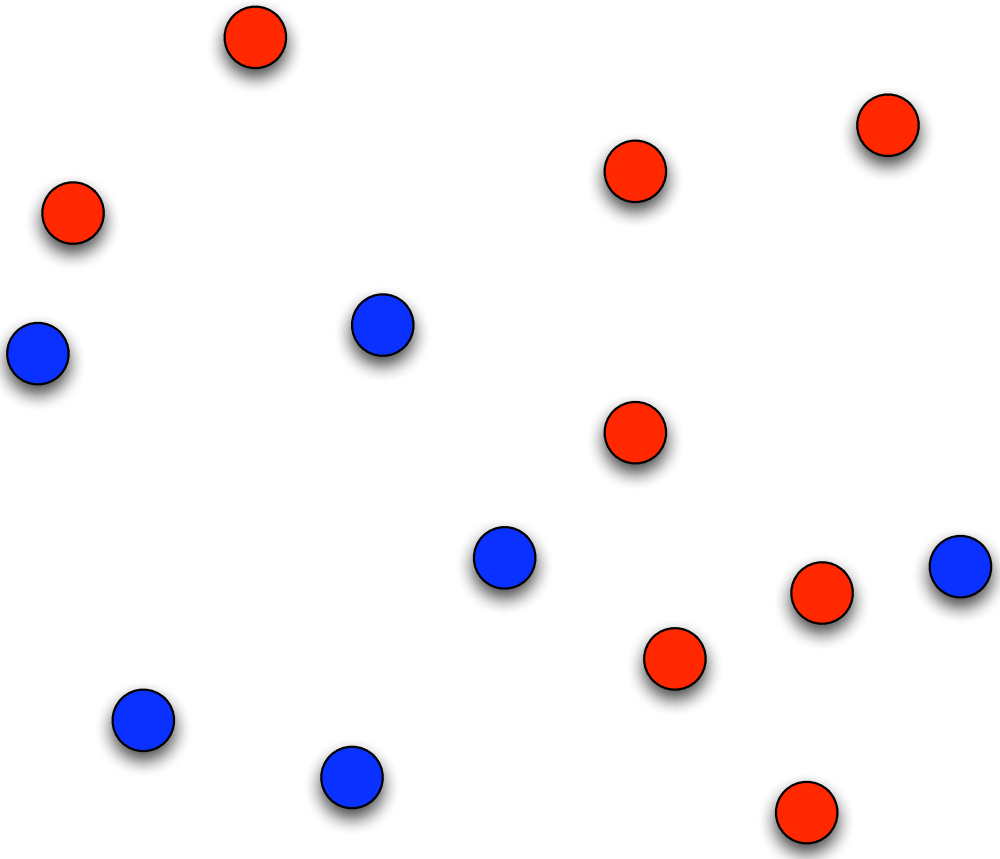
# Estimating Labels from Label Proportions

