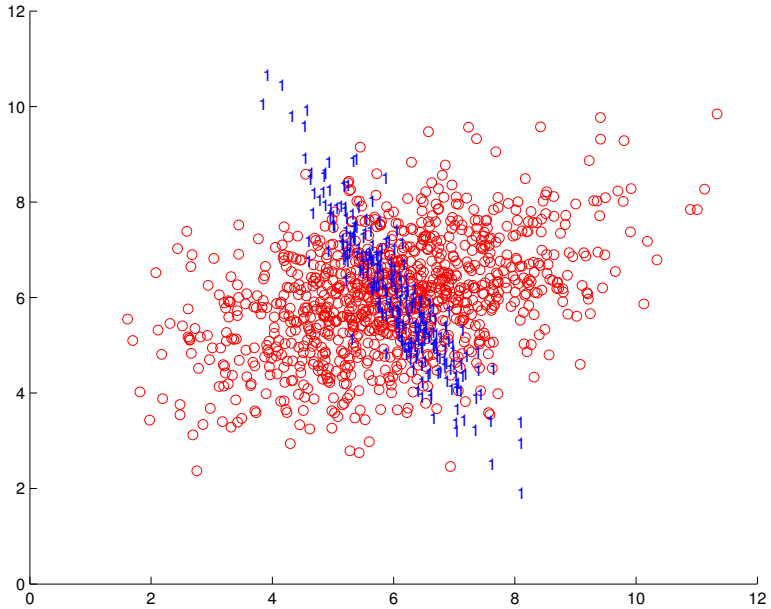
Semi-Supervised Novelty Detection

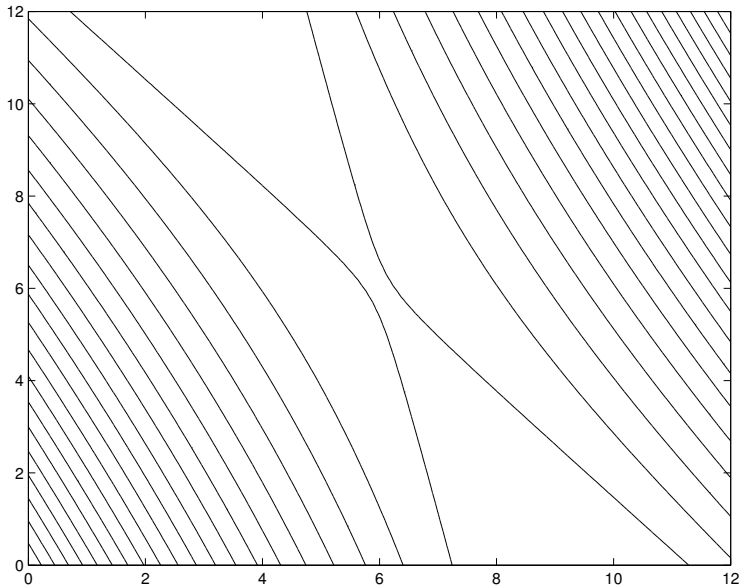
Clayton Scott

EECS and Statistics
University of Michigan

Collaborators: Gilles Blanchard, Gyemin Lee

Classification



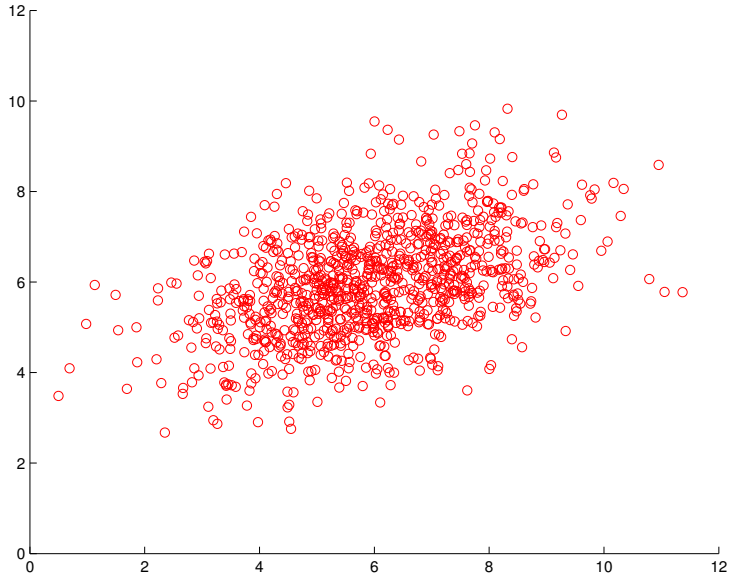


Many nonparametric methods:
Nearest neighbors, decision trees,
support vector machines, etc.

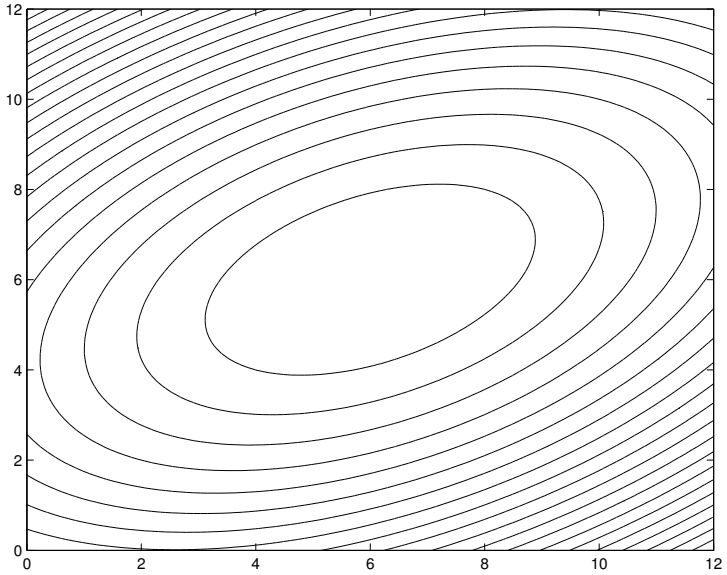
Optimal classifier:

