Learning from Corrupted Binary Labels via Class-Probability Estimation

Aditya Krishna Menon[*] ADITYA.MENON@NICTA.COM.AU
Brendan van Rooyen[†] BRENDAN.VANROOYEN@NICTA.COM.AU
Cheng Soon Ong[*] CHENGSOON.ONG@NICTA.COM.AU
Robert C. Williamson[*] BOB.WILLIAMSON@NICTA.COM.AU
[*] National ICT Australia and The Australian National University, Canberra
[†] The Australian National University and National ICT Australia, Canberra

Abstract

Many supervised learning problems involve learning from samples whose labels are *corrupted* in some way. For example, each label may be flipped with some constant probability (learning with *label noise*), or one may have a pool of unlabelled samples in lieu of negative samples (learning from *positive and unlabelled data*). This paper uses class-probability estimation to study these and other corruption processes belonging to the *mutually contaminated distributions* framework (Scott et al., 2013), with three conclusions. First, one can optimise balanced error and AUC *without knowledge* of the corruption parameters. Second, given estimates of the corruption parameters, one can minimise a range of classification risks. Third, one can estimate corruption parameters via a class-probability estimator (e.g. kernel logistic regression) *trained solely on corrupted data*. Experiments on label noise tasks corroborate our analysis.

1. Learning from corrupted binary labels

In many practical scenarios involving learning from binary labels, one observes samples whose labels are *corrupted* versions of the actual ground truth. For example, in learning from *class-conditional label noise* (*CCN learning*), the labels are flipped with some constant probability (Angluin & Laird, 1988). In *positive and unlabelled learning* (*PU learning*), we have access to some positive samples, but in lieu of negative samples only have a pool of samples whose label is unknown (Denis, 1998). More generally, suppose there is a notional *clean* distribution D over instances and labels. We say a problem involves learning from *corrupted binary labels* if we observe training samples drawn from some *corrupted* distribution D_{corr} such that the observed labels do not represent those we would observe under D.

A fundamental question is whether one can minimise a given performance measure with respect to D, *given access only to samples from* D_{corr}. Intuitively, in general this requires knowledge of the parameters of the corruption process that determines D_{corr}. This yields two further questions: are there measures for which knowledge of these corruption parameters is *unnecessary*, and for other measures, can we *estimate* these parameters?

In this paper, we consider corruption problems belonging to the *mutually contaminated distributions* framework (Scott et al., 2013). We then study the above questions through the lens of class-probability estimation, with three conclusions. First, optimising balanced error (BER) as-is on corrupted data equivalently optimises BER on clean data, and similarly for the area under the ROC curve (AUC). That is, these measures can be optimised *without knowledge of the corruption process parameters*; further, we present evidence that these are essentially the *only* measures with this property. Second, given estimates of the corruption parameters, a range of classification measures can be minimised by thresholding corrupted class-probabilities. Third, under some assumptions, these corruption parameters may be estimated from the range of the corrupted class-probabilities.

For all points above, observe that learning requires *only corrupted data*. Further, corrupted class-probability estimation can be seen as *treating the observed samples as if they were uncorrupted*. Thus, our analysis gives justification (under some assumptions) for this apparent heuristic in problems such as CCN and PU learning.

While some of our results are known for the special cases of CCN and PU learning, our interest is in determining to what extent they generalise to other label corruption problems. This is a step towards a unified treatment of these problems. We now fix notation and formalise the problem.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

2. Background and problem setup

Fix an instance space $\mathcal{X}$. We denote by D some distribution over $\mathcal{X} \times \{\pm 1\}$, with $(\mathsf{X}, \mathsf{Y}) \sim D$ a pair of random variables. Any D may be expressed via the *class-conditional distributions* $(P, Q) = (\mathbb{P}(\mathsf{X} \mid \mathsf{Y} = 1), \mathbb{P}(\mathsf{X} \mid \mathsf{Y} = -1))$ and *base rate* $\pi = \mathbb{P}(\mathsf{Y} = 1)$, or equivalently via *marginal distribution* $M = \mathbb{P}(\mathsf{X})$ and *class-probability function* $\eta \colon x \mapsto \mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x)$. When referring to these constituent distributions, we write D as $D_{P,Q,\pi}$ or $D_{M,\eta}$.

2.1. Classifiers, scorers, and risks

A *classifier* is any function $f \colon \mathcal{X} \to \{\pm 1\}$. A *scorer* is any function $s \colon \mathcal{X} \to \mathbb{R}$. Many learning methods (e.g. SVMs) output a scorer, from which a classifier is formed by thresholding about some $t \in \mathbb{R}$. We denote the resulting classifier by $\text{thresh}(s, t) \colon x \mapsto \text{sign}(s(x) - t)$.

The *false positive* and *false negative rates* of a classifier f are denoted $\text{FPR}^D(f), \text{FNR}^D(f)$, and are defined by $\underset{\mathsf{X} \sim Q}{\mathbb{P}}(f(\mathsf{X}) = 1)$ and $\underset{\mathsf{X} \sim P}{\mathbb{P}}(f(\mathsf{X}) = -1)$ respectively.

Given a function $\Psi \colon [0,1]^3 \to [0,1]$, a *classification performance measure* $\text{Class}_\Psi^D \colon \{\pm 1\}^{\mathcal{X}} \to [0,1]$ assesses the performance of a classifier f via (Narasimhan et al., 2014)

$$\text{Class}_\Psi^D(f) = \Psi(\text{FPR}^D(f), \text{FNR}^D(f), \pi).$$

A canonical example is the *misclassification error*, where $\Psi \colon (u, v, p) \mapsto p \cdot v + (1 - p) \cdot u$. Given a scorer s, we use $\text{Class}_\Psi^D(s; t)$ to refer to $\text{Class}_\Psi^D(\text{thresh}(s, t))$.

The Ψ*-classification regret* of a classifier $f \colon \mathcal{X} \to \{\pm 1\}$ is

$$\text{regret}_\Psi^D(f) = \text{Class}_\Psi^D(f) - \inf_{g \colon \mathcal{X} \to \{\pm 1\}} \text{Class}_\Psi^D(g).$$

A *loss* is any function $\ell \colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$. Given a distribution D, the ℓ*-risk* of a scorer s is defined as

$$\mathbb{L}_\ell^D(s) = \underset{(\mathsf{X}, \mathsf{Y}) \sim D}{\mathbb{E}} \left[\ell(\mathsf{Y}, s(\mathsf{X})) \right]. \tag{1}$$

The ℓ-regret of a scorer, regret_ℓ^D, is as per the Ψ-regret.

