Probabilistic Approach to One Class SVM

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

MODE 2016
March 25, 2016
Presentation Outline

1. A probabilistic approach to binary classification
   - Problem formulation
   - Toward the SVM formulation

2. Other elements to take into account
   - Some other ingredients
   - Large scale problem
   - Kernelization

3. Numerica results and comparisons
   - Some recalls
   - Small experiment: protein classification
   - Large experiment: text classification

4. Conclusion
2 Contents

1 A probabilistic approach to binary classification
   - Problem formulation
   - Toward the SVM formulation

2 Other elements to take into account
   - Some other ingredients
   - Large scale problem
   - Kernelization

3 Numerica results and comparisons
   - Some recalls
   - Small experiment : protein classification
   - Large experiment : text classification

4 Conclusion
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 interested 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 $x$.
- We want a classifier that truly classify each positive point, and minimize the probability of a false negative, i.e.

$$\max_{w,b} \mathbb{P}(w^T x - b \leq 0),$$

subject to

$$w^T x_i - b \geq 0, \quad \forall i \in I^+. $$

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

$$\max_{w,b} \inf_{x \sim (\bar{x}, \Sigma)} \mathbb{P}(w^T x - b \leq 0),$$

subject to

$$w^T x_i - b \geq 0, \quad \forall i \in I^+. $$
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.
  \[
  \max_{\mathbf{w},b} \mathbb{P}(\mathbf{w}^\top \mathbf{x} - b \leq 0),
  \]
  \[
  \text{s.t.} \quad \mathbf{w}^\top \mathbf{x}_i - b \geq 0, \quad \forall i \in I^+.
  \]

- 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 $\Sigma$ are known.
  \[
  \max_{\mathbf{w},b} \inf_{\mathbf{x} \sim (\bar{\mathbf{x}},\Sigma)} \mathbb{P}(\mathbf{w}^\top \mathbf{x} - b \leq 0),
  \]
  \[
  \text{s.t.} \quad \mathbf{w}^\top \mathbf{x}_i - b \geq 0, \quad \forall i \in I^+.
  \]
A probabilistic approach to binary classification

1. Problem formulation
2. Toward the SVM formulation

Other elements to take into account

- Some other ingredients
- Large scale problem
- Kernelization

Numerica results and comparisons

- Some recalls
- Small experiment: protein classification
- Large experiment: text classification

Conclusion
Toward an SVM-like representation 1/3

The condition

$$\inf_{x \sim (\bar{x}, \Sigma)} \mathbb{P}(x^T w - b \leq 0) \geq \alpha,$$

holds if and only if

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

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

Hence, our problem reads

$$\max_{\alpha, w, b} \alpha$$

s.t. $$b - \bar{x}^T w \geq \kappa(\alpha) \sqrt{w^T \Sigma w},$$

$$x_i^T w - b \geq 0, \quad \forall i \in l^+.$$
The condition

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

holds if and only if

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

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

Hence, our problem reads

$$\max_{\alpha, w, b} \alpha$$

s.t. $$b - \bar{x}^T w \geq \kappa(\alpha) \sqrt{w^T \Sigma w},$$

$$x_i^T w - b \geq 0, \quad \forall i \in I^+. $$
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:

$$\begin{align*}
\max_{\kappa, w, b} & \quad \kappa \\
\text{s.t.} & \quad b - \bar{x}^T w \geq \kappa \sqrt{w^T \Sigma w}, \\
& \quad x_i^T w - b \geq 0, \quad \forall i \in I^+.
\end{align*}$$

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

$$\begin{align*}
\max_{w, b} & \quad \frac{1}{\sqrt{w^T \Sigma w}} \\
\text{s.t.} & \quad b - \bar{x}^T w \geq 1, \\
& \quad x_i^T w - b \geq 0, \quad \forall i \in I^+.
\end{align*}$$
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:

$$\begin{align*}
\max_{\kappa, w, b} & \quad \kappa \\
\text{s.t.} & \quad b - \bar{x}^T w \geq \kappa \sqrt{w^T \Sigma w}, \\
& \quad x_i^T w - b \geq 0, \quad \forall i \in I^+. 
\end{align*}$$