$$\lambda \gtrless \frac{f_1(x)}{f_0(x)}$$

Novelty Detection



Nominal data only

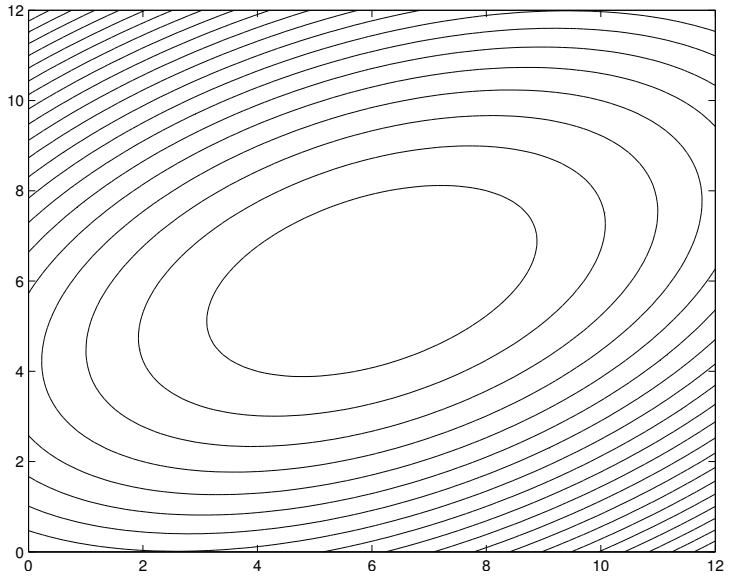


Typical approach: estimate a **level set** of the nominal density

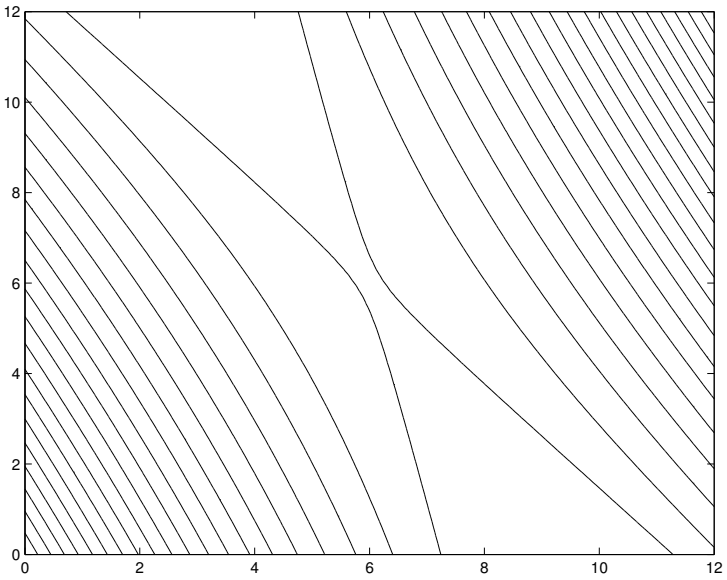
$$\lambda \gtrless f_0(x)$$

Nonparametric methods:
thresholded KDE, one-class SVM

Problem with Level Set Approach



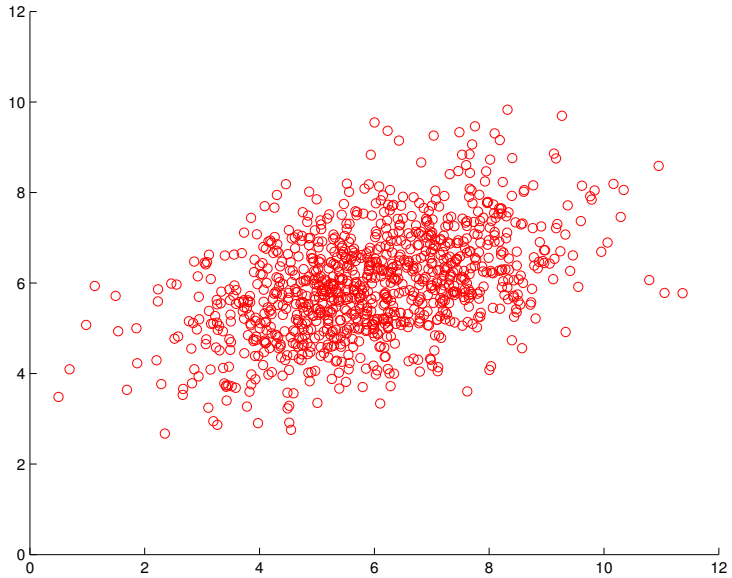
≠



The more f_1 overlaps f_0, the bigger the problem

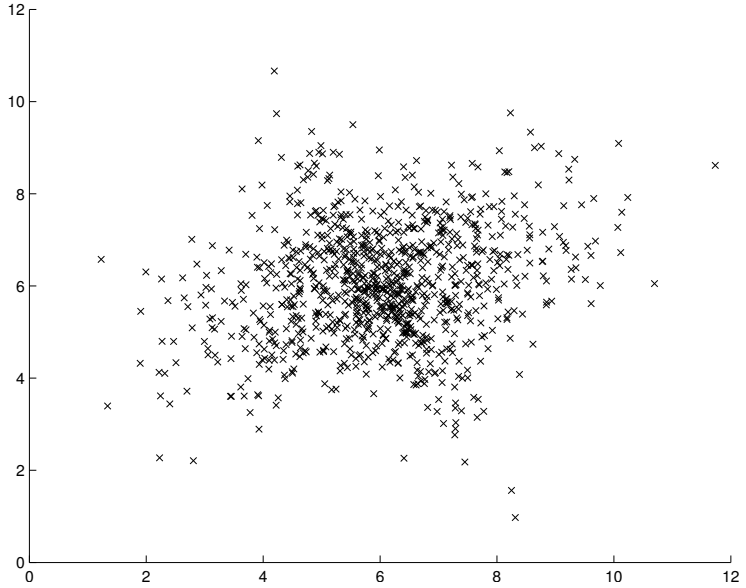
Semi-Supervised Novelty Detection

Suppose you observe



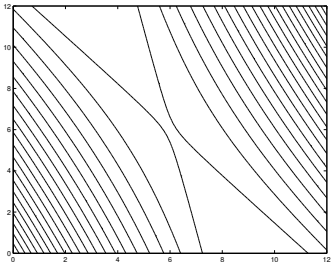
Nominal data

and



Unlabeled data

Claim: We can approach



Applications:

- Document classification
- Manufacturing (quality control)
- Proteomics

Problem Formulation

- $X_1, \ldots, X_m \sim f_0(x)$

Nominal component

Novel component

- $X_{m+1}, \ldots, X_{m+n} \sim f_\times(x) = (1-\pi) f_0(x) + \pi f_1(x)$
- $X_i \in \mathbf{R}^d$
- No prior knowledge on f_0, f_1, $f_\times$, π
- **Goal**: Predict label of $\begin{cases} X_{m+1}, \ldots, X_{m+n} \\ \text{arbitrary } X \end{cases}$ with preformance approaching

$$\lambda \gtrless \frac{f_1(x)}{f_0(x)}$$

as $m, n \to \infty$.

If Distributions Are Known (Let's Suppose)

Problem of interest

$$\begin{aligned} H_0 : X &\sim f_0 \\ H_1 : X &\sim f_1 \end{aligned} \quad \Longrightarrow \quad \lambda \gtrless \frac{f_1(x)}{f_0(x)}$$

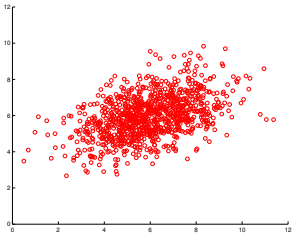
Surrogate problem

$$\begin{aligned} H_0 : X &\sim f_0 \\ H_\times : X &\sim f_\times \end{aligned} \quad \Longrightarrow \quad \lambda \gtrless \frac{f_\times(x)}{f_0(x)} = \frac{(1-\pi) f_0(x) + \pi f_1(x)}{f_0(x)}$$

