Nonparametric estimation of class prior

Shantanu Jain

January 2016

Outline

- Introduction and motivation
- Problem statement
- Identifiability
- Nonparametric algorithm (*AlphaMax*)
- Handling multivariate data
- Results
- ~~Parametric approach with skew normals~~
- ~~Open issues and future work~~

Selection bias

L: Labeled data
U : Unlabeled data
L and U don't come from the same distribution.
Consequence:

- Errors computed on the test data don't generalize on the unlabeled data.
- Classifier learnt from training data might give suboptimal performance on unlabeled data.

Population with two groups

Notation

$x \in \mathcal{X}$	an object in the population
$y \in \{0, 1\}$	group of an object
$p(x)$	distribution of the population
$p(x\|y = 0)$	distribution of group 0 subpopulation (**class conditional**)
$p(x\|y = 1)$	distribution of group 1 subpopulation (**class conditional**)
$p(y = 1)$	proportion of group 1 objects (**class prior**)
$p(y = 0)$	proportion of group 0 objects (**class prior**)
$p(y = 1\|x)$	**posterior**
$p(y = 0\|x)$	**posterior**

$p(x)$ as two component mixture

$$\boxed{p(x) = p(y = 1)p(x|y = 1) + (1 - p(y = 1))p(x|y = 0)}$$

Binary classification

Labeled dataset

$L = (x_i, y_i)$

$x_i \in \mathcal{X}, y_i \in \mathcal{Y} = \{0, 1\}$

Prediction $\hat{y}$

$$\hat{y} = \begin{cases} 1 & s(x) \geq t \\ 0 & s(x) < t \end{cases}$$

Unbeatable s(x)

$s(x) = p(y = 1|x)$

For a given FP achieves smallest FN

Not unique

$s(x)$ that ranks points as $p(y = 1|x)$ is also unbeatable.

Predicted	Actual +	Actual -
Y	True positives	False positives
N	False negatives	True negatives

Figure: Confusion matrix

Probabilistic classifier

Direct estimation of the posterior

Naive Bayes

Logistic Regression

Neural Network

Indirect estimation of posterior

Convert scores to posteriors using:

Platt scaling

Isotonic Regression

Class prior

- Related to posterior:

$$\boxed{p(y=1|x) = \frac{p(x|y=1)}{p(x)} p(y=1).}$$

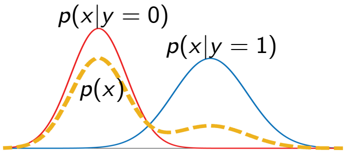
- Interesting quantity in itself:
  Proprtion of enzymes in proteins.
  Proportion of Facebook users liking a page.

Supervised learning: Traditional classifier

Given

$$L = \begin{bmatrix} \boldsymbol{X_0} & 0 \\ \boldsymbol{X_1} & 1 \end{bmatrix}$$

$\boldsymbol{X_0}$: sample from $p(x|y=0)$,
$\boldsymbol{X_1}$: sample from $p(x|y=1)$,
$\boldsymbol{X_0} \cup \boldsymbol{X_1}$: sample from $p(x)$.



Goal

Use x to predict y $\boxed{p(y=1|x)}$.

Estimate class prior

$$\boxed{p(y=1) \approx \hat{\alpha} = \frac{|\boldsymbol{X_1}|}{|\boldsymbol{X_0}| + |\boldsymbol{X_1}|}}.$$

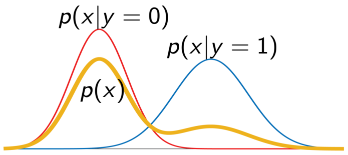
Semi-supervised learning: Selection Bias

Given

$$L = \begin{bmatrix} \boldsymbol{X_0} & 0 \\ \boldsymbol{X_1} & 1 \end{bmatrix}$$

$\boldsymbol{X_0}$: sample from $p(x|y=0)$,
$\boldsymbol{X_1}$: sample from $p(x|y=1)$,
$\boldsymbol{X_0} \cup \boldsymbol{X_1}$: **not** a sample from $p(x)$:

$$\boxed{p_L(y=1) \neq p(y=1).}$$



Learning

$$\boxed{p_L(y=1|x) = \frac{\overbrace{p_L(x|y=1)}^{=p(x|y=1)}}{\underbrace{p_L(x)}_{\neq p(x)}} \overbrace{p_L(y=1)}^{\neq p(y=1)} \neq p(y=1|x).}$$

Solution

Use unlabeled dataset U, a sample from $p(x)$.

