \to \mathbb{R}$ using regret:

$$\text{regret}(s; D, \ell) = \mathbb{L}(s; D, \ell) - \min_{s^* \in \mathbb{R}^{\mathcal{X}}} \mathbb{L}(s^*; D, \ell)$$

- excess risk over best (Bayes-optimal) scorer
- calibrated losses ℓ have surrogate regret bounds:

$$\text{regret}(s; D, \ell^{01}) \leq \varphi_\ell(\text{regret}(s; D, \ell)).$$

Can we relate regret on clean D and noisy $\bar{D}$?

Classification regret bound

Lemma

For any $D = (M, \eta)$, $\rho : \mathcal{X} \to [0, \rho_{max}]$, *and scorer* $s : \mathcal{X} \to \mathbb{R}$,

$$\mathrm{regret}(s; D, \ell^{01}) \leq \frac{1}{1 - 2 \cdot \rho_{max}} \cdot \mathrm{regret}(s; \bar{D}, \ell^{01}).$$

Classification regret bound

Lemma

For any $D = (M, \eta)$, $\rho : \mathcal{X} \to [0, \rho_{max}]$, *and scorer* $s : \mathcal{X} \to \mathbb{R}$,

$$\text{regret}(s; D, \ell^{01}) \leq \frac{1}{1 - 2 \cdot \rho_{max}} \cdot \text{regret}(s; \bar{D}, \ell^{01}).$$

Consistent classification from noisy samples alone

- can be ensured with calibrated surrogate minimisation

Classification regret bound

Lemma

For any $D = (M, \eta)$, $\rho : \mathcal{X} \to [0, \rho_{max}]$, *and scorer* $s : \mathcal{X} \to \mathbb{R}$,

$$\mathrm{regret}(s; D, \ell^{01}) \leq \frac{1}{1 - 2 \cdot \rho_{max}} \cdot \mathrm{regret}(s; \bar{D}, \ell^{01}).$$

Consistent classification from noisy samples alone

- can be ensured with calibrated surrogate minimisation

For $\rho_{\mathrm{max}} \approx \frac{1}{2}$, large constant penalty

- can trade-off dependence on ρ_{max} and on noisy regret

spare me the details!

Classification regret bound: proof

Proof.

Suppose $w(x) = \frac{1}{1-2\cdot\rho(x)}$, and $w_{\text{max}} = \max_x w(x)$. Then:

□

Classification regret bound: proof

Proof.

Suppose $w(x) = \frac{1}{1-2\cdot\rho(x)}$, and $w_{\mathsf{max}} = \max_x w(x)$. Then:

$$\mathrm{regret}(s; D, \ell^{01}) = \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[\left|\eta(\mathsf{X}) - \frac{1}{2}\right| [\![(2\cdot\eta(\mathsf{X}) - 1)\cdot s(\mathsf{X}) < 0]\!]\right]$$

□

Classification regret bound: proof

Proof.

Suppose $w(x) = \frac{1}{1 - 2 \cdot \rho(x)}$, and $w_{\text{max}} = \max_x w(x)$. Then:

$$\begin{aligned}
\text{regret}(s; D, \ell^{01}) &= \mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[\left| \eta(\mathsf{X}) - \frac{1}{2} \right| [\![(2 \cdot \eta(\mathsf{X}) - 1) \cdot s(\mathsf{X}) < 0]\!] \right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[\left| \eta(\mathsf{X}) - \frac{1}{2} \right| [\![(2 \cdot \bar{\eta}(\mathsf{X}) - 1) \cdot s(\mathsf{X}) < 0]\!] \right]
\end{aligned}$$

□

Classification regret bound: proof

Proof.