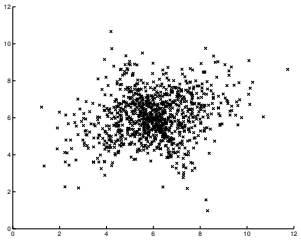
$$= (1-\pi) + \pi \frac{f_1(x)}{f_0(x)}$$

$$\Longrightarrow \lambda' \gtrless \frac{f_1(x)}{f_0(x)}$$

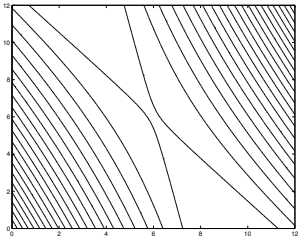
Surrogate LR is monotone function of optimal test statistic → UMP test



vs.



→



Data-Based Approaches

- f_0, $f_\times$ unknown, but we have *random samples*

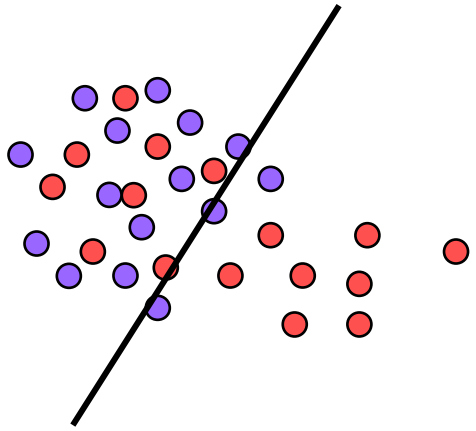
$$\begin{aligned} H_0 &: X_1, \ldots, X_m \sim f_0 \\ H_\times &: X_{m+1}, \ldots, X_{m+n} \sim f_\times \end{aligned}$$

- **Plug-in approach**: Estimate densities and plug in to likelihood ratio test:

$$\lambda \gtrless \frac{\widehat{f}_\times(x)}{\widehat{f}_0(x)}$$

  Choose λ to achieve desired false positive rate

- **Classification approach**: View $X_1, \ldots, X_m$ and $X_{m+1}, \ldots, X_{m+n}$ as input for a binary classification algorithm



Neyman-Pearson Classification

Problem statement: Given training data

$$\begin{aligned} X_1, \ldots, X_m &\sim f_0 \\ X_{m+1}, \ldots, X_{m+n} &\sim f_\times \end{aligned}$$

and α, $0 < \alpha < 1$, construct a classifier g such that

$$P_{FP}(g) := Pr\{g(X) = 1 \,|\, X \sim f_0\} \leq \alpha$$

and

$$P^{\times}_{FN}(g) := Pr\{g(X) = 0 \,|\, X \sim f_\times\}$$

is as small as possible

Learning Reduction

Notation:

$$\begin{aligned} P_{FP}(g) &= Pr\{g(X) = 1 \,|\, X \sim f_0\} \\ P^1_{FN}(g) &= Pr\{g(X) = 0 \,|\, X \sim f_1\} \\ P^\times_{FN}(g) &= Pr\{g(X) = 0 \,|\, X \sim f_\times\} \end{aligned}$$