Estimating class prior

Define

$$\mathbf{P} = \begin{bmatrix} p(\hat{y}=1|y=1) & p(\hat{y}=1|y=0) \\ p(\hat{y}=0|y=1) & p(\hat{y}=0|y=0) \end{bmatrix}$$

Define

$$q = [p(\hat{y}=1), p(\hat{y}=0)]'$$
$$p = [p(y=1), p(y=0)]'$$

It follows that

$$\boxed{q = \mathbf{P}p.}$$

Procedure

- Use $\hat{P} = \begin{bmatrix} TP/|\mathbf{X_1}| & FP/|\mathbf{X_0}| \\ FN/|\mathbf{X_1}| & TN/|\mathbf{X_0}| \end{bmatrix}$ as estimate of P. (The estimation is justified because $\hat{y}$ is a deterministic function of x and consequently, $p_L(\hat{y}|y) = p(\hat{y}|y)$.)
- Apply $\hat{y}$ on U to get $\hat{q}$
- Use $\boxed{\hat{p} = \hat{P}^{-1}\hat{q}}$ to estimate p.

Semi-supervised learning: Estimating class prior and posterior

U: an unlabeled sample from $p(x)$.

EM-algorithm[a]

Initialize:

$$\boxed{\begin{aligned}\hat{p}(y=1|x) &\leftarrow \hat{p}_L(y=1|x)\\ \hat{\alpha} \leftarrow \alpha_L &= \frac{|\boldsymbol{X_1}|}{|\boldsymbol{X_0}|+|\boldsymbol{X_1}|}.\end{aligned}}$$

Update:

$$\boxed{\begin{aligned}\hat{p}(y=1|x) &\leftarrow \frac{\frac{\hat{\alpha}}{\alpha_L}\hat{p}(y=1|x)}{\frac{\hat{\alpha}}{\alpha_L}\hat{p}(y=1|x)+\frac{(1-\hat{\alpha})}{(1-\alpha_L)}(1-\hat{p}(y=1|x))}\\ \hat{\alpha} &\leftarrow \frac{1}{|U|}\sum_{x\in U}\hat{p}(y=1|x).\end{aligned}}$$

[a] P. Latinne, M. Saerens, and C. Decaestecker. "Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification

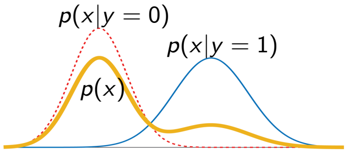
Semi-supervised learning: positive and unlabeld learning/ novelty detection

Given

$L = \left[\begin{array}{cc} \boldsymbol{X_1} & 1 \end{array} \right]$

$\boldsymbol{X_1}$: sample from $p(x|y=1)$,

U: unlabeled sample from $p(x)$.



Questions

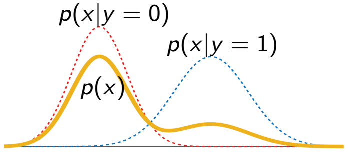
$\boxed{\textit{How to estimate } p(y=1|x)?}$

$\boxed{\textit{How to estimate } p(y=1)?}$

Unsupervised learning: clustering

Given

X: unlabeled sample from $p(x)$.
$L = \emptyset$.



Estimation recipe

Model U by parametrized two component mixture.
Use EM to estimate $p(y=1|x), p(y=1)$.

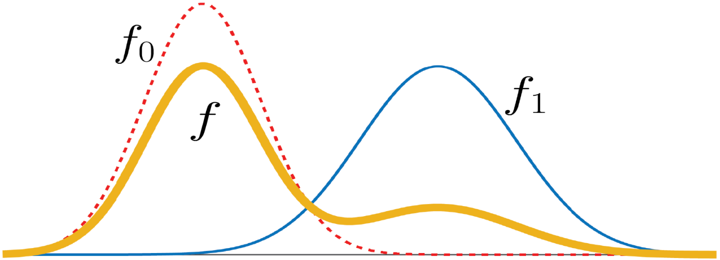
Mixture

f is a mixture:

$$\boxed{f(x) = \alpha \cdot f_1(x) + (1-\alpha) \cdot f_0(x).}$$

where

$$f(x) = p(x),$$
$$f_1(x) = p(x|Y=1),$$
$$f_0(x) = p(x|Y=0),$$
$$\alpha = p(Y=1).$$



Problem definition

Estimate α:

$$\boxed{\boldsymbol{X}, \boldsymbol{X_1} :\rightarrow \alpha}$$

where

$$\boldsymbol{X} = \{x_i\}_1^n, \qquad x_i \sim f,$$
$$\boldsymbol{X_1} = \{x_{1i}\}_1^m, \quad x_{1i} \sim f_1.$$

