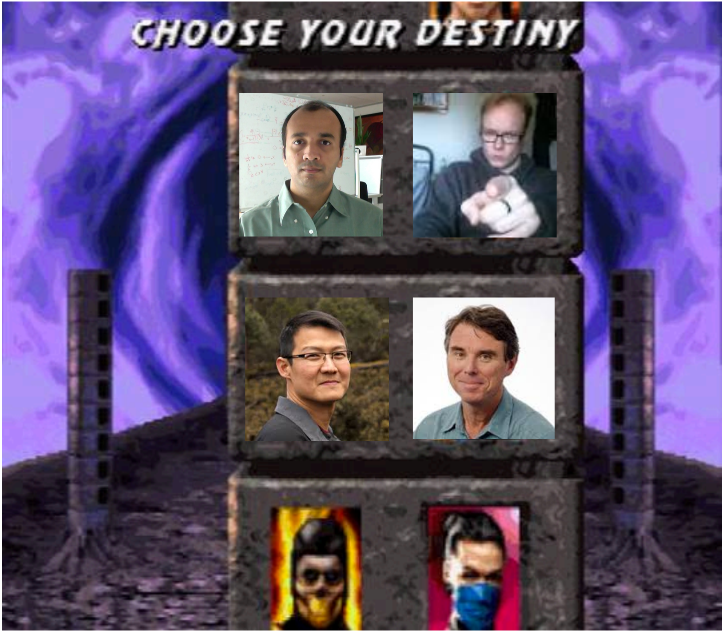
Learning from Corrupted Binary Labels via Class-Probability Estimation

