Weierstraß-Institut für Angewandte Analysis und Stochastik

Gilles Blanchard

(joint work with C. Scott and G. Lee, U. of Michigan)

From semi-supervised novelty detection to multiple testing

Plan

Is binary classification *exactly the same* as MT?

- each test decision can be seen as classification and vice-versa
- the popular iid mixture model for p-values: $h_i \sim B(1-\pi_0)$,

$$p_i \sim \begin{cases} U([0,1]) & \text{if } h_i = 0\,, \\ P_1 & \text{if } h_i = 1\,, \end{cases}$$

is exactly the generating model of the classification setting with observations (p_i) and labels (h_i).

Is binary classification *exactly the same* as MT?

- each test decision can be seen as classification and vice-versa
- the popular iid mixture model for p-values: $h_i \sim B(1-\pi_0)$,

$$p_i \sim \begin{cases} U([0,1]) & \text{if } h_i = 0\,, \\ P_1 & \text{if } h_i = 1\,, \end{cases}$$

is exactly the generating model of the classification setting with observations (p_i) and labels (h_i).

Is binary classification *exactly the same* as MT?

- in MT: classes do not play a symmetric role
- in MT: P_0 is (generally) assumed to be known ($U([0,1]$ for p-values)
- in MT: no direct observations of the alternate class (data is unlabeled)
- performance measures differ (misclassification error vs. type I error constraint + power optimization)

A common ground

In both settings, and with all common quality criteria in use, the optimal decision regions are always given by superlevel sets of the likelihood ratio function

$$\frac{dP_1}{dP_0}(x)$$

In the special case of p-values, $P_0 = U([0,1])$, decision regions of the form $\mathbf{1}\{p \leq t\}$ are optimal if the above likelihood ratio is decreasing, i.e. the cdf F_1 of P_1 is concave.

A common ground

In both settings, and with all common quality criteria in use, the optimal decision regions are always given by superlevel sets of the likelihood ratio function

$$\frac{dP_1}{dP_0}(x)$$

In the special case of p-values, $P_0 = U([0,1])$, decision regions of the form $\mathbf{1}\{p \leq t\}$ are optimal if the above likelihood ratio is decreasing, i.e. the cdf F_1 of P_1 is concave.

Classification performance scores

- Classical performance measure for classification: misclassification error $P(f(X) \neq Y)$.
- 'Neyman-Pearson classification' (Scott and Nowak, 2005): constraint on error rate on class 0 (i.e. type I error rate or false positive rate) + error optimization on class 1
- (When learning score functions) optimizing the area under the ROC (= sensitivity/specificity curve)

Novelty detection (1-class classification)

- observed: $(X_1, \ldots, X_n)$ iid $\sim P_0$
- goal: detection function $f : \mathcal{X} \to \{0, 1\}$ detecting 'atypical' data in new incoming observations
- constraint (in general): false positive rate level α
- in practice: most methods explicitly or implicitly estimate sublevel sets of $dP_0/d\lambda$ for a reference measure λ (e.g. Lebesgue if $\mathcal{X}$ compact subset of $\mathbb{R}^d$)
- equivalent to implicitly assuming that 'novelties' are distributed according to λ
- in that setting: P_0 unknown, P_1 'known'.

Novelty detection (1-class classification)

- observed: $(X_1, \ldots, X_n)$ iid $\sim P_0$
- goal: detection function $f : \mathcal{X} \to \{0, 1\}$ detecting 'atypical' data in new incoming observations
- constraint (in general): false positive rate level α
- in practice: most methods explicitly or implicitly estimate sublevel sets of $dP_0/d\lambda$ for a reference measure λ (e.g. Lebesgue if $\mathcal{X}$ compact subset of $\mathbb{R}^d$)
- equivalent to implicitly assuming that 'novelties' are distributed according to λ
- in that setting: P_0 unknown, P_1 'known'.

Novelty detection (1-class classification)

- observed: $(X_1, \ldots, X_n)$ iid $\sim P_0$
- goal: detection function $f : \mathcal{X} \to \{0, 1\}$ detecting 'atypical' data in new incoming observations
- constraint (in general): false positive rate level α
- in practice: most methods explicitly or implicitly estimate sublevel sets of $dP_0/d\lambda$ for a reference measure λ (e.g. Lebesgue if $\mathcal{X}$ compact subset of $\mathbb{R}^d$)
- equivalent to implicitly assuming that 'novelties' are distributed according to λ
- in that setting: P_0 unknown, P_1 'known'.

Plan

The Semi-supervised novelty detection (SSND) setting

In a nutshell: novelty detection setup where in addition to a reference sample, a second sample containing some novelties is available at learn time.

Observed:

- a nominal iid sample $X_1, \ldots, X_m \sim P_0$
- a contaminated iid sample $X_{m+1}, \ldots, X_{m+n} \sim P_X = \pi_0 P_0 + (1 - \pi_0) P_1$.
- for convenience we assume an unobserved label Y_i indicating for each example X_i whether it was drawn from P_0 or P_1.