We say ℓ is *strictly proper composite* (Reid & Williamson, 2010) if $\text{argmin}_s \mathbb{L}_\ell^D(s)$ is some strictly monotone transformation ψ of η, i.e. we can recover class-probabilities from the optimal prediction via the *link function* ψ. We call *class-probability estimation* (*CPE*) the task of minimising Equation 1 for some strictly proper composite ℓ.

The conditional Bayes-risk of a strictly proper composite ℓ is $L_\ell \colon \eta \mapsto \eta \ell_1(\psi(\eta)) + (1 - \eta) \ell_{-1}(\psi(\eta))$. We call ℓ *strongly proper composite* with modulus λ if L_ℓ is λ-strongly concave (Agarwal, 2014). Canonical examples of such losses are the logistic and exponential loss, as used in logistic regression and AdaBoost respectively.

Quantity	Clean	Corrupted
Joint distribution	D	$\text{Corr}(D, \alpha, \beta, \pi_{\text{corr}})$ or D_{corr}
Class-conditionals	P, Q	$P_{\text{corr}}, Q_{\text{corr}}$
Base rate	π	π_{corr}
Class-probability	η	η_{corr}
Ψ-optimal threshold	t_Ψ^D	$t_{\text{corr},\Psi}^D$

Table 1. Common quantities on clean and corrupted distributions.

2.2. Learning from contaminated distributions

Suppose $D_{P,Q,\pi}$ is some "clean" distribution where performance will be assessed. (We do *not* assume that D is separable.) In MC learning (Scott et al., 2013), we observe samples from some *corrupted* distribution $\text{Corr}(D, \alpha, \beta, \pi_{\text{corr}})$ over $\mathcal{X} \times \{\pm 1\}$, for some *unknown* noise parameters $\alpha, \beta \in [0, 1]$ with $\alpha + \beta < 1$; where the parameters are clear from context, we occasionally refer to the corrupted distribution as D_{corr}. The corrupted class-conditional distributions $P_{\text{corr}}, Q_{\text{corr}}$ are

$$\begin{aligned} P_{\text{corr}} &= (1 - \alpha) \cdot P + \alpha \cdot Q \\ Q_{\text{corr}} &= \beta \cdot P + (1 - \beta) \cdot Q, \end{aligned} \tag{2}$$

and the corrupted base rate π_{corr} in general has *no relation* to the clean base rate π. (If $\alpha + \beta = 1$, then $P_{\text{corr}} = Q_{\text{corr}}$, making learning impossible, whereas if $\alpha + \beta > 1$, we can swap $P_{\text{corr}}, Q_{\text{corr}}$.) Table 1 summarises common quantities on the clean and corrupted distributions.

From (2), we see that none of $P_{\text{corr}}, Q_{\text{corr}}$ or π_{corr} contain any information about π in general. Thus, estimating π from $\text{Corr}(D, \alpha, \beta, \pi_{\text{corr}})$ is *impossible* in general. The parameters α, β are also non-identifiable, but can be estimated under some assumptions on D (Scott et al., 2013).

2.3. Special cases of MC learning

Two special cases of MC learning are notable. In learning from *class-conditional label noise* (*CCN learning*) (Angluin & Laird, 1988), positive samples have labels flipped with probability ρ_+, and negative samples with probability ρ_-. This can be shown to reduce to MC learning with

$$\alpha = \pi_{\text{corr}}^{-1} \cdot (1 - \pi) \cdot \rho_- \,, \beta = (1 - \pi_{\text{corr}})^{-1} \cdot \pi \cdot \rho_+, \tag{3}$$

and the corrupted base rate $\pi_{\text{corr}} = (1 - \rho_+) \cdot \pi + \rho_- \cdot (1 - \pi)$. (See Appendix C for details.)

In learning from *positive and unlabelled data* (*PU learning*) (Denis, 1998), one has access to *unlabelled* samples in lieu of negative samples. There are two subtly different settings: in the *case-controlled* setting (Ward et al., 2009), the unlabelled samples are drawn from the marginal distribution M, corresponding to MC learning with $\alpha = 0, \beta = \pi$, and π_{corr} arbitrary. In the *censoring* setting (Elkan & Noto, 2008), observations are drawn from D followed by a *label censoring* procedure. This is in fact a special of CCN (and hence MC) learning with $\rho_- = 0$.

3. BER and AUC are immune to corruption

We first show that optimising balanced error and AUC on corrupted data is *equivalent* to doing so on clean data. Thus, with a suitably rich function class, one can optimise balanced error and AUC from corrupted data *without* knowledge of the corruption process parameters.

3.1. BER minimisation is immune to label corruption

The *balanced error* (*BER*) (Brodersen et al., 2010) of a classifier is simply the mean of the per-class error rates,

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

This is a popular measure in imbalanced learning problems (Cheng et al., 2002; Guyon et al., 2004) as it penalises sacrificing accuracy on the rare class in favour of accuracy on the dominant class. The negation of the BER is also known as the AM (arithmetic mean) metric (Menon et al., 2013).

The BER-optimal classifier thresholds the class-probability function at the base rate (Menon et al., 2013), so that:

$$\underset{f\colon \mathcal{X}\to\{\pm1\}}{\text{argmin}}\ \text{BER}^D(f) = \text{thresh}(\eta, \pi) \tag{4}$$

$$\underset{f\colon \mathcal{X}\to\{\pm1\}}{\text{argmin}}\ \text{BER}^{D_{\text{corr}}}(f) = \text{thresh}(\eta_{\text{corr}}, \pi_{\text{corr}}), \tag{5}$$

where η_{corr} denotes the corrupted class-probability function. As Equation 4 depends on π, it may appear that one must know π to minimise the clean BER from corrupted data. Surprisingly, the BER-optimal classifiers in Equations 4 and 5 *coincide*. This is because of the following relationship between the clean and corrupted BER.

Proposition 1. *Pick any D and $\text{Corr}(D, \alpha, \beta, \pi_{\text{corr}})$. Then, for any classifier $f : \mathcal{X} \to \{\pm1\}$,*

$$\text{BER}^{D_{\text{corr}}}(f) = (1 - \alpha - \beta) \cdot \text{BER}^D(f) + \frac{\alpha + \beta}{2}, \tag{6}$$

and so the minimisers of the two are identical.

Thus, when BER is the desired performance metric, we *do not need to estimate the noise parameters, or the clean base rate*: we can (approximately) optimise the BER on the corrupted data using estimates $\hat{\eta}_{\text{corr}}, \hat{\pi}_{\text{corr}}$, from which we build a classifier $\text{thresh}(\hat{\eta}_{\text{corr}}, \hat{\pi}_{\text{corr}})$. Observe that this approach effectively *treats the corrupted samples as if they were clean*, e.g. in a PU learning problem, we treat the unlabelled samples as negative, and perform CPE as usual.