Suppose $w(x) = \frac{1}{1-2\cdot\rho(x)}$, and $w_{\mathsf{max}} = \max_x w(x)$. Then:

$$\begin{aligned}
\mathrm{regret}(s; D, \ell^{01}) &= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[\left|\eta(\mathsf{X}) - \frac{1}{2}\right| [\![(2\cdot\eta(\mathsf{X}) - 1)\cdot s(\mathsf{X}) < 0]\!]\right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[\left|\eta(\mathsf{X}) - \frac{1}{2}\right| [\![(2\cdot\bar{\eta}(\mathsf{X}) - 1)\cdot s(\mathsf{X}) < 0]\!]\right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[w(\mathsf{X})\cdot\left|\bar{\eta}(\mathsf{X}) - \frac{1}{2}\right| [\![(2\cdot\bar{\eta}(\mathsf{X}) - 1)\cdot s(\mathsf{X}) < 0]\!]\right]
\end{aligned}$$

□

Classification regret bound: proof

Proof.

Suppose $w(x) = \frac{1}{1 - 2 \cdot \rho(x)}$, and $w_{\mathrm{max}} = \max_x w(x)$. Then:

$$\begin{aligned}
\mathrm{regret}(s; D, \ell^{01}) &= \mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[\left| \eta(\mathsf{X}) - \frac{1}{2} \right| [\![(2 \cdot \eta(\mathsf{X}) - 1) \cdot s(\mathsf{X}) < 0]\!] \right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[\left| \eta(\mathsf{X}) - \frac{1}{2} \right| [\![(2 \cdot \bar{\eta}(\mathsf{X}) - 1) \cdot s(\mathsf{X}) < 0]\!] \right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[w(\mathsf{X}) \cdot \left| \bar{\eta}(\mathsf{X}) - \frac{1}{2} \right| [\![(2 \cdot \bar{\eta}(\mathsf{X}) - 1) \cdot s(\mathsf{X}) < 0]\!] \right] \\
&\leq w_{\mathrm{max}} \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[\left| \bar{\eta}(\mathsf{X}) - \frac{1}{2} \right| [\![(2 \cdot \bar{\eta}(\mathsf{X}) - 1) \cdot s(\mathsf{X}) < 0]\!] \right]
\end{aligned}$$

□

Classification regret bound: proof

Proof.

Suppose $w(x) = \frac{1}{1-2\cdot\rho(x)}$, and $w_{\text{max}} = \max_x w(x)$. Then:

$$
\begin{aligned}
\text{regret}(s; D, \ell^{01}) &= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[\left|\eta(\mathsf{X}) - \frac{1}{2}\right| [\![(2\cdot\eta(\mathsf{X}) - 1)\cdot s(\mathsf{X}) < 0]\!]\right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[\left|\eta(\mathsf{X}) - \frac{1}{2}\right| [\![(2\cdot\bar{\eta}(\mathsf{X}) - 1)\cdot s(\mathsf{X}) < 0]\!]\right] \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[w(\mathsf{X})\cdot\left|\bar{\eta}(\mathsf{X}) - \frac{1}{2}\right| [\![(2\cdot\bar{\eta}(\mathsf{X}) - 1)\cdot s(\mathsf{X}) < 0]\!]\right] \\
&\leq w_{\text{max}}\cdot\mathop{\mathbb{E}}_{\mathsf{X}\sim M}\left[\left|\bar{\eta}(\mathsf{X}) - \frac{1}{2}\right| [\![(2\cdot\bar{\eta}(\mathsf{X}) - 1)\cdot s(\mathsf{X}) < 0]\!]\right] \\
&= w_{\text{max}}\cdot\text{regret}(s; \bar{D}, \ell^{01}).
\end{aligned}
$$

□

Roadmap

To ensure robustness, either

- pick a "good" loss ℓ
- pick a "good" scoring class $\mathcal{S}$

	Noise			
	Symmetric	Class-conditional	Instance	Instance and label
Loss ℓ	Unhinged	Weighted unhinged	Calibrated	?
Scorer $\mathcal{S}$	Arbitrary	Arbitrary	$\mathbb{R}^{\mathcal{X}}$	?