Also called 'Learning from positive and unlabeled examples' (LPUE) in a part of the classification literature.

The Semi-supervised novelty detection (SSND) setting

In a nutshell: novelty detection setup where in addition to a reference sample, a second sample containing some novelties is available at learn time.

Observed:

- a nominal iid sample $X_1, \ldots, X_m \sim P_0$
- a contaminated iid sample $X_{m+1}, \ldots, X_{m+n} \sim P_X = \pi_0 P_0 + (1 - \pi_0) P_1$.
- for convenience we assume an unobserved label Y_i indicating for each example X_i whether it was drawn from P_0 or P_1.

Also called 'Learning from positive and unlabeled examples' (LPUE) in a part of the classification literature.

Notation for error criteria

- for a classification function $f : \mathcal{X} \to \{0, 1\}$ define

$$\mathcal{E}_0(f) = P_0(f(X) = 1) = P(f(X) = 1 | Y = 0)$$
$$\mathcal{E}_1(f) = P_1(f(X) = 0) = P(f(X) = 0 | Y = 1)$$

  the false positive and false negative rates, and

$$\mathcal{E}_X(f) = P_X(f = 0) = \pi_0(1 - \mathcal{E}_0(f)) + (1 - \pi_0)\mathcal{E}_1(f)\,.$$

- Goal: least FNR under a FPR constraint,

$$\mathcal{E}^*_{1,\alpha} := \inf_f \mathcal{E}_1(f) \text{ s.t. } \mathcal{E}_0(f) \leq \alpha\,.$$

- the above is the *semi-supervised* setting;
- *transductive* setting: replace $\mathcal{E}_1(f)$ by its empirical realization on the contaminated sample (requires label information, hence not directly accessible to learner).

Notation for error criteria

- for a classification function $f : \mathcal{X} \to \{0, 1\}$ define

$$\mathcal{E}_0(f) = P_0(f(X) = 1) = P(f(X) = 1|Y = 0)$$
$$\mathcal{E}_1(f) = P_1(f(X) = 0) = P(f(X) = 0|Y = 1)$$

  the false positive and false negative rates, and

$$\mathcal{E}_X(f) = P_X(f = 0) = \pi_0(1 - \mathcal{E}_0(f)) + (1 - \pi_0)\mathcal{E}_1(f)\,.$$

- Goal: least FNR under a FPR constraint,

$$\mathcal{E}^*_{1,\alpha} := \inf_f \mathcal{E}_1(f) \text{ s.t. } \mathcal{E}_0(f) \leq \alpha\,.$$

- the above is the *semi-supervised* setting;
- *transductive* setting: replace $\mathcal{E}_1(f)$ by its empirical realization on the contaminated sample (requires label information, hence not directly accessible to learner).