Theorem: Assume $\pi > 0$. If g is any classifier such that

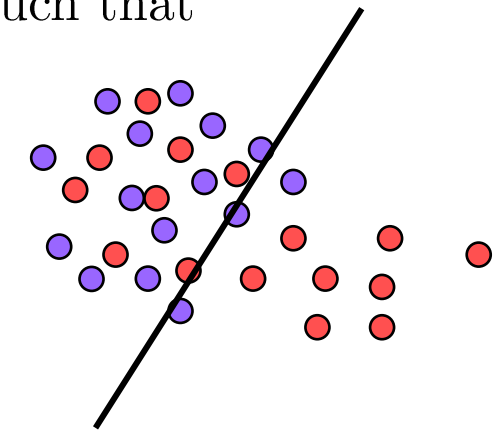
$$P_{FP}(g) \leq \alpha + \epsilon_0$$

and

$$P^\times_{FN}(g) \leq P^\times_{FN}(g^*_\alpha) + \epsilon_\times,$$

then

$$P^1_{FN}(g) \leq P^1_{FN}(g^*_\alpha) + \frac{1}{\pi}(\epsilon_\times + (1-\pi)\epsilon_0)$$



optimal likelihood ratio test (from Neyman-Pearson lemma)

Neyman-Pearson Classification Theory

Illustrative result: Let $\mathcal{G} = \{g_1, \ldots, g_M\}$ be a set of classifiers. Fix $\epsilon_0, \epsilon_\times > 0$. Define

$$\begin{aligned} \widehat{g} = \arg\min_{g \in \mathcal{G}} \quad & \widehat{P}^{\times}_{FN}(g) \\ \text{s.t.} \quad & \widehat{P}_{FP}(g) \leq \alpha + \tfrac{1}{2}\epsilon_0 \end{aligned}$$

With probability at least $1 - 2M(e^{-m\epsilon_0^2/2} + e^{-n\epsilon_\times^2/2})$,

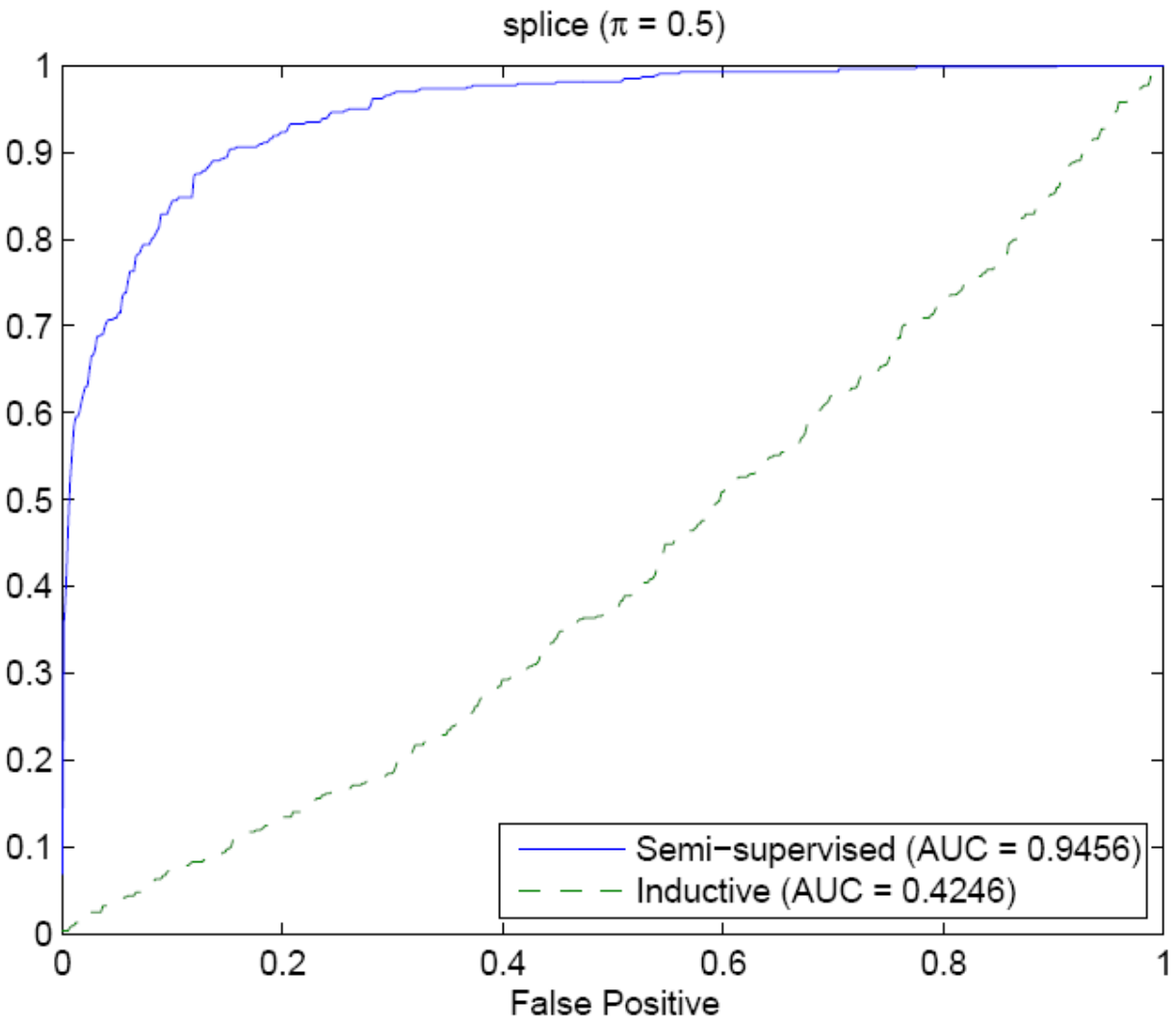
$$P_{FP}(\widehat{g}) \leq \alpha + \epsilon_0$$

