Classification Performance Measures and Weakly Supervised Learning

Clay Scott

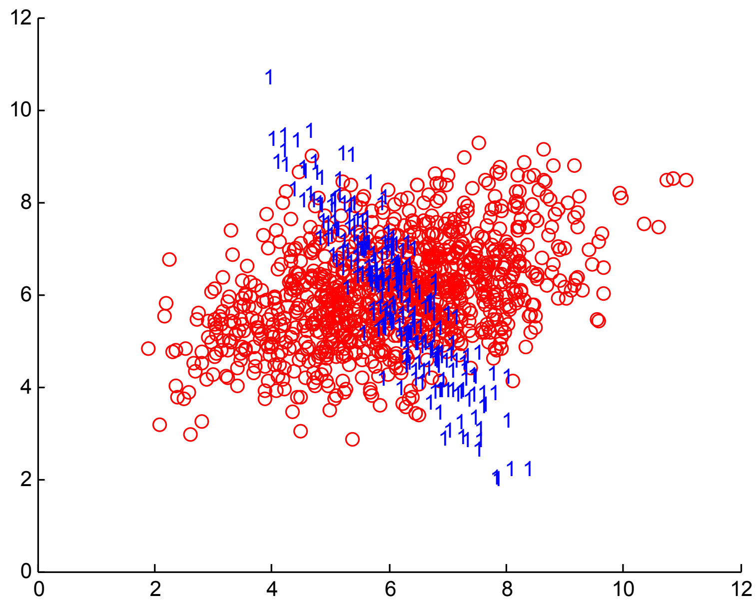
Electrical and Computer Engineering & Statistics
University of Michigan

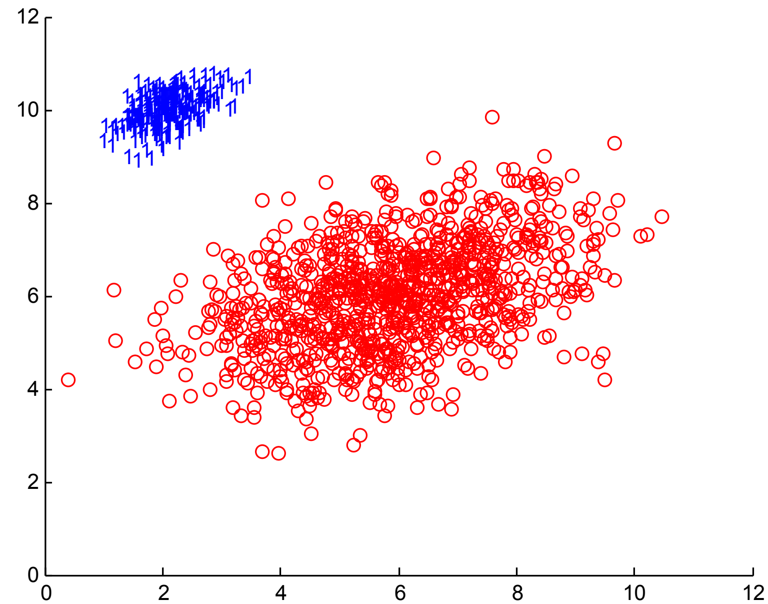
Outline

Part 1: Performance measures for classification

Part 2: Weakly supervised learning

Classification





Many nonparametric methods:
Nearest neighbors, decision trees,
support vector machines, neural
networks, etc.

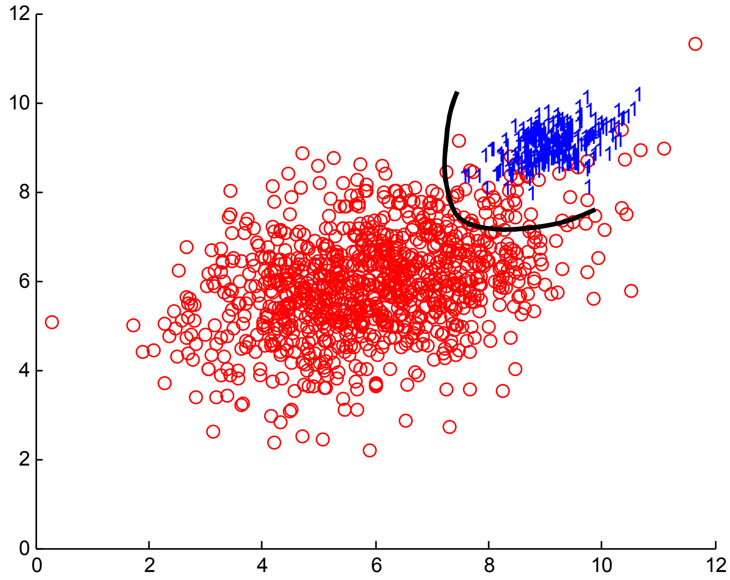
Probability of Error

- $X \in \mathbb{R}^d$ = pattern of interest
- $Y \in \{0, 1\}$ = label
- Classifier:

$$f : \mathbb{R}^d \to \{0, 1\}$$
$$f(x) = 1_{\{h(x) > 0\}}$$

- Probability of error

$$R(f) = P(f(X) \neq Y)$$



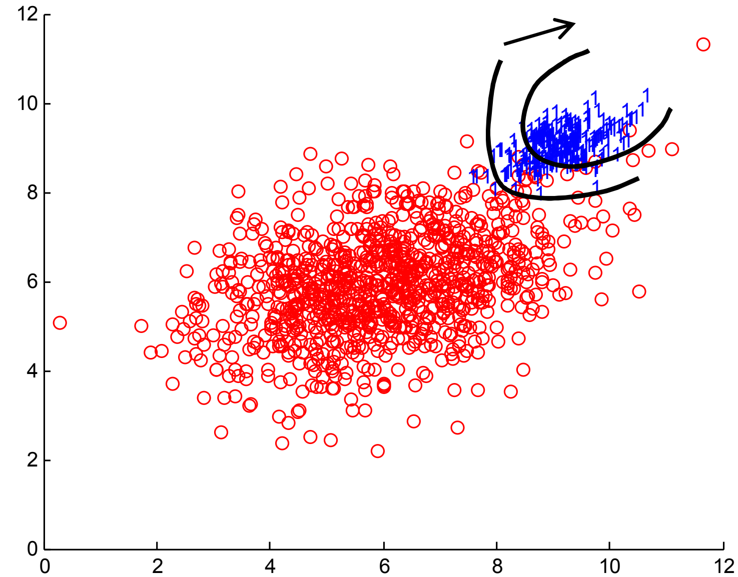
Cost-Sensitive Risk

- Misclassification rate can be expressed

$$
\begin{aligned}
R(f) &= P(Y=1, f(X)=0) + P(Y=0, f(X)=1) \\
&= \pi_1 R_1(f) + \pi_0 R_0(f)
\end{aligned}
$$

where

$$
\begin{aligned}
\pi_0 &= P(Y=0) \\
\pi_1 &= P(Y=1) \\
R_0(f) &= P(f(X)=1 \mid Y=0) \\
R_1(f) &= P(f(X)=0 \mid Y=1)
\end{aligned}
$$