**Novi Quadrianto**
Novi.Quad@gmail.com

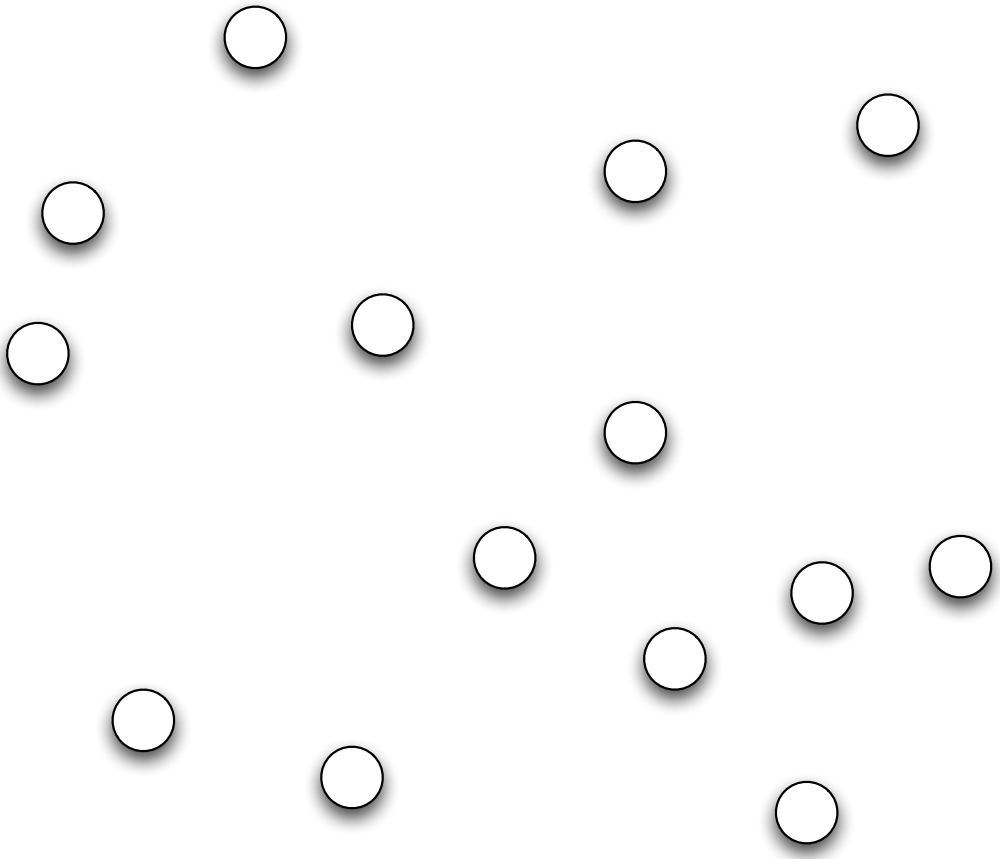
The Australian National University, Australia
NICTA, Statistical Machine Learning Program, Australia