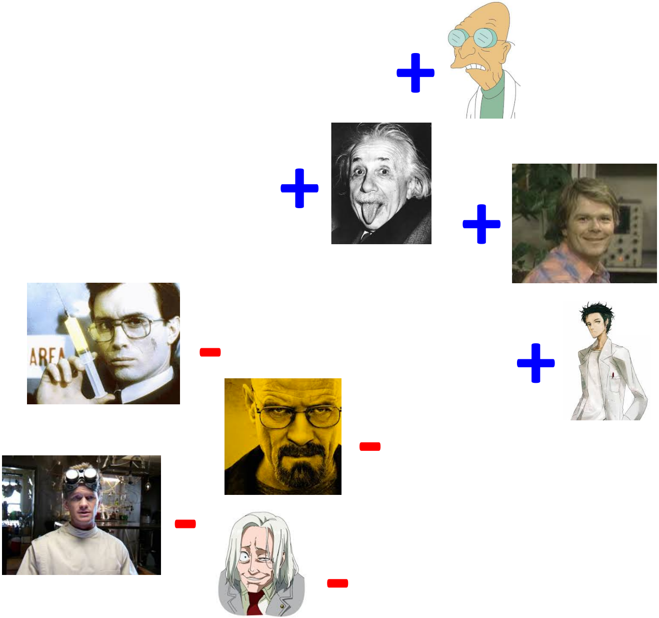
Aditya Krishna Menon

Aug 6th, 2015

National ICT Australia and The Australian National University

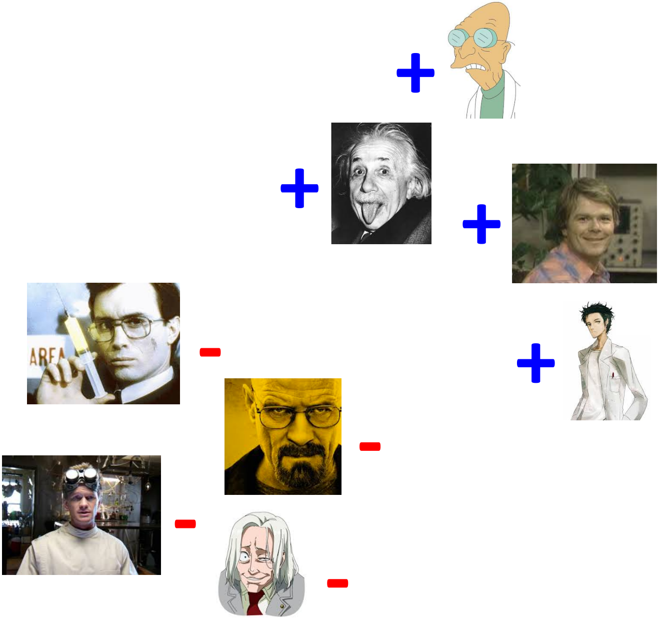

CHOOSE YOUR DESTINY

Learning from binary labels

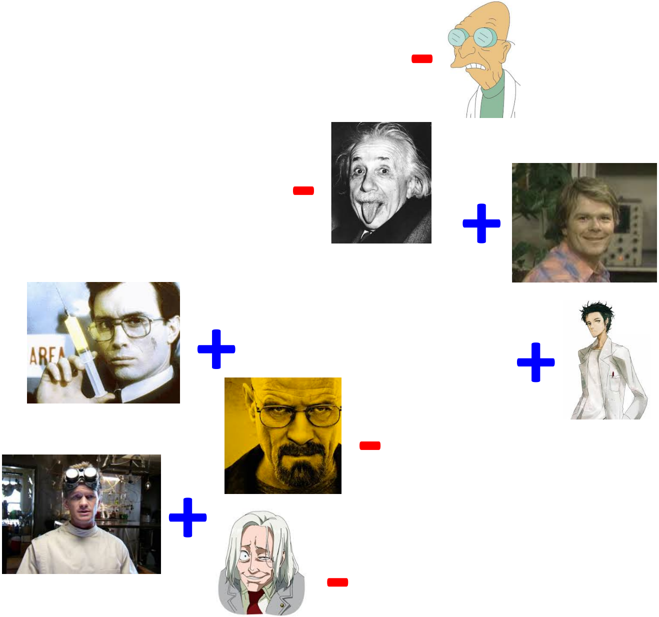


Learning from binary labels

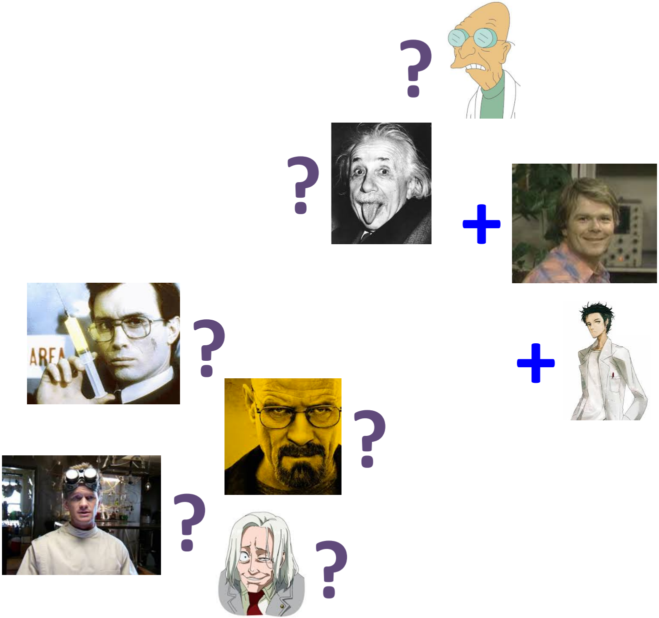
Learning from binary labels



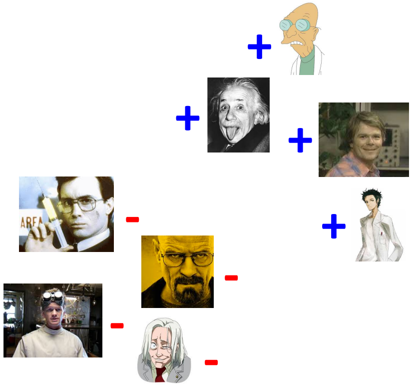
Learning from noisy labels

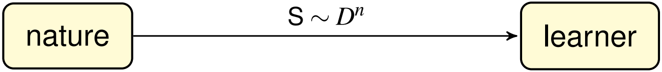


Learning from positive and unlabelled data



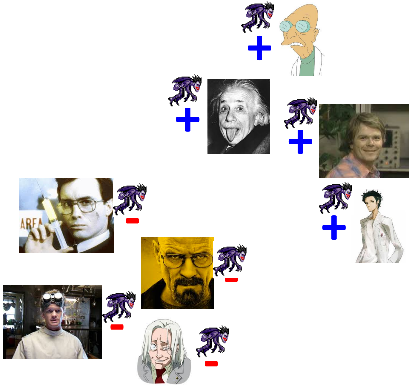
Learning from binary labels





Goal: good classification wrt distribution D

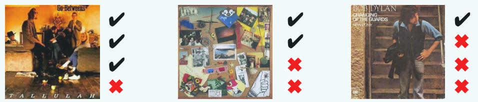
Learning from corrupted labels



Goal: good classification wrt (unobserved) distribution D

Learning from corrupted labels: applications

Learning from noisy annotators



Implicit feedback recommendation

Habitat modelling

Talk summary

Can we learn a good classifier from corrupted samples?

Talk summary

Can we learn a good classifier from corrupted samples?

Yes, if we make assumptions on:

Talk summary

Can we learn a good classifier from corrupted samples?

Yes, if we make assumptions on:

- the corruption process
- (optionally) the true distribution

Solution strategy

What we do:

1. write down the distribution we want to observe samples from
2. compare to distribution we actually observe samples from
3. agree upon measure of performance
4. figure out how to correct for discrepancy between (1) and (2)

Solution strategy

What we do:

1. write down the distribution we want to observe samples from
2. compare to distribution we *actually* observe samples from
3. agree upon measure of performance
4. figure out how to correct for discrepancy between (1) and (2)
5. 

Solution sneak peek

What we suggest:

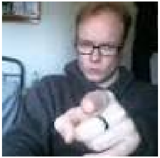
1. treat corrupted labels as if they were uncorrupted
2. train class-probability estimator (e.g. logistic regression)
3. threshold predictions appropriately

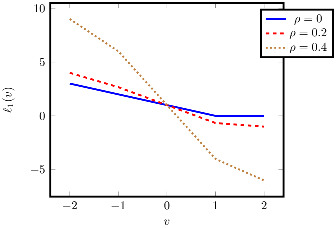
Comment: why not be unhinged?

Precursor to unhinged learning work for label noise

Here, we consider a broader class of corruptions

- some results similar in spirit to “noise immunity”





Binary classification and class-probability estimation

Learning from binary labels: distributions

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

Underlying distribution D over $\mathcal{X} \times \{\pm 1\}$

Constituent components of D:

$$(P(x), Q(x), \pi) = (\mathbb{P}[\mathsf{X} = x | \mathsf{Y} = 1], \mathbb{P}[\mathsf{X} = x | \mathsf{Y} = -1], \mathbb{P}[\mathsf{Y} = 1])$$

Learning from binary labels: distributions

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

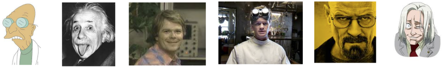
Underlying distribution D over $\mathcal{X} \times \{\pm 1\}$

Constituent components of D:

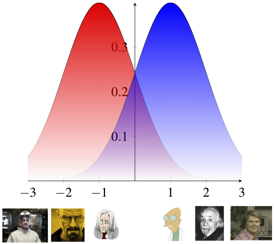
$$(P(x), Q(x), \pi) = (\mathbb{P}[\mathsf{X} = x | \mathsf{Y} = 1], \mathbb{P}[\mathsf{X} = x | \mathsf{Y} = -1], \mathbb{P}[\mathsf{Y} = 1])$$
$$(M(x), \eta(x)) = (\mathbb{P}[\mathsf{X} = x], \mathbb{P}[\mathsf{Y} = 1 | \mathsf{X} = x])$$

Learning from binary labels: example

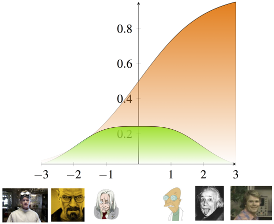
$$\mathcal{X} =$$



$$(P, Q) =$$



$$(M, \eta) =$$



Class-probability estimation

Classification: estimate $\mathrm{sign}(\eta(x) - \frac{1}{2})$

- Bayes-optimal decision boundary
- returned by e.g. SVM with universal kernel

Class-probability estimation

Classification: estimate $\operatorname{sign}(\eta(x) - \frac{1}{2})$

- Bayes-optimal decision boundary
- returned by e.g. SVM with universal kernel

Class-probability estimation: estimate $\phi \circ \eta$ for invertible ϕ

- e.g. logistic regression: $\phi : z \mapsto \frac{1}{1+e^{-z}}$
- e.g. AdaBoost: $\phi : z \mapsto \frac{1}{1+e^{-2z}}$

Class-probability estimation useful when going beyond 0-1 error

Classification performance measures

General classification performance measure expressible as (Narasimhan et al., 2014):

$$\Psi(\mathrm{FNR}^D(f), \mathrm{FPR}^D(f), \pi)$$

where

$$\mathrm{FNR}^D(f) = \mathbb{P}_{\mathsf{X} \sim P}(f(\mathsf{X}) = -1)$$
$$\mathrm{FPR}^D(f) = \mathbb{P}_{\mathsf{X} \sim Q}(f(\mathsf{X}) = 1)$$

Classification performance measures

General classification performance measure expressible as (Narasimhan et al., 2014):

$$\Psi(\mathrm{FNR}^D(f), \mathrm{FPR}^D(f), \pi)$$

where

$$\mathrm{FNR}^D(f) = \mathbb{P}_{\mathsf{X}\sim P}(f(\mathsf{X}) = -1)$$
$$\mathrm{FPR}^D(f) = \mathbb{P}_{\mathsf{X}\sim Q}(f(\mathsf{X}) = 1)$$

Examples:

- 0-1 error $\to \Psi$: $(u,v,p) \to p \cdot u + (1-p) \cdot v$
- Balanced error $\to \Psi$: $(u,v,p) \to (u+v)/2$
- F-score $\to \Psi$: $(u,v,p) \to \frac{2\cdot p\cdot(1-u)}{p+p\cdot(1-u)+(1-p)\cdot v}$

Class-probabilities and classification

Most "reasonable" performance measures Ψ optimised by

$$f^* : x \mapsto \operatorname{sign}(\eta(x) - t)$$

- 0-1 error $\rightarrow t = \frac{1}{2}$
- Balanced error $\rightarrow t = \pi$
- F-score $\rightarrow$ optimal t depends on D
  - (Lipton et al., 2014, Koyejo et al., 2014)

Class-probabilities and classification

Most “reasonable” performance measures Ψ optimised by

$$f^*: x \mapsto \operatorname{sign}(\eta(x) - t)$$

- 0-1 error $\to t = \frac{1}{2}$
- Balanced error $\to t = \pi$
- F-score $\to$ optimal t depends on D
  - (Lipton et al., 2014, Koyejo et al., 2014)

Can optimise such Ψ using a class-probability estimator

Optimising performance measures

Simple algorithm to optimise performance measure Ψ:

- compute class-probability estimates $\hat{\eta}$ (e.g. by logistic regression)
- tune threshold $\hat{t}$ to optimise Ψ on validation set
- return classifier

$$\hat{f} : x \mapsto \mathrm{sign}(\hat{\eta}(x) - \hat{t})$$

Resulting classifier $\hat{f}$ is consistent (Narasimhan et al., 2014)

- surrogate regret bounds also exist (Kotlowski & Dembczynski, 2015)

Assumed corruption model

Learning from binary labels

$$\text{nature} \xrightarrow{\mathsf{S} \sim D^n} \text{learner}$$

Samples from clean distribution $D = (P, Q, \pi)$

Goal: good classification wrt distribution D

Learning from corrupted binary labels

Samples from corrupted distribution $\bar{D} = (\bar{P}, \bar{Q}, \bar{\pi})$

Goal: good classification wrt (unobserved) distribution D

Learning from corrupted binary labels

nature $\xrightarrow{\mathsf{S} \sim D^n}$ corruptor $\xrightarrow{\bar{\mathsf{S}} \sim \bar{D}^n}$ learner

Samples from corrupted distribution $\bar{D} = (\bar{P}, \bar{Q}, \bar{\pi})$, where

$$\bar{P} = (1-\alpha) \cdot P + \alpha \cdot Q$$

$$\bar{Q} = \beta \cdot P + (1-\beta) \cdot Q$$

and $\bar{\pi}$ is arbitrary

- α, β are noise rates
- mutually contaminated distributions (Scott et al., 2013)