- For $\rho \in (0,1)$, define the **cost-sensitive risk**

$$R_\rho(f) := \rho \pi_0 R_0(f) + (1-\rho)\pi_1 R_1(f)$$

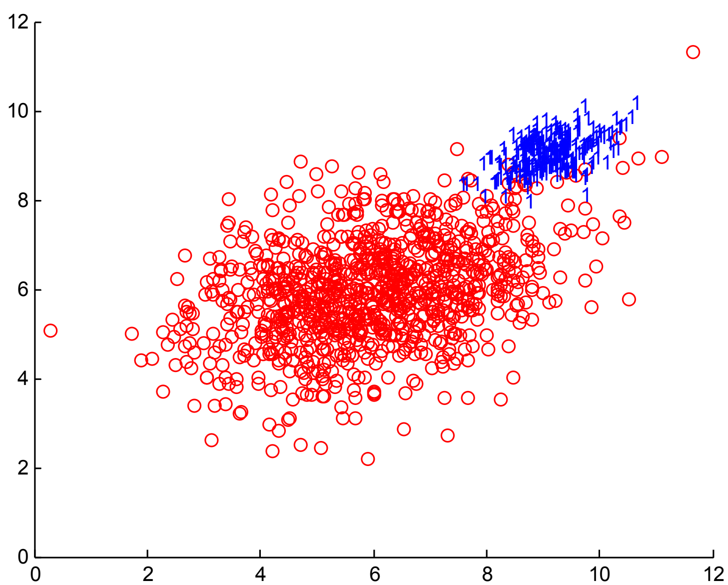
Optimal Classifiers

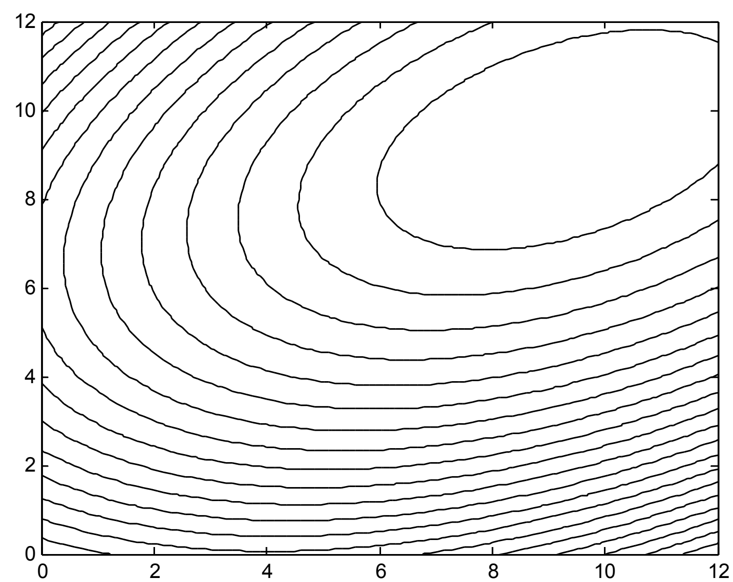
The optimal classifier for any cost-sensitive risk is a **likelihood ratio test**

$$\frac{p_1(x)}{p_0(x)} \gtrless \lambda$$

for some $\lambda > 0$, where

$$p_1(x) = \text{probability density of } X \text{ given } Y = 1$$
$$p_0(x) = \text{probability density of } X \text{ given } Y = 0$$





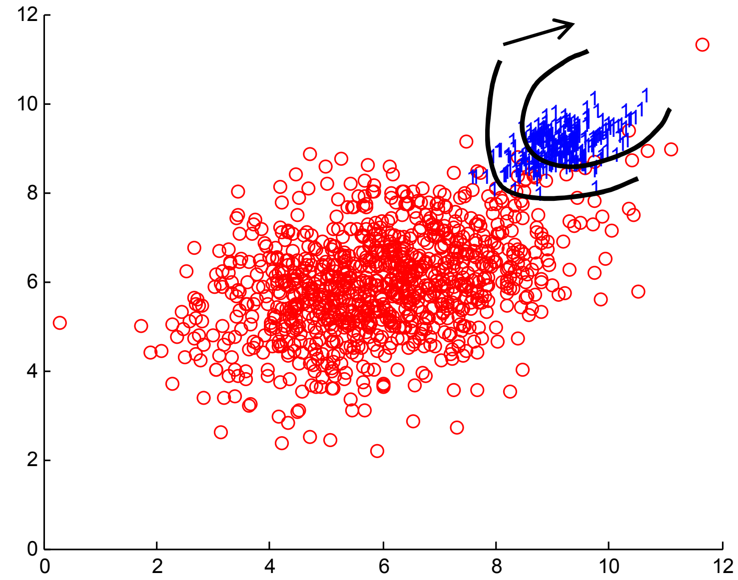
Neyman-Pearson

- False positive/negative rates:

$$R_0(f) = P(f(X) = 1 \,|\, Y = 0)$$
$$R_1(f) = P(f(X) = 0 \,|\, Y = 1)$$

- Given $\alpha \in (0,1)$, the **Neyman-Pearson classifier** solves

$$\begin{aligned} \min \quad & R_1(f) \\ \text{s.t.} \quad & R_0(f) \leq \alpha \end{aligned}$$



- Solution also a likelihood ratio test
- Advantages:
  - Class proportions in test and training data need not be the same
  - Imbalanced data

Other Frequentist Criteria

- Min-max

$$R_{\text{mm}}(f) = \max\{R_0(f), R_1(f)\}$$

- Balanced error

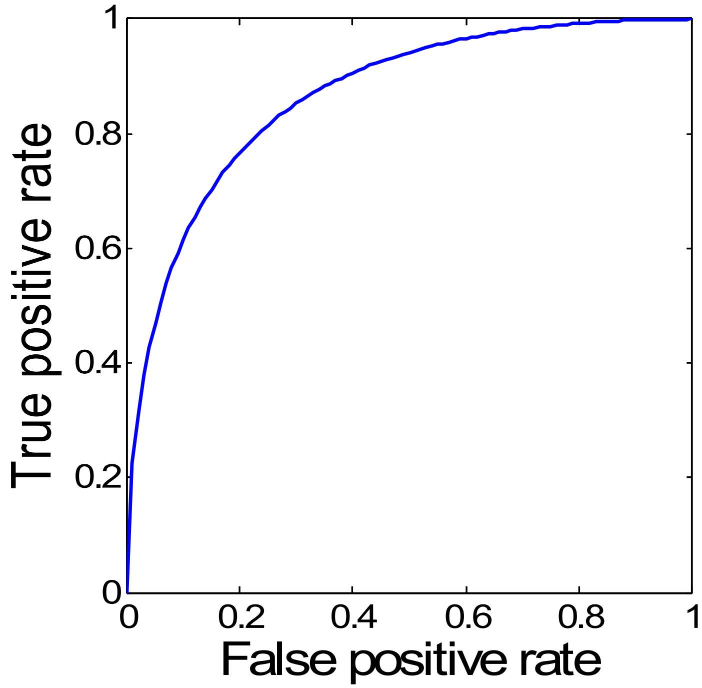
$$R_{\text{bal}}(f) = \frac{R_0(f) + R_1(f)}{2}$$

- Weighted error

$$\rho R_0(f) + (1 - \rho) R_1(f)$$

- Optimal classifiers are again likelihood ratio tests

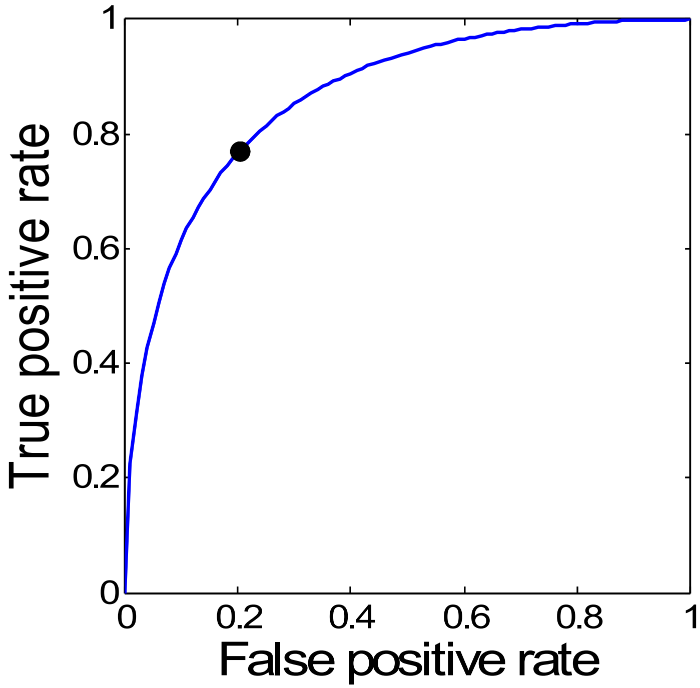
Area Under ROC Curve



- Again optimized by the family of likelihood ratio tests

Algorithms

- Since all criteria are solved by likelihood ratio tests, it suffices to minimize the cost-sensitive risk $R_\rho(f)$, where ρ is chosen according to the desired criterion.
- Therefore, can apply existing algorithms, which can easily be adapted to minimize the cost-sensitive (empirical) risk