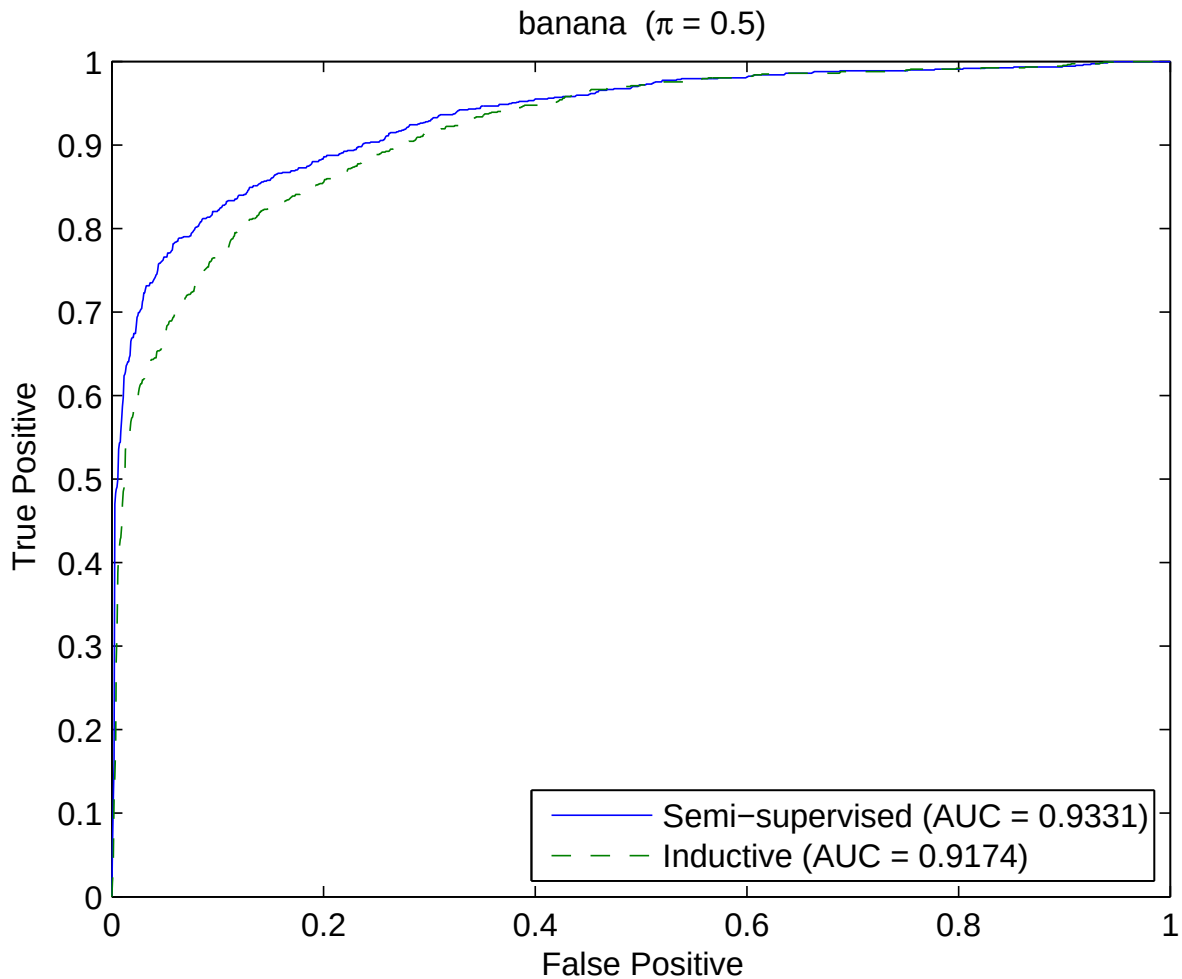
A simple observation

- as usual the optimal decision set is given by a superlevel set of the likelihood ratio $dP_1/dP_0(x)$.
- observe

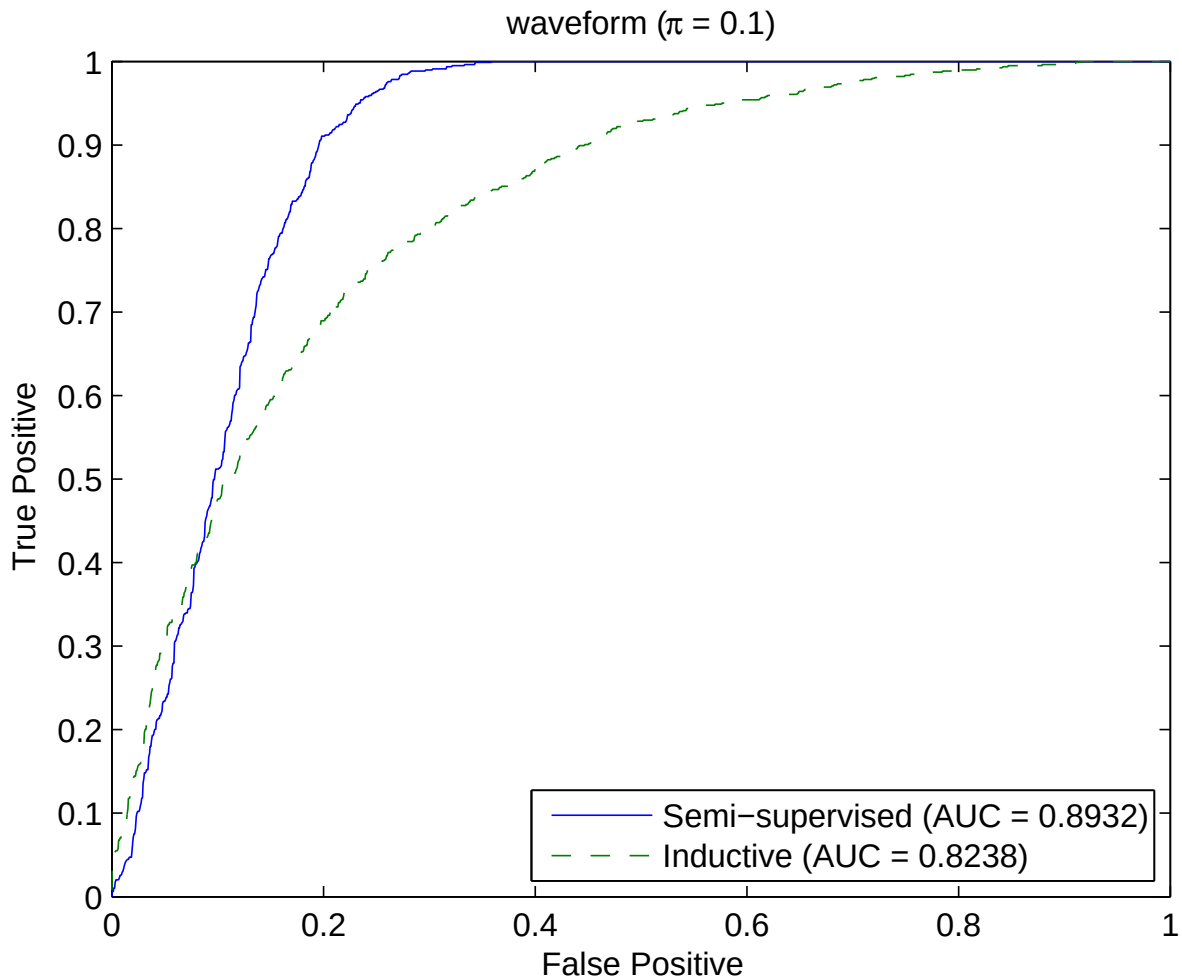
$$dP_X/dP_0(x) = \pi_0 + (1-\pi_0)dP_1/dP_0(x),$$