Goal: good classification wrt (unobserved) distribution D

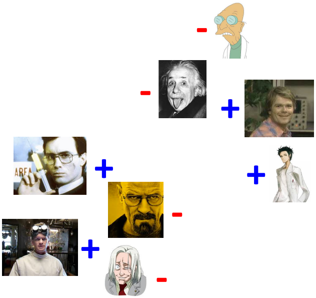
Special case: label noise

Labels flipped with probability ρ_+, ρ_-

$$\bar{\pi} = (1 - \rho_+ - \rho_-) \cdot \pi + \rho_+ + \rho_-$$

$$\alpha = \bar{\pi}^{-1} \cdot (1 - \pi) \cdot \rho_-$$

$$\beta = (1 - \bar{\pi})^{-1} \cdot \pi \cdot \rho_+$$



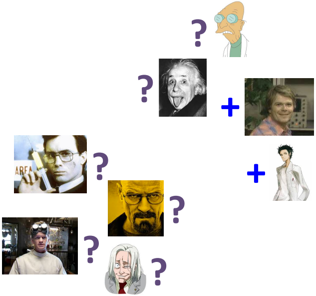
Special case: PU learning

Observe M instead of Q

$\bar{\pi} =$ arbitrary

$\bar{P} = 1 \cdot P + 0 \cdot Q$

$$\bar{Q} = M$$
$$= \pi \cdot P + (1 - \pi) \cdot Q$$



Caution: two faces of PU learning

Can also cast PU learning as specific case of asymmetric label noise (Elkan and Noto, 2008)

- +'ves flipped with censoring probability c
- -'ves flipped with probability 0

"Case-controlled" versus "censoring" versions of the problem

Corrupted class-probabilities

Structure of corrupted class-probabilities underpins analysis

Corrupted class-probabilities

Structure of corrupted class-probabilities underpins analysis

Proposition

For any $D, \bar{D}$*,*

$$\bar{\eta}(x) = \phi_{\alpha,\beta,\pi}(\eta(x))$$

where $\phi_{\alpha,\beta,\pi}$ *is strictly monotone for fixed* α, β, π*.*

Corrupted class-probabilities

Structure of corrupted class-probabilities underpins analysis

Proposition

For any $D, \bar{D}$,

$$\bar{\eta}(x) = \phi_{\alpha,\beta,\pi}(\eta(x))$$

where $\phi_{\alpha,\beta,\pi}$ is strictly monotone for fixed α, β, π.

Follows from Bayes' rule:

$$\frac{\bar{\eta}(x)}{1-\bar{\eta}(x)} = \frac{\bar{\pi}}{1-\bar{\pi}} \cdot \frac{\bar{P}(x)}{\bar{Q}(x)}$$

Corrupted class-probabilities

Structure of corrupted class-probabilities underpins analysis

Proposition

For any $D, \bar{D}$,

$$\bar{\eta}(x) = \phi_{\alpha,\beta,\pi}(\eta(x))$$

where $\phi_{\alpha,\beta,\pi}$ is strictly monotone for fixed α, β, π.

Follows from Bayes' rule:

$$\frac{\bar{\eta}(x)}{1-\bar{\eta}(x)} = \frac{\bar{\pi}}{1-\bar{\pi}} \cdot \frac{\bar{P}(x)}{\bar{Q}(x)} = \frac{\bar{\pi}}{1-\bar{\pi}} \cdot \frac{(1-\alpha)\cdot\frac{P(x)}{Q(x)} + \alpha}{\beta\cdot\frac{P(x)}{Q(x)} + (1-\beta)}.$$

Corrupted class-probabilities: special cases

Label noise

$\bar{\eta}(x) = (1 - \rho_+ - \rho_-) \cdot \eta(x) + \rho_-$

ρ_+, ρ_- unknown

(Natarajan et al., 2013)

PU learning

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

π unknown

(Ward et al., 2009)

Corrupted class-probabilities: comments

Form of $\bar{\eta}$ implies suitable choice of function class

e.g. if $\eta : x \mapsto \frac{1}{1+e^{-s(x)}}$, then neural network is well-specified for $\bar{\eta}$

- if you can't be unhinged, be neurotic

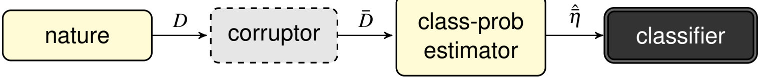
Label noise

$$\bar{\eta}(x) = a \cdot \frac{1}{1+e^{-s(x)}} + b$$

PU learning

$$\bar{\eta}(x) = \frac{1}{a+b \cdot e^{-s(x)}}$$

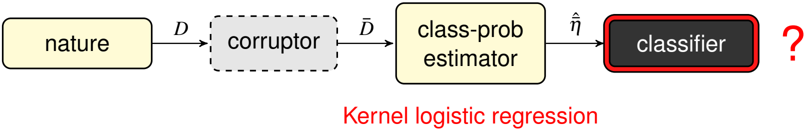
Roadmap



Kernel logistic regression

Roadmap

Exploit monotone relationship between η and $\bar{\eta}$



Classification with noise rates

Recap: class-probabilities and classification

Most "reasonable" performance measures optimised by

$$f^*: x \mapsto \text{sign}(\eta(x) - t)$$

- 0-1 error $\rightarrow t = \frac{1}{2}$
- Balanced error $\rightarrow t = \pi$
- F-score $\rightarrow$ optimal t depends on D
  - (Lipton et al., 2014, Koyejo et al., 2014)

Recap: class-probabilities and classification

Most "reasonable" performance measures optimised by

$$f^* : x \mapsto \mathrm{sign}(\eta(x) - t)$$

- 0-1 error $\rightarrow t = \frac{1}{2}$
- Balanced error $\rightarrow t = \pi$
- F-score $\rightarrow$ optimal t depends on D
  - (Lipton et al., 2014, Koyejo et al., 2014)

We can relate this to thresholding of $\bar{\eta}$!