Cost-Insensitive Learning

Given training data $(x_1, y_1), \ldots, (x_n, y_n)$, $y_i \in \{-1, 1\}$, solve

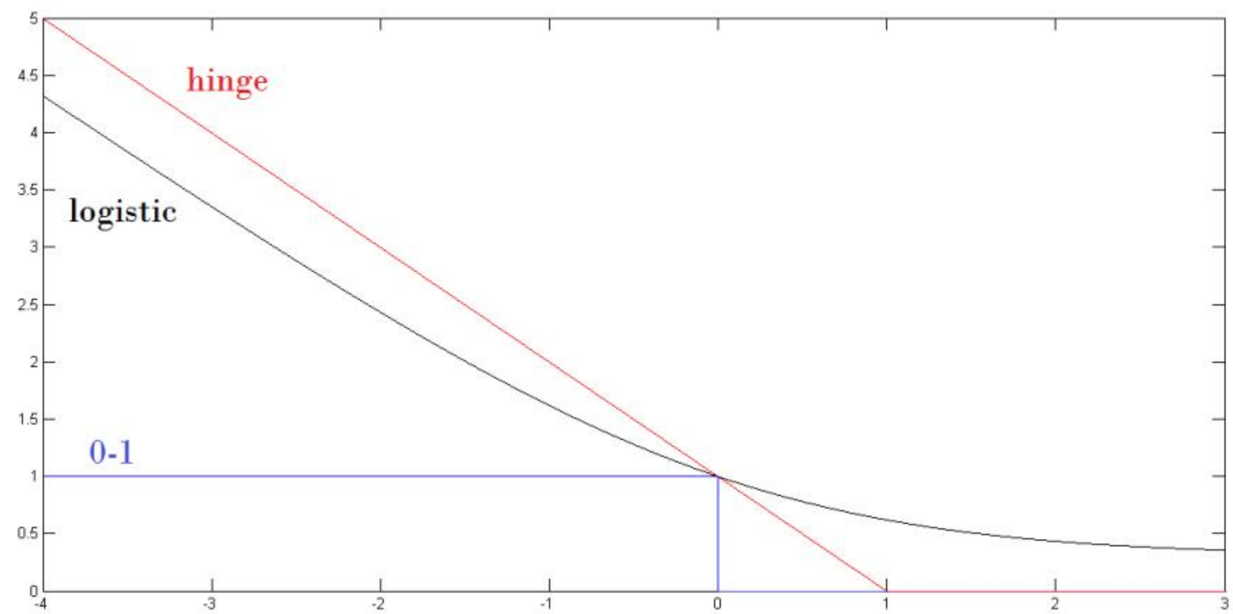
$$\widehat{h} = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \phi(y_i h(x_i))$$

$$\longrightarrow \quad f(x) = \text{sign}(h(x))$$

where

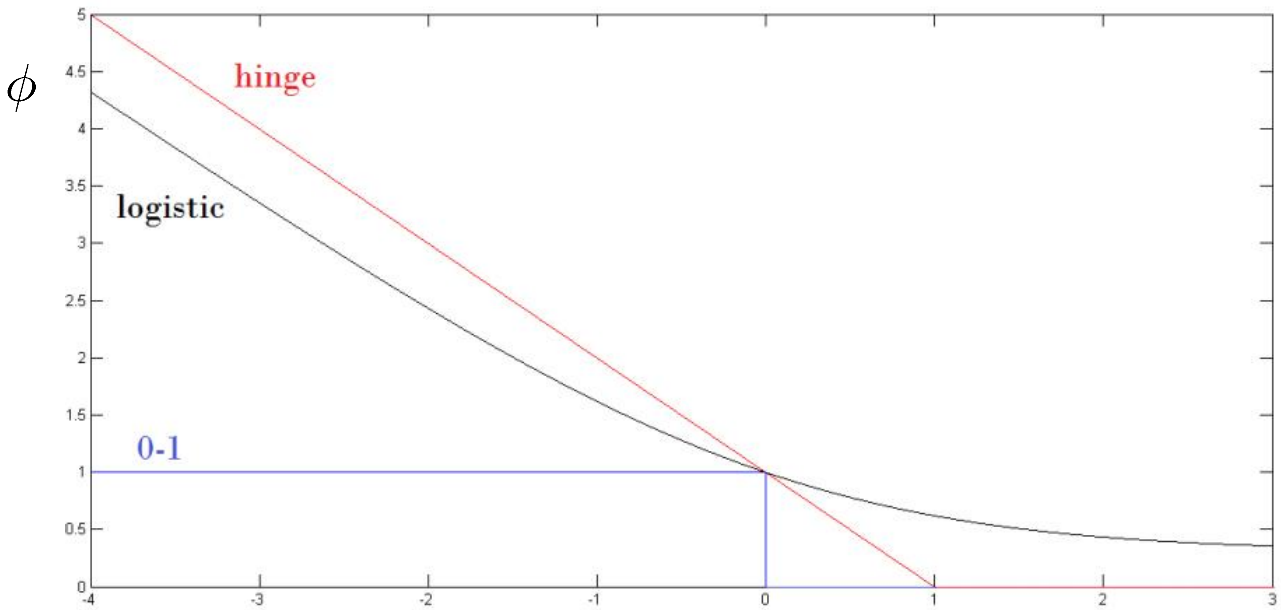
- $\mathcal{H}$ is a function class
- ϕ is a loss

ϕ



Cost-Sensitive Learning

$$\widehat{h}_\rho = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \left[(1-\rho) 1_{\{y_i=1\}} \phi(h(x_i)) + \rho 1_{\{y_i=-1\}} \phi(-h(x_i)) \right]$$



Summary of Part 1

- Frequentist performance measures are not affected when the training class proportions and testing class proportions differ (the simplest form of domain adaptation)
- Frequentist performance measures can be optimized by cost-sensitive learning, although ρ becomes an additional tuning parameter
- For neural networks, how does feature representation depend on performance measure?

Outline

Part 1: Performance measures for classification

→ Part 2: Weakly supervised learning

Weakly Supervised Learning

Definition: Weakly supervised learning (WSL) = supervised learning where some or all labels are corrupted, contaminated, or missing

Weakly supervised

Fully supervised — Unsupervised

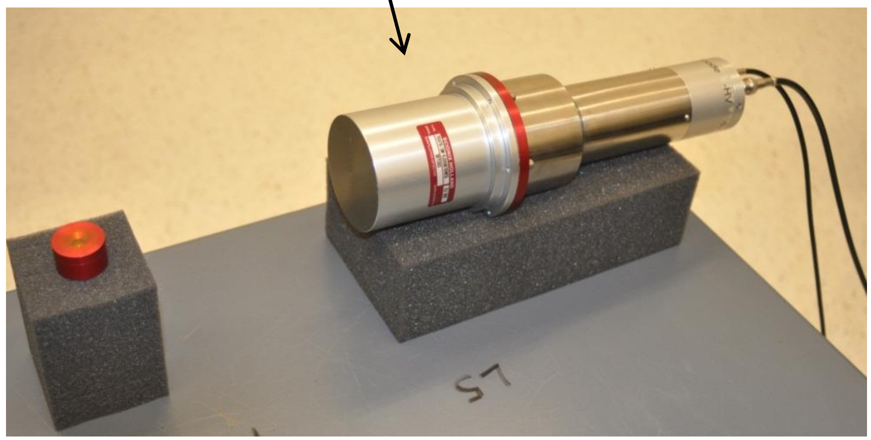
Important theme: Many WSL problems are easier to solve for certain performance measures

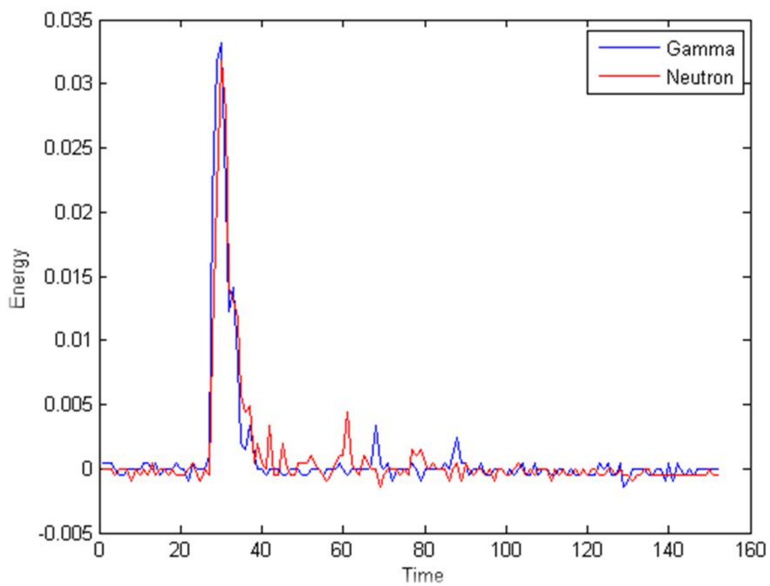
Nuclear Nonproliferation

- Radioactive sources are characterized by distribution of neutron energies
- <u>Organic scintillation detectors</u>: prominent technology for neutron detection

