

Cost-sensitive learning based on Bregman divergences

Raúl Santos-Rodríguez¹, Alicia Guerrero-Curieses², Rocío Alaiz-Rodríguez³ and Jesús Cid-Sueiro¹

¹ Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés (Madrid), Spain
`{rsrodriguez,jcid}@tsc.uc3m.es`

² Department of Signal Theory and Communications, Universidad Rey Juan Carlos, Fuenlabrada (Madrid), Spain
`alicia.guerrero@urjc.es`

³ Department of Electrical and Electronic Engineering, Universidad de León, León, Spain
`rocio.alaiz@unileon.es`

September 9, 2009

- 1 Introduction
- 2 Posterior probability estimation
- 3 Designing Bregman Divergences
- 4 Towards a maximum margin classifier
- 5 Conclusions

- **Cost-sensitive learning**: designing decision or regression machines that take into account the costs involved in the whole decision/estimation process.
- Decision: 3-class problem.

Hypothesis $\mathbf{u}_1 = (1, 0, 0)$
 $\mathbf{u}_2 = (0, 1, 0)$
 $\mathbf{u}_3 = (0, 0, 1)$

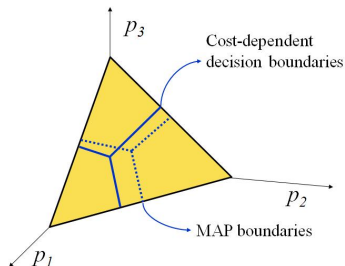
Decision costs $\rightarrow c(\hat{\mathbf{d}}, \mathbf{d})$
 $\hat{\mathbf{d}}$: decision (based on sample x)
 \mathbf{d} : true hypothesis.

Goal minimizing the mean risk $E\{c(\hat{\mathbf{d}}, \mathbf{d})\}$

- The minimum is reached by taking, for every sample \mathbf{x} , class $\hat{\mathbf{d}}^*$ such that

$$\hat{\mathbf{d}}^* = \arg \min_{\hat{\mathbf{d}}} \left\{ \sum_{j=1}^L E\{c(\hat{\mathbf{d}}, \mathbf{u}_j) | \mathbf{x}\} p_j \right\} \quad (1)$$

where $p_j = P\{\mathbf{d} = \mathbf{u}_j | \mathbf{x}\}$ is the posterior probability of class j given sample \mathbf{x} .



- General approaches have been proposed to deal with multiclass cost-sensitive problems:
 - ① Methods based on modifying the training data set.
 - ② Methods that change the learning process in order to build a binary cost-sensitive classifier.
 - ③ **Methods based on estimating posterior probabilities.**
- Accurate probability estimates are only required near the decision boundaries: Bregman divergences.

Posterior probability estimation

Classical discriminative approach:

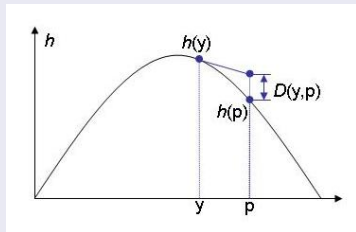
- estimating a posterior probability map $\mathbf{y} = \mathbf{f}_{\mathbf{w}}(\mathbf{x})$.

Definition (Bregman 1967)

(Bregman Divergence)

Given a concave function h , the *Bregman divergence* D_h is defined as

$$D_h(\mathbf{p}, \mathbf{y}) = h(\mathbf{y}) - h(\mathbf{p}) + (\mathbf{p} - \mathbf{y})^T \nabla_{\mathbf{y}} h(\mathbf{y}). \quad (2)$$



h could be thought to be an entropy.

Theorem (Savage 1971, Cid-Sueiro 1999, Banerjee 2005, ...)

Let (\mathbf{x}, \mathbf{d}) be a pair of random variables with arbitrary joint distribution $p(\mathbf{x}, \mathbf{d})$, and let \mathbf{p} be the posterior probability map given by $p_i = P\{\mathbf{d} = \mathbf{u}_i | \mathbf{x}\}$. The divergence measure D satisfies

$$\arg \min_{\mathbf{y}} E\{D(\mathbf{d}, \mathbf{y}) | \mathbf{x}\} = \arg \min_{\mathbf{y}} E\{D(\mathbf{p}, \mathbf{y}) | \mathbf{x}\} \quad (3)$$

for any distribution $p(\mathbf{x}, \mathbf{d})$ if and only if D is a Bregman divergence for some entropy measure h .

- Probability estimates minimizing the mean divergence can be found by minimizing $E\{D(\mathbf{d}, \mathbf{y})\}$. It can be estimated from samples.
- Since $\arg \min_{\mathbf{y}} E\{D(\mathbf{p}, \mathbf{y}) | \mathbf{x}\} = \mathbf{p}$, the posterior class probability vector is the minimizer of the expected divergence.

Our approach: estimation of posterior class probabilities by minimizing divergence sums given by

$$O(\mathbf{w}) = \sum_{k=1}^K D(\mathbf{d}^k, \mathbf{y}^k) \quad (4)$$

where $\mathbf{y}^k = \mathbf{f}_{\mathbf{w}}(\mathbf{x}^k)$.

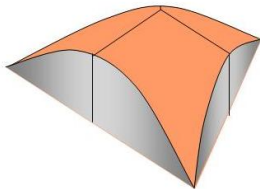
We aim to:

- optimize Bregman divergences very sensitive to deviations of \mathbf{y} from values of \mathbf{p} close to the decision boundaries and in the direction orthogonal to the boundary.
- design specific divergence measures for each decision problem.