and

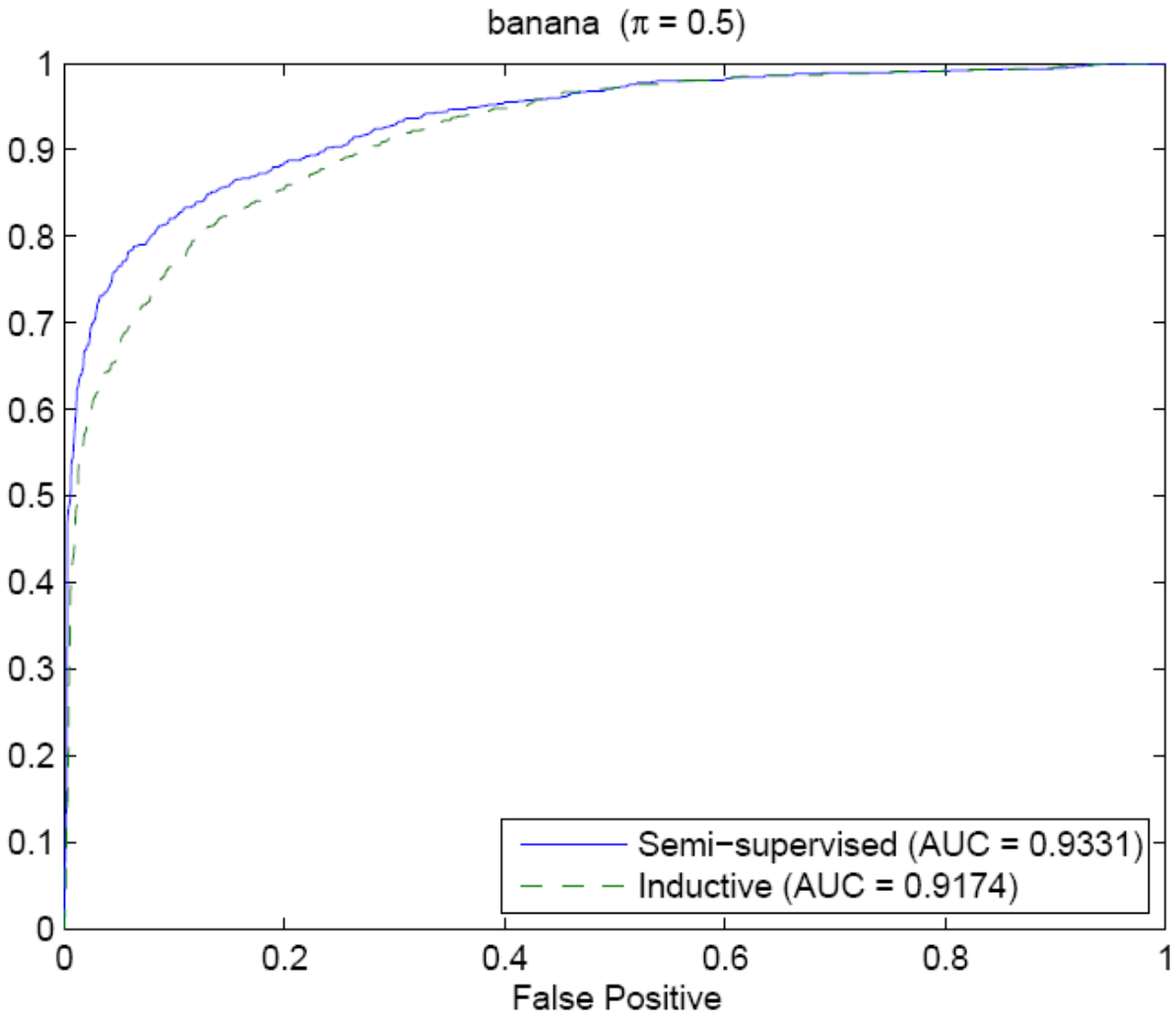
$$P^{\times}_{FN}(\widehat{g}) \leq \min_{\substack{g \in \mathcal{G} \\ P_{FP}(g) \leq \alpha}} P^{\times}_{FN}(g) + \epsilon_\times$$

Extensions: VC classes, consistency, rates of convergence (Cannon et al., 2002, Scott and Nowak, 2005)