Joint work with Alex Smola, Tiberio Caetano, and Quoc Le

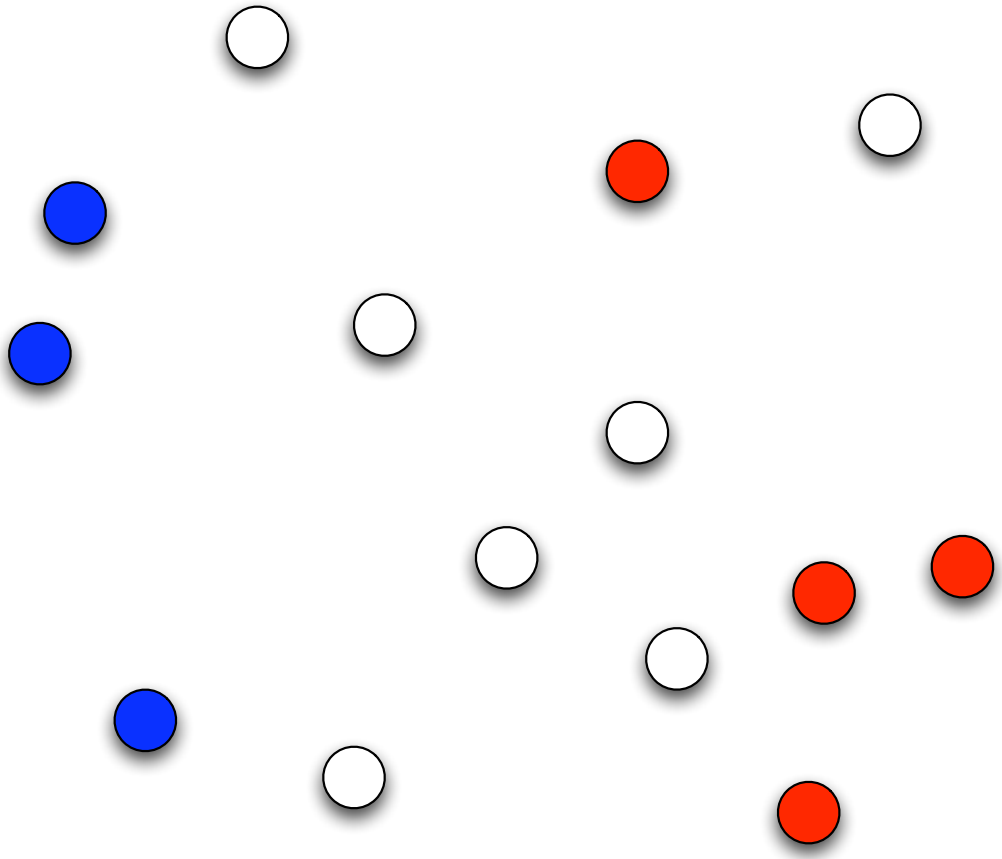
# Supervised Learning



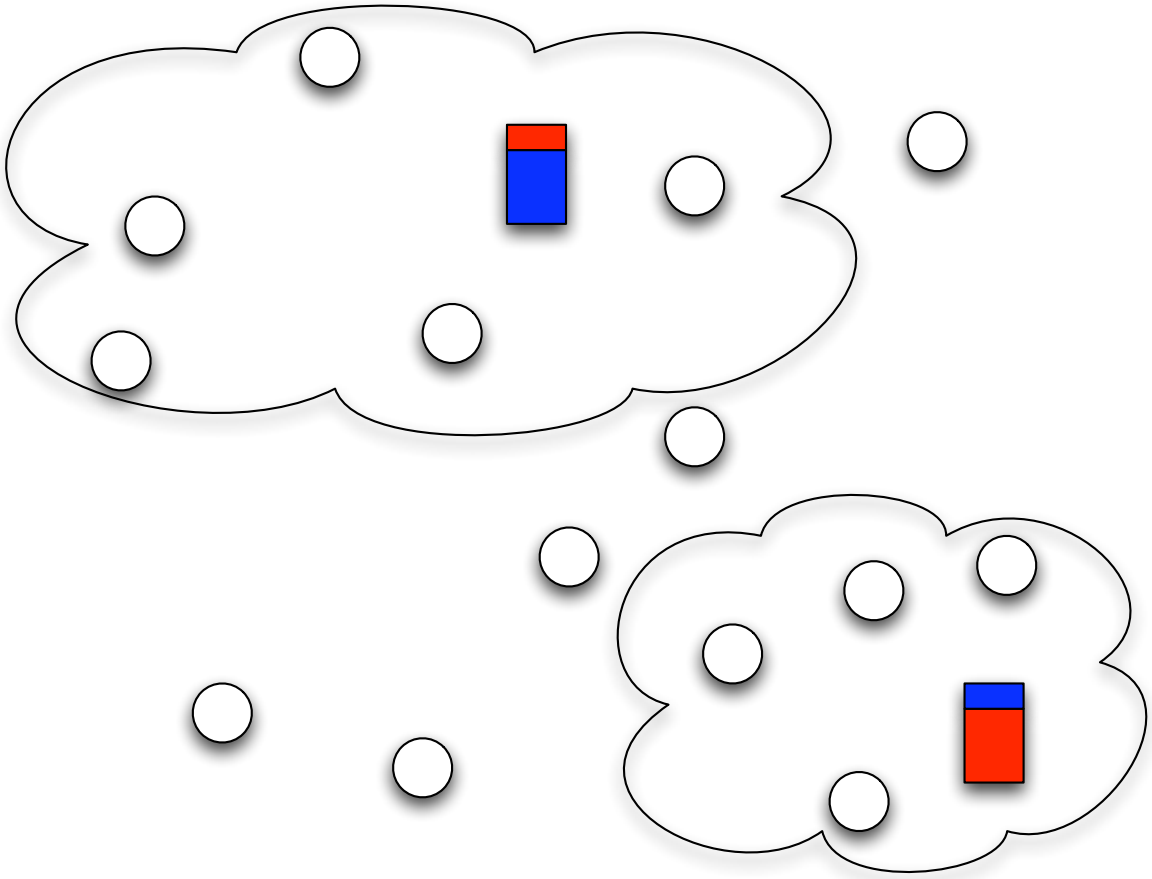
# Unsupervised Learning



# Semi-supervised Learning



# Learning from Proportions



# 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?

# An example application

## 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

| | Buy | Doesn't Buy |
|---|---|---|
| Exp. 1 : Given Coupon | A $\cup$ C | N |
| Exp. 2 : Not Given Coupon | A | N $\cup$ C |