Organic Scintillation Detector

Source material

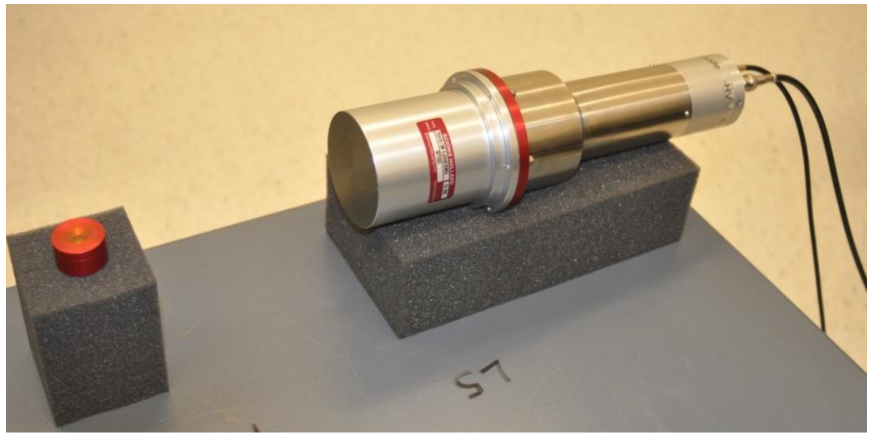




- Detects both neutrons and gamma rays
- Need to classify neutrons and gamma rays

Nuclear Particle Classification

Source material



- $X \in \mathbb{R}^d$, $d =$ signal length
- Training data:

$$X_1, \ldots, X_m \overset{iid}{\sim} P_0 \qquad \text{(from gamma ray source, e.g. Na-22)}$$

$$X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} P_1 \qquad \text{(from neutron source, e.g. Cf-252)}$$

- $P_0, P_1 =$ class-conditional distributions; don't want to model

Reality: No Pure Neutron Sources

- Contamination model for training data:

$$X_1, \ldots, X_m \overset{iid}{\sim} P_0$$
$$X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} \tilde{P}_1 = (1-\pi)P_1 + \pi P_0$$

- π unknown
- P_0, P_1 may have overlapping supports (nonseparable problem)
- Problem known as "learning with negative and unlabeled examples" or "classification with one-sided label noise"

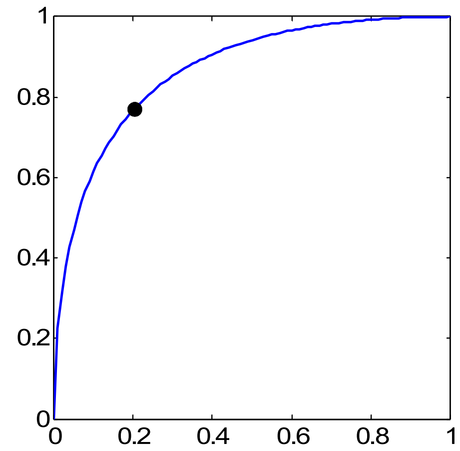
Training On Contaminated Data

- Train a binary classifer on

$$X_1, \ldots, X_m \overset{iid}{\sim} P_0$$
$$X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} \tilde{P}_1 = (1-\pi)P_1 + \pi P_0$$

- "Contaminated" likelihood ratio

$$\frac{\tilde{p}_1(x)}{p_0(x)} = \frac{(1-\pi)p_1(x) + \pi p_0(x)}{p_0(x)}$$
$$= (1-\pi)\frac{p_1(x)}{p_0(x)} + \pi$$



- **Key insights:**
  - True and contaminated LRs have same ROC
  - For Neyman-Pearson criterion, can set threshold because class 0 is uncontaminated

More Reality: Both Classes Contaminated

- Gammas and neutrons from background radiation

- Contaminated training data:

$$X_1, \ldots, X_m \overset{iid}{\sim} \tilde{P}_0 = (1 - \pi_0)P_0 + \pi_0 P_1$$
$$X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} \tilde{P}_1 = (1 - \pi_1)P_1 + \pi_1 P_0$$

- π_0, π_1 **unknown**
- "Classification with (two-sided) label noise"
- Label noise is **in addition to** the usual noise that is present in binary classification (i.e., $y|x$ is random)
- **Random** label noise, as opposed to adversarial, or feature-dependent

Understanding Label Noise

- Assume P_0, P_1 have densities $p_0(x), p_1(x)$
- Then $\tilde{P}_0, \tilde{P}_1$ have densities

$$\tilde{p}_0(x) = (1 - \pi_0)p_0(x) + \pi_0 p_1(x)$$
$$\tilde{p}_1(x) = (1 - \pi_1)p_1(x) + \pi_1 p_0(x)$$

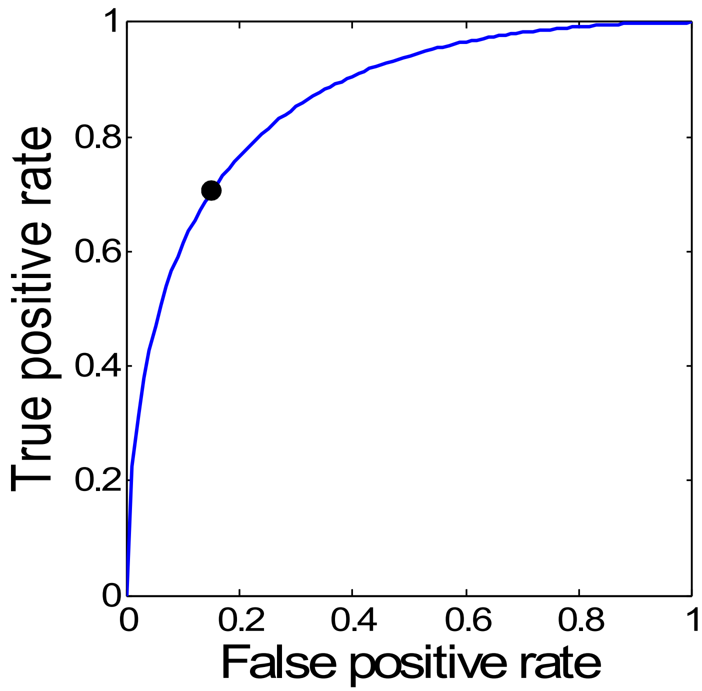
- Simple algebra:

$$\frac{p_1(x)}{p_0(x)} > \gamma \iff \frac{\tilde{p}_1(x)}{\tilde{p}_0(x)} > \lambda,$$

where

$$\lambda = \frac{\pi_1 + \gamma(1 - \pi_1)}{1 - \pi_0 + \gamma\pi_0}.$$

- **Balanced error** immune to label noise (Menon et al., 2015)



Cost-Sensitive Approach

- If π_0 and π_1 are known (or can be estimated), can optimize a performance measure of interest by performing cost-sensitive classification with an appropriate cost parameter.