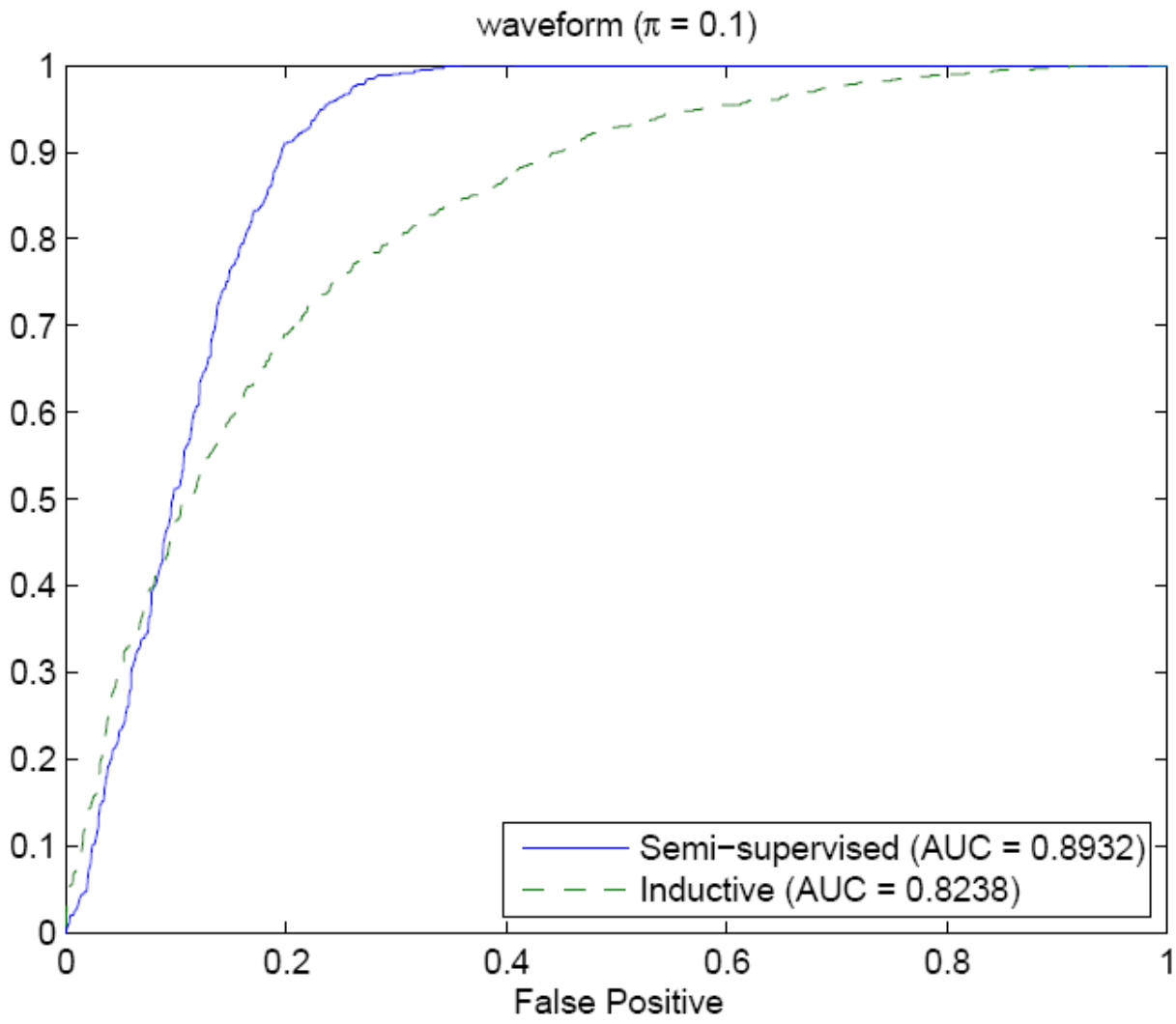
Benchmark Data



Benchmark Data



Benchmark Data



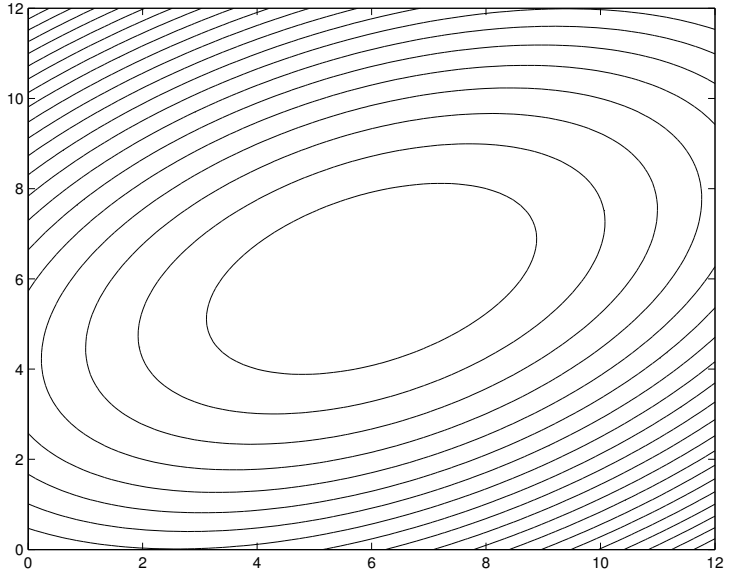
Inductive vs. Semi-Supervised

Inductive: Implicitly assumes novelties are **uniform**

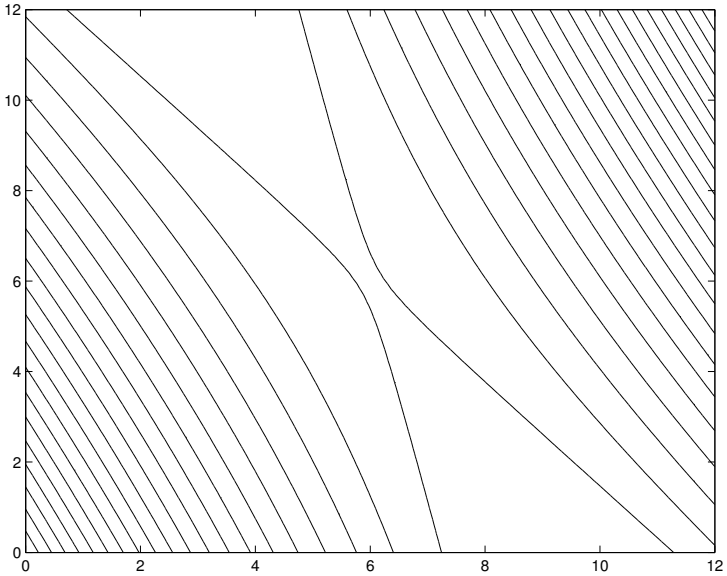
$$\lambda \geqslant f_0(x) \Leftrightarrow \frac{1}{\lambda} \leqslant \frac{1}{f_0(x)}$$