With a suitably rich function class, surrogate regret bounds quantify the efficacy of thresholding approximate class-probability estimates. Suppose we know the corrupted base rate[1] π_{corr}, and suppose that s is a scorer with low ℓ-regret on the *corrupted* distribution for some proper composite loss ℓ with link ψ i.e. $\psi^{-1}(s)$ is a good estimate of η_{corr}. Then, the classifier resulting from thresholding this scorer will attain low BER on the *clean* distribution D.

Proposition 2. *Pick any D and $\text{Corr}(D, \alpha, \beta, \pi_{\text{corr}})$. Let ℓ be a strongly proper composite loss with modulus λ and link function ψ. Then, for any scorer $s\colon \mathcal{X} \to \mathbb{R}$,*

$$\text{regret}^D_{\text{BER}}(f) \le \frac{C(\pi_{\text{corr}})}{1 - \alpha - \beta} \cdot \sqrt{\frac{2}{\lambda}} \cdot \sqrt{\text{regret}^{D_{\text{corr}}}_{\ell}(s)},$$

where $f = \text{thresh}(s, \psi(\pi_{\text{corr}}))$ and $C(\pi_{\text{corr}}) = (2 \cdot \pi_{\text{corr}} \cdot (1 - \pi_{\text{corr}}))^{-1}$.

Thus, good estimates of the *corrupted* class-probabilities let us minimise the *clean* BER. Of course, learning from corrupted data comes at a price: compared to the regret bound obtained if we could minimise ℓ on the *clean* distribution D, we have an extra penalty of $(1 - \alpha - \beta)^{-1}$. This matches our intuition that for high-noise regimes (i.e. $\alpha + \beta \approx 1$), we need more corrupted samples to learn effectively with respect to the clean distribution; confer van Rooyen & Williamson (2015) for lower and upper bounds on sample complexity for a range of corruption problems.

3.2. AUC maximisation is immune to label corruption

Another popular performance measure in imbalanced learning scenarios is the *area under the ROC curve* (*AUC*). The AUC of a scorer, $\text{AUC}^D(s)$, is the probability of a random positive instance scoring higher than a random negative instance (Agarwal et al., 2005):

$$\underset{\mathsf{X}\sim P, \mathsf{X}'\sim Q}{\mathbb{E}}\left[[\![s(\mathsf{X}) > s(\mathsf{X}')]\!] + \frac{1}{2} [\![s(\mathsf{X}) = s(\mathsf{X}')]\!] \right].$$

We have a counterpart to Proposition 1 by rewriting the AUC as an average of BER across a range of thresholds ((Flach et al., 2011); see Appendix A.5):

$$\text{AUC}^D(s) = \frac{3}{2} - 2 \cdot \mathbb{E}_{\mathsf{X}\sim P}[\text{BER}^D(s; s(\mathsf{X}))]. \tag{7}$$

Corollary 3. *Pick any $D_{P,Q,\pi}$ and $\text{Corr}(D, \alpha, \beta, \pi_{\text{corr}})$. Then, for any scorer $s\colon \mathcal{X} \to \mathbb{R}$,*

$$\text{AUC}^{D_{\text{corr}}}(s) = (1 - \alpha - \beta) \cdot \text{AUC}^D(s) + \frac{\alpha + \beta}{2}. \tag{8}$$

Thus, like the BER, optimising the AUC with respect to the corrupted distribution optimises the AUC with respect

[1] Surrogate regret bounds may also be derived for an empirically chosen threshold (Kotłowski & Dembczyński, 2015).

to the clean one. Further, via recent bounds on the AUC-regret (Agarwal, 2014), we can show that a good corrupted class-probability estimator will have good clean AUC.

Corollary 4. *Pick any D and $\mathrm{Corr}(D, \alpha, \beta, \pi_{\mathrm{corr}})$. Let ℓ be a strongly proper composite loss with modulus λ. Then, for every scorer $s \colon \mathcal{X} \to \mathbb{R}$,*

$$\mathrm{regret}^D_{\mathrm{AUC}}(s) \leq \frac{C(\pi_{\mathrm{corr}})}{1-\alpha-\beta} \cdot \sqrt{\frac{2}{\lambda}} \cdot \sqrt{\mathrm{regret}^{D_{\mathrm{corr}}}_{\ell}(s)},$$

where $C(\pi_{\mathrm{corr}}) = (\pi_{\mathrm{corr}} \cdot (1-\pi_{\mathrm{corr}}))^{-1}$.

What is special about the BER (and consequently the AUC) that lets us avoid estimation of the corruption parameters? To answer this, we more carefully study the structure of η_{corr} to understand why Equation 4 and 5 coincide, and whether any *other* measures have this property.

Relation to existing work For the special case of CCN learning, Proposition 1 was shown in Blum & Mitchell (1998, Section 5), and for case-controlled PU learning, in (Lee & Liu, 2003; Zhang & Lee, 2008). None of these works established surrogate regret bounds.

4. Corrupted and clean class-probabilities

The equivalence between a specific thresholding of the clean and corrupted class-probabilities (Equations 4 and 5) hints at a relationship between the two functions. We now make this relationship explicit.

Proposition 5. *For any $D_{M,\eta}$ and $\mathrm{Corr}(D, \alpha, \beta, \pi_{\mathrm{corr}})$,*

$$(\forall x \in \mathcal{X})\, \eta_{\mathrm{corr}}(x) = T(\alpha, \beta, \pi, \pi_{\mathrm{corr}}, \eta(x)) \qquad (9)$$

where, for $\phi \colon z \mapsto \frac{z}{1+z}$, $T(\alpha, \beta, \pi, \pi_{\mathrm{corr}}, t)$ is given by

$$\phi\left(\frac{\pi_{\mathrm{corr}}}{1-\pi_{\mathrm{corr}}} \cdot \frac{(1-\alpha)\cdot\frac{1-\pi}{\pi}\cdot\frac{t}{1-t}+\alpha}{\beta\cdot\frac{1-\pi}{\pi}\cdot\frac{t}{1-t}+(1-\beta)}\right). \qquad (10)$$

It is evident that η_{corr} is a strictly monotone increasing transform of η. This is useful to study classifiers based on thresholding η, as per Equation 4. Suppose we want a classifier of the form $\mathrm{thresh}(\eta, t)$. The structure of η_{corr} means that this is equivalent to a *corrupted classifier* $\mathrm{thresh}(\eta_{\mathrm{corr}}, T(\alpha, \beta, \pi, \pi_{\mathrm{corr}}, t))$, where the function T (as per Equation 10) tells us how to modify the threshold t on corrupted data. We now make this precise.

