Probabilistic Approach to One Class SVM

V. Leclère (ENPC), L. El Ghaoui, E. Graves

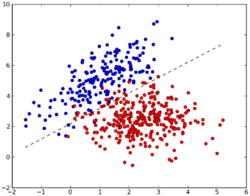
MODE 2016
March 25, 2016

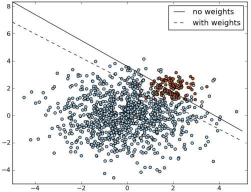
Presentation Outline

Contents

Binary classification

- Consider that you have a set of training data points $\{x_i\}_{i \in I}$.
- Each data is either labeled as a positive or negative point, through $y_i \in \{-1, +1\}$.
- We are looking for a linear classifier (w, b) such that
  - $w^T x_i - b \geq 0$, if $y_i = 1$
  - $w^T x_i - b \leq 0$, if $y_i = -1$
- We are intereseted in imbalanced classification where there are a lot more negative points than positive points.





Probabilistic formulation

- We represent the negative points as a random variable $\mathbf{x}$.
- We want a classifier that truly classify each positive point, and minimize the probability of a false negative, i.e.

$$\begin{aligned} \max_{\mathbf{w},b} \quad & \mathbb{P}(\mathbf{w}^\top \mathbf{x} - b \leq 0), & \\ s.t. \quad & \mathbf{w}^\top \mathbf{x}_i - b \geq 0, & \forall i \in I^+. \end{aligned}$$

- However, as we do not know the probability distribution of the negative points, we consider a robust approach where only the mean $\bar{\mathbf{x}}$ and covariance Σ are known.

$$\begin{aligned} \max_{\mathbf{w},b} \quad & \inf_{\mathbf{x} \sim (\bar{\mathbf{x}}, \Sigma)} \mathbb{P}(\mathbf{w}^\top \mathbf{x} - b \leq 0), & \\ s.t. \quad & \mathbf{w}^\top \mathbf{x}_i - b \geq 0, & \forall i \in I^+. \end{aligned}$$

Probabilistic formulation

- We represent the negative points as a random variable $\mathbf{x}$.
- We want a classifier that truly classify each positive point, and minimize the probability of a false negative, i.e.

$$\begin{aligned} \max_{\mathbf{w},b} \quad & \mathbb{P}(\mathbf{w}^\top \mathbf{x} - b \leq 0), \\ s.t. \quad & \mathbf{w}^\top \mathbf{x}_i - b \geq 0, \qquad \forall i \in I^+. \end{aligned}$$

- However, as we do not know the probability distribution of the negative points, we consider a robust approach where only the mean $\bar{\mathbf{x}}$ and covariance Σ are known.

$$\begin{aligned} \max_{\mathbf{w},b} \quad & \inf_{\mathbf{x}\sim(\bar{\mathbf{x}},\boldsymbol{\Sigma})} \mathbb{P}(\mathbf{w}^\top \mathbf{x} - b \leq 0), \\ s.t. \quad & \mathbf{w}^\top \mathbf{x}_i - b \geq 0, \qquad \forall i \in I^+. \end{aligned}$$

Contents

Toward an SVM-like representation 1/3

- The condition

$$\inf_{\mathbf{x}\sim(\bar{\mathbf{x}},\boldsymbol{\Sigma})} \mathbb{P}\left(\mathbf{x}^\top \mathbf{w} - b \leq 0\right) \geq \alpha,$$

holds if and only if

$$b - \bar{\mathbf{x}}^\top \mathbf{w} \geq \kappa(\alpha)\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}},$$

where $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$.

- Hence, our problem reads

$$\begin{aligned} \max_{\alpha,\mathbf{w},b} \quad & \alpha \\ \text{s.t.} \quad & b - \bar{\mathbf{x}}^\top \mathbf{w} \geq \kappa(\alpha)\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}, \\ & \mathbf{x}_i^\top \mathbf{w} - b \geq 0, \qquad \forall i \in I^+. \end{aligned}$$

Toward an SVM-like representation 1/3

- The condition

$$\inf_{\mathbf{x}\sim(\bar{\mathbf{x}},\boldsymbol{\Sigma})} \mathbb{P}\left(\mathbf{x}^\top \mathbf{w} - b \leq 0\right) \geq \alpha,$$

holds if and only if

$$b - \bar{\mathbf{x}}^\top \mathbf{w} \geq \kappa(\alpha)\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}},$$

where $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$.