- Assumption: no defiers
- Fact: we don't have a pure sample of $C$, and we want $p(C|\text{customer profile})$

# An example application

- 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 $X_i = \{x_1^i, \ldots, x_{m_i}^i\}$ of respective sample sizes $m_i$ as calibration sets
- a set $X = \{x_1, \ldots, x_m\}$ as a test set
- fractions $\pi_{iy}$ 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 $X$

## What we want

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

# Gaussian process solution

## Conditional exponential likelihood model

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

Some details

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

## Gaussian prior

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

## Posterior

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

# Optimization

## Optimization

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

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

- This is a convex optimization problem
- So is our job done?

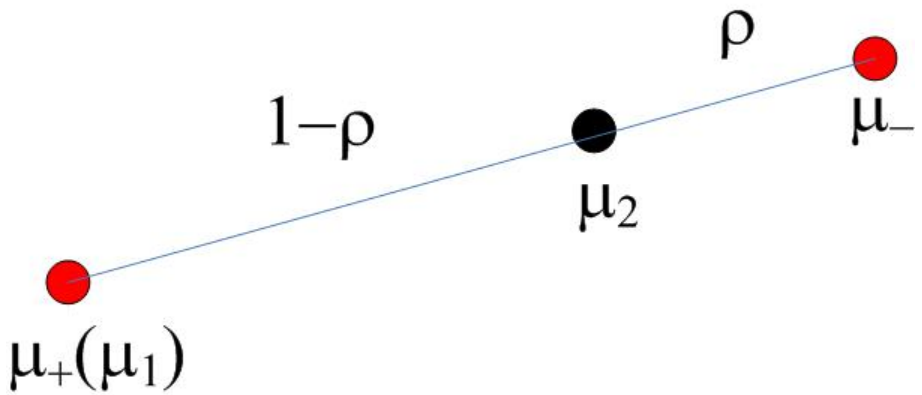
Convergence of empirical means (Bartlett & Mandelson 2002):

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

# Intuition

## Binary classification

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




$$\mu_+ := \mathbf{E}_{(x)\sim p(x|y=+1)}[\phi(x,y)]$$
$$\mu_1 := \mathbf{E}_{(x)\sim p(x|\text{set }1)}[\phi(x,y)]$$

# Re-calibrated sufficient statistics

## Binary classification

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

$$\Downarrow$$

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

$$\Downarrow$$

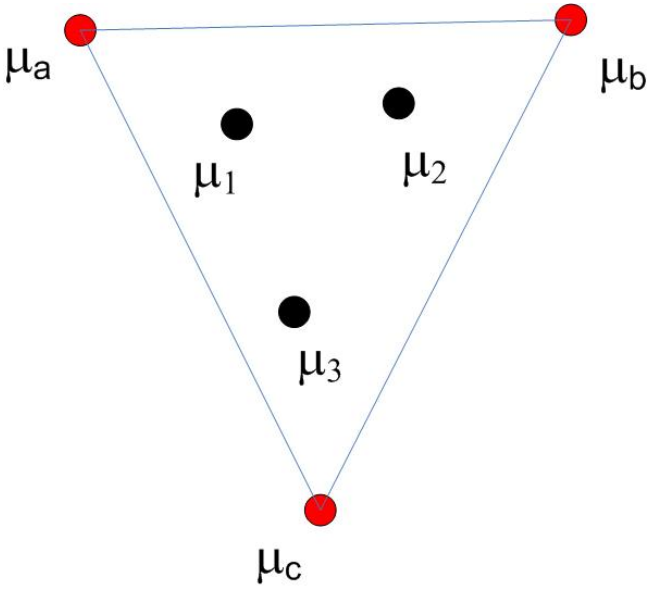
$$\begin{bmatrix} \mu_+ \\ \mu_- \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{-\rho}{1-\rho} & \frac{1}{1-\rho} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\Downarrow$$