Corollary 6. *Pick any $D_{M,\eta}$ and $\mathrm{Corr}(D, \alpha, \beta, \pi_{\mathrm{corr}})$. Then, $\forall x \in \mathcal{X}$ and $\forall t \in [0, 1]$,*

$$\eta(x) > t \iff \eta_{\mathrm{corr}}(x) > T(\alpha, \beta, \pi, \pi_{\mathrm{corr}}, t)$$

where T is as defined in Equation 10.

By viewing the minimisation of a general classification measure in light of the above, we now return to the issue of why BER can avoid estimating corruption parameters.

Relation to existing work In PU learning, Proposition 5 has been shown in both the case-controlled (McCullagh & Nelder, 1989, pg. 113), (Phillips et al., 2009; Ward et al., 2009) and censoring settings (Elkan & Noto, 2008, Lemma 1). In CCN learning, Proposition 5 is used in Natarajan et al. (2013, Lemma 7). Corollary 6 is implicit in Scott et al. (2013, Proposition 1), but the explicit form for the corrupted threshold is useful for subsequent analysis.

5. Classification from corrupted data

Consider the problem of optimising a classification measure $\mathrm{Class}^D_\Psi(f)$ for some $\Psi \colon [0,1]^3 \to [0,1]$. For a range of Ψ, the optimal classifier is $f = \mathrm{thresh}(\eta, t^D_\Psi)$ (Koyejo et al., 2014; Narasimhan et al., 2014), for some *optimal threshold* t^D_Ψ. For example, by Equation 4, the BER-optimal classifier thresholds class-probabilities at the base rate; other examples of such Ψ are those corresponding to misclassification error, and the F-score. But by Corollary 6, $\mathrm{thresh}(\eta, t^D_\Psi) = \mathrm{thresh}(\eta_{\mathrm{corr}}, t^D_{\mathrm{corr},\Psi})$, where

$$t^D_{\mathrm{corr},\Psi} = T(\alpha, \beta, \pi, \pi_{\mathrm{corr}}, t^D_\Psi) \qquad (11)$$

is the corresponding *optimal corrupted threshold*. Based on this, we now look at two approaches to minimising $\mathrm{Class}^D_\Psi(f)$. For the purposes of description, we shall assume that α, β, π are known (or can be estimated). We then study the practically important question of when these approaches can be applied *without* knowledge of α, β, π.

5.1. Classification when t^D_Ψ is known

Suppose that t^D_Ψ has some closed-form expression; for example, for misclassification risk, $t^D_\Psi = 1/2$. Then, there is a simple strategy for minimising Class^D_Ψ: compute estimates $\hat{\eta}_{\mathrm{corr}}$ of the corrupted class probabilities, and threshold them via $t^D_{\mathrm{corr},\Psi}$ computed from Equation 11. Standard cost-sensitive regret bounds may then be invoked. For concreteness, consider the misclassification risk, where plugging in $t^D_\Psi = 1/2$ into Equation 10 gives

$$t^D_{\mathrm{corr},\Psi} = \phi\left(\frac{\pi_{\mathrm{corr}}}{1-\pi_{\mathrm{corr}}} \cdot \frac{(1-\alpha)\cdot\frac{1-\pi}{\pi}+\alpha}{\beta\cdot\frac{1-\pi}{\pi}+(1-\beta)}\right), \qquad (12)$$

for $\phi \colon z \mapsto z/(1+z)$. We have the following.

Proposition 7. *Pick any D and $\mathrm{Corr}(D, \alpha, \beta, \pi_{\mathrm{corr}})$. Let ℓ be a strongly proper composite loss with modulus λ and link function ψ. Then, for any scorer $s \colon \mathcal{X} \to \mathbb{R}$,*

$$\mathrm{regret}^D_{\mathrm{ERR}}(f) \leq \gamma \cdot \sqrt{\frac{2}{\lambda}} \cdot \sqrt{\mathrm{regret}^{D_{\mathrm{corr}}}_{\ell}(s)},$$

where $f = \mathrm{thresh}(s, \psi(t^D_{\mathrm{corr},\Psi}))$, $t^D_{\mathrm{corr},\Psi}$ is as per Equation 12, and γ is a constant depending on $\alpha, \beta, \pi, \pi_{\mathrm{corr}}$.

5.2. Classification when t^D_Ψ is unknown

For some Ψ, t^D_Ψ does not have a simple closed-form expression, rendering the above approach inapplicable[2]. For example, the optimal threshold for F-score does not have a closed form (Koyejo et al., 2014), and is typically computed by a grid search. In such cases, we can make progress by re-expressing $\mathrm{Class}^D_\Psi(f)$ as an equivalent measure $\mathrm{Class}^{D_{\mathrm{corr}}}_{\Psi_{\mathrm{corr}}}(f)$ on the corrupted distribution, and then tune thresholds on $\hat{\eta}_{\mathrm{corr}}$ to optimise the latter. Here, Ψ_{corr} is *not* the same as Ψ in general, but rather is the result of re-expressing the clean false positive and negative rates in terms of the corrupted ones, as per Scott et al. (2013):

$$\begin{bmatrix} \mathrm{FPR}^D(f) \\ \mathrm{FNR}^D(f) \end{bmatrix} = \begin{bmatrix} 1-\beta & -\beta \\ -\alpha & 1-\alpha \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{FPR}^{D_{\mathrm{corr}}}(f)-\beta \\ \mathrm{FNR}^{D_{\mathrm{corr}}}(f)-\alpha \end{bmatrix}.$$

Thus, for example, for $\Psi \colon (u,v,p) = u$, we would have $\Psi_{\mathrm{corr}} \colon (u,v,p) \mapsto \frac{(1-\alpha)}{1-\alpha-\beta} \cdot (u-\beta) + \frac{\beta}{1-\alpha-\beta} \cdot (v-\alpha)$.

In general, both of the above approaches will require knowledge of α, β, π. For the approach in §5.1, $t^D_{\mathrm{corr},\Psi}$ may clearly depend on these parameters. For the approach in §5.2, the corrupted measure Ψ_{corr} may similarly depend on these parameters, as in the example of $\Psi(u,v,p) = u$. We now provide strong evidence that for both approaches, BER is essentially the only measure that obviates the need for estimation of the corruption parameters.

5.3. BER threshold is uniquely corruption-immune

One way of interpreting the immunity of BER is that the corrupted threshold function (Equation 10) sheds all dependence on α, β, π when instantiated with a threshold of π:

$$(\forall \alpha, \beta, \pi, \pi_{\mathrm{corr}})\, T(\alpha, \beta, \pi, \pi_{\mathrm{corr}}, \pi) = \pi_{\mathrm{corr}}.$$

Is $t \colon \pi \mapsto \pi$ the *only* threshold whose corrupted counterpart does not depend on α, β, π? As stated, the answer is trivially "no"; we can set the corrupted threshold to be any function of π_{corr}, and invert Equation 10 to get an equivalent threshold t for η. However, this t will depend on α, β, and it is unreasonable for the performance measure to depend on the exogenous corruption process. Refining the question to ask whether π is the only threshold independent of α, β such that T is independent of α, β, π, the answer is "yes". We can formalise this as follows.

