Estimating Labels from Label Proportions

Novi Quadrianto
Novi.Quad@gmail.com

The Australian National University, Australia
NICTA, Statistical Machine Learning Program, Australia

Joint work with Alex Smola, Tiberio Caetano, and Quoc Le
Supervised Learning
Semi-supervised Learning
An example application

Promotional coupon
Apple Inc. decides to distribute the following coupon:

To whom this coupon should be mailed?

- every college students in the world?
- selected college students?
Selection criteria

- Some people would *always* buy Mac, even without coupon
- Some other people will *never* buy Mac anyway
- Others will buy Mac *if and only if* they receive the coupon
An example application

Four types of customers: A - Always buyers, N - Never buyers, C - Compliers (buy iff coupon), D - Defiers (buy iff no coupon).

Four data aggregates

<table>
<thead>
<tr>
<th></th>
<th>Buy</th>
<th>Doesn’t Buy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. 1: Given Coupon</td>
<td>A ∪ C</td>
<td>N</td>
</tr>
<tr>
<td>Exp. 2: Not Given Coupon</td>
<td>A</td>
<td>N ∪ C</td>
</tr>
</tbody>
</table>

Assumption: no defiers

Fact: we don’t have a pure sample of C, and we want $p(C|\text{customer profile})$
We know the proportions $p(A)$ and $p(N)$ from the random assignment experiment.

Therefore we know $p(C)$.

Therefore we know all the proportions.
Problem formulation

What we have

- \( n \) sets of observations \( X_i = \{x^i_1, \ldots, x^i_{m_i}\} \) of respective sample sizes \( m_i \) as calibration sets
- a set \( X = \{x_1, \ldots, x_m\} \) as a test set
- fractions \( \pi_{i y} \) of patterns of labels \( y \in Y (|Y| \leq n) \) contained in each set \( X_i \)
- marginal probability \( p(y) \) of the test set \( X \)

What we want

- conditional class probability estimates \( p(y|x) \)
Gaussian process solution

Conditional exponential likelihood model

\[ p(y|x, \theta) = \exp \left( \langle \phi(x, y), \theta \rangle - g(\theta|x) \right) \]  with

\[ g(\theta|x) = \log \sum_{y \in \mathcal{Y}} \exp \langle \phi(x, y), \theta \rangle \]

Some details
- \( \phi(x, y) \) is the sufficient statistics
- \( g(\theta|x) \) is the log-partition function

Gaussian prior

\[ - \log p(\theta) \propto \lambda \| \theta \|^2 \]

Posterior

\[ - \log p(Y|X, \theta)p(\theta) = \sum_{i=1}^{m} \left[ g(\theta|x_i) - \langle \phi(x_i, y_i), \theta \rangle \right] + \lambda \| \theta \|^2 \]
\[ \theta^* = \arg\min_{\theta} \left[ \sum_{i=1}^{m} g(\theta|x_i) - m \left< \mu_{XY}, \theta \right> + \lambda \| \theta \|^2 \right] \]

with

\[ \mu_{XY} := \frac{1}{m} \sum_{i=1}^{m} \phi(x_i, y_i) \]

This is a convex optimization problem

So is our job done?

Convergence of empirical means (Bartlett & Mandelson 2002):

\[ \mu_{xy} := \sum_{y \in Y} p(y) \mathbb{E}_{x \sim p(x|y)}[\phi(x,y)] \]

\[ \mu_{XY} \quad \text{sample} \quad \leftarrow \quad \mu_{xy} \quad \text{population} \]
Binary classification

- Dataset 1 contains class +1
- Dataset 2 contains class +1 and -1 with proportions $p(+1) := \rho$ and $p(-1) = 1 - \rho$

\[
\mu_+ := \mathbb{E}_{(x) \sim p(x|y=+1)}[\phi(x, y)]
\]
\[
\mu_1 := \mathbb{E}_{(x) \sim p(x|\text{set 1})}[\phi(x, y)]
\]
Re-calibrated sufficient statistics