Corrupted class-probabilities and classification

By monotone relationship,

$$\eta(x) > t \iff \bar{\eta}(x) > \phi_{\alpha,\beta,\pi}(t).$$

Threshold $\bar{\eta}$ at $\phi_{\alpha,\beta,\pi}(t) \rightarrow$ optimal classification on D

Optimal classifiers for 0-1 error: special cases

Label noise

$$\operatorname{sign}\left(\bar{\eta}(x) - \frac{1-\rho_+ + \rho_-}{2}\right)$$

PU learning

$$\operatorname{sign}\left(\bar{\eta}(x) - \frac{\bar{\pi}}{\bar{\pi}+2\cdot(1-\bar{\pi})\cdot\pi}\right)$$

Thresholding at $\frac{1}{2}$ is in general not optimal

- using standard binary classifier will fail
- but changing the threshold overcomes this

Optimising performance measures from corrupted samples

Simple algorithm to optimise performance measure Ψ:

- compute corrupted class-probability estimates $\hat{\bar{\eta}}$ (e.g. by logistic regression)
- tune threshold $\hat{t}$ to optimise Ψ on validation set
- return classifier

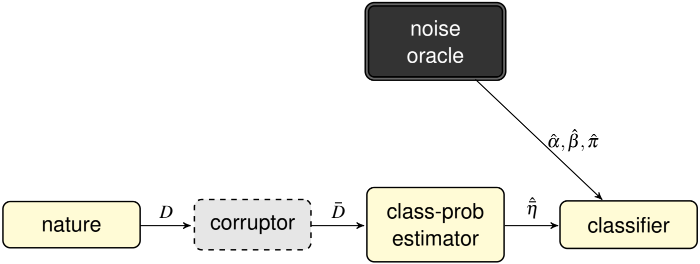
$$\hat{f} : x \mapsto \operatorname{sign}(\hat{\bar{\eta}}(x) - \hat{t})$$

Can derive surrogate regret bounds as before

Story so far

Classification scheme requires:

- $\bar{\eta}$
- t
- α, β, π

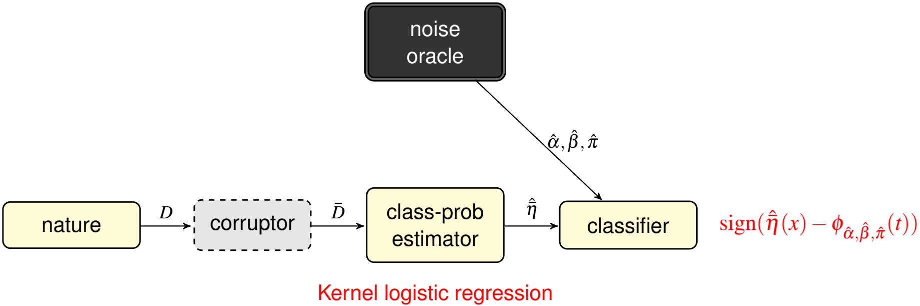


$\mathrm{sign}(\hat{\bar{\eta}}(x) - \phi_{\hat{\alpha},\hat{\beta},\hat{\pi}}(t))$

Story so far

Classification scheme requires:

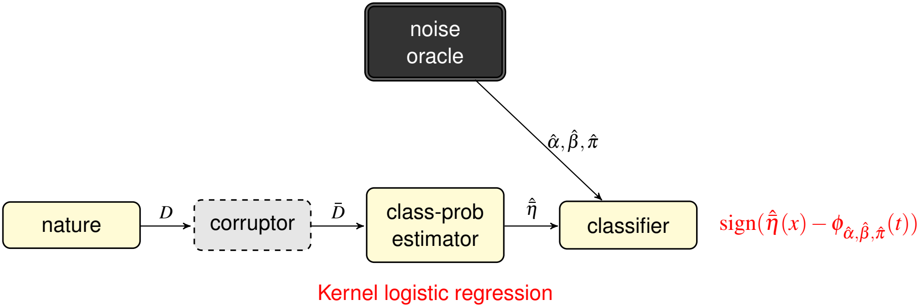
- $\bar{\eta} \rightarrow$ class-probability estimation
- t
- α, β, π



Story so far

Classification scheme requires:

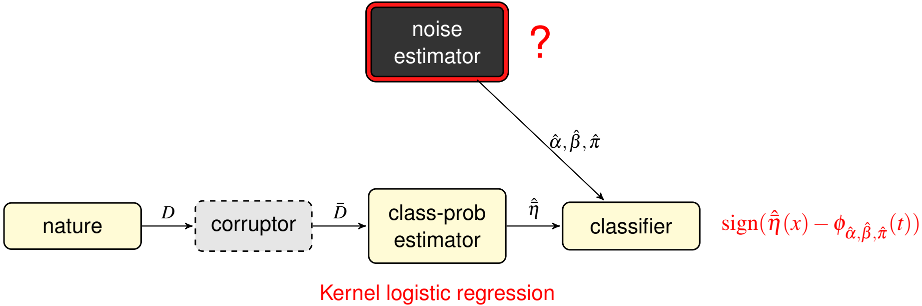
- $\bar{\eta} \rightarrow$ class-probability estimation
- $t \rightarrow$ constant, or using $\bar{\Psi}$
- α, β, π



Story so far

Classification scheme requires:

- $\bar{\eta} \rightarrow$ class-probability estimation
- $t \rightarrow$ constant, or using $\bar{\Psi}$
- $\alpha, \beta, \pi \rightarrow$ **can we estimate these?**



Estimating noise rates: some bad news

π strongly non-identifiable!

- $\bar{\pi}$ allowed to be arbitrary (e.g. PU learning)

α, β non-identifiable without assumptions (Scott et al., 2013)

Estimating noise rates: some bad news

π strongly non-identifiable!