Proposition 8. *Pick any* Ψ*. Then, there exists* $F \colon (0,1) \to (0,1)$ *such that* Class^D_Ψ *has a unique minimiser of the form* $x \mapsto \mathrm{sign}(\eta_{\mathrm{corr}}(x) - F(\pi_{\mathrm{corr}}))$ *for every* D, D_{corr} *if and only if* $\Psi \colon (u,v,p) \mapsto (u+v)/2$ *corresponds to the BER.*

Thus, for measures other than BER which are uniquely optimised by thresholding η[3], we must know one of α, β, π to find the optimal corrupted threshold. But as it is impossible in general to estimate π, it will similarly be impossible to compute this threshold to optimally classify.

While this seems disheartening, two qualifications are in order. First, in the special cases of CCN and PU learning, π *can* be estimated (see §6.2). Second, Proposition 8 is concerned with immunity to *arbitrary* corruption, where α, β, π may be chosen independently. But in special cases where these parameters are tied, other measures may have a threshold independent of these parameters; e.g., in CCN learning, the misclassification error threshold is (Natarajan et al., 2013, Theorem 9)

$$t^D_{\mathrm{corr},\Psi} = \frac{1-\rho_+ + \rho_-}{2}. \tag{13}$$

So, when $\rho_+ = \rho_-$, $t^D_{\mathrm{corr},\Psi} = \frac{1}{2}$; i.e. *for symmetric label noise, we do not need to know the noise parameters.* Appendix G discusses this issue further.

5.4. BER is uniquely affinely related

Another way of interpreting the immunity of the BER is that, for $\Psi \colon (u,v,p) \mapsto (u+v)/2$, the corresponding corrupted performance measure Ψ_{corr} is simply an affine transformation of Ψ (Proposition 6). Thus, for this measure, $\mathrm{Class}^{D_{\mathrm{corr}}}_{\Psi_{\mathrm{corr}}}$ may be minimised without knowing α, β, π. More generally, we seek Ψ for which there exist f, g such that the corresponding Ψ_{corr} is expressible as

$$\Psi_{\mathrm{corr}}(u,v,p) = f(\alpha,\beta,\pi) \cdot \Psi(u,v,p) + g(\alpha,\beta,\pi). \tag{14}$$

While we do not have a general characterisation of all Ψ satisfying Equation 14, we can show that BER is the only *linear* combination of the false positive and negative rates with an affine relationship between Ψ and Ψ_{corr}. The key is that $(1,1)$ is the only noise-agnostic eigenvector of the row-stochastic matrix implicit in Equation 2.

Proposition 9. *The set of* Ψ *of the form* $\Psi \colon (u,v,p) \mapsto w_1(p) \cdot u + w_2(p) \cdot v$ *where, for every* D, D_{corr}, f, $\mathrm{Class}^D_\Psi(f)$ *is an affine transformation of* $\mathrm{Class}^{D_{\mathrm{corr}}}_\Psi(f)$ *is* $\{\Psi \colon (u,v,p) \mapsto w(p) \cdot (u+v) \mid w \colon [0,1] \to \mathbb{R}\}$*, corresponding to a scaled version of the BER.*

In special cases, other Ψ may have such an affine relationship. In the censoring version of PU learning, Lee & Liu (2003) gave one example, the product of Precision and Recall; Appendix G discusses others.

Relation to existing work Scott (2015, Corollary 1) established an analogue of Proposition 7 for CCN learning.

[2] Recent work has shown how for F-score, we can employ a *series* of thresholds (Parambath et al., 2014); studying this approach in our framework would be of interest.

[3] This rules out degenerate cases such $\Psi \equiv 0$, where there is a *set* of optimal classifiers (i.e. all of them).

Scott et al. (2013) used the approach in §5.2 of rewriting clean in terms of corrupted rates to minimise the minimax risk on D. We are unaware of prior study on conditions where estimation of corruption parameters is unnecessary.

6. Estimating noise rates from corrupted data

We have seen that given estimates of α, β, π, a range of classification measures can be minimised by corrupted class-probability estimation. We now show that under mild assumptions on D, corrupted class-probability estimation lets us estimate α, β, and in special cases, π as well.

6.1. Estimating α, β from η_{corr}

An interesting consequence of Equation 9 is that the range of η_{corr} will be a strict subset of $[0, 1]$ in general. This is because each instance has a nonzero chance of being assigned to either the positive or negative corrupted class; thus, one cannot be *sure* as to its corrupted label.

The precise range of η_{corr} depends on $\alpha, \beta, \pi_{\mathrm{corr}}$, and the range of η. We can thus compute α, β from the range of η_{corr}, with the proviso that we impose the following *weak separability* assumption on D:

$$\inf_{x \in \mathcal{X}} \eta(x) = 0 \text{ and } \sup_{x \in \mathcal{X}} \eta(x) = 1. \tag{15}$$

This does not require D to be separable (i.e. $(\forall x)\, \eta(x) \in \{0, 1\}$), but instead stipulates that *some* instance is "perfectly positive", and another "perfectly negative". This assumption is equivalent to the "mutually irreducible" condition of Scott et al. (2013) (see Appendix H).

Equipped with this assumption, and defining

$$\eta_{\min} = \inf_{x \in \mathcal{X}} \eta_{\mathrm{corr}}(x) \text{ and } \eta_{\max} = \sup_{x \in \mathcal{X}} \eta_{\mathrm{corr}}(x),$$

we can compute the corruption parameters as follows.

Proposition 10. *Pick any $D_{M,\eta}$ satisfying Equation 15. Then, for any $\mathrm{Corr}(D, \alpha, \beta, \pi_{\mathrm{corr}})$,*

$$\begin{aligned} \alpha &= \frac{\eta_{\min} \cdot (\eta_{\max} - \pi_{\mathrm{corr}})}{\pi_{\mathrm{corr}} \cdot (\eta_{\max} - \eta_{\min})} \\ \beta &= \frac{(1 - \eta_{\max}) \cdot (\pi_{\mathrm{corr}} - \eta_{\min})}{(1 - \pi_{\mathrm{corr}}) \cdot (\eta_{\max} - \eta_{\min})}. \end{aligned} \tag{16}$$

The right hand sides above involve quantities that can be estimated given *only corrupted data*. Thus, plugging in estimates of $\hat{\eta}_{\min}, \hat{\eta}_{\max}, \hat{\pi}_{\mathrm{corr}}$ into Equation 16, we obtain estimates $\hat{\alpha}, \hat{\beta}$ of α, β. (Without the weak separability assumption, the right hand sides would depend on the unknown minimal and maximal values of η.)