- A cost-sensitive family of entropies:

$$h_R(\mathbf{y}) = -\|\mathbf{z}(\mathbf{y})\|_R \quad (5)$$

$$\begin{aligned} \mathbf{z}(\mathbf{y}) &= \mathbf{s} - \mathbf{C}\mathbf{y}; \\ \mathbf{s} &= \max_j \{\mathbf{c}_{ij}\}. \end{aligned}$$



- A cost-sensitive Bregman Divergence:

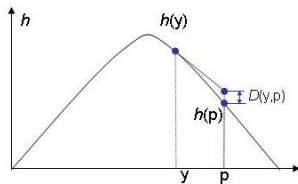
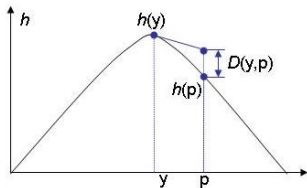
$$D_{h_R}(\mathbf{p}, \mathbf{y}) = h_R(\mathbf{y}) - h_R(\mathbf{p}) + (\mathbf{p} - \mathbf{y})^T \nabla_{\mathbf{y}} h_R(\mathbf{y}) \quad (6)$$

Designing Bregman Divergences: Sensitivity analysis

- Far from the boundary: when $R \rightarrow \infty$, then $\|z\|_R \rightarrow \max_i \{z_i\}$ and

$$\text{sensitivity} \rightarrow 0 \quad (7)$$

- At the boundary: at each point \mathbf{y} in the boundary between several decision regions, the sensitivity to directions along the boundary tends to zero, while it tends to ∞ for any orthogonal direction.

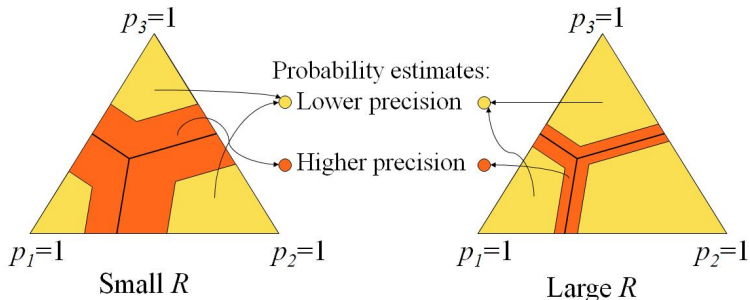


Designing Bregman Divergences: Sensitivity analysis

- Far from the boundary: when $R \rightarrow \infty$, then $\|z\|_R \rightarrow \max_i \{z_i\}$ and

$$\text{sensitivity} \rightarrow 0 \quad (8)$$

- At the boundary: at each point \mathbf{y} in the boundary between several decision regions, the sensitivity to directions along the boundary tends to zero, while it tends to ∞ for any orthogonal direction.



- Non-separable data

$$\lim_{R \rightarrow \infty} \sum_{k=1}^K D_{h_R}(\mathbf{d}^k, \mathbf{y}^k) = \sum_{k=1}^K \left(c_{\hat{m}^k m^k} - \min_i c_{i m^k} \right) \quad (9)$$

where m^k and \hat{m}^k represent the index of the true class and the assigned class for sample \mathbf{x}^k , respectively.

- The divergence converges to the difference in the total classification cost and the minimum achievable cost.

MAP case \rightarrow number of decision errors.

Designing Bregman Divergences: Experiments

Table: Average error rate. Exponential map: $y_i = \exp(\mathbf{w}_i^T \mathbf{x}) / (\sum_m \mathbf{w}_m^T \mathbf{x})$.

Dataset	Error rate (Test)			
	<i>BD</i>	<i>CE</i>	<i>Oversampling</i>	<i>Th.Moving</i>
German	0.232 ± 0.032	0.247 ± 0.031	0.244 ± 0.061	0.253 ± 0.041
Heart	0.144 ± 0.049	0.187 ± 0.053	0.193 ± 0.044	0.226 ± 0.060

Table: Average cost. Exponential map: $y_i = \exp(\mathbf{w}_i^T \mathbf{x}) / (\sum_m \mathbf{w}_m^T \mathbf{x})$.

Dataset	Cost (Test)			
	<i>BD</i>	<i>CE</i>	<i>Oversampling</i>	<i>Th.Moving</i>
German	43.2 ± 1.7	57.9 ± 3.3	45.9 ± 4.1	47.7 ± 1.5
Heart	9.1 ± 0.9	12.1 ± 1.3	11.7 ± 2.1	12.3 ± 1.4

In the Maximum A Posteriori (MAP) case and using an exponential probability map:

$$O(\mathbf{W}) \approx q^\ell \exp(-R \|\mathbf{w}_{n^\ell} - \mathbf{w}_{m^\ell}\|_2 d(\mathbf{x}^\ell, P_{n^\ell, m^\ell})) \quad (10)$$

where ℓ is the index of the sample in the training set that minimizes the negative of the exponent,

$$\ell = \arg \max_k \left\{ \|\mathbf{w}_{n^k} - \mathbf{w}_{m^k}\|_2 d(\mathbf{x}^k, P_{n^k, m^k}) \right\} \quad (11)$$

Imposing some constraints on the size of \mathbf{W} , the minimum of $O(\mathbf{W})$ is obtained by maximizing the distances from samples to decision boundaries: for large R , the classifier optimizing $O(\mathbf{W})$ tends to behave as a **maximum margin classifier**.

- We have proposed a parametric family of Bregman divergences that can be tuned to a specific cost matrix to estimate posterior probabilities.
- Sensitivity analysis: very sensitive to deviations of \mathbf{y} from values of \mathbf{p} close to the decision boundaries and in the direction orthogonal to the boundary.
- Asymptotic analysis: the optimization is equivalent to minimize the overall cost regret in non-separable problems.
- The error/cost results obtained are lower than those given by the cross entropy solely or combined with some well-known cost-sensitive algorithms.

Drawback Optimization stage

Alternative Maximum margin in separable problems.