  one-to-one correspondence between superlevel sets of dP_X/dP_0 and that of dP_1/dP_0
- simply consider the problem of testing P_0 against P_X!
- naive approach: use density estimators for P_0, P_X and the corresponding plug-in superlevel sets.

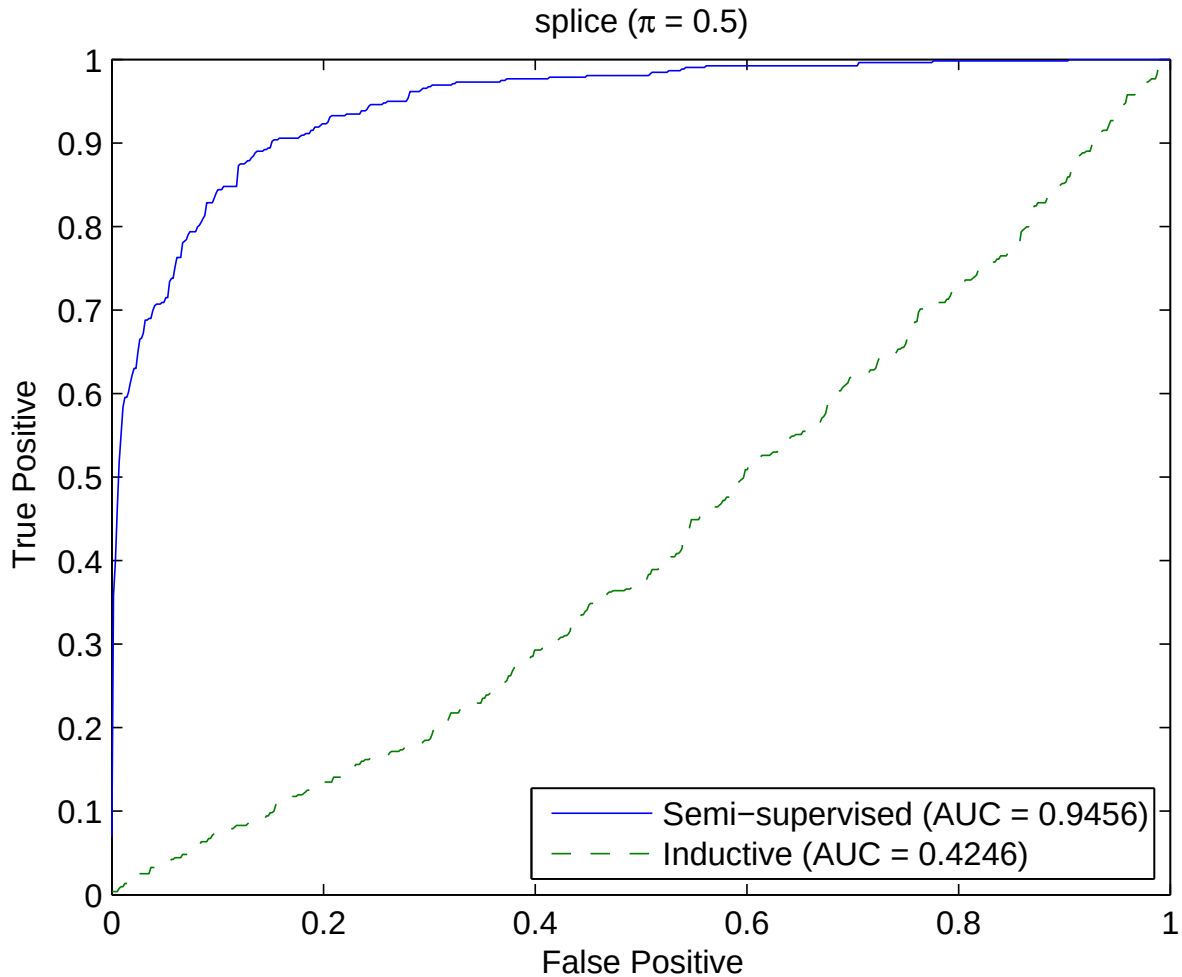
Semisupervised vs inductive



Semisupervised vs inductive



Semisupervised vs inductive



The learning theoretic point of view

- estimating/modelling densities = generative approach
- learning theoretic point of view concentrates directly on decision boundaries (predictive approach)
- finite sample control bounds?
- complexity control/model selection?