Roadmap

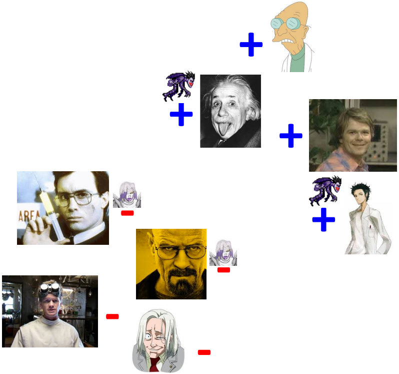
To ensure robustness, either

- pick a "good" loss ℓ
- pick a "good" scoring class $\mathcal{S}$

	Noise			
	Symmetric	Class-conditional	Instance	Instance and label
Loss ℓ	Unhinged	Weighted unhinged	Calibrated	**?**
Scorer $\mathcal{S}$	Arbitrary	Arbitrary	$\mathbb{R}^{\mathcal{X}}$	**?**

Instance- and label-dependent noise

Labels flipped with instance- and label-dependent probability



Does rich $\mathcal{S}$ help here?

Comment: instance- and label-dependent noise

Bad news: no longer have 0-1 consistency

Comment: instance- and label-dependent noise

Bad news: no longer have 0-1 consistency

Worse news: balancing doesn't help!

Comment: instance- and label-dependent noise

Bad news: no longer have 0-1 consistency

Worse news: balancing doesn't help!

Why is this so?

Relating clean- and corrupted- probabilities

Lemma

For any $D = (M, \eta)$ *and* $\rho_{\pm 1} : \mathcal{X} \to [0, 1/2)$, $\bar{D} = \mathsf{ILN}(D, \rho_{\pm 1})$ *has*

$$(\forall x \in \mathcal{X})\ \bar{\eta}(x) = (1 - \rho_{+}(x) - \rho_{-}(x)) \cdot \eta(x) + \rho_{-}(x)$$

Here, $\mathsf{ILN}(D, \rho_{\pm 1})$ means D with instance- and label-dependent noise

Relating clean- and corrupted- probabilities

Lemma

For any $D = (M, \eta)$ *and* $\rho_{\pm 1} : \mathcal{X} \to [0, 1/2)$, $\bar{D} = \mathrm{ILN}(D, \rho_{\pm 1})$ *has*

$$(\forall x \in \mathcal{X})\, \bar{\eta}(x) = (1 - \rho_+(x) - \rho_-(x)) \cdot \eta(x) + \rho_-(x)$$

Here, $\mathrm{ILN}(D, \rho_{\pm 1})$ means D with instance- and label-dependent noise

As a result, we find

$$\underset{s \in \{\pm 1\}^{\mathcal{X}}}{\operatorname{argmin}}\, \mathbb{L}(s; D, \ell) \neq \underset{s \in \{\pm 1\}^{\mathcal{X}}}{\operatorname{argmin}}\, \mathbb{L}(s; \bar{D}, \ell).$$

- cannot preserve thresholds of η

Relating clean- and corrupted- probabilities

Lemma

For any $D = (M, \eta)$ *and* $\rho_{\pm 1} : \mathcal{X} \to [0, 1/2)$, $\bar{D} = \mathrm{ILN}(D, \rho_{\pm 1})$ *has*

$$(\forall x \in \mathcal{X})\ \bar{\eta}(x) = (1 - \rho_+(x) - \rho_-(x)) \cdot \eta(x) + \rho_-(x)$$

Here, $\mathrm{ILN}(D, \rho_{\pm 1})$ means D with instance- and label-dependent noise

As a result, we find

$$\underset{s \in \{\pm 1\}^{\mathcal{X}}}{\operatorname{argmin}}\ \mathbb{L}(s; D, \ell) \neq \underset{s \in \{\pm 1\}^{\mathcal{X}}}{\operatorname{argmin}}\ \mathbb{L}(s; \bar{D}, \ell).$$

- cannot preserve thresholds of η
- what about ordering of η?