- Hence, our problem reads

$$\begin{aligned} \max_{\alpha,\mathbf{w},b} \quad & \alpha \\ \text{s.t.} \quad & b - \bar{\mathbf{x}}^\top \mathbf{w} \geq \kappa(\alpha)\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}, \\ & \mathbf{x}_i^\top \mathbf{w} - b \geq 0, \qquad \forall i \in I^+. \end{aligned}$$

Toward an SVM-like representation 2/3

- $\kappa : \alpha \mapsto \sqrt{\frac{\alpha}{1-\alpha}}$ is increasing on $[0,1[$, this problem is equivalent to the program:

$$\begin{aligned} \max_{\kappa,\mathbf{w},b} \quad & \kappa \\ \text{s.t.} \quad & b - \bar{\mathbf{x}}^\top \mathbf{w} \geq \kappa\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}, \\ & \mathbf{x}_i^\top \mathbf{w} - b \geq 0, \qquad \forall i \in I^+. \end{aligned}$$

- Note that $w \neq 0$, hence we can impose $\kappa\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}} = 1$.
- Leading to

$$\begin{aligned} \max_{\mathbf{w},b} \quad & \frac{1}{\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}} \\ \text{s.t.} \quad & b - \bar{\mathbf{x}}^\top \mathbf{w} \geq 1, \\ & \mathbf{x}_i^\top \mathbf{w} - b \geq 0, \qquad \forall i \in I^+. \end{aligned}$$

Toward an SVM-like representation 2/3

- $\kappa : \alpha \mapsto \sqrt{\frac{\alpha}{1-\alpha}}$ is increasing on $[0, 1[$, this problem is equivalent to the program:

$$\begin{aligned} \max_{\kappa, \mathbf{w}, b} \quad & \kappa \\ \text{s.t.} \quad & b - \bar{\mathbf{x}}^\top \mathbf{w} \geq \kappa \sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}, \\ & \mathbf{x}_i^\top \mathbf{w} - b \geq 0, \qquad \forall i \in I^+. \end{aligned}$$

- Note that $w \neq 0$, hence we can impose $\kappa \sqrt{\mathbf{w}^\top \Sigma \mathbf{w}} = 1$.
- Leading to

$$\begin{aligned} \max_{\mathbf{w}, b} \quad & \frac{1}{\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}} \\ \text{s.t.} \quad & b - \bar{\mathbf{x}}^\top \mathbf{w} \geq 1, \\ & \mathbf{x}_i^\top \mathbf{w} - b \geq 0, \qquad \forall i \in I^+. \end{aligned}$$

Toward an SVM-like representation 2/3

- $\kappa : \alpha \mapsto \sqrt{\frac{\alpha}{1-\alpha}}$ is increasing on $[0,1[$, this problem is equivalent to the program:

$$\begin{aligned} \max_{\kappa,\mathbf{w},b} \quad & \kappa \\ \text{s.t.} \quad & b - \bar{\mathbf{x}}^\top \mathbf{w} \geq \kappa\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}, \\ & \mathbf{x}_i^\top \mathbf{w} - b \geq 0, \qquad \forall i \in I^+. \end{aligned}$$

- Note that $w \neq 0$, hence we can impose $\kappa\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}} = 1$.
- Leading to

$$\begin{aligned} \max_{\mathbf{w},b} \quad & \frac{1}{\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}} \\ \text{s.t.} \quad & b - \bar{\mathbf{x}}^\top \mathbf{w} \geq 1, \\ & \mathbf{x}_i^\top \mathbf{w} - b \geq 0, \qquad \forall i \in I^+. \end{aligned}$$

Toward an SVM-like representation 3/3

- Finally, since the function $x \mapsto 1/\sqrt{x}$ is decreasing on $\mathbb{R}_+^*$, we obtain the equivalent program:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^\top \Sigma \mathbf{w} \\ \text{s.t.} \quad & b - \bar{\mathbf{x}}^\top \mathbf{w} \geq 1, \\ & \mathbf{x}_i^\top \mathbf{w} - b \geq 0, \qquad \forall i \in I^+. \end{aligned}$$

- Which is a Support Vector Machine formulation with two differences :
  - instead of minimizing the ℓ_2-norm of **w**, we minimize the Mahalanobis norm corresponding to the covariance matrix of the negative class distribution,
  - the negative class contains only $\bar{\mathbf{x}}$.