Restriction over a specific class

- Consider a class $\mathcal{F}$ of decision rules of limited complexity (eg. finite VC dimension)
- Define the goal over class $\mathcal{F}$:

$$\mathcal{E}^*_{1,\alpha}(\mathcal{F}) := \inf_{f \in \mathcal{F}} \mathcal{E}_1(f) \text{ s.t. } \mathcal{E}_0(f) \leq \alpha \,.$$

(the optimal decision rule in $\mathcal{F}$ is not any longer based on the likelihood ratio)

- Assume the following condition satisfied:

$(\mathbf{A})$ For any $\alpha \in (0,1)$ there exists $f^*_\alpha \in \mathcal{F}$ s.t.

$$\mathcal{E}_1(f^*_\alpha) = \mathcal{E}^*_{1,\alpha}(\mathcal{F}) \text{ and } \mathcal{E}_0(f^*_\alpha) = \alpha$$

Restriction over a specific class

- Consider a class $\mathcal{F}$ of decision rules of limited complexity (eg. finite VC dimension)
- Define the goal over class $\mathcal{F}$:

$$\mathcal{E}^*_{1,\alpha}(\mathcal{F}) := \inf_{f \in \mathcal{F}} \mathcal{E}_1(f) \text{ s.t. } \mathcal{E}_0(f) \leq \alpha \,.$$

(the optimal decision rule in $\mathcal{F}$ is not any longer based on the likelihood ratio)

- Assume the following condition satisfied:

$(\mathbf{A})$ For any $\alpha \in (0,1)$ there exists $f^*_\alpha \in \mathcal{F}$ s.t.

$$\mathcal{E}_1(f^*_\alpha) = \mathcal{E}^*_{1,\alpha}(\mathcal{F}) \text{ and } \mathcal{E}_0(f^*_\alpha) = \alpha$$

Comparison of excess losses: learning reduction

Proposition

Assume **(A)** *satisfied and* $\pi_0 > 0$.
For any $f \in \mathcal{F}$ *, if* $\mathcal{E}_0(f) \leq \alpha + \varepsilon$ *, then*

$$(\mathcal{E}_1(f) - \mathcal{E}^*_{1,\alpha}(\mathcal{F})) \leq (1-\pi_0)^{-1}((\mathcal{E}_X(f) - \mathcal{E}^*_{X,\alpha}(f)) + \pi_0\varepsilon)$$

Conclusion: also on a restricted class we can reduce the unobservable testing problem of P_0 vs P_1 to the observable P_0 vs P_X.

Consider the estimator

$$\widehat{f}_\tau = \operatorname*{Arg\,Min}_{f\in\mathcal{F}} \widehat{\mathcal{E}}_X(f)$$

$$\text{s.t.} \qquad \widehat{\mathcal{E}}_0(f) \leq \alpha + \tau$$

Comparison of excess losses: learning reduction

Proposition

Assume **(A)** *satisfied and* $\pi_0 > 0$.
For any $f \in \mathcal{F}$ *, if* $\mathcal{E}_0(f) \leq \alpha + \varepsilon$ *, then*

$$(\mathcal{E}_1(f) - \mathcal{E}^*_{1,\alpha}(\mathcal{F})) \leq (1-\pi_0)^{-1}((\mathcal{E}_X(f) - \mathcal{E}^*_{X,\alpha}(f)) + \pi_0\varepsilon)$$

Conclusion: also on a restricted class we can reduce the unobservable testing problem of P_0 vs P_1 to the observable P_0 vs P_X.

Consider the estimator

$$\widehat{f}_\tau = \operatorname*{Arg\,Min}_{f\in\mathcal{F}} \widehat{\mathcal{E}}_X(f)$$

$$\text{s.t.} \qquad \widehat{\mathcal{E}}_0(f) \leq \alpha + \tau$$

Uniform control of the excess error

Assume the VC dimension of $\mathcal{F}$ is V. Let

$$\varepsilon_k(V,\delta) = 3\frac{\sqrt{V\log k - \log\delta}}{\sqrt{k}}\,.$$

Proposition

For some constants c, c', if $\tau = c\varepsilon_n$, it holds with prob. at least $(1-\delta)^2$:

$$\begin{aligned}
\mathcal{E}_0(\widehat{f}_\tau) - \alpha &\le c'\varepsilon_n \\
\mathcal{E}_1(\widehat{f}_\tau) - \mathcal{E}^*_{1,\alpha}(\mathcal{F}) &\le c'(1-\pi_0)^{-1}(\varepsilon_n + \varepsilon_m) \\
\mathcal{R}_i(\widehat{f}_\tau) - \mathcal{R}_i(f^*_\alpha) &\le \frac{c'}{P_X(f^*_\alpha(X)=i)}(\varepsilon_n + \varepsilon_m)\,,
\end{aligned}$$

where $\mathcal{R}_i(f) = P_X(Y \neq i | f(X) = i)$.