Difficulty: Identifiability

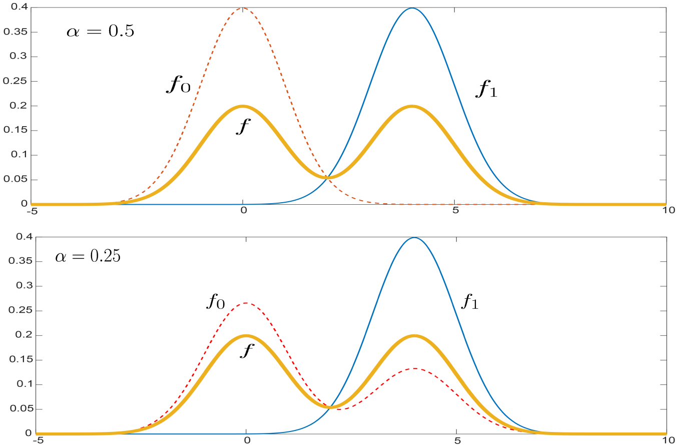


Figure: Unidentifiability

Identifiability

Set of valid α

$$A(f, f_1, \mathcal{P}_{\mathbf{0}}^{\text{all}}),$$

where

$\mathcal{P}_0^{\text{all}}$: set of all pdfs, except f_1,
$A(f, f_1, \mathcal{P}_0) = \{\alpha \in (0,1) : f = \alpha f_1 + (1-\alpha) f_0, f_0 \in \mathcal{P}_0\}$.

Relation between f_0 and α

$$\boxed{f_0(x) = \frac{f(x) - \alpha f_1(x)}{1-\alpha},}$$

where

f is fixed,
f_1 is fixed.

$f_0 \leftrightarrow \alpha$.

f_0 not always a density.

$f_0 \in \mathcal{P}_0^{\text{all}}$
$\Leftrightarrow \alpha \in A(f, f_1, \mathcal{P}_0^{\text{all}})$.

Condition for density:
$f - \alpha f_1 \geq 0$ or $\alpha \leq f/f_1$

$\alpha \in A(f, f_1, \mathcal{P}_0^{\text{all}})$
$\Rightarrow (0, \alpha] \subseteq A(f, f_1, \mathcal{P}_0^{\text{all}})$

Identifiability[1]

α^*: maximum proportion

$$\boxed{\alpha^* = \inf R(f, f_1),}$$

$$\boxed{A(f, f_1, \mathcal{P}_0^{\text{all}}) = (0, \alpha^*],}$$

where

$$R(f, f_1) = \{f(x)/f_1(x) : x \in \mathcal{X},\ f_1(x) \neq 0\}.$$

f_0^* does not contain f_1

$$\inf R(f_0^*, f_1) = 0,$$

where

f_0^* : f_0 corresponding to α^*.

Make the problem identifiable by estimating α^*

[1] **blanchard2010semi**.

Estimation[23]

AlphaMax

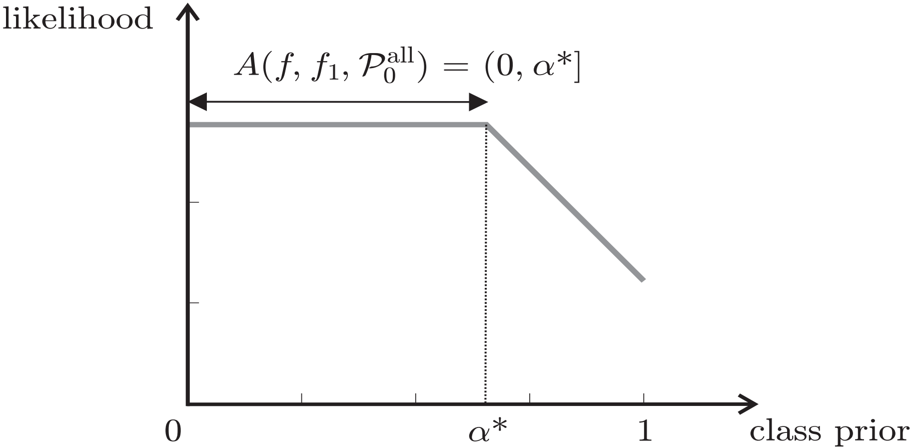


Figure: Theoretical log likelihood versus α.