- For example, if the performance measure of interest is the probability of error, take

$$\rho = \frac{\frac{1}{2} - \pi_0}{1 - \pi_0 - \pi_1}$$

Feature-Dependent Label Noise

- Unobserved: $(X_1, Y_1), \ldots, (X_n, Y_n)$
- Observed: $(X_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Y}_n)$. Y_i flips with probability depending on X_i
- Under a certain condition, the contaminated and true likelihood ratios are monontonically equivalent
- That assumptions essentially states that the more a 0 looks like a 1, the more probable a label of 0 is to flip to a 1 (and vice versa)
- The following criterion is immune to feature-dependent label noise:

$$\begin{aligned} \min \quad & R_1(f) \\ \text{s.t.} \quad & P(f(X) = 1) \leq \alpha \end{aligned}$$

- Constraint on the "discovery rate"

Learning From Label Proportions

- (B_i, π_i), $i = 1, 2, \ldots$
- $B_i =$ collection of feature vectors, iid from a mixture of P_0 and P_1
- $\pi_i =$ proportion of class 1 in B_i
- Recent applications to HEP: Dery, Nachman, Rubbo, Schwartzman (1702:00414), Cohen, Freytsis, Ostdiek (1706:09451), classification of jets

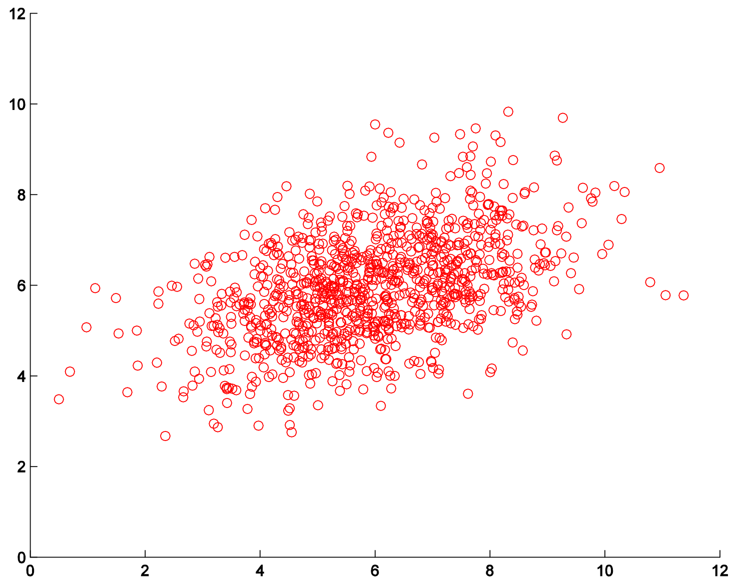
LLP as Classification with Noisy Labels

- Consider two bags
- Suppose:

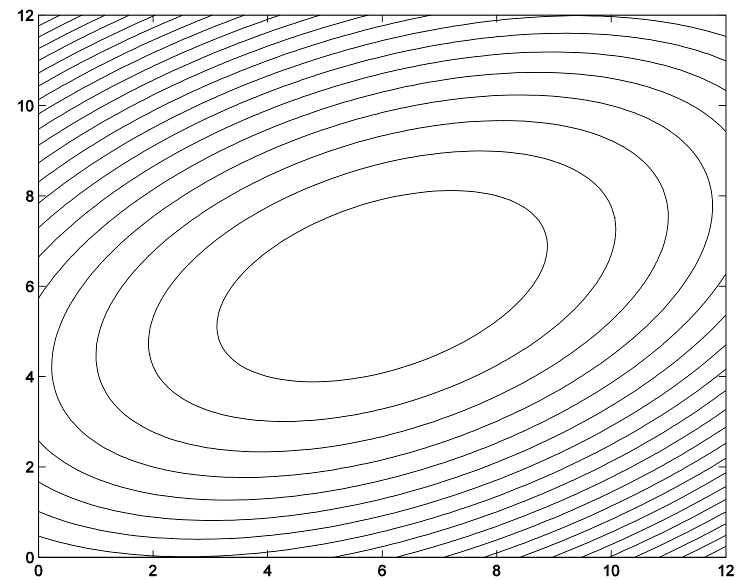
$$\begin{array}{lrl} \text{Bag 1:} & x_1, \ldots, x_m & \sim (1-\pi_1)P_0 + \pi_1 P_1,\ \pi_1 < \dfrac{1}{2} \\ \text{Bag 2:} & x_{m+1}, \ldots, x_{m+n} & \sim (1-\pi_2)P_0 + \pi_2 P_1,\ \pi_2 > \dfrac{1}{2} \end{array}$$

- This is a classification with label noise problem. Since the label proportions are given, can define appropriate cost-sensitive loss

Novelty Detection



Background data only

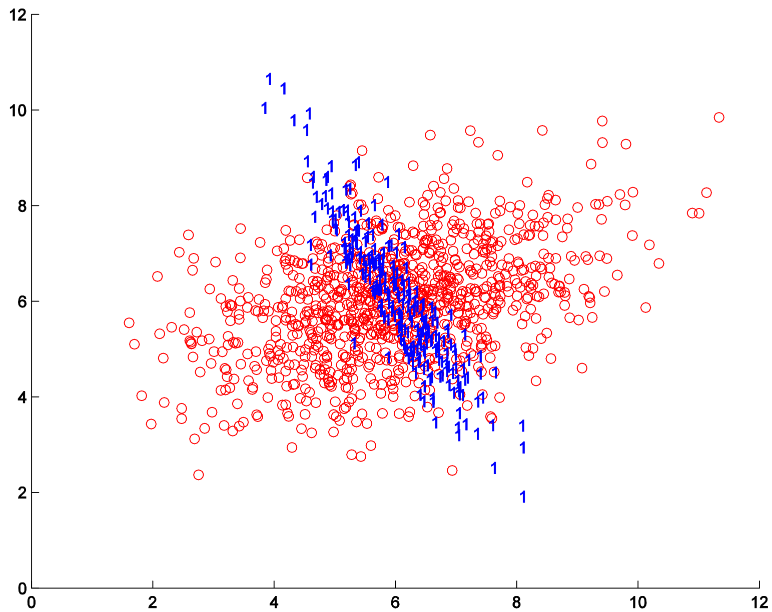


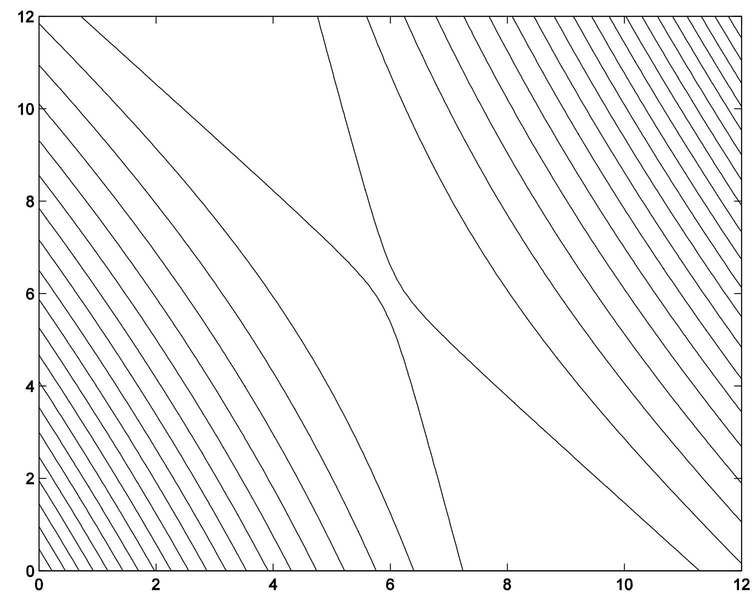
Typical approach: estimate a **level set** of the background density