The formulae for the noise rates simplify in special cases; e.g., in CCN learning (see Appendix D),

$$\rho_+ = 1 - \eta_{\max} \text{ and } \rho_- = \eta_{\min}. \tag{17}$$

Thus, corrupted class-probability estimation gives a simple means of estimating noise rates for CCN problems.

6.2. Estimating π from η_{corr} in special cases

Unlike the general case, in both CCN and PU learning, π may be estimated. This is because in each case, some information about π is present in $(P_{\mathrm{corr}}, Q_{\mathrm{corr}})$ or π_{corr}. For example, in CCN learning (see Appendix E),

$$\pi = \frac{\pi_{\mathrm{corr}} - \eta_{\min}}{\eta_{\max} - \eta_{\min}},$$

while for the case-controlled PU setting,

$$\pi = \frac{\pi_{\mathrm{corr}}}{1 - \pi_{\mathrm{corr}}} \cdot \frac{1 - \eta_{\max}}{\eta_{\max}}.$$

Estimating π may be of inherent interest beyond its use in computing classification thresholds, as e.g. in case-controlled PU learning scenarios, it lets us assess how prevalent a characteristic is in the underlying population.

6.3. Practical considerations

Equation 16 is an asymptotic identity. In practice, we typically employ estimates $\hat{\eta}_{\min}, \hat{\eta}_{\max}$ computed from a finite sample. We note several points related to this estimation.

First, it is crucial that one employs a rich model class (e.g. Gaussian kernel logistic regression, or single-layer neural network with large number of hidden units). With a misspecified model, it is impossible to determine whether the observed range reflects that of η_{corr}, or simply arises from an inability to model η_{corr}. For example, with a linear logistic regression model $\hat{\eta}_{\mathrm{corr}}(x) = \sigma(\langle w, x \rangle + b)$ applied to instances from $\mathbb{R}^d$, our estimated $\hat{\eta}_{\max}$ may be arbitrarily close to 1 *regardless* of α, β. This is because $\hat{\eta}_{\mathrm{corr}}(N \cdot \mathrm{sign}(w)) = \sigma(N||w|| + b) \to 1$ as $N \to \infty$.

Second, when constructing $\hat{\eta}_{\mathrm{corr}}$, one will often have to choose certain hyper-parameters (e.g. strength of regularisation). Motivated by our regret bounds, these can be chosen to yield the best corrupted class-probability estimates $\hat{\eta}_{\mathrm{corr}}$, as measured by some strictly proper loss. Thus, one can tune parameters by cross-validation *on the corrupted data*; clean samples are *not* required.

Third, for statistical purposes, it is ideal to compute $\hat{\eta}_{\min}, \hat{\eta}_{\max}$ from a fresh sample not used for constructing probability estimates $\hat{\eta}_{\mathrm{corr}}$. These range estimates may even be computed on *unlabelled test* instances, as they do not require ground truth labels. (This does not constitute overfitting to the test set, as the underlying model for $\hat{\eta}_{\mathrm{corr}}$ is learned purely from corrupted training data.)

Fourth, the sample maximum and minimum are clearly susceptible to outliers. Therefore, it may be preferable to employ e.g. the 99% and 1% quantiles as a robust alternative.

Alternately, one may perform some form of aggregation (e.g. the bootstrap) to smooth the estimates.

Finally, to compute a suitable threshold for classification, noisy estimates of α, β may be sufficient. For example, in CCN learning, we only need the estimated difference $\hat{\rho}_+ - \hat{\rho}_-$ to be comparable to the true difference $\rho_+ - \rho_-$ (by Equation 13). du Plessis et al. (2014) performed such an analysis for the case-controlled PU learning setting.

Relation to existing work The estimator in Equation 16 may be seen as a generalisation of that proposed by Elkan & Noto (2008) for the censoring version of PU learning. For CCN learning, in independent work, Liu & Tao (2014, Theorem 4) proposed the estimators in Equation 17.

Scott et al. (2013) proposed a means of estimating the noise parameters, based on a reduction to the problem of mixture proportion estimation. By an interpretation provided by Blanchard et al. (2010), the noise parameters can be seen as arising from the derivative of the right hand side of the optimal ROC curve on $\mathrm{Corr}(D, \alpha, \beta, \pi_{\mathrm{corr}})$. Sanderson & Scott (2014); Scott (2015) explored a practical estimator along these lines. As the optimal ROC curve for D_{corr} is produced by any strictly monotone transformation of η_{corr}, class-probability estimation is implicit in this approach, and so our estimator is simply based on a different perspective. (See Appendix I.) The class-probability estimation perspective shows that a single approach can both estimate the corruption parameters and be used to classify optimally for a range of performance measures.

7. Experiments

We now present experiments that aim to validate our analysis[4] via three questions. First, can we optimise BER and AUC from corrupted data without knowledge of the noise parameters? Second, can we accurately estimate corruption parameters? Third, can we optimise other classification measures using estimates of corruption parameters?

We focus on CCN learning with label flip probabilities $\rho_+, \rho_- \in \{0, 0.1, 0.2, 0.3, 0.4, 0.49\}$; recall that $\rho_- = 0$ is the censoring version of PU learning. For this problem, a number of approaches have been proposed to answer the third question above, e.g. (Stempfel & Ralaivola, 2007; 2009; Natarajan et al., 2013). To our knowledge, all of these operate in the setting where the noise parameters are *known*. It is thus possible to use the noise estimates from class-probability estimation as inputs to these approaches, and we expect such a fusion will be beneficial. We leave such a study for future work, as our aim here is merely to illustrate that with corrupted class-probability estimation, we can answer all three questions in the affirmative.

We report results on a range of UCI datasets. For each dataset, we construct a random 80% – 20% train-test split. For fixed ρ_+, ρ_-, we inject label noise into the training set. The learner estimates class-probabilities from these noisy samples, with predictions on the clean test samples used to estimate $\hat{\eta}_{\min}, \hat{\eta}_{\max}$ if required. We summarise performance across τ independent corruptions of the training set.

Observe that if $D_{M,\eta}$ can be modelled by a linear scorer, so that $\eta \colon x \mapsto \sigma(\langle w, x \rangle + b)$, then $\eta_{\mathrm{corr}} \colon x \mapsto (1 - \rho_+ - \rho_-) \cdot \sigma(\langle w, x \rangle + b) + \rho_-$; i.e. , a neural network with a single hidden sigmoidal unit, bias term, and identity output link is well-specified. Thus, in all experiments, we use as our base model a neural network with a sigmoidal hidden layer, trained to minimise squared error[5] with ℓ_2 regularisation. The regularisation parameter for the model was tuned by cross-validation (on the *corrupted* data) based on squared error. We emphasise that both learning and parameter tuning is *solely on corrupted data*.