Binary classification

\[
\begin{bmatrix}
\mu_1 \\
\mu_2
\end{bmatrix} = \begin{bmatrix}
1 & 0 \\
\rho & 1 - \rho
\end{bmatrix} \begin{bmatrix}
\mu_+ \\
\mu_-
\end{bmatrix}
\]

\[\downarrow\]

\[\pi = \begin{bmatrix}
1 & 0 \\
\rho & 1 - \rho
\end{bmatrix} \Rightarrow \pi^{-1} = \begin{bmatrix}
1 & 0 \\
\frac{-\rho}{1-\rho} & \frac{1}{1-\rho}
\end{bmatrix}\]

\[\downarrow\]

\[
\begin{bmatrix}
\mu_+ \\
\mu_-
\end{bmatrix} = \begin{bmatrix}
1 & 0 \\
\frac{-\rho}{1-\rho} & \frac{1}{1-\rho}
\end{bmatrix} \begin{bmatrix}
\mu_1 \\
\mu_2
\end{bmatrix}
\]

\[\downarrow\]

\[
\hat{\mu}_{XY} = \rho \mu_1 - (1 - \rho) \left[\frac{-\rho}{1-\rho} \mu_1 + \frac{1}{1-\rho} \mu_2\right]
\]
Three class classification

\[
\begin{bmatrix}
\mu_1 \\
\mu_2 \\
\mu_3
\end{bmatrix} =
\begin{bmatrix}
\alpha & \beta & 1 - (\alpha + \beta) \\
\eta & \xi & 1 - (\eta + \xi) \\
\sigma & \lambda & 1 - (\lambda + \sigma)
\end{bmatrix}
\begin{bmatrix}
\mu_a \\
\mu_b \\
\mu_c
\end{bmatrix}
\]
The algorithm

Algorithm 1

Input datasets $X$, $\{X_i\}$, probabilities $\pi_{iy}$ and $p(y)$
for $i = 1$ to $n$ and $y' \in Y$ do
    Compute empirical means $\mu^\text{set}[i, y']$
end for
Compute $\hat{\mu}_x^\text{class} = (\pi^T \pi)^{-1} \pi^T \mu^\text{set}_X$
Compute $\hat{\mu}_{XY} = \sum_{y \in Y} p(y) \hat{\mu}_x^\text{class}[y, y]$
Solve the minimization problem

$$\hat{\theta}^* = \arg\min_{\theta} \left[ \sum_{i=1}^{m} g(\theta|x_i) - m \langle \hat{\mu}_{XY}, \theta \rangle + \lambda ||\theta||^2 \right]$$

Return $\hat{\theta}^*$.

$\mu^\text{set}_X \xrightarrow{} \hat{\mu}_x^\text{class} \xrightarrow{} \hat{\mu}_{XY}$
Binary classification, $\phi(x, y) = y\psi(x)$ and $X_2 = X$

**Theorem 1** With probability $1 - \delta$ the following bound holds:

$$\|\hat{\mu}_{XY} - \mu_{XY}\| \leq 2\rho \left[ 2 + \sqrt{\log 2/\delta} \right] \left[ m_1^{-\frac{1}{2}} + m_+^{-\frac{1}{2}} \right]$$

**Some details**

- $m_1$ is the number of observations in $X_1$
- $m_+$ is the number of observations with $y = +1$ in $X_2$
Bound on the minimizer of the log-posterior (Altun & Smola 2006)

$$\| \theta^* - \hat{\theta}^* \| \leq \lambda^{-1} \| \mu - \hat{\mu} \|$$

Bound on the log-posterior (Altun & Smola 2006)

$$L(\hat{\theta}^*, \hat{\mu}) - L(\theta^*, \mu) \leq \| \hat{\theta}^* - \theta^* \| \| \hat{\mu} - \mu \| = \lambda^{-1} \| \mu - \hat{\mu} \|^2$$

Some details

- $\theta^*$ is the minimizer of $L(\theta, \mu)$
- $\hat{\theta}^*$ is the minimizer of $L(\hat{\theta}, \hat{\mu})$
Alternative Solutions

Reduction to binary

- a binary classifier between set $X_1$ and $X_2$
- label thresholding according to the known proportions