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

Leading to

$$\begin{align*}
\max_{w, b} & \quad \frac{1}{\sqrt{w^T \Sigma w}} \\
\text{s.t.} & \quad b - \bar{x}^T w \geq 1, \\
& \quad x_i^T w - b \geq 0, \quad \forall i \in I^+. 
\end{align*}$$
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:

$$\max_{\kappa, w, b} \kappa$$

s.t. $b - \bar{x}^T w \geq \kappa \sqrt{w^T \Sigma w}$,

$x_i^T w - b \geq 0, \quad \forall i \in I^+.$

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

$$\max_{w, b} \frac{1}{\sqrt{w^T \Sigma w}}$$

s.t. $b - \bar{x}^T w \geq 1$,

$x_i^T w - b \geq 0, \quad \forall i \in I^+.$
Finally, since the function $x \mapsto 1/\sqrt{x}$ is decreasing on $\mathbb{R}^*_+$, we obtain the equivalent program:

$$\begin{align*}
\min_w & \quad w^\top \Sigma w \\
\text{s.t.} & \quad b - \bar{x}^\top w \geq 1, \\
& \quad x_i^\top w - b \geq 0, \quad \forall i \in I^+.
\end{align*}$$

Which is a Support Vector Machine formulation with two differences:

- instead of minimizing the $\ell_2$-norm of $w$, we minimize the Mahalanobis norm corresponding to the covariance matrix of the negative class distribution,
- the negative class contains only $\bar{x}$.
Link to the One-Class SVM

- In the previous formulation we see that the optimal $b$ can be derived as
  \[ b^\# = \min_{i \in I^+} x_i^T w^\# = 1 + \overline{x}^T w^\# \]

- Hence, we have the formulation
  \[
  \min_w w^T \Sigma w \\
  \text{s.t. } (x_i - \overline{x})^T w \geq 1, \quad \forall i \in I^+.
  \]

- 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:
\[ \tilde{x}_i = \Sigma^{-1/2}(x_i - \overline{x}). \]
Contents

1 A probabilistic approach to binary classification
   • Problem formulation
   • Toward the SVM formulation

2 Other elements to take into account
   • Some other ingredients
   • Large scale problem
   • Kernelization

3 Numerica results and comparisons
   • Some recalls
   • Small experiment : protein classification
   • Large experiment : text classification

4 Conclusion
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 $\xi_i$, which are penalized.
- The formulation reads:

$$\min_{w, b, \xi \geq 0} w^\top \Sigma w + \frac{1}{\nu |I^+|} \sum_{i \in I^+} \xi_i$$

s.t.  
- $b - \bar{x}^\top w \geq 1$,
- $x_i^\top w - b + \xi_i \geq 0, \quad \forall i \in I^+$. 
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

\[
\min_{w, b, \xi \geq 0} \quad w^\top \Sigma w + \frac{1}{\nu |I^+|} \sum_{i \in I^+} \xi_i + \eta \| w \|_1 \\
\text{s.t.} \quad b - \bar{x}^\top w \geq 1, \\
\quad \bar{x}_i^\top w - b + \xi_i \geq 0, \quad \forall i \in I^+.
\]
1 A probabilistic approach to binary classification
   - Problem formulation
   - Toward the SVM formulation

2 Other elements to take into account
   - Some other ingredients
   - Large scale problem
   - Kernelization

3 Numerica results and comparisons
   - Some recalls
   - Small experiment: protein classification
   - Large experiment: text classification

4 Conclusion
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 $k \ll n$.

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

1 A probabilistic approach to binary classification
   - Problem formulation
   - Toward the SVM formulation

2 Other elements to take into account
   - Some other ingredients
   - Large scale problem
   - Kernelization

3 Numerica results and comparisons
   - Some recalls
   - Small experiment: protein classification
   - Large experiment: text classification

4 Conclusion
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 $\varphi$ 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.
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

1 A probabilistic approach to binary classification
   - Problem formulation
   - Toward the SVM formulation

2 Other elements to take into account
   - Some other ingredients
   - Large scale problem
   - Kernelization

3 Numerica results and comparisons
   - Some recalls
   - Small experiment : protein classification
   - Large experiment : text classification

4 Conclusion
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 penalty $C$ 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 $w$. 