7.1. Are BER and AUC immune to corruption?

We first assess how effectively we can optimise BER and AUC from corrupted data *without* knowledge or estimates of the noise parameters. For a fixed setting of ρ_+, ρ_-, and each of $\tau = 100$ corruption trials, we learn a class-probability estimator from the corrupted training set. We use this to predict class-probabilities for instances on the *clean* test set. We measure the AUC of the resulting class-probabilities, as well as the BER resulting from thresholding these probabilities about the *corrupted* base rate.

Table 2 summarises the results for a selection of datasets and noise rates ρ_+, ρ_-. (Appendix J contains a full set of results.) We see that in general the BER and AUC in the noise-free case ($\rho_+ = \rho_- = 0$) and in the noisy cases are commensurate. This is in agreement with our analysis on the immunity of BER and AUC. For smaller datasets and higher levels of noise, we see a greater degradation in performance. This matches our regret bounds (Proposition 2), which indicated a penalty in high-noise regimes.

7.2. Can we reliably estimate noise rates?

We now study the viability of learning label flip probabilities ρ_+, ρ_-. As above, we compute corrupted class-probability estimates, and use these to compute label flip probability estimates $\hat{\rho}_+, \hat{\rho}_-$ as per the approach in §6.

Figure 1 presents violin plots (Hintze & Nelson, 1998) of the signed errors in the estimate $\hat{\rho}_+$, for symmetric ground-truth ρ_+, ρ_-, on three of the UCI datasets. (For plots of

[4] Sample scripts are available at http://users.cecs.anu.edu.au/~akmenon/papers/corrupted-labels/index.html.

[5] Using log-loss requires explicitly constraining the range of the bias and hidden → output term, else the loss is undefined.

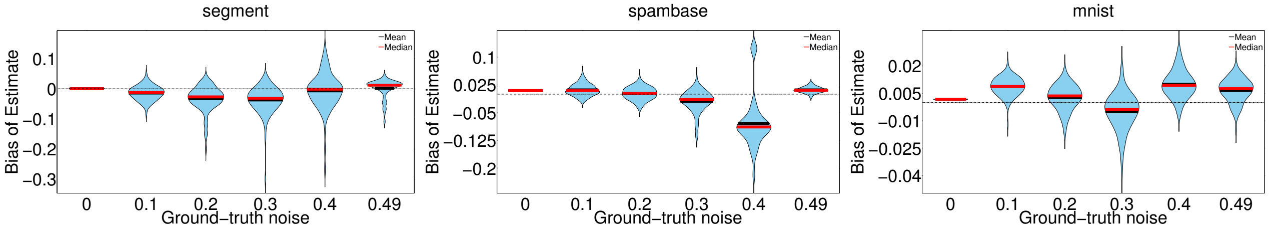

Figure 1. Violin plots of bias in estimate $\hat{\rho}_+$ over $\tau = 100$ trials on Segment (L), Spambase (M) and MNIST (R).

Dataset	**Noise**	1 - **AUC** (%)	**BER** (%)	**ERR**$_{\max}$ (%)	**ERR**$_{\text{oracle}}$ (%)
segment	None	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.0)$	0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.2)$	0.02 ± 0.01	0.90 ± 0.08	0.31 ± 0.05	0.30 ± 0.05
	$(\rho_+, \rho_-) = (0.2, 0.4)$	0.03 ± 0.01	3.24 ± 0.20	0.31 ± 0.06	0.27 ± 0.06
spambase	None	2.49 ± 0.00	6.93 ± 0.00	6.52 ± 0.00	6.52 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.0)$	2.67 ± 0.02	7.10 ± 0.03	6.88 ± 0.03	6.89 ± 0.03
	$(\rho_+, \rho_-) = (0.1, 0.2)$	3.01 ± 0.03	7.66 ± 0.05	7.51 ± 0.05	7.48 ± 0.05
	$(\rho_+, \rho_-) = (0.2, 0.4)$	4.91 ± 0.09	10.52 ± 0.13	10.82 ± 0.31	10.26 ± 0.12
mnist	None	0.92 ± 0.00	3.63 ± 0.00	3.63 ± 0.00	3.63 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.0)$	0.95 ± 0.01	3.56 ± 0.01	3.55 ± 0.01	3.55 ± 0.01
	$(\rho_+, \rho_-) = (0.1, 0.2)$	0.97 ± 0.01	3.63 ± 0.02	3.62 ± 0.02	3.62 ± 0.02
	$(\rho_+, \rho_-) = (0.2, 0.4)$	1.17 ± 0.02	4.06 ± 0.03	4.06 ± 0.03	4.05 ± 0.03

Table 2. Mean and standard error (standard deviation scaled by $\sqrt{\tau}$) of performance measures on UCI datasets injected with random label noise $\tau = 100$ times. The case $\rho_- = 0$ corresponds to the censoring version of PU learning. ERR$_{\max}$ and ERR$_{\text{oracle}}$ are the misclassification errors of the classifiers formed by thresholding using $\hat{\rho}_+, \hat{\rho}_-$, and by the ground-truth ρ_+, ρ_- respectively.

$\hat{\rho}_-$, see Appendix J.) These plots show the distribution of signed errors across the noise trials; concentration about zero is ideal. For lower noise rates, the estimates are generally only mildly biased, and possess low mean squared error. As previously, we see a greater spread in the error distribution for higher ground-truth noise rates.

7.3. Can other classification measures be minimised?

We finally study the misclassification error[6] of a classifier learned from noisy data. As above, we learn a corrupted class-probability estimator, and compute noise estimates ρ_+, ρ_- as per §6. We then threshold predictions based on Equation 13 to form a classifier. We also include the results of an oracle that has exact knowledge of ρ_+, ρ_-, but only access to the noisy data. The performance of this method illustrates whether increased classification error is due to inexact estimates of ρ_+, ρ_-, or inexact estimates of η_{corr}.

Table 2 illustrates that while compared to BER and AUC, we see slightly higher levels of degradation, in general the misclassification rate can be effectively minimised even in high noise regimes. As previously, we find that under higher levels of ground-truth noise, there is in general a slight decrease in accuracy. Interestingly, this is so *even for the oracle estimator*, again corroborating our regret bounds which indicate a penalty in high-noise regimes.

In summary, class-probability estimation lets us both estimate the parameters of the contamination process, as well as minimise a range of classification measures.