$$\lambda \gtrless p_0(x)$$

Nonparametric methods:
thresholded KDE, one-class SVM

Underlying Classification Problem

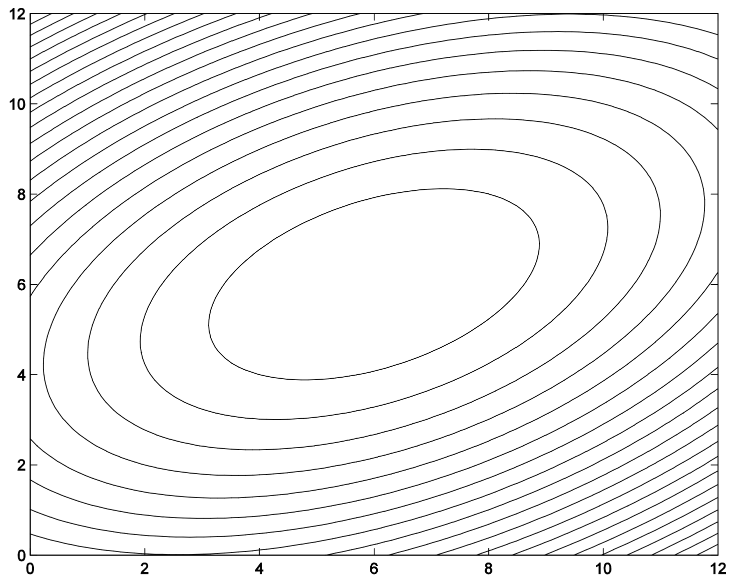




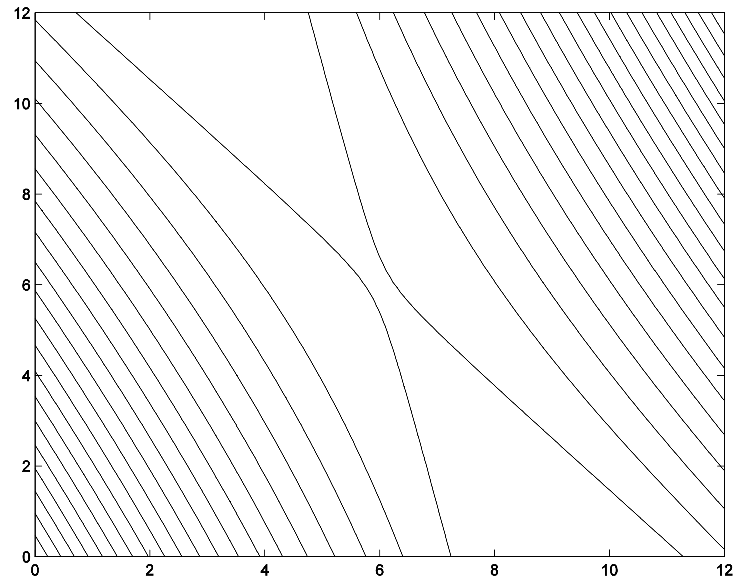
Optimal classifier:

$$\lambda \gtrless \frac{p_1(x)}{p_0(x)}$$

Problem with Level Set Approach



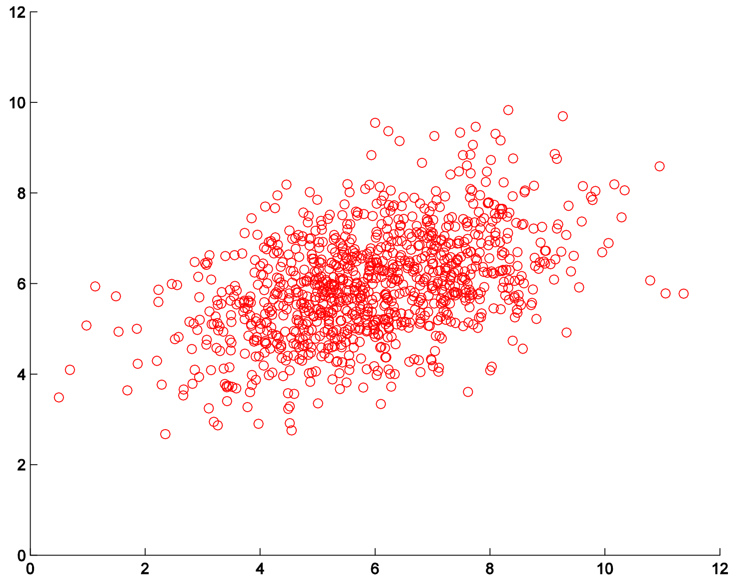
$\neq$



The more p_1 overlaps p_0, the bigger the problem

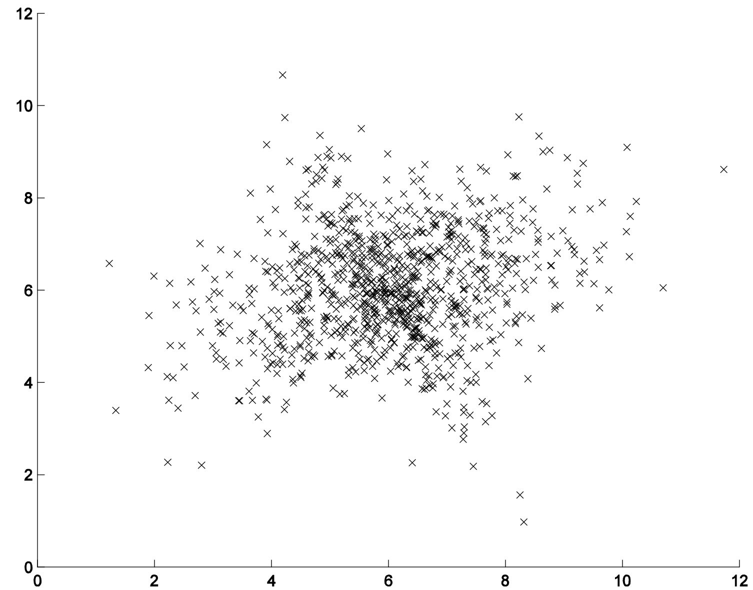
Semi-Supervised Novelty Detection

Suppose you observe



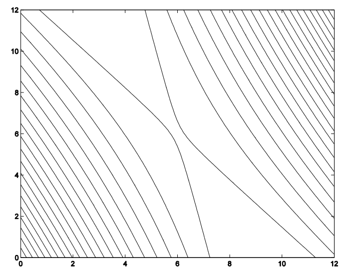
Background data

and

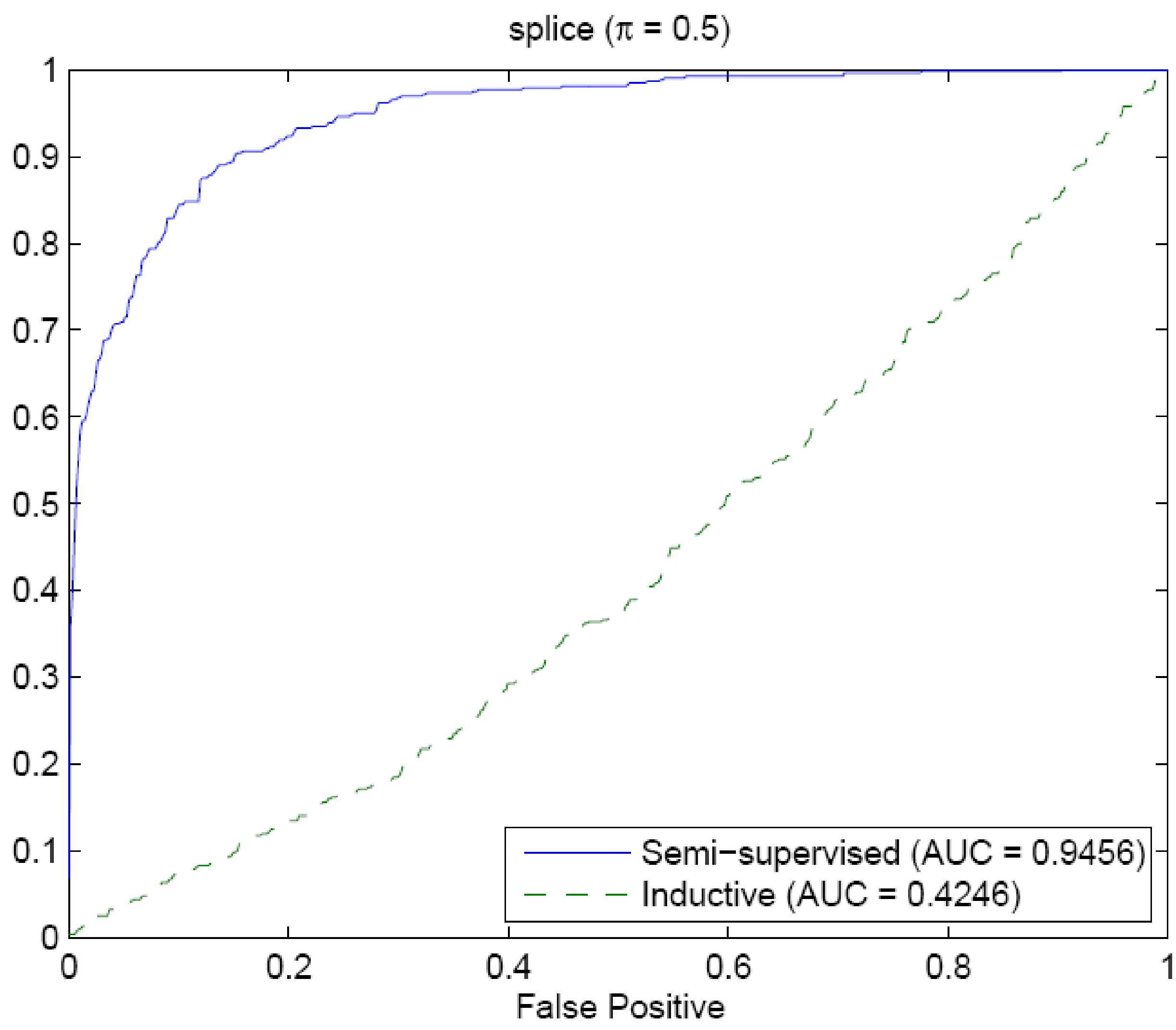


Unlabeled test data

Claim: We can achieve



Benchmark Data



Estimating Performance

- Even if we can find an optimal classifier in a WSL problem by choosing an appropriate performance measure, we can't necessarily estimate its performance.