Probabilistically consistent noise

Suppose the noise is probabilistically consistent:

$$\rho_{\pm 1}(x) = f_{\pm 1}(\eta(x))$$

Probabilistically consistent noise

Suppose the noise is probabilistically consistent:

$$\rho_{\pm 1}(x) = f_{\pm 1}(\eta(x))$$

where $f_{\pm 1}$ are increasing on $[0, 1/2)$ and decreasing on $(1/2, 1]$

- higher inherent uncertainty $\rightarrow$ higher noise
- could model annotator noise

Further assume $f_1(z) - f_{-1}(z)$ is non-increasing

- trivially satisfied for label-independent noise

Probabilistically consistent noise

Suppose the noise is probabilistically consistent:

$$\rho_{\pm 1}(x) = f_{\pm 1}(\eta(x))$$

where $f_{\pm 1}$ are increasing on $[0, 1/2)$ and decreasing on $(1/2, 1]$

- higher inherent uncertainty $\rightarrow$ higher noise
- could model annotator noise

Further assume $f_1(z) - f_{-1}(z)$ is non-increasing

- trivially satisfied for label-independent noise

Lemma

For probabilistically consistent noise, $\bar{\eta}$ is monotone transform of η.

Efficiently learning under ILN

Suppose we assume D has $\eta(x) = u(\langle w^*, x \rangle)$

- u known $\rightarrow$ generalised linear model (GLM)
- u unknown $\rightarrow$ single index model (SIM)

Efficiently learning under ILN

Suppose we assume D has $\eta(x) = u(\langle w^*, x \rangle)$

- u known $\rightarrow$ generalised linear model (GLM)
- u unknown $\rightarrow$ single index model (SIM)

Under probabilistically consistent noise,

$$\bar{\eta}(x) = \bar{u}(\langle w^*, x \rangle)$$

- different, but still monotone, transform
- even if u known, $\bar{u}$ will be unknown

The Isotron algorithm

Can learn generic SIMs using Isotron

- akin to standard GLM, but additional step to estimate link function

The Isotron algorithm

Can learn generic SIMs using Isotron

- akin to standard GLM, but additional step to estimate link function

Input: Samples $\{(x_i, y_i)\}_{i=1}^{m}$, iterations T

Output: Link function u_T, weight vector w_T

$w_0 \leftarrow 0$

$u_0 \leftarrow z \mapsto \min(\max(0, 2 \cdot z + 1), 1)$

`for` $t = 1, 2, \ldots$

$$w_t \leftarrow w_{t-1} + \frac{1}{m} \sum_{i=1}^{m} (y_i - u_{t-1}(\langle w_{t-1}, x_i \rangle)) \cdot x_i$$

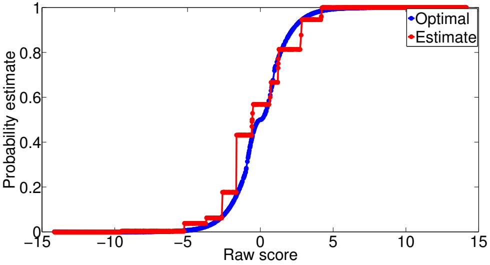
$$u_t \leftarrow \text{IsotonicRegression}(\{\langle w_t, x_i \rangle, y_i\})$$

The Isotron and ILN noise

For probabilistically consistent noise, can estimate $\bar{\eta}$ via Isotron!

Do not need to know flip functions

- only need to know noise is probabilistically consistent



Isotron illustration

Instance-dependent noise with $f_{\pm 1}(z) = (1 + e^{|\langle w^*, x\rangle|/\alpha})^{-1}$ on USPS 0v9 and MNIST 6v7

α	Flip %	Ridge ACC	Isotron ACC
$\frac{1}{8}$	0.03 $\pm$ 0.01	0.9940 $\pm$ 0.0003	0.9974 $\pm$ 0.0002
$\frac{1}{4}$	0.17 $\pm$ 0.01	0.9947 $\pm$ 0.0004	0.9974 $\pm$ 0.0003
$\frac{1}{2}$	2.15 $\pm$ 0.09	0.9944 $\pm$ 0.0004	0.9937 $\pm$ 0.0006
1	11.84 $\pm$ 0.17	0.9853 $\pm$ 0.0012	0.9700 $\pm$ 0.0021
2	26.57 $\pm$ 0.22	0.8988 $\pm$ 0.0053	0.9239 $\pm$ 0.0050
4	37.65 $\pm$ 0.24	0.7410 $\pm$ 0.0072	0.7863 $\pm$ 0.0138
8	43.76 $\pm$ 0.25	0.6185 $\pm$ 0.0078	0.6467 $\pm$ 0.0405