Density estimation

- density estimation for each dataset $X_i$
- re-calibration to get $p(x|y)$ via $\sum_i \left( \pi^{-1} \right) y_i p(x, y|i)$
- compute posterior probabilities

MCMC (Kück & de Freitas 2005)

- explicitly generate mixing proportions per group by hierarchical probabilistic model
- use sampling to generate samples of model posterior distribution
### Experiments

#### Table 1. Classification error on the UCI/LibSVM database

Errors are reported in % with standard error. (%) ± SE. The best result and those results not significantly worse than it, are highlighted in red. We used a one-sided paired Welch t-test with 95% confidence level as reference.

<table>
<thead>
<tr>
<th>Data</th>
<th>MM</th>
<th>KDE</th>
<th>DS</th>
<th>MCMC</th>
<th>BA</th>
</tr>
</thead>
<tbody>
<tr>
<td>iono</td>
<td>18.4±3.2</td>
<td>17.5±3.2</td>
<td>12.2±2.6</td>
<td>18.0±2.1</td>
<td>35.8</td>
</tr>
<tr>
<td>iris</td>
<td>10.0±3.6</td>
<td>16.8±3.4</td>
<td>15.4±1.1</td>
<td>21.1±3.6</td>
<td>29.9</td>
</tr>
<tr>
<td>optd</td>
<td>1.8±0.5</td>
<td>0.7±0.4</td>
<td>9.8±1.2</td>
<td>2.0±0.4</td>
<td>49.1</td>
</tr>
<tr>
<td>page</td>
<td>3.8±2.3</td>
<td>7.1±2.8</td>
<td>18.5±5.6</td>
<td>5.4±2.8</td>
<td>43.9</td>
</tr>
<tr>
<td>pima</td>
<td>27.5±3.0</td>
<td>34.8±0.6</td>
<td>34.4±1.7</td>
<td>23.8±1.8</td>
<td>34.8</td>
</tr>
<tr>
<td>tic</td>
<td>31.0±1.5</td>
<td>34.6±0.5</td>
<td>26.1±1.5</td>
<td>31.3±2.5</td>
<td>34.6</td>
</tr>
<tr>
<td>yeast</td>
<td>9.3±1.5</td>
<td>6.5±1.3</td>
<td>25.6±3.6</td>
<td>10.4±1.9</td>
<td>39.9</td>
</tr>
<tr>
<td>wine</td>
<td>7.4±3.0</td>
<td>12.1±4.4</td>
<td>18.8±6.4</td>
<td>8.7±2.9</td>
<td>40.3</td>
</tr>
<tr>
<td>wdbc</td>
<td>7.8±1.3</td>
<td>5.9±1.2</td>
<td>10.1±2.1</td>
<td>15.5±1.3</td>
<td>37.2</td>
</tr>
<tr>
<td>sonar</td>
<td>24.2±3.5</td>
<td>35.2±3.5</td>
<td>31.4±4.0</td>
<td>39.8±2.8</td>
<td>44.5</td>
</tr>
<tr>
<td>heart</td>
<td>30.0±4.0</td>
<td>38.1±3.8</td>
<td>28.4±2.8</td>
<td>33.7±4.7</td>
<td>44.9</td>
</tr>
<tr>
<td>brea</td>
<td>5.3±0.8</td>
<td>14.2±1.6</td>
<td>3.5±1.3</td>
<td>4.8±2.0</td>
<td>34.5</td>
</tr>
<tr>
<td>aust</td>
<td>17.0±1.7</td>
<td>33.8±2.5</td>
<td>15.8±2.9</td>
<td>30.8±1.8</td>
<td>44.4</td>
</tr>
<tr>
<td>svm3</td>
<td>20.4±0.9</td>
<td>27.2±1.3</td>
<td>25.5±1.5</td>
<td>24.2±0.8</td>
<td>23.7</td>
</tr>
<tr>
<td>adult</td>
<td>18.9±1.2</td>
<td>24.5±1.3</td>
<td>22.1±1.4</td>
<td>18.7±1.2</td>
<td>24.6</td>
</tr>
<tr>
<td>cleve</td>
<td>19.1±3.6</td>
<td>35.9±4.5</td>
<td>23.4±2.9</td>