Model selection

- Consider a sequence of sets of decision functions ("models") $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_K$ of VC-dim $V_1 \leq \ldots \leq V_K$; put $\varepsilon_{j,k} = \varepsilon_j(V_k, k^{-2}\delta)$.
- Define

$$\widehat{f}_k = \operatorname*{Arg\,Min}_{f \in \mathcal{F}_k} \widehat{\mathcal{E}}_X(f)$$

$$\text{s.t.} \qquad \widehat{\mathcal{E}}_0(f) \leq \alpha + c\varepsilon_{n,k}$$

- Select the model using

$$\widehat{k} = \operatorname*{Arg\,Min}_{1 \leq k \leq K} \left(\widehat{\mathcal{E}}_X(\widehat{f}_k) + c\varepsilon_{m,k} \right)$$

- With prob. $(1-2\delta)^2$,

$$\mathcal{E}_0(\widehat{f}_{\widehat{k}}) \leq \alpha + c'\varepsilon_{n,K}\,;$$

$$\mathcal{E}_1(\widehat{f}_{\widehat{k}}) - \mathcal{E}^*_{1,\alpha} \leq \min_{1 \leq k \leq K} \left(\mathcal{E}^*_{1,\alpha}(\mathcal{F}_k) - \mathcal{E}^*_{1,\alpha} + (1-\pi_0)^{-1} \left(\varepsilon_{n,k} + \varepsilon_{m,k} \right) \right)$$

Distribution-free upper confidence bound on π_0

- Define the estimator

$$\widehat{\pi}_0^+(\mathcal{F},\delta) = \inf_{f\in\mathcal{F}} \frac{\mathcal{E}_X(f) + \varepsilon_n(\delta)}{(1-\mathcal{E}_0(f) - \varepsilon_m(\delta))_+} \,.$$

Proposition

1) For any fixed n and any data generating distribution, $\widehat{\pi}_0^+$ is a $(1-\delta)$ upper confidence bound on π_0 .

2) There exists no non-trivial distribution-free lower bound on π_0.

Consistency of the estimate for π_0

Call P_1 a proper novelty distribution if there exists no decomposition

$$P_1 = (1-\gamma)Q + \gamma P_0\,, \qquad \gamma > 0.$$

Proposition

1) For any contaminated distribution P_X and nominal distribution P_0, there exists a unique proper novelty distribution P_1 and $\pi_0^* \geq 0$ s.t.

$$P_X = \pi_0^* P_0 + (1-\pi_0^*) P_1.$$

2) Assume $(\mathcal{F}_k)$ is a sequence of decision sets of finite VC dimension having the universal approximation property. Then

$$\bar{\pi}_0 = \inf_k \widehat{\pi}_0^+(\mathcal{F}_k, (mnk)^{-2})$$

is a universally consistent estimator of π_0^*.

Consistency of the estimate for π_0

Call P_1 a proper novelty distribution if there exists no decomposition

$$P_1 = (1-\gamma)Q + \gamma P_0\,, \qquad \gamma > 0.$$

Proposition

1) For any contaminated distribution P_X and nominal distribution P_0, there exists a unique proper novelty distribution P_1 and $\pi_0^* \geq 0$ s.t.

$$P_X = \pi_0^* P_0 + (1-\pi_0^*) P_1.$$

2) Assume $(\mathcal{F}_k)$ is a sequence of decision sets of finite VC dimension having the universal approximation property. Then

$$\bar{\pi}_0 = \inf_k \hat{\pi}_0^+(\mathcal{F}_k, (mnk)^{-2})$$

is a universally consistent estimator of π_0^*.

Plan

Random effect models is a specification of SSND

The standard i.i.d. mixture model for p-values in multiple testing is the SSND model under the following specific assumptions:

1. Observation space = $[0,1]$;
2. Nominal distribution is known with $P_0 = U([0,1])$. (Equivalently, nominal sample has infinite size)
3. Class $\mathcal{F}$ of novelty detectors considered = intervals indicators $\mathbf{1}\{[0,t]\}, t \in [0,1]$.

Under these assumptions in particular the upper bound $\widehat{\pi}_0^+$ recovers the bound proposed by Genovese and Wassermann (2004), and proper novelty distribution recovers their notion of “purity”.

Putting into questions the usual MT assumptions(1)

- P_0 is known. . . or is it really?
- More realistic for certain applications: reference sample under controlled experimental conditions.

Putting into questions the usual MT assumptions(1)

- P_0 is known. . . or is it really?
- More realistic for certain applications: reference sample under controlled experimental conditions.