Semi-supervised: Learns novelty distribution from unlabeled data

$$\gamma \leqslant \frac{f_1(x)}{f_0(x)}$$



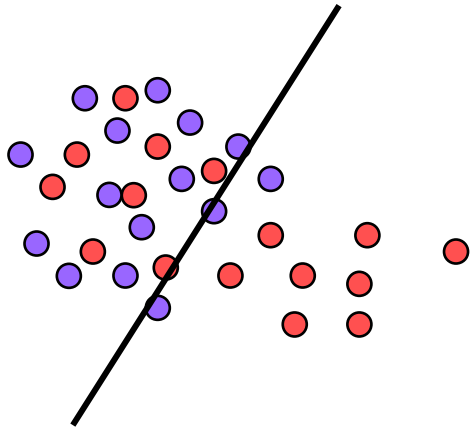
$\neq$



Benefits of Unlabeled Data

In binary classification: Not needed for consistency, rates, bounds. In some special cases, rates can be improved.

In novelty detection: Makes consistency, rates, tight bounds possible



What about π?

- **Testing** $\pi = 0$: Are there any novelties present?
- **Estimating** π: How many novelties are present?

Identifiability assumption: In the representation

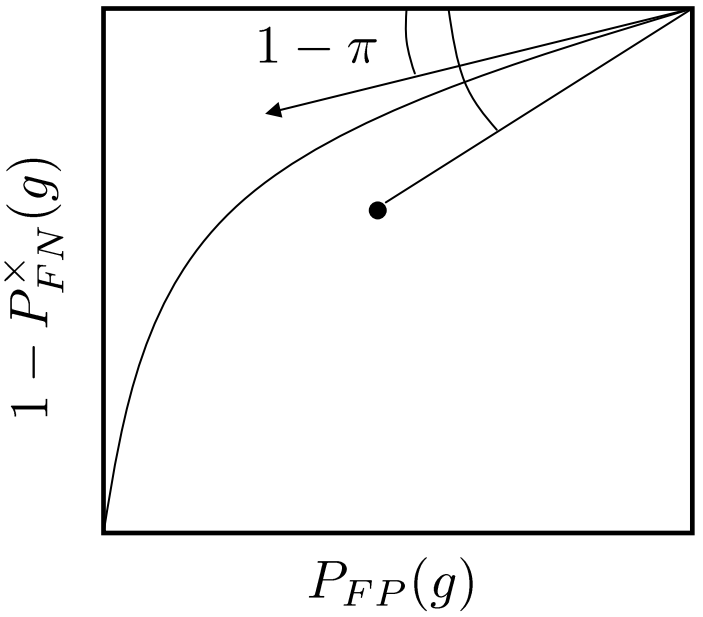
$$P_{\times} = (1 - \pi)P_0 + \pi P_1,$$

π is **minimal**.

Estimating π

Claim: For all classifiers g,

$$1 - \pi \leq \frac{P_{FN}^{\times}(g)}{1 - P_{FP}(g)}$$



$$\begin{aligned} 1 - P_{FN}^{\times}(g) = P_{\times}(g(X) = 1) &= (1-\pi)P_0(g(X) = 1) + \pi P_1(g(X) = 1) \\ &= (1-\pi)P_{FP}(g) + \pi(1 - P_{FN}^{1}(g)) \\ &\leq (1-\pi)P_{FP}(g) + \pi \end{aligned}$$

Furthermore,

$$\pi = \max_{\text{all } g} \left[1 - \frac{P_{FN}^{\times}(g)}{1 - P_{FP}(g)} \right]$$

Estimating π: Two approaches

- For the optimal receiver operating characterstic (for P_0 vs. $P_\times$),

$$\pi = 1 - \left(\frac{\partial}{\partial \alpha} ROC(\alpha) \right) \Big|_{\alpha=1-}$$

  Under certain conditions, ROC is concave $\Longrightarrow$ similar to monotone density estimation

- Fix a set of classifiers $\mathcal{G}$ and estimate

$$\widehat{\pi} = \max_{g \in \mathcal{G}} \; 1 - \frac{P^{\times}_{FN}(g)}{1 - P_{FP}(g)}$$

Estimation: One-sided Confidence Interval

Illustrative result: Let $\mathcal{G} = \{g_1, \ldots, g_M\}$ be a set of classifiers. Fix $\epsilon_0, \epsilon_\times > 0$. Define