- $\bar{\pi}$ allowed to be arbitrary (e.g. PU learning)

α, β non-identifiable without assumptions (Scott et al., 2013)

Can we estimate α, β under assumptions?

Weak separability assumption

Assume that D is "weakly separable":

$$\min_{x \in \mathcal{X}} \eta(x) = 0$$

$$\max_{x \in \mathcal{X}} \eta(x) = 1$$

- i.e. $\exists$ deterministically +'ve and -'ve instances
- weaker than full separability

Weak separability assumption

Assume that D is "weakly separable":

$$\min_{x \in \mathcal{X}} \eta(x) = 0$$

$$\max_{x \in \mathcal{X}} \eta(x) = 1$$

- i.e. $\exists$ deterministically +'ve and -'ve instances
- weaker than full separability

Assumed range of η constrains observed range of $\bar{\eta}$!

Estimating noise rates

Proposition

Pick any weakly separable D. Then, for any $\bar{D}$,

$$\alpha = \frac{\eta_{\min} \cdot (\eta_{\max} - \bar{\pi})}{\bar{\pi} \cdot (\eta_{\max} - \eta_{\min})} \text{ and } \beta = \frac{(1 - \eta_{\max}) \cdot (\bar{\pi} - \eta_{\min})}{(1 - \bar{\pi}) \cdot (\eta_{\max} - \eta_{\min})}$$

where

$$\eta_{\min} = \min_{x \in \mathcal{X}} \bar{\eta}(x)$$

$$\eta_{\max} = \max_{x \in \mathcal{X}} \bar{\eta}(x)$$

α, β **can be estimated from corrupted data alone**

Estimating noise rates: special cases

Label noise

$$\rho_+ = 1 - \eta_{\max}$$
$$\rho_- = \eta_{\min}$$
$$\pi = \frac{\bar{\pi} - \eta_{\min}}{\eta_{\max} - \eta_{\min}}$$

(Elkan and Noto, 2008),
(Liu and Tao, 2014)

PU learning

$$\alpha = 0$$
$$\beta = \pi = \frac{1 - \eta_{\max}}{\eta_{\max}} \cdot \frac{\bar{\pi}}{1 - \bar{\pi}}$$

In these cases, π can be estimated as well

Estimating noise rates: comments

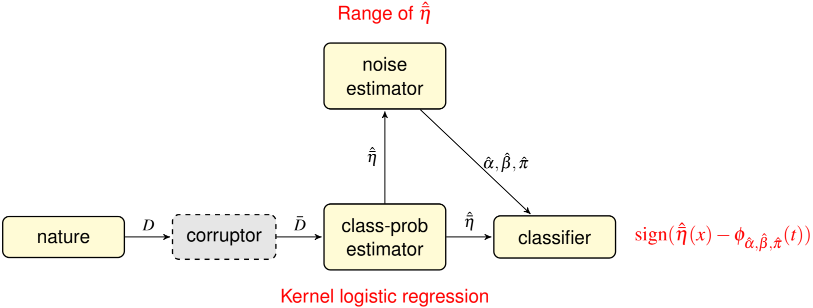
Given estimates $\hat{\hat{\eta}}$, can use plugin versions of $\eta_{\min}, \eta_{\max}$

Estimating order statistics not ideal

- estimates of e.g. π will be sensitive to errors in $\eta_{\min}, \eta_{\max}$
- under stronger assumptions on D, more well-behaved estimators possible, e.g.
$$\rho_+ = \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\eta(\mathsf{X})]$$

Story so far

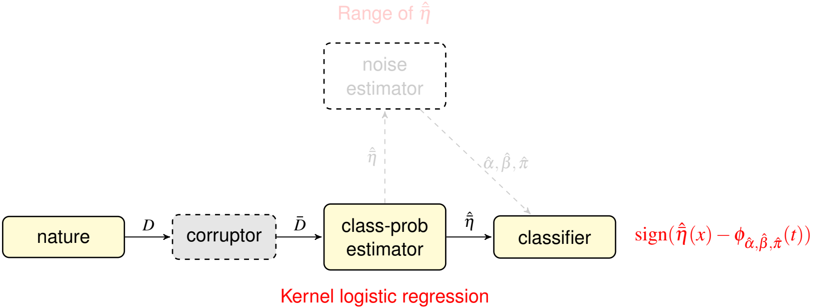
Optimal classification in general requires α, β, π



Story so far

Optimal classification in general requires α, β, π

- **when does $\phi_{\alpha,\beta,\pi}(t)$ not depend on α, β, π?**



Classification without noise rates

Balanced error (BER) of classifier

Balanced error (BER) of a classifier $f\colon \mathcal{X} \to \{\pm 1\}$ is:

$$\mathrm{BER}^D(f) = \frac{\mathrm{FPR}^D(f) + \mathrm{FNR}^D(f)}{2}$$

for false positive and negative rates $\mathrm{FPR}^D(f), \mathrm{FNR}^D(f)$

- average classification performance on each class
- favoured when classes are imbalanced

BER "immunity" under corruption

Proposition (c.f. (Zhang and Lee, 2008))

For any $D, \bar{D}$, and any classifier $f \colon \mathcal{X} \to \{\pm 1\}$,

$$\mathrm{BER}^{\bar{D}}(f) = (1 - \alpha - \beta) \cdot \mathrm{BER}^{D}(f) + \frac{\alpha + \beta}{2}$$

BER “immunity” under corruption

Proposition (c.f. (Zhang and Lee, 2008))

For any $D, \bar{D}$, and any classifier $f \colon \mathcal{X} \to \{\pm 1\}$,

$$\mathrm{BER}^{\bar{D}}(f) = (1 - \alpha - \beta) \cdot \mathrm{BER}^{D}(f) + \frac{\alpha + \beta}{2}$$

Minimising corrupted BER minimises clean BER!