α	Flip %	Ridge ACC	Isotron ACC
$\frac{1}{8}$	0.04 $\pm$ 0.00	0.9958 $\pm$ 0.0001	0.9984 $\pm$ 0.0001
$\frac{1}{4}$	0.44 $\pm$ 0.01	0.9958 $\pm$ 0.0001	0.9979 $\pm$ 0.0001
$\frac{1}{2}$	4.25 $\pm$ 0.04	0.9953 $\pm$ 0.0002	0.9966 $\pm$ 0.0003
1	15.97 $\pm$ 0.05	0.9871 $\pm$ 0.0005	0.9864 $\pm$ 0.0007
2	29.97 $\pm$ 0.09	0.9446 $\pm$ 0.0012	0.9565 $\pm$ 0.0013
4	39.49 $\pm$ 0.08	0.8262 $\pm$ 0.0022	0.8768 $\pm$ 0.0041
8	44.63 $\pm$ 0.08	0.6872 $\pm$ 0.0024	0.8088 $\pm$ 0.0291

Isotron illustration

Instance-dependent noise with $f_{\pm 1}(z) = (1 + e^{|\langle w^*, x \rangle|/\alpha})^{-1}$ on USPS 0v9 and MNIST 6v7

α	Flip %	Ridge ACC	Isotron ACC
$\frac{1}{8}$	0.03 ± 0.01	0.9940 ± 0.0003	0.9974 ± 0.0002
$\frac{1}{4}$	0.17 ± 0.01	0.9947 ± 0.0004	0.9974 ± 0.0003
$\frac{1}{2}$	2.15 ± 0.09	0.9944 ± 0.0004	0.9937 ± 0.0006
1	11.84 ± 0.17	0.9853 ± 0.0012	0.9700 ± 0.0021
2	26.57 ± 0.22	0.8988 ± 0.0053	0.9239 ± 0.0050
4	37.65 ± 0.24	0.7410 ± 0.0072	0.7863 ± 0.0138
8	43.76 ± 0.25	0.6185 ± 0.0078	0.6467 ± 0.0405

α	Flip %	Ridge ACC	Isotron ACC
$\frac{1}{8}$	0.04 ± 0.00	0.9958 ± 0.0001	0.9984 ± 0.0001
$\frac{1}{4}$	0.44 ± 0.01	0.9958 ± 0.0001	0.9979 ± 0.0001
$\frac{1}{2}$	4.25 ± 0.04	0.9953 ± 0.0002	0.9966 ± 0.0003
1	15.97 ± 0.05	0.9871 ± 0.0005	0.9864 ± 0.0007
2	29.97 ± 0.09	0.9446 ± 0.0012	0.9565 ± 0.0013
4	39.49 ± 0.08	0.8262 ± 0.0022	0.8768 ± 0.0041
8	44.63 ± 0.08	0.6872 ± 0.0024	0.8088 ± 0.0291

Thresholding still problematic for label-dependent noise...

Ranking and area under ROC

Area under ROC curve (AUC) is probability of random positive scoring higher than random negative

$$\mathrm{AUC}(s;D) = \mathbb{P}_{\mathsf{X}|\mathsf{Y}=+1,\mathsf{X}'|\mathsf{Y}=-1}\left(s(\mathsf{X}) > s(\mathsf{X}')\right).$$

- assesses ranking performance of s

Ranking and area under ROC

Area under ROC curve (AUC) is probability of random positive scoring higher than random negative

$$\mathrm{AUC}(s; D) = \mathbb{P}_{\mathsf{X}|\mathsf{Y}=+1, \mathsf{X}'|\mathsf{Y}=-1}\left(s(\mathsf{X}) > s(\mathsf{X}')\right).$$

- assesses ranking performance of s

Classical result:

$$\underset{s \colon \mathcal{X} \to \mathbb{R}}{\operatorname{argmin}}\, 1 - \mathrm{AUC}(s; D) = \phi \circ \eta$$

for any monotone increasing ϕ

Ranking and area under ROC

Area under ROC curve (AUC) is probability of random positive scoring higher than random negative

$$\mathrm{AUC}(s; D) = \mathbb{P}_{\mathsf{X}|\mathsf{Y}=+1, \mathsf{X}'|\mathsf{Y}=-1}\left(s(\mathsf{X}) > s(\mathsf{X}')\right).$$