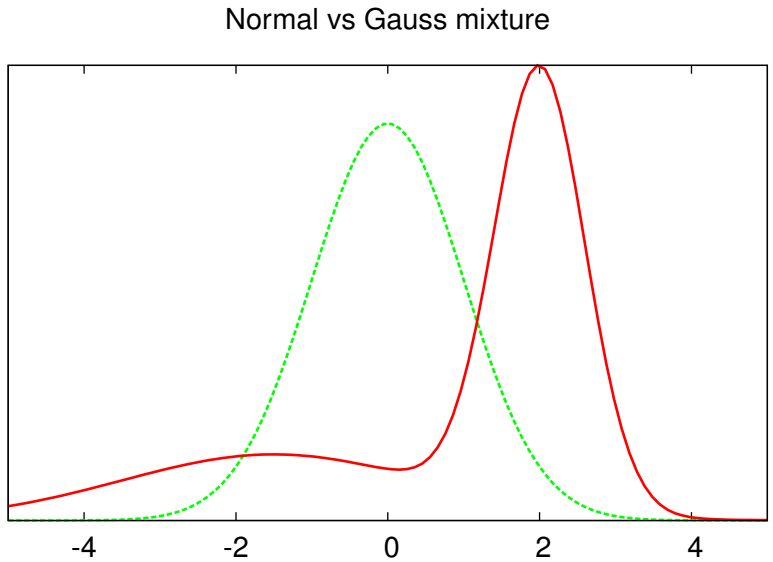
Putting into questions the usual MT assumptions(2)

- Rejection regions: $\mathbf{1}\{[0, t]\}\ldots$

- Specific methods for finding more complex rejection regions:
  - Sun and Cai (2007) (estimate density)
  - Chi (2007) union of intervals

Putting into questions the usual MT assumptions(2)

- Rejection regions: $\mathbf{1}\{[0, t]\}\ldots$



- Specific methods for finding more complex rejection regions:
  - Sun and Cai (2007) (estimate density)
  - Chi (2007) union of intervals

Putting into questions the usual MT assumptions(3)

- observation space is $[0,1]$ (or univariate statistic): why?
- why not consider several statistics at the same time (or possibly the original multivariate data)
- Chi (2008) proposed considering multivariate p-values in $[0,1]^d$.
- (The assumption of known uniform distribution unfer the null is all the more questionable...)

Putting into questions the usual MT assumptions(3)

- observation space is $[0, 1]$ (or univariate statistic): why?
- why not consider several statistics at the same time (or possibly the original multivariate data)
- Chi (2008) proposed considering multivariate p-values in $[0, 1]^d$.
- (The assumption of known uniform distribution unfer the null is all the more questionable. . .)

What about the FDR?

The proposed FPR constraint in the SSND setting is interpreted as a per-comparison error rate in the MT setting.

Theorem

The following inequalities hold with probability at least $(1-\delta)^2 \geq 1-2\delta$ *(over the draw of the nominal and unlabeled samples) :*

$$\forall f \in \mathcal{F} \;\; \mathrm{mFDR}(f) = P_X(H=0|X=1) \leq \frac{(\widehat{\mathcal{E}}_0(f)+\varepsilon_m)\widehat{\pi}_0^+(\mathcal{F},\delta)}{(1-\widehat{\mathcal{E}}_X(f)-\varepsilon_n)_+},$$

and

$$\forall f \in \mathcal{F} \;\; \mathrm{FDP}(f) \leq \frac{(\widehat{\mathcal{E}}_0(f)+\varepsilon_m)\widehat{\pi}_0^+(\mathcal{F},\delta)+\varepsilon_n}{(1-\widehat{\mathcal{E}}_X(f))},$$

What about the FDR?

The proposed FPR constraint in the SSND setting is interpreted as a per-comparison error rate in the MT setting.