- Example: Learning from negative and unlabeled data

$$X_1, \ldots, X_m \stackrel{iid}{\sim} P_0$$

$$X_{m+1}, \ldots, X_{m+n} \stackrel{iid}{\sim} \tilde{P}_1 = (1-\pi)P_1 + \pi P_0$$

- Need to know π to estimate $R_1(f)$

Mixture Proportion Estimation

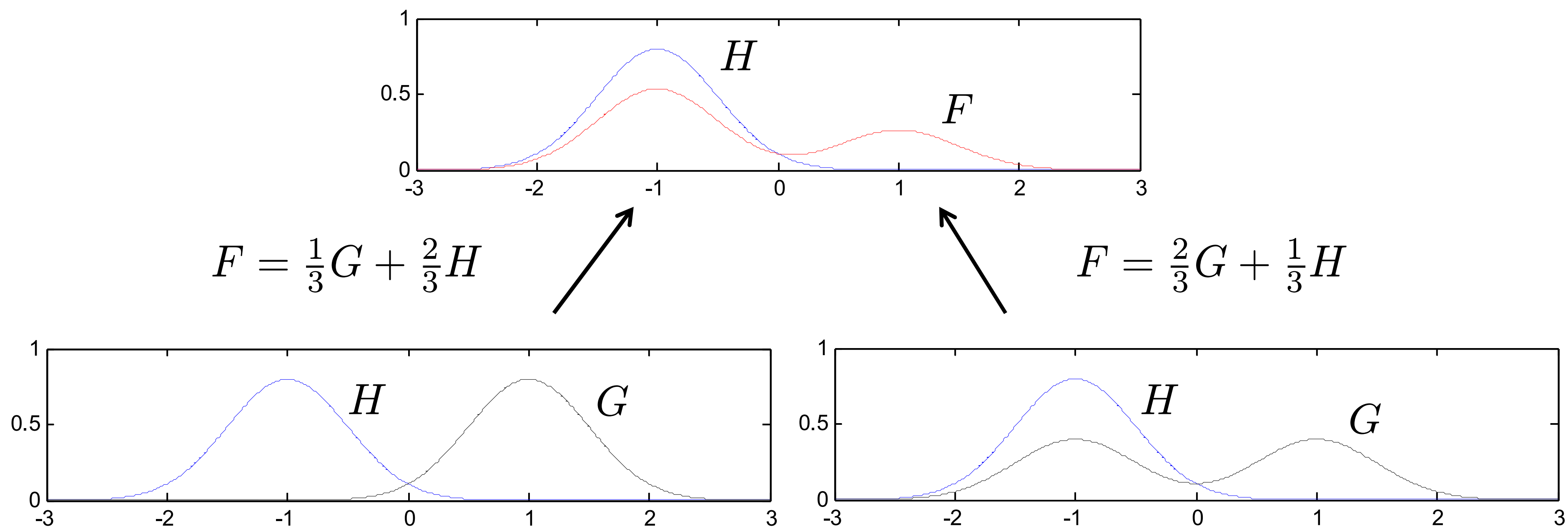
- Consider

$$Z_1, \ldots, Z_m \overset{iid}{\sim} H$$

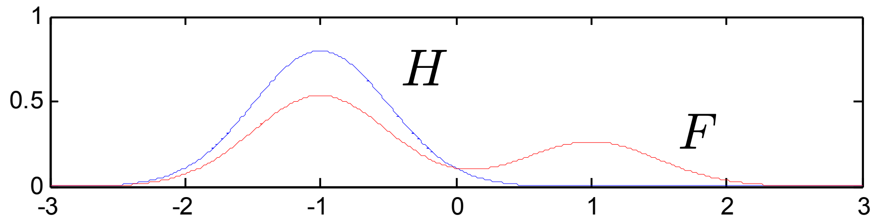
$$Z_{m+1}, \ldots, Z_{m+n} \overset{iid}{\sim} F = (1-\kappa)G + \kappa H$$

- Need consistent estimate of κ

- Note: κ not identifiable in general



Mixture Proportion Estimation



- Given two distributions F, H, define

$$\kappa^*(F|H) = \max\{\alpha \in [0,1] : \exists G' \text{ s.t. } F = (1-\alpha)G' + \alpha H\}$$

- κ^* can be estimated – stay tuned
- When is $\kappa = \kappa^*(F|H)$?

Identifiability Condition

- If

$$F = (1-\kappa)G + \kappa H$$

then

$$\kappa = \kappa^*(F\,|H) \Longleftrightarrow \boxed{\kappa^*(G\,|H) = 0}$$

- Apply to LNUE

$$X_1, \ldots, X_m \overset{iid}{\sim} P_0$$

$$X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} \tilde{P}_1 = (1-\pi)P_1 + \pi P_0$$

- Need

$$\kappa^*(P_1\,|P_0) = 0$$

In words: Can't write P_1 as a (nontrivial) mixture of P_0 and some other distribution

Label Noise Proportion Estimation

- Recall contaminaton model:

$$X_1, \ldots, X_m \overset{iid}{\sim} \tilde{P}_0 = (1-\pi_0)P_0 + \pi_0 P_1$$
$$X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} \tilde{P}_1 = (1-\pi_1)P_1 + \pi_1 P_0$$

- **Proposition:** If $\pi_0 + \pi_1 < 1$ and $P_0 \neq P_1$, then

$$\tilde{P}_0 = (1-\tilde{\pi}_0)P_0 + \tilde{\pi}_0 \tilde{P}_1$$
$$\tilde{P}_1 = (1-\tilde{\pi}_1)P_1 + \tilde{\pi}_1 \tilde{P}_0$$

where

$$\tilde{\pi}_0 = \frac{\pi_0}{1-\pi_1}, \quad \tilde{\pi}_1 = \frac{\pi_1}{1-\pi_0}$$

MPE for Label Noise

- Modified contamination model

$$X_1, \ldots, X_m \overset{iid}{\sim} \tilde{P}_0 = (1 - \tilde{\pi}_0) P_0 + \tilde{\pi}_0 \tilde{P}_1$$

$$X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} \tilde{P}_1 = (1 - \tilde{\pi}_1) P_1 + \tilde{\pi}_1 \tilde{P}_0$$

- Need consistent estimates of $\tilde{\pi}_0$, $\tilde{\pi}_1$ $\Rightarrow$ MPE