Link to the One-Class SVM

- In the previous formulation we see that the optimal b can be derived as

$$b^\sharp = \min_{i \in I^+} x_i^T w^\sharp = 1 + \bar{\mathbf{x}}^\top \mathbf{w}^\sharp$$

- Hence, we have the formulation

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^\top \Sigma \mathbf{w} \\ \text{s.t.} \quad & (\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{w} \geq 1, \qquad \forall i \in I^+. \end{aligned}$$

- This is almost a one-class SVM :
  - we separate in the Mahalanobis norm,
  - we separate from the mean of the negative class instead of the origin.
- It is equivalent to apply classical one class SVM to preprocessed positive datapoints:

$$\tilde{x}_i = \Sigma^{-1/2}(x_i - \bar{x}).$$

Contents

Non-separability

- Our formulation require that the mean of the negative class is not in the convex hull of the positive points.
- To relax this strong assumption we add slack variable ξ_i, which are penalized.
- The formulation reads

$$\begin{aligned} \min_{\mathbf{w},b,\xi\geq 0} \quad & \mathbf{w}^\top \Sigma \mathbf{w} + \frac{1}{\nu|I^+|} \sum_{i\in I^+} \xi_i \\ \text{s.t.} \quad & b - \bar{\mathbf{x}}^\top \mathbf{w} \geq 1, \\ & \mathbf{x}_i^\top \mathbf{w} - b + \xi_i \geq 0, \qquad \forall i \in I^+. \end{aligned}$$

Sparsity

- We might want the optimal classifier w to be sparse to be able to interpret it, and to regularize the solution.
- A classical way of asking for Sparsity consists in adding a penalization of the L_1 norm of the solution w.
- Leading to

$$\begin{aligned} \min_{\mathbf{w},b,\xi\geq 0} \quad & \mathbf{w}^\top \Sigma \mathbf{w} + \frac{1}{\nu|I^+|}\sum_{i\in I^+}\xi_i + \eta\|w\|_1 \\ \text{s.t.} \quad & b - \bar{\mathbf{x}}^\top \mathbf{w} \geq 1, \\ & \mathbf{x}_i^\top \mathbf{w} - b + \xi_i \geq 0, \qquad \forall i \in I^+. \end{aligned}$$

Contents

Factor model of covariance matrix

- To reduce the computation cost and to reduce the effect of noise in the data we look to a factor model of the covariance matrix.
- More precisely we look for matrices D and F such that

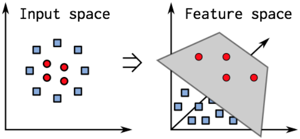
$$\Sigma \approx D + FF^T,$$

  where D is diagonal and F is $n \times k$ with $k << n$.
- This approximation can be obtained by use of svd decomposition of the centered dataset.
- Note that if the data is sparse the svd decomposition of the centered data can be done efficiently.

Contents

Kernelization

- It is classical for different application to train svm with kernel.
- Basically a kernel approach consists in looking for non linear separation of the datapoints by considering a linear separation in a bigger space:
  - consider a feature function φ such that for any (x, y) we have $\varphi(x)^T\varphi(y) = K(x, y)$;
  - apply linear SVM to the points $\varphi(x_i)$.



- Practically it only requires to replace the scalar product $x_i^T x_j$ by another symmetric function $K(x_i, x_j)$ satisfying some properties.

Kernelization

- We have seen that our approach is equivalent to applying one-class SVM to some preprocessed data. So why not applying kernel to this preprocessed data ?
- We don't want to preprocess the datapoints (consume time and lose sparsity).
- However applying classical kernels (polynomial, RBF and sigmoidal) applied to the preprocessed data are equivalent to customized kernels applied to the original positive data.

Contents

Other method available

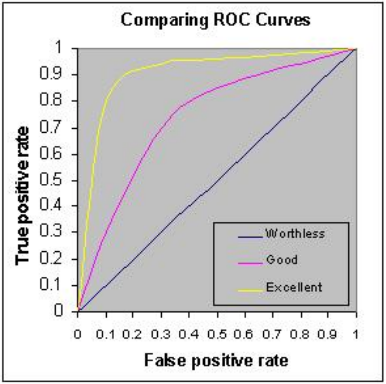
Standard binary classification on an imbalanced subset gives poor results for the small class. In particular most classifier have a tendency to always classify points as element of the big class.

Here are two classical ways of dealing with this problem

- Subsampling: randomly selectionning a subsample of the negative class and consider it as the whole negative class.
- Differential costs: assign different penality C to each class, penalizing more any misclassification of the small class.

Area Under the Curve Metric

- linear classifier: $w^T x_i + b$.
- Parameter b depend on your willingness to have false positive or false negative.
- The ROC curve is a curve in the true positive / false positive plane.
- Interpretation : take randomly a positive and a negative point. AUC is the probability of finding the positive with classifier w.