8. Conclusion

We have used class-probability estimation to study learning from corrupted binary labels. In particular, we have shown that for optimising the balanced error and AUC, the corruption process may be ignored; given estimates of the corruption parameters, several classification measures can be minimised; and that such estimates may be obtained by the range of the class-probabilities.

In future work, we aim to study the impact of corruption on estimation rates of class-probabilities; study ranking risks beyond the AUC; and study potential extensions of our results to more general corruption problems.

[6] While BER is more apposite on imbalanced data, we simply aim to assess the feasibility of minimising misclassification risk.

Acknowledgments. NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

References

Agarwal, Shivani. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15:1653–1674, 2014.

Agarwal, Shivani, Graepel, Thore, Herbrich, Ralf, Har-Peled, Sariel, and Roth, Dan. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425, December 2005.

Angluin, Dana and Laird, Philip. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Blanchard, Gilles, Lee, Gyemin, and Scott, Clayton. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, December 2010.

Blum, Avrim and Mitchell, Tom. Combining labeled and unlabeled data with co-training. In *Conference on Computational Learning Theory (COLT)*, pp. 92–100, New York, NY, USA, 1998. ACM. ISBN 1-58113-057-0.

Brodersen, Kay H., Ong, Cheng Soon, Stephan, Klaas E., and Buhmann, Joachim M. The balanced accuracy and its posterior distribution. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pp. 3121–3124, 2010.

Cheng, Jie, Hatzis, Christos, Hayashi, Hisashi, Krogel, Mark-A., Morishita, Shinchi, Page, David, and Sese, Jun. KDD Cup 2001 report. *ACM SIGKDD Explorations Newsletter*, 3(2):47–64, 2002.

Clémençon, Stéphan, Lugosi, Gábor, and Vayatis, Nicolas. Ranking and Empirical Minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, April 2008.

Denis, François. PAC learning from positive statistical queries. In *Algorithmic Learning Theory (ALT)*, volume 1501 of *Lecture Notes in Computer Science*, pp. 112–126. Springer Berlin Heidelberg, 1998. ISBN 978-3-540-65013-3.

du Plessis, Marthinus C, Niu, Gang, and Sugiyama, Masashi. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 703–711. Curran Associates, Inc., 2014.

du Plessis, Marthinus Christoffel and Sugiyama, Masashi. Class prior estimation from positive and unlabeled data. *IEICE Transactions*, 97-D(5):1358–1362, 2014.

Elkan, Charles and Noto, Keith. Learning classifiers from only positive and unlabeled data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 213–220, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-193-4.

Flach, Peter, Hernândez-Orallo, Josè, and Ferri, Cësar. A coherent interpretation of AUC as a measure of aggregated classification performance. In *International Conference on Machine Learning (ICML)*, 2011.

Guyon, Isabelle, Hur, Asa Ben, Gunn, Steve, and Dror, Gideon. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 545–552. MIT Press, 2004.

Hintze, Jerry L. and Nelson, Ray D. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998.

Kotłowski, Wojciech and Dembczyński, Krzysztof. Surrogate regret bounds for generalized classification performance metrics. In *Conference on Learning Theory (COLT)*, 2015.

Koyejo, Oluwasanmi O, Natarajan, Nagarajan, Ravikumar, Pradeep K, and Dhillon, Inderjit S. Consistent binary classification with generalized performance metrics. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2744–2752. Curran Associates, Inc., 2014.

Krzanowski, Wojtek J. and Hand, David J. *ROC Curves for Continuous Data*. Chapman & Hall/CRC, 1st edition, 2009. ISBN 1439800219, 9781439800218.

Lee, Wee Sun and Liu, Bing. Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression. In *International Conference on Machine Learning (ICML)*, 2003.

Liu, T. and Tao, D. Classification with Noisy Labels by Importance Reweighting. *ArXiv e-prints*, November 2014. URL `http://arxiv.org/abs/1411.7718`.

McCullagh, Peter and Nelder, John Ashworth. *Generalized linear models (Second edition)*. London: Chapman & Hall, 1989.

Menon, Aditya Krishna, Narasimhan, Harikrishna, Agarwal, Shivani, and Chawla, Sanjay. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning (ICML)*, pp. 603–611, 2013.

Narasimhan, Harikrishna, Vaish, Rohit, and Agarwal, Shivani. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In

Advances in Neural Information Processing Systems (NIPS), pp. 1493–1501. Curran Associates, Inc., 2014.

Natarajan, Nagarajan, Dhillon, Inderjit S., Ravikumar, Pradeep D., and Tewari, Ambuj. Learning with noisy labels. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1196–1204, 2013.

Parambath, Shameem A. Puthiya, Usunier, Nicolas, and Grandvalet, Yves. Optimizing F-measures by cost-sensitive classification. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2123–2131, 2014.

Phillips, Steven J., Dudik, Miroslav, Elith, Jane, Graham, Catherine H., Lehmann, Anthony, Leathwick, John, and Ferrier, Simon. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19 (1):181–197, 2009.

Reid, Mark D. and Williamson, Robert C. Composite binary losses. *Journal of Machine Learning Research*, 11: 2387–2422, December 2010.

Sanderson, Tyler and Scott, Clayton. Class proportion estimation with application to multiclass anomaly rejection. In *AISTATS*, pp. 850–858, 2014.

Scott, Clayton. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *To appear in International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Scott, Clayton, Blanchard, Gilles, and Handy, Gregory. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference on Learning Theory (COLT)*, volume 30 of *JMLR Proceedings*, pp. 489–511, 2013.

Stempfel, Guillaume and Ralaivola, Liva. Learning kernel perceptrons on noisy data using random projections. In Hutter, Marcus, Servedio, Rocco., and Takimoto, Eiji (eds.), *Algorithmic Learning Theory*, volume 4754 of *Lecture Notes in Computer Science*, pp. 328–342. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-75224-0.

Stempfel, Guillaume and Ralaivola, Liva. Learning SVMs from sloppily labeled data. In Alippi, Cesare, Polycarpou, Marios, Panayiotou, Christos, and Ellinas, Georgios (eds.), *Artificial Neural Networks (ICANN)*, volume 5768 of *Lecture Notes in Computer Science*, pp. 884–893. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-04273-7.

van Rooyen, Brendan and Williamson, Robert C. Learning in the Presence of Corruption. *ArXiv e-prints*, March 2015. URL http://arxiv.org/abs/1504.00091.

Ward, Gill, Hastie, Trevor, Barry, Simon, Elith, Jane, and Leathwick, John R. Presence-only data and the EM algorithm. *Biometrics*, 65(2):554–563, 2009.

Zhang, Dell and Lee, Wee Sun. Learning classifiers without negative examples: A reduction approach. In *Third International Conference on Digital Information Management (ICDIM)*, pp. 638–643, 2008.