- assesses ranking performance of s

Classical result:

$$\underset{s \colon \mathcal{X} \to \mathbb{R}}{\operatorname{argmin}}\, 1 - \mathrm{AUC}(s; D) = \phi \circ \eta$$

for any monotone increasing ϕ

Thus, for probabilistically consistent noise,

$$\underset{s \colon \mathcal{X} \to \mathbb{R}}{\operatorname{argmin}}\, 1 - \mathrm{AUC}(s; D) = \underset{s \colon \mathcal{X} \to \mathbb{R}}{\operatorname{argmin}}\, 1 - \mathrm{AUC}(s; \bar{D})$$

AUC regret bound

We can similarly obtain an AUC regret bound

Lemma

For any D *and* $\bar{D} = \mathsf{ILN}(D, f_{-1} \circ \eta, f_1 \circ \eta)$ *where* (f_{-1}, f_1) *are probabilistically consistent, and for any scorer* $s \colon \mathcal{X} \to \mathbb{R}$*,*

$$\mathrm{regret}_{\mathrm{AUC}}(s; D) \leq \frac{C}{1 - 2 \cdot \rho_{\max}} \cdot \mathrm{regret}_{\mathrm{AUC}}(s; \bar{D})$$

for constant $C > 0$ *and*

$$\rho_{\max} = \frac{1}{2} \cdot \max_{x \in \mathcal{X}} (\rho_1(x) + \rho_{-1}(x)).$$

AUC regret bound

We can similarly obtain an AUC regret bound

Lemma

For any D *and* $\bar{D} = \mathsf{ILN}(D, f_{-1} \circ \eta, f_1 \circ \eta)$ *where* (f_{-1}, f_1) *are probabilistically consistent, and for any scorer* $s\colon \mathcal{X} \to \mathbb{R}$*,*

$$\text{regret}_{\text{AUC}}(s; D) \leq \frac{C}{1 - 2 \cdot \rho_{\max}} \cdot \text{regret}_{\text{AUC}}(s; \bar{D})$$

for constant $C > 0$ *and*

$$\rho_{\max} = \frac{1}{2} \cdot \max_{x \in \mathcal{X}} (\rho_1(x) + \rho_{-1}(x)).$$

Can guarantee $\text{regret}_{\text{AUC}}(s; D) \to 0$ by minimising a proper loss

- fundamental losses of class-probability estimation

The final picture

To ensure robustness, either

- pick a "good" loss ℓ
- pick a "good" scoring class $\mathcal{S}$

	Noise			
	Symmetric	Class-conditional	Instance	Instance and label*
Loss ℓ	Unhinged	Weighted unhinged	Calibrated	Proper
Scorer $\mathcal{S}$	Arbitrary	Arbitrary	$\mathbb{R}^{\mathcal{X}}$	$\mathbb{R}^{\mathcal{X}}$

Conclusion

Talk recap

Can we learn a good classifier from noisy samples?

Yes, by either:

- choosing a suitably robust loss function
- choosing a suitably rich function class

For another day

More to be said about coping with noise:

- optimising more complex performance measures
- procedure for estimating noise rates
- application to positive and unlabelled learning
- ...

The rat pack

Brendan van Rooyen

Bob Williamson

Cheng Soon Ong

Nagarajan Natarajan

Further reading

Learning with symmetric label noise: the importance of being unhinged. Brendan van Rooyen, Aditya Krishna Menon and Robert C. Williamson. NIPS 2015.

Learning from corrupted binary labels via class-probability estimation. Aditya Krishna Menon, Brendan van Rooyen, Cheng Soon Ong and Robert C. Williamson. ICML 2015.

Learning from binary labels with instance-dependent corruption. Aditya Krishna Menon, Brendan van Rooyen and Nagarajan Natarajan. `https://arxiv.org/abs/1605.00751`.

Thanks!