Contents

Dataset presentation

Dataset	# positive	# negative	ratio
PhosS	613	10,798	17
PhosT	140	9,051	64
PhosY	136	5,103	37
CaM	942	17,974	19

Table : Basic statistics about the different datasets.

Available here :
www.informatics.indiana.edu/predrag/publications.htm

Results comparison

	This work	Cost-sensitive	Sampling
PhosS	$77.2^{\dagger} \pm 0.7$	76.8 ± 0.8	74.3 ± 1.1
PhosT	$77.4^{\dagger} \pm 1.7$	73.0 ± 2.0	72.0 ± 1.5
PhosY	$76.2^{\dagger} \pm 1.5$	72.8 ± 1.7	70.1 ± 2.1
CaM	78.2 ± 0.5	78.1 ± 0.5	75.3 ± 0.4

Table : Areas under the ROC curve (with confidence intervals), averaged over twenty experiments. † indicates that our method is significantly better than the two others, (with p-value $p < 0.01$).

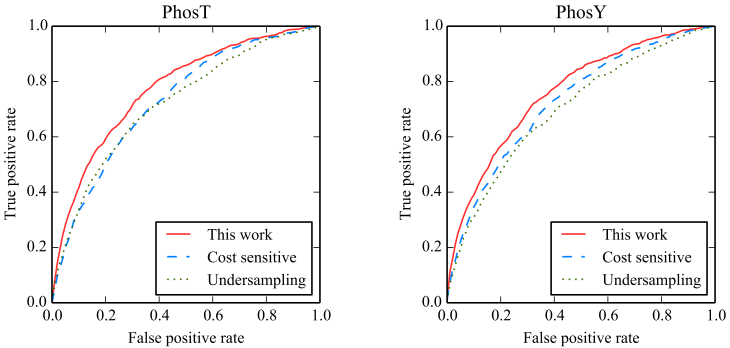


Figure : ROC curves, averaged over twenty experiments, on the PhosT and PhosY datasets.

Contents

The reuter dataset

- About 200'000 documents, with 50'000 features, classified in 40 classes.
- Available on Liblinear website.
- We select one-class to be positive all the other are negative.

Reuters results

Topic	This work	Cost-sensitive	Sampling
2	89.7 ± 1.0	89.9 ± 1.4	87.7 ± 1.2
9	96.1 ± 0.7	96.3 ± 0.8	94.1 ± 1.3
25	95.1 ± 0.8	94.3 ± 1.6	93.7 ± 1.2
33	96.0 ± 0.4	95.7 ± 0.6	93.9 ± 0.7
59	96.1 ± 0.4	95.9 ± 1.4	95.0 ± 0.6
84	96.9 ± 0.8	96.4 ± 1.5	96.3 ± 0.9

Table : Areas under the ROC curve (with confidence intervals), averaged over ten experiments. Differences between our moment-based imbalanced binary classifier and subsampling results are statistically significant (with p-value $p < 0.01$).

Reuters results II

Topic	This work	Cost-sensitive	Speed-up
2	33	1088	33×
9	49	1451	29×
25	56	1211	21×
33	74	1788	24×
59	62	1299	21×
84	56	2056	36×

Table : Computational times, in milliseconds, required to solve one problem, averaged over ten experiments.

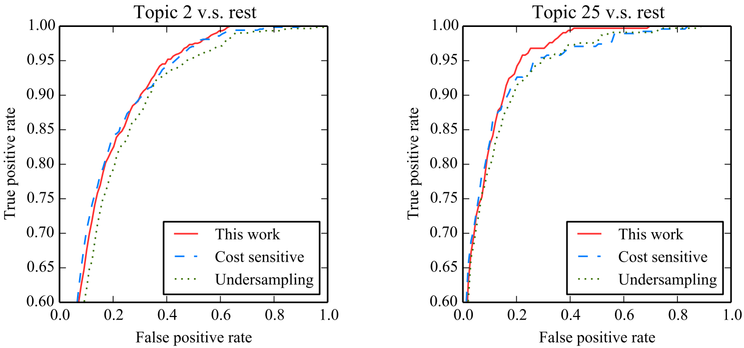


Figure : ROC curves, averaged over ten experiments, on the REUTERS RCV1 dataset.

Conclusion

- We give a theoretical interpretation for the one-class SVM method.
- We show how to adapt the one-class SVM in the case where we have first and second order information over the negative class.
- We apply this approach to imbalanced classification with good results both in precision and in computational speed.
- We apply this approach to large-scale imbalanced classification with significative speed improvement.

The end

Thank you for your attention !