- can ignore corruption process

BER "immunity" under corruption

Proposition (c.f. (Zhang and Lee, 2008))

For any $D, \bar{D}$, and any classifier $f \colon \mathcal{X} \to \{\pm 1\}$,

$$\mathrm{BER}^{\bar{D}}(f) = (1 - \alpha - \beta) \cdot \mathrm{BER}^{D}(f) + \frac{\alpha + \beta}{2}$$

Minimising corrupted BER minimises clean BER!

- can ignore corruption process

Trivially, we also have

$$\mathrm{regret}^{D}_{\mathrm{BER}}(f) = (1 - \alpha - \beta)^{-1} \cdot \mathrm{regret}^{\bar{D}}_{\mathrm{BER}}(f).$$

i.e. **good corrupted BER** $\implies$ **good clean BER**

BER "immunity" & class-probability estimation

Can optimise corrupted BER via class-probability estimation:

- compute corrupted class-probability estimates $\hat{\bar{\eta}}$
- threshold $\hat{\bar{\eta}}$ around corrupted base rate $\bar{\pi}$

BER "immunity" & class-probability estimation

Can optimise corrupted BER via class-probability estimation:

- compute corrupted class-probability estimates $\hat{\bar{\eta}}$
- threshold $\hat{\bar{\eta}}$ around corrupted base rate $\bar{\pi}$

For strongly proper composite ℓ, and scorer $s \colon \mathcal{X} \to \mathbb{R}$,

$$\text{regret}^{\bar{D}}_{\text{BER}}(f_s) \leq C_{\ell,\pi} \cdot \sqrt{\text{regret}^{\bar{D}}_{\ell}(s)}.$$

i.e. can make $\text{regret}^{D}_{\text{BER}}(f) \to 0$ by class-probability estimation

BER "immunity" under corruption: proof

From (Scott et al., 2013),

$$\left[\mathrm{FPR}^{\bar{D}}(f) \quad \mathrm{FNR}^{\bar{D}}(f)\right]^T = \left[\mathrm{FPR}^{D}(f) \quad \mathrm{FNR}^{D}(f)\right]^T \cdot \begin{bmatrix} 1-\beta & -\alpha \\ -\beta & 1-\alpha \end{bmatrix} + \left[\beta \quad \alpha\right]^T,$$

BER "immunity" under corruption: proof

From (Scott et al., 2013),

$$\begin{bmatrix}\mathrm{FPR}^{\bar{D}}(f) & \mathrm{FNR}^{\bar{D}}(f)\end{bmatrix}^T = \begin{bmatrix}\mathrm{FPR}^{D}(f) & \mathrm{FNR}^{D}(f)\end{bmatrix}^T \cdot \begin{bmatrix}1-\beta & -\alpha \\ -\beta & 1-\alpha\end{bmatrix} + \begin{bmatrix}\beta & \alpha\end{bmatrix}^T,$$

and $\begin{bmatrix}1 \\ 1\end{bmatrix}$ is an eigenvector of $\begin{bmatrix}1-\beta & -\alpha \\ -\beta & 1-\alpha\end{bmatrix}$

BER “immunity” under corruption: comments

Results do not rely on weak separability assumption for D

Regret relation does not rely on model being well-specified

- close to best corrupted BER in class $\mathcal{H} \rightarrow$ close to best clean BER in class $\mathcal{H}$

Corollary: AUC "immunity" under corruption

Area under ROC curve (AUC) of a scorer $s \colon \mathcal{X} \to \mathbb{R}$:

$$\mathrm{AUC}^D(s) = \mathop{\mathbb{E}}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[[\![s(\mathsf{X}) > s(\mathsf{X}')]\!] + \frac{1}{2} [\![s(\mathsf{X}) = s(\mathsf{X}')]\!] \right]$$

- probability of random +'ve scoring higher than random -'ve

Corollary: AUC “immunity” under corruption

Area under ROC curve (AUC) of a scorer $s \colon \mathcal{X} \to \mathbb{R}$:

$$\mathrm{AUC}^D(s) = \mathop{\mathbb{E}}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q} \left[[\![s(\mathsf{X}) > s(\mathsf{X}')]\!] + \frac{1}{2} [\![s(\mathsf{X}) = s(\mathsf{X}')]\!] \right]$$

- probability of random +’ve scoring higher than random -’ve

Corollary

For any $D, \bar{D}$, and scorer $s \colon \mathcal{X} \to \mathbb{R}$,

$$\mathrm{AUC}^{\bar{D}}(s) = (1 - \alpha - \beta) \cdot \mathrm{AUC}^D(s) + \frac{\alpha + \beta}{2}$$

Pairwise ranking $\to$ can ignore corruption process

Are other measures "immune"?