![Comparing ROC Curves](image)
Contents

1. A probabilistic approach to binary classification
   - Problem formulation
   - Toward the SVM formulation

2. Other elements to take into account
   - Some other ingredients
   - Large scale problem
   - Kernelization

3. Numerica results and comparisons
   - Some recalls
   - **Small experiment : protein classification**
   - Large experiment : text classification

4. Conclusion
Dataset presentation

<table>
<thead>
<tr>
<th>Dataset</th>
<th># positive</th>
<th># negative</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhoSS</td>
<td>613</td>
<td>10,798</td>
<td>17</td>
</tr>
<tr>
<td>PhoST</td>
<td>140</td>
<td>9,051</td>
<td>64</td>
</tr>
<tr>
<td>PhoSY</td>
<td>136</td>
<td>5,103</td>
<td>37</td>
</tr>
<tr>
<td>CAM</td>
<td>942</td>
<td>17,974</td>
<td>19</td>
</tr>
</tbody>
</table>

Table: Basic statistics about the different datasets.

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

<table>
<thead>
<tr>
<th></th>
<th>This work</th>
<th>Cost-sensitive</th>
<th>Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhosS</td>
<td>77.2(\dagger) ± 0.7</td>
<td>76.8 ± 0.8</td>
<td>74.3 ± 1.1</td>
</tr>
<tr>
<td>PhosT</td>
<td>77.4(\dagger) ± 1.7</td>
<td>73.0 ± 2.0</td>
<td>72.0 ± 1.5</td>
</tr>
<tr>
<td>PhosY</td>
<td>76.2(\dagger) ± 1.5</td>
<td>72.8 ± 1.7</td>
<td>70.1 ± 2.1</td>
</tr>
<tr>
<td>CaM</td>
<td>78.2 ± 0.5</td>
<td>78.1 ± 0.5</td>
<td>75.3 ± 0.4</td>
</tr>
</tbody>
</table>

Table: Areas under the ROC curve (with confidence intervals), averaged over twenty experiments. \(\dagger\) indicates that our method is significantly better than the two others, (with \(p\)-value \(p < 0.01\)).
A probabilistic approach to binary classification
Other elements to take into account
Numerical results and comparisons
Conclusion

Some recalls
Small experiment : protein classification
Large experiment : text classification

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

1. A probabilistic approach to binary classification
   • Problem formulation
   • Toward the SVM formulation

2. Other elements to take into account
   • Some other ingredients
   • Large scale problem
   • Kernelization

3. Numerica results and comparisons
   • Some recalls
   • Small experiment: protein classification
   • Large experiment: text classification

4. Conclusion
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

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

**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

<table>
<thead>
<tr>
<th>Topic</th>
<th>This work</th>
<th>Cost-sensitive</th>
<th>Speed-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>33</td>
<td>1088</td>
<td>33\times</td>
</tr>
<tr>
<td>9</td>
<td>49</td>
<td>1451</td>
<td>29\times</td>
</tr>
<tr>
<td>25</td>
<td>56</td>
<td>1211</td>
<td>21\times</td>
</tr>
<tr>
<td>33</td>
<td>74</td>
<td>1788</td>
<td>24\times</td>
</tr>
<tr>
<td>59</td>
<td>62</td>
<td>1299</td>
<td>21\times</td>
</tr>
<tr>
<td>84</td>
<td>56</td>
<td>2056</td>
<td>36\times</td>
</tr>
</tbody>
</table>

**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.
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.
Thank you for your attention!