Nonparametric Estimation

f as k-comp mixture

$$\boxed{\hat{f}(\cdot) = \sum_1^k w_i \kappa_i(\cdot),}$$

where

κ_i: component densities,
$w_i \in (0,1], \sum w_i = 1.$

β parametrized densities

$$\boxed{\begin{aligned} \boldsymbol{h_1}(\cdot|\boldsymbol{\beta}) &= \frac{\sum_1^k \beta_i w_i \kappa_i(\cdot)}{\sum_1^k \beta_i w_i}, \\ h_0(\cdot|\boldsymbol{\beta}) &= \frac{\sum_1^k (1-\beta_i) w_i \kappa_i(\cdot)}{\sum_1^k (1-\beta_i) w_i}, \end{aligned}}$$

where

$\boldsymbol{\beta} = [\beta_i],\ \beta_i \in (0,1].$

Re-express $\hat{f}$

$$\boxed{\hat{f}(\cdot) = \overbrace{(\textstyle\sum \beta_i w_i)}^{\alpha}\overbrace{h_1(\cdot|\boldsymbol{\beta})}^{f_1} + \overbrace{(1-\textstyle\sum \beta_i w_i)}^{1-\alpha}\overbrace{h_0(\cdot|\boldsymbol{\beta})}^{f_0}}$$

Substitute $\hat{f}_1$

$$\boxed{h(\cdot|\boldsymbol{\beta}) = \overbrace{(\textstyle\sum \beta_i w_i)}^{\alpha}\hat{f}_1(\cdot) + \overbrace{(1-\textstyle\sum \beta_i w_i)}^{1-\alpha}\overbrace{h_0(\cdot|\boldsymbol{\beta})}^{f_0}}$$

Nonparametric Estimation

Log-likelihood

$$\boxed{\mathcal{L}(\beta|\boldsymbol{X},\boldsymbol{X_1}) = \textstyle\sum_{x\in\boldsymbol{X}} \log h(x|\beta) + \sum_{x\in\boldsymbol{X_1}} \log h_1(x|\beta).}$$

Optimization problem

$$\boxed{\begin{array}{ll} \underset{\boldsymbol{\beta}}{\text{maximize}} & \mathcal{L}(\beta|\boldsymbol{X},\boldsymbol{X_1}) \\ \text{subject to} & \sum \beta_i w_i{=}\alpha. \end{array}}$$

~~Convexity~~

$$\cancel{\log h(\cdot|\beta) = \log}\left(\sum \beta_i \cancel{w_i(\hat{f}_1(\cdot) - \kappa_i(\cdot)) + \hat{f}(\cdot)}\right),$$

$$\cancel{\log h_1(\cdot|\beta) = \log}\left(\sum \beta_i \cancel{w_i\kappa_i(\cdot)}\right) \underbrace{\cancel{-\log}\left(\sum \beta_i \cancel{w_i}\right)}_{\textit{constant}}.$$

Estimated log-likelihood curve

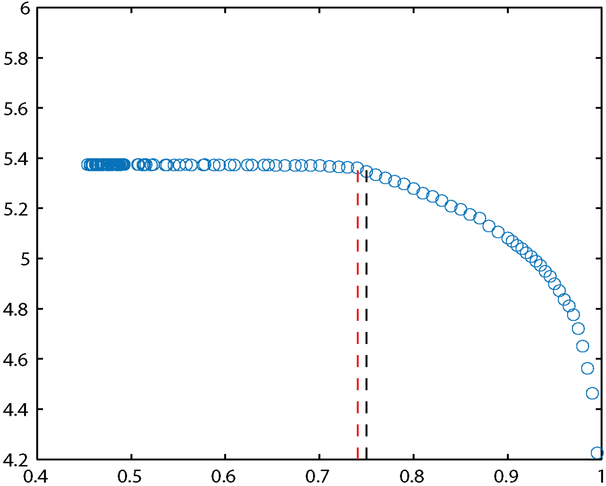


Figure: *Gaussian*$\Delta\mu = 4$

To convert multivariate data to univariate data

α^* preserving transform

Use the output of a positive versus unlabeld classifier.

$$\boxed{\tau(x) = p(s = 1 | x, s \in \{0, 1\})}$$

where

s is the selection variable:

$$s = \begin{cases} 0 & \Rightarrow \text{add } x \text{ to } \boldsymbol{X}, \\ 1 & \Rightarrow \text{add } x \text{ to } \boldsymbol{X_1}, \\ 2 & \Rightarrow \text{throw away } x. \end{cases}$$

Not all transforms guarantee preservation of α^*

~~**Implied assumptions on s:**~~

$$\boxed{\begin{aligned} &\cancel{p(x|s = 1) = p(x|y = 1)}, \\ &\cancel{p(x|s = 0) = p(x)}. \end{aligned}}$$

Learning the true posterior

$$p(y = 1|x) = \overbrace{\left(\frac{p(s = 0)}{p(s = 1)}\right)}^{\approx |\mathbf{X}|/|\mathbf{X_1}|} \overbrace{P(Y = 1)}^{\alpha^*} \left(\frac{\tau(x)}{1 - \tau(x)}\right).$$