Estimating Labels from Label Proportions

Novi Quadrianto, Alex Smola
Tiberio Caetano, Quoc Le

Statistical Machine Learning - NICTA , Australian National University, Stanford University

Abstract

Consider the following problem: given sets of unlabeled observations, each set with known label proportions, predict the labels of another set of observations, also with known label proportions. This problem appears in areas like e-commerce, spam filtering and improper content detection. We present consistent estimators which can reconstruct the correct labels with high probability in a uniform convergence sense. Experiments show that our method works well in practice.

Problem Statement

Assume you want to entice customers to purchase a product. You have the choice of handing out a coupon to entice users to buy. It is your goal to determine who should be sent a coupon to encourage a purchase and who would buy the product regardless. However, the only information available is the set of customers which purchased the product when no discount was offered, and another (independent) set of customers which purchased the product with the discount, without explicit information as to whether they would have made the purchase if the discount had not been offered.

(Formal) problem formulation:
Given

- n sets of datasets $X_i = \{x_1^i, \ldots, x_{m_i}^i\}$ of respective sample sizes m_i as calibration sets
- a set $X = \{x_1, \ldots, x_m\}$ as a test set
- fractions π_{iy} of patterns of labels $y \in \mathcal{Y}$ ($|\mathcal{Y}| \leq n$) contained in each set X_i
- marginal probability $p(y)$ of the test set X

Find

- conditional class probability estimates $p(y|x)$

Gaussian Process Solution

- Conditional exponential likelihood model

$$p(y|x,\theta) = \exp\left(\langle\phi(x,y),\theta\rangle - g(\theta|x)\right) \text{ with}$$
$$g(\theta|x) = \log \sum_{y\in\mathcal{Y}} \exp\langle\phi(x,y),\theta\rangle$$

- Gaussian prior

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

- Posterior

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

- Convex optimization problem

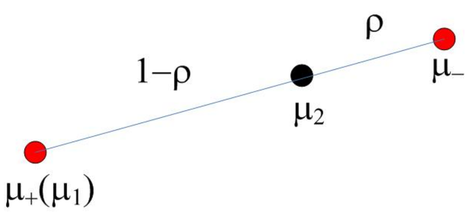
$$\theta^* = \operatorname*{argmin}_\theta \left[\sum_{i=1}^{m} g(\theta|x_i) - m\langle\mu_{XY},\theta\rangle + \lambda\|\theta\|^2\right] \text{ with}$$
$$\mu_{XY} := \frac{1}{m}\sum_{i=1}^{m}\phi(x_i,y_i)$$

However, we do not have y_i, label for each observation.

Re-calibration of Sufficient Statistics - Intuition

Convergence of empirical means:

$$\underset{sample}{\mu_{XY}} \longleftarrow \underset{population}{\mu_{xy} := \sum_{y\in\mathcal{Y}} p(y)\mathbf{E}_{x\sim p(x|y)}[\phi(x,y)]} \longleftarrow \underset{population}{\mu_x^{\text{set}}} \longleftarrow \underset{sample}{\mu_X^{\text{set}}}$$



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

$$\hat{\mu}_{XY} = \rho\mu_1 - (1-\rho)\left[\tfrac{-\rho}{1-\rho}\mu_1 + \tfrac{1}{1-\rho}\mu_2\right]$$

Same procedure applies for multiclass setting .

Performance Guaranteed!

- Bound on the mean operator
  With probability $1-\delta$ the following bound holds:

$$\|\mu_{XY} - \hat{\mu}_{XY}\| \leq \left[2 + \sqrt{\log((n+1)/\delta)}\right] \times$$
$$\left[m^{-\frac{1}{2}} + \left[\sum_{i,j} m_i^{-\frac{1}{2}} m_j^{-\frac{1}{2}} \left|\left[\pi^{-1}\right]^\top K^{y,p} \pi^{-1}\right|_{ij}\right]^{\frac{1}{2}}\right]$$

  For binary classification, the bound simplifies to

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

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

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

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

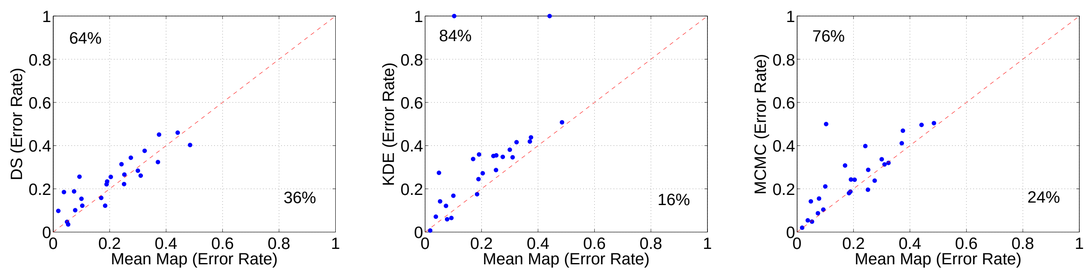
$$L(\hat{\theta}^*, \hat{\mu}) - L(\theta^*, \mu) \leq \left\|\hat{\theta}^* - \theta^*\right\| \|\hat{\mu} - \mu\| = \lambda^{-1}\|\mu - \hat{\mu}\|^2$$

Alternative Solutions

- Reduction to binary (DS)
  - a binary classifier between set X_1 and X_2
  - label thresholding according to the known proportions
- Density estimation (KDE)
  - density estimation for each dataset X_i
  - re-calibration to get $p(x|y)$ via $\sum_i \left[\pi^{-1}\right]_{yi} p(x,y|i)$
  - finally invoke Bayes' theorem to compute posterior probabilities.
- MCMC (Kück & de Freitas 2005)
  - explicitly generate mixing proportions per group by hierarchical probabilistic graphical model
  - use sampling to generate samples of model posterior distribution

Experiments

- Binary



- 3-class

Data	Mean Map	KDE	DS	MCMC	BA
protein A	44.6±0.3	60.2±0.1	N/A	65.3±1.9	61.2
protein B	45.7±0.6	61.2±0.0	N/A	67.7±1.8	61.2
dna A	16.6±1.0	30.7±0.8	N/A	37.7±0.8	40.5
dna B	29.1±1.0	33.0±0.7	N/A	40.5±0.0	40.5
senseit A	19.8±0.1	43.1±0.0	N/A	‡	43.2
senseit B	21.0±0.1	43.1±0.0	N/A	‡	43.2

‡: Program as implemented fails (large datasets)

Extensions

- 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 term of ℓ_∞ norm recovers sparse coding ℓ_1 (dual connection)

Summary

- A new problem formulation 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!