Theorem

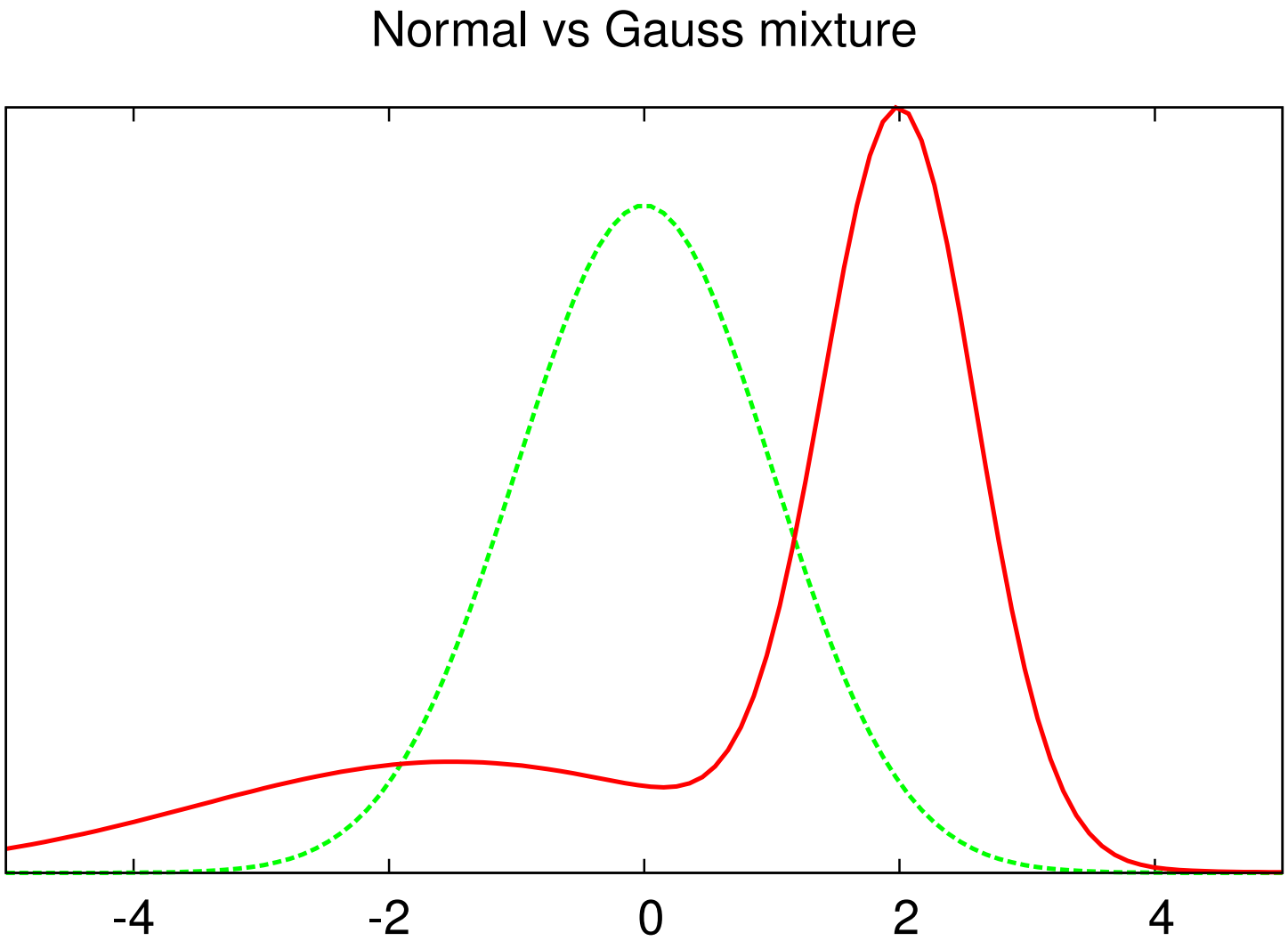
The following inequalities hold with probability at least $(1-\delta)^2 \geq 1-2\delta$ *(over the draw of the nominal and unlabeled samples) :*

$$\forall f \in \mathcal{F} \;\; \mathrm{mFDR}(f) = P_X(H=0|X=1) \leq \frac{(\widehat{\mathcal{E}}_0(f)+\varepsilon_m)\widehat{\pi}_0^+(\mathcal{F},\delta)}{(1-\widehat{\mathcal{E}}_X(f)-\varepsilon_n)_+},$$

and

$$\forall f \in \mathcal{F} \;\; \mathrm{FDP}(f) \leq \frac{(\widehat{\mathcal{E}}_0(f)+\varepsilon_m)\widehat{\pi}_0^+(\mathcal{F},\delta)+\varepsilon_n}{(1-\widehat{\mathcal{E}}_X(f))},$$

Revisiting a simple case



Towards new algorithms?

- let $\mathcal{F}_k$=rejection regions made of k intervals
- goal: optimize $\widehat{\mathcal{E}}_X(f)$ under FDP constraint
- need to know: for each fixed value of $\widehat{\mathcal{E}}_X(f)$, best possible value of $\mathcal{E}_0(f)$
- dynamic programming feasible ($\mathcal{O}(kn^3)$)

Outlook

- advocated here: not a cure-all, but the suggestion that applying classification/learning theoretic approach to MT can provide an alternative point of view, still relatively unexplored
- if you do not believe bounds of learning theory to give useful practical results. . .
- . . . you can still be pragmatic and use cross-validation.

Outlook

- advocated here: not a cure-all, but the suggestion that applying classification/learning theoretic approach to MT can provide an alternative point of view, still relatively unexplored
- if you do not believe bounds of learning theory to give useful practical results...
- ...you can still be pragmatic and use cross-validation.


C. Scott, G. Blanchard
Novelty detection: unlabeled data definitely help.
AISTATS 2009, *JMLR Workshop and Conference Proceedings* 5:464-471, 2009.


G. Blanchard, G. Lee, C. Scott
Semi-supervised novelty detection
WIAS Preprint 1471, 2009.