<td>24.3±3.1</td>
<td>22.7</td>
</tr>
<tr>
<td>derm</td>
<td>4.9±1.4</td>
<td>27.4±2.6</td>
<td>4.7±1.9</td>
<td>14.2±2.8</td>
<td>30.5</td>
</tr>
<tr>
<td>musk</td>
<td>25.1±2.3</td>
<td>28.7±2.6</td>
<td>22.2±1.8</td>
<td>19.6±2.8</td>
<td>43.5</td>
</tr>
<tr>
<td>ger</td>
<td>32.4±1.8</td>
<td>41.6±2.9</td>
<td>37.6±1.9</td>
<td>32.0±0.6</td>
<td>32.0</td>
</tr>
<tr>
<td>cove</td>
<td>37.1±2.5</td>
<td>41.9±1.7</td>
<td>32.4±1.8</td>
<td>41.1±2.2</td>
<td>45.9</td>
</tr>
<tr>
<td>spli</td>
<td>25.2±2.0</td>
<td>35.5±1.5</td>
<td>26.6±1.7</td>
<td>28.8±1.6</td>
<td>48.4</td>
</tr>
<tr>
<td>giss</td>
<td>10.3±0.9</td>
<td>†</td>
<td>12.2±0.8</td>
<td>50.0±0.0</td>
<td>50.0</td>
</tr>
<tr>
<td>made</td>
<td>44.1±1.5</td>
<td>†</td>
<td>46.0±2.0</td>
<td>49.6±0.2</td>
<td>50.0</td>
</tr>
<tr>
<td>cmc</td>
<td>37.5±1.4</td>
<td>43.8±0.7</td>
<td>45.1±2.3</td>
<td>46.9±2.6</td>
<td>49.9</td>
</tr>
<tr>
<td>bupa</td>
<td>48.5±2.9</td>
<td>50.8±5.1</td>
<td>40.3±4.9</td>
<td>50.4±0.8</td>
<td>49.7</td>
</tr>
<tr>
<td>protA</td>
<td>44.6±0.3</td>
<td>60.2±0.1</td>
<td>N/A</td>
<td>65.3±1.9</td>
<td>61.2</td>
</tr>
<tr>
<td>protB</td>
<td>45.7±0.6</td>
<td>61.2±0.0</td>
<td>N/A</td>
<td>67.7±1.8</td>
<td>61.2</td>
</tr>
<tr>
<td>dnaA</td>
<td>16.6±1.0</td>
<td>30.7±0.8</td>
<td>N/A</td>
<td>37.7±0.8</td>
<td>40.5</td>
</tr>
<tr>
<td>dnaB</td>
<td>29.1±1.0</td>
<td>33.0±0.7</td>
<td>N/A</td>
<td>40.5±0.0</td>
<td>40.5</td>
</tr>
<tr>
<td>sensA</td>
<td>19.8±0.1</td>
<td>43.1±0.0</td>
<td>N/A</td>
<td>†</td>
<td>43.2</td>
</tr>
<tr>
<td>sensB</td>
<td>21.0±0.1</td>
<td>43.1±0.0</td>
<td>N/A</td>
<td>†</td>
<td>43.2</td>
</tr>
</tbody>
</table>

- **MM**: Mean Map (ours)
- **KDE**: Kernel Density Estimation
- **DS**: Discriminative Sorting
- **MCMC**: Sampling Method
- **BA**: Baseline
Zooming in (binary results)
Design parameters:

- **Entropy and regularization**: choosing various Csiszar and Bregman distances will produce a range of diverse estimators.

- **Function space**: measuring the deviation in moment matching in terms of $\ell_\infty$ norm recovers sparse coding $\ell_1$ (dual connection).
Take home messages

- A new problem formulation which has not been solved and quite relevant in many aspects

- Our estimator can be easily implemented

- Our estimator enjoys the same rates of convergence as what can be expected from building an estimator with a fully labeled sample

- Our solution can be easily extended to other learning frameworks

- Our estimator works well in practice!