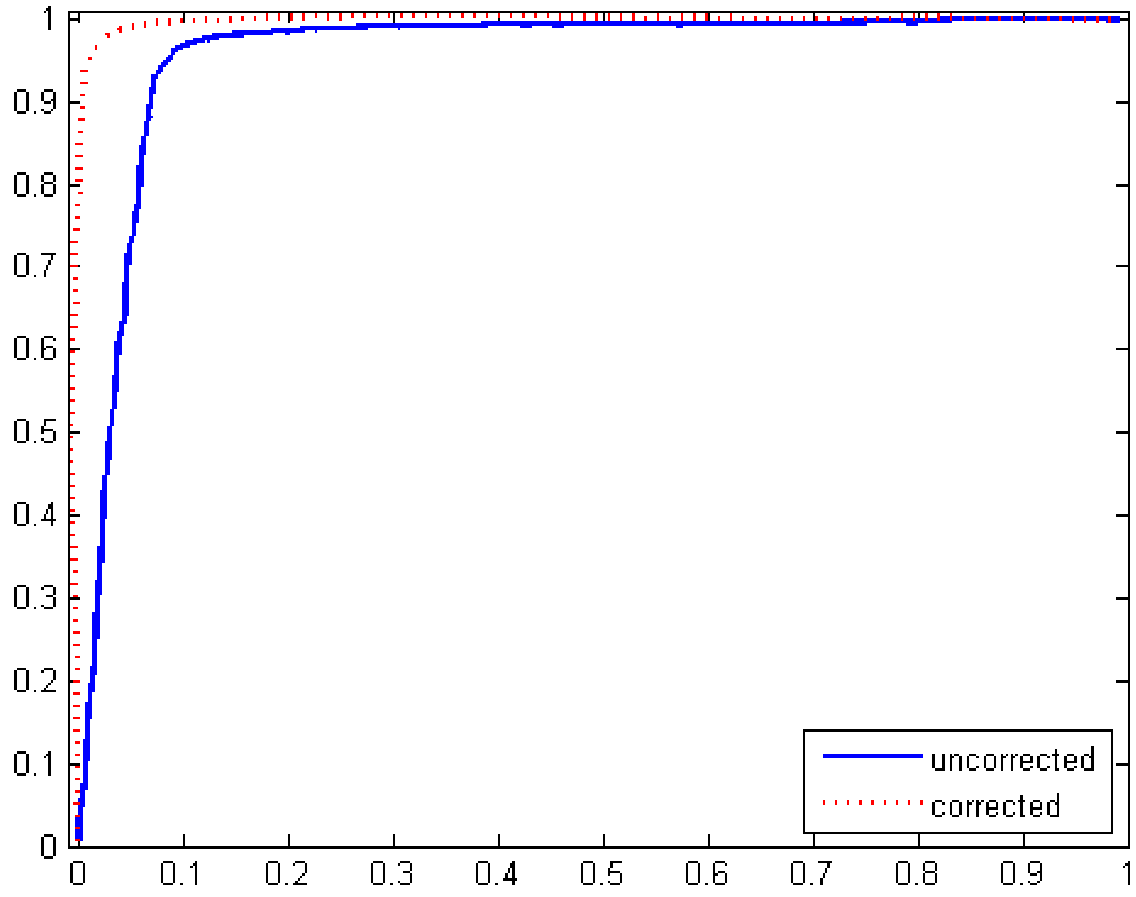
- Identifiability: Need

$$\kappa^*(P_0 \,|\, \tilde{P}_1) = 0 \text{ and } \kappa^*(P_1 \,|\, \tilde{P}_0) = 0$$

or equivalently (it can be shown)

$$\kappa^*(P_0 \,|\, P_1) = 0 \text{ and } \kappa^*(P_1 \,|\, P_0) = 0$$

Effect on Performance Estimate



Approaches to Mixture Prop. Est.

- Plug-in
- ROC slope
- Class probability estimation
- Kernel mean embedding

MPE: Density Ratio Formulation

- Key observation: For any F, H

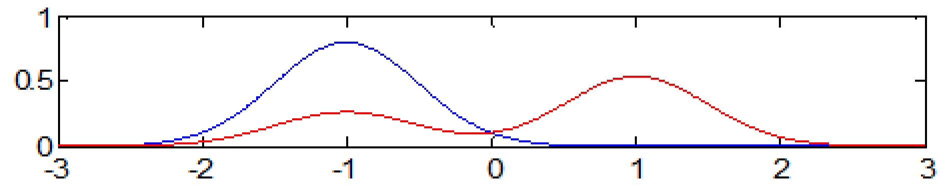
$$\kappa^*(F \mid H) = \inf_{A : H(A) > 0} \frac{F(A)}{H(A)}$$

- Proof: κ^* is the largest κ such that

$$G = \frac{F - \kappa H}{1 - \kappa}$$

is a distribution.

- Similarly, if F and H have densities f and h, then



$$\kappa^*(F \mid H) = \operatorname*{ess\,inf}_{x : h(x) > 0} \frac{f(x)}{h(x)}$$

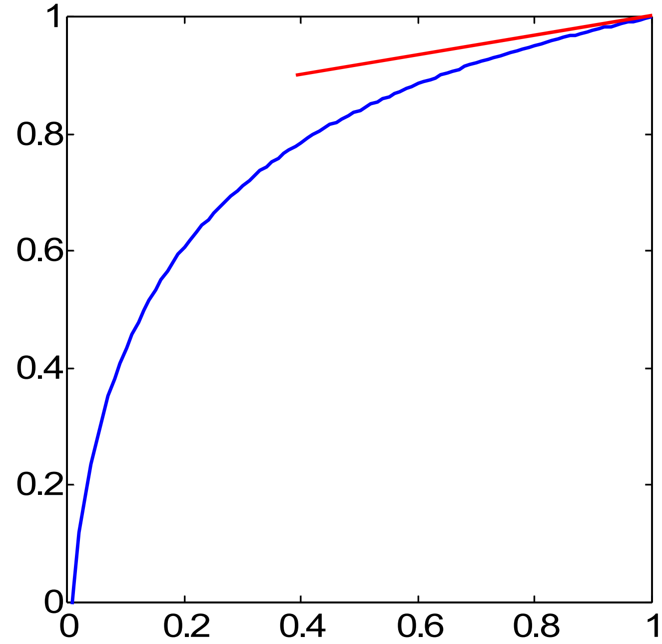
- Universally consistent estimator established by Blanchard et al. (2010)

ROC Method

- Rewrite previous identity as (substituting $A \to A^c$)

$$\kappa^*(F \mid H) = \inf_{A: H(A) < 1} \frac{1 - F(A)}{1 - H(A)}$$

- Slope of ROC at its right endpoint



Class Probability Estimation

- Assume joint distribution on (X, Y), $Y = 0, 1$, where

$$
\begin{aligned}
X|Y = 1 &\sim F \\
X|Y = 0 &\sim H
\end{aligned}
$$

- Prior / posterior class probabilities

$$
\begin{aligned}
\theta &:= \Pr(Y = 1) \\
\eta(x) &:= \Pr(Y = 1 \,|\, X = x)
\end{aligned}
$$

- By a simple application of Bayes rule,

$$
\eta_{\max} := \sup_x \eta(x) = \frac{1}{1 + \frac{1-\theta}{\theta}\kappa^*(F \,|\, H)}
$$

- Menon et al. (2015), Liu and Tao (2016).

Additional WSL Problems

- Multiclass extensions of the preceding
- Classification with reject option
- Learning with partial labels
- Multiple instance learning
- Semi-supervised learning (reduces to classification with label noise under co-training assumption)
- . . .

Summary of Part 2

- Some performance measures are ideally suited to certain WSL problems
- To actually estimate the performance can require additional work
- Are some performance measures well-suited for more general types of domain adaptation?
- **Bottom Line**: For many WSL problems, we can do as well as in the fully supervised setting

Some Related Work

LNUE: Liu et al. (2002), Denis et al. (2005), Elkan and Noto (2008), Ward et al. (2009), Smola et al. (2009), Goernitz et al. (2013)

MPE: du Plessis and Sugiyama (2013, 2015), Jain et al. (2016)

Label noise: Long and Servido (2010), Natarajan et al. (2013), Menon et al. (2015), Liu and Tao (2016), van Rooyen et al. (2015), Patrini et al. (2016)

Multiple hypothesis testing: Genovese and Wasserman (2004)

Feature-dependent label noise: Urner, Ben-David and Shamir (2012)

Multiple instance learning: Sabato and Tishby (2012)

Learning from label proportions: Patrini et al. (2014)

Some of My Papers

Neyman-Pearson Classification: Trans. IT 2006

Semi-supervised novelty detection: JMLR 2010

Cost-sensitive loss functions: Electronic J. Statistics 2012

Classification with Label Noise: COLT 2013, AISTATS 2014, AISTATS 2015, Electronic J. Statistics 2016

Mixture proportion estimation: ICML 2016

Collaborators

- Gilles Blanchard
- Gregory Handy, Tyler Sanderson
- Marek Flaska, Sara Pozzi
- Harish Ramaswamy, Ambuj Tewari

Supported in part by NSF