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

# Generalization

## 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** $y' \in \mathcal{Y}$ **do**
  Compute empirical means $\mu_X^{\text{set}}[i, y']$
**end for**
Compute $\hat{\mu}_x^{\text{class}} = (\pi^\top \pi)^{-1} \pi^\top \mu_X^{\text{set}}$
Compute $\hat{\mu}_{XY} = \sum_{y \in \mathcal{Y}} p(y) \hat{\mu}_x^{\text{class}}[y, y]$
Solve the minimization problem

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

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

$$\mu_X^{\text{set}} \longrightarrow \hat{\mu}_x^{\text{class}} \longrightarrow \hat{\mu}_{XY}$$

# Performance guaranteed!

**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 $X_2$

# Performance guaranteed!

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

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

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

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

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 $\sum_i \left[\pi^{-1}\right]_{yi} p(x, y|i)$
- 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

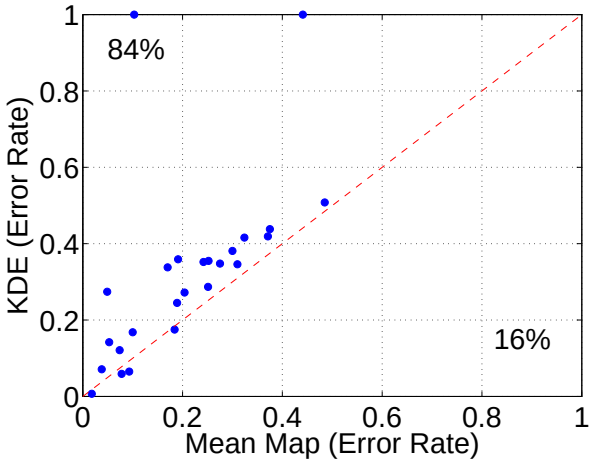
*Table 1.* Classification error on the UCI/LibSVM database

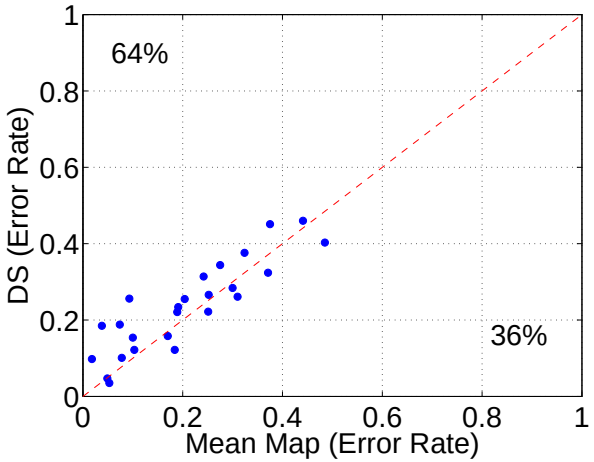
Errors are reported in % with standard error. (%) $\pm$ SE. The best result and those results not significantly worse than it, are highlighted in red. We used a one-sided paired Welch t-test with 95% confidence level as reference.

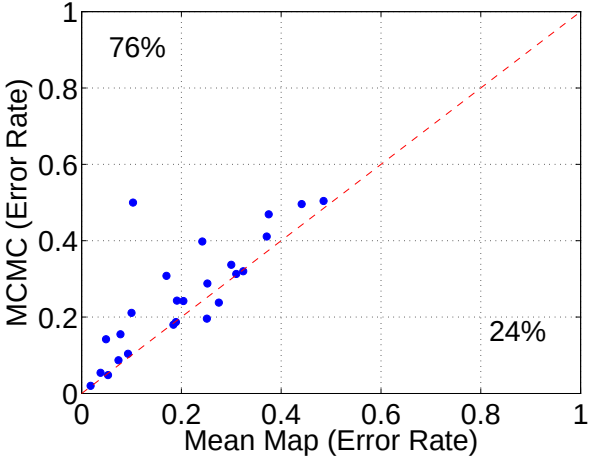
- `MM`: Mean Map (ours)
- `KDE`: Kernel Density Estimation
- `DS`: Discriminative Sorting
- `MCMC`: Sampling Method
- `BA`: Baseline

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

# Zooming in (binary results)



# Extensions

## 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 term of $\ell_\infty$ norm recovers sparse coding $\ell_1$ (dual connection)

# Summary

## 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!