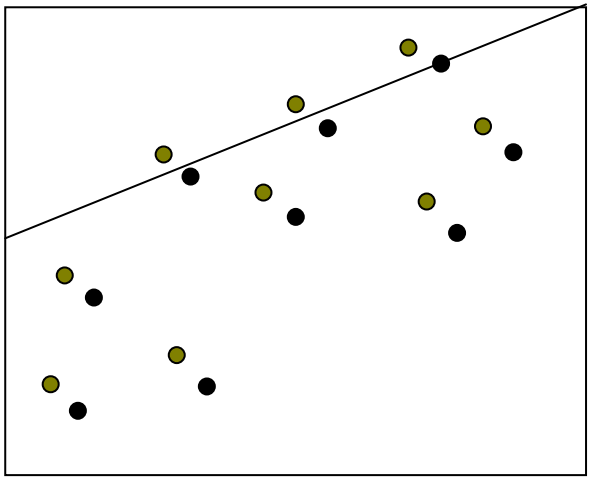
$$\widehat{\pi} = \max_{g \in \mathcal{G}} \frac{1 - \widehat{P}_{FP}(g) - \epsilon_0 - \widehat{P}^{\times}_{FN}(g) - \epsilon_\times}{(1 - \widehat{P}_{FP}(g) - \epsilon_0)_+}$$

With probability at least $1 - 2M(e^{-m\epsilon_0^2/2} + e^{-n\epsilon_\times^2/2})$,

$$\widehat{\pi} \leq \pi$$

Universal consistency: If $\mathcal{G} = \mathcal{G}_n$ such that $\{\mathcal{G}_n\}$ has the universal approximation property, and $\log |\mathcal{G}_n|/n \to 0$, then

$$\widehat{\pi} \to \pi \quad a.s.$$



Testing π = 0

Corollary: With high probability, if $\widehat{\pi} > 0$, then $\pi > 0$.

This can also be viewed as a solution to the two sample problem: Given

$$\begin{aligned} X_1, \ldots, X_m &\sim P_0 \\ X_{m+1}, \ldots, X_{m+n} &\sim P_\times \end{aligned}$$

Decide

$$\begin{aligned} H_0 &: \pi = 0 \quad [P_\times = P_0] \\ H_\times &: \pi > 0 \quad [P_\times = (1-\pi)P_0 + \pi P_1] \end{aligned}$$

Multiple Testing

- Null hypotheses $H_1, H_2, \ldots, H_n$
- Compute test statistics $Y_1, Y_2, \ldots, Y_n$
- Transform to p-values $p_1, p_2, \ldots, p_n$
- Threshold (possibly data dependent)

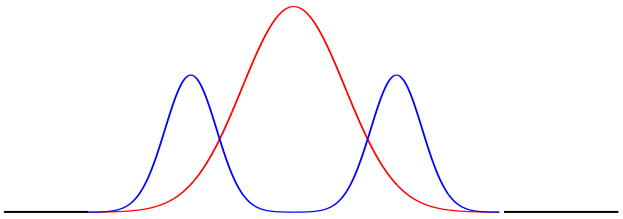
$$p_i \gtrless \widehat{T}$$

Multiple Testing and SSND

- If H_i true, then $p_i \sim U[0,1]$
- *Random effects* model:

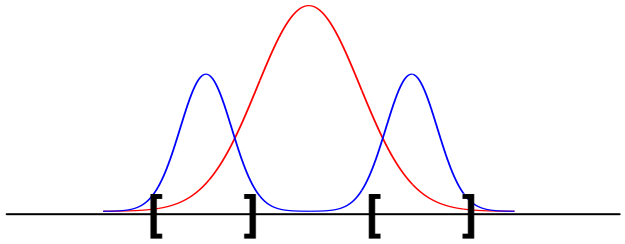
$$p_i \overset{iid}{\sim} (1-\pi)U[0,1] + \pi P_1$$

- This is a specification of the SSND model where
  - The observation space is $[0,1]$
  - $P_0 = U[0,1]$ is known exactly
  - Classifiers based on intervals of the form $[0,t]$



Only optimal if P_1 monotone decreasing
(easy to find counterexamples)

Multiple Testing and SSND



- With SSND, we can
  - Allow more general classifiers (e.g. union of disjoint intervals)
  - Directly use multidimensional test statistics
  - Leverage examples from P_0 if available

  and inheret the aforementioned theoretical guarantees
- Related work: Sun and Cai (2007), Chi (2007, 2008)
- Open questions: Practical implementation, performance guarantees for FDR

Adaptation to Unknown π

In theory: Test $\pi = 0$

- $\widehat{\pi} > 0 \Longrightarrow$ semi-supervised approach
- $\widehat{\pi} \leq 0 \Longrightarrow$ inductive approach

In practice: Bounds are too loose

Hybrid Inductive/Semi-Supervised Approach

- **Recall**: Inductive approach $\Longleftrightarrow$ classification of nominal sample against a uniform sample
- **Basic idea**: Append a uniformly distributed sample to the unlabeled data

$$\underbrace{X_{m+1}, \ldots, X_{m+n}}_{\text{unlabeled data}}, \underbrace{X_{m+n+1}, \ldots, X_{m+n+k}}_{\text{uniform data}}$$

Now classifiy

$$X_1, \ldots, X_m$$

against

$$X_{m+1}, \ldots, X_{m+n}, X_{m+n+1}, \ldots, X_{m+n+k}$$

Hybrid Inductive/Semi-Supervised Approach

Theorem: If $k \to \infty, k/n \to 0$ as $m, n \to \infty$, then

$$\text{hybrid method} \to \begin{cases} \text{inductive method} & \text{if } \pi = 0 \\ \text{semi-supervised method} & \text{if } \pi > 0 \end{cases}$$

Conclusions

- How does unlabeled data impact novelty detection?
- Inductive approach:
  - reduces to density level set estimation
  - assumes novelties are uniformly distributed
- Semi-supervised approach:
  - reduces to Neyman-Pearson classification
  - learns correct novelty distribution from unlabeled data
- SSND offers general framework for two-sample problem and multiple testing