BER is only (non-trivial) performance measure for which:

- corrupted risk = affine transform of clean risk
  - because of eigenvector interpretation
- corrupted threshold is independent of α, β, π
  - because of nature of $\phi_{\alpha,\beta,\pi}$

Other performance measures $\rightarrow$ need (one of) α, β, π

Experiments

Experimental setup

Injected label noise on UCI datasets

Estimate corrupted class-probabilities via neural network

- well-specified if D linearly separable:

$$\eta(x) = \sigma(\langle w, x \rangle) \implies \bar{\eta}(x) = a \cdot \sigma(\langle w, x \rangle) + b$$

Evaluate:

- BER performance on clean test set
  - corrupted data used for training and validation
- 0-1 performance on clean test set
- reliability of noise estimates

Experimental results: BER immunity

Generally, low observed degradation in BER

Dataset	**Noise**	1 - **AUC** (%)	**BER** (%)
segment	None	0.00 ± 0.00	0.00 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.0)$	0.00 ± 0.00	0.01 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.2)$	0.02 ± 0.01	0.90 ± 0.08
	$(\rho_+, \rho_-) = (0.2, 0.4)$	0.03 ± 0.01	3.24 ± 0.20
spambase	None	2.49 ± 0.00	6.93 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.0)$	2.67 ± 0.02	7.10 ± 0.03
	$(\rho_+, \rho_-) = (0.1, 0.2)$	3.01 ± 0.03	7.66 ± 0.05
	$(\rho_+, \rho_-) = (0.2, 0.4)$	4.91 ± 0.09	10.52 ± 0.13
mnist	None	0.92 ± 0.00	3.63 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.0)$	0.95 ± 0.01	3.56 ± 0.01
	$(\rho_+, \rho_-) = (0.1, 0.2)$	0.97 ± 0.01	3.63 ± 0.02
	$(\rho_+, \rho_-) = (0.2, 0.4)$	1.17 ± 0.02	4.06 ± 0.03

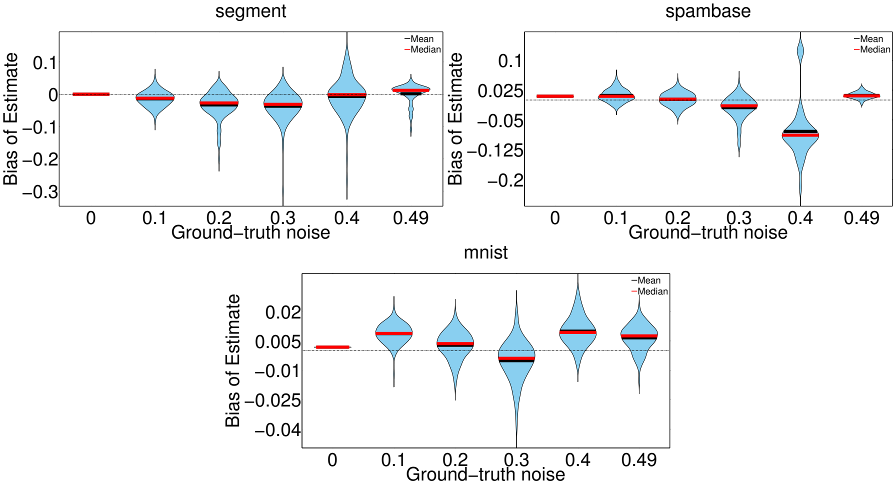
Experimental results: 0-1 error

0-1 error with estimated noise rates $\sim$ using oracle noise rates

Dataset	Noise	$\mathbf{ERR}_{\text{est}}(\%)$	$\mathbf{ERR}_{\text{oracle}}(\%)$
segment	None	0.00 ± 0.00	0.00 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.0)$	0.01 ± 0.00	0.01 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.2)$	0.31 ± 0.05	0.30 ± 0.05
	$(\rho_+, \rho_-) = (0.2, 0.4)$	0.31 ± 0.06	0.27 ± 0.06
spambase	None	6.52 ± 0.00	6.52 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.0)$	6.88 ± 0.03	6.89 ± 0.03
	$(\rho_+, \rho_-) = (0.1, 0.2)$	7.51 ± 0.05	7.48 ± 0.05
	$(\rho_+, \rho_-) = (0.2, 0.4)$	10.82 ± 0.31	10.26 ± 0.12
mnist	None	3.63 ± 0.00	3.63 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.0)$	3.55 ± 0.01	3.55 ± 0.01
	$(\rho_+, \rho_-) = (0.1, 0.2)$	3.62 ± 0.02	3.62 ± 0.02
	$(\rho_+, \rho_-) = (0.2, 0.4)$	4.06 ± 0.03	4.05 ± 0.03

Experimental results: noise rates

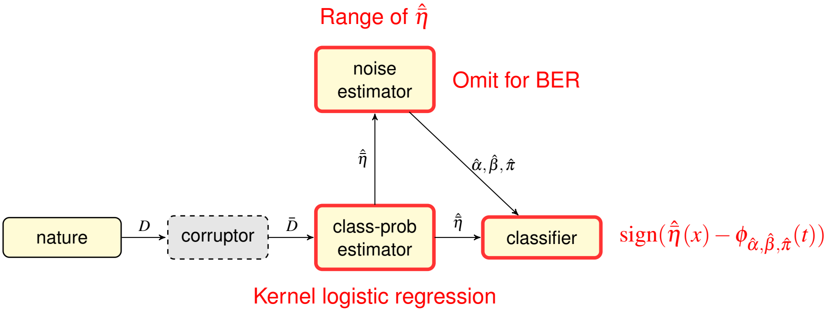
Estimated noise rates are generally reliable



Conclusion

Learning from corrupted binary labels

Monotone relationship $\bar{\eta}(x) = \phi_{\alpha,\beta,\pi}(\eta(x))$ facilitates:



Future work - I

Better noise estimators?

- c.f. (Elkan and Noto, 2008) when D separable

More general noise estimators?

- e.g. learning from partial labels, multi-class corruption, ...
- see formulation of (van Rooyen & Williamson, 2015)

Future work - II

Alternatives to neural network for class-probabilities?

- choice of being unhinged versus neurotic
- for linearly separable D, Isotron (Kalai and Sastry, 2009)

Fusion with “loss transfer” (Natarajan et al., 2013) approach

- better for misspecified models
- assumes noise rates known

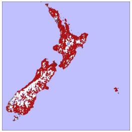
Future work - III

Applications:

Bike crashes

Implicit feedback

Eels' habitats



Thanks!