Support Vector Machines

Classifying data is a common task in machine learning. Suppose some given data points each belong to one of two classes, and the goal is to decide which class a new data point will be in. The training set consists of $n$ data points (observations) in dimension $d$:

$$\{ (x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\} \mid i = 1, 2, \ldots, n \}$$  \hspace{1cm} (1)

where $y_i$ is the class label of the $i$-th observation (point in $\mathbb{R}^d$). Once an hyperplane is computed ($w_0, b_0$), we can determine the following classifier:

$$\mathbf{x} \mapsto \text{sgn}(w_0 \cdot \mathbf{x} + b_0)$$

SVM Models for Linearly non-separable datasets

• The SVM with Hinge loss

Introducing a set of continuous variables $\xi_i$ ($i = 1, 2, \ldots, n$) with the following meaning:

$$\xi_i = \begin{cases} 0 & \text{if } y_i(w \cdot x_i + b) \geq 1 \\ 1 - y_i(w \cdot x_i + b) & \text{if } y_i(w \cdot x_i + b) < 1 \end{cases}$$

The model reads as follows:

$$\min \frac{1}{2} \|w\|^2 + C \left( \sum_{i=1}^{n} \xi_i \right)$$  \hspace{1cm} (2)

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \hspace{1cm} i = 1, 2, \ldots, n$$  \hspace{1cm} (3)

$$\xi_i \geq 0 \hspace{1cm} i = 1, 2, \ldots, n$$  \hspace{1cm} (4)

• The SVM with Ramp loss

Introducing a set of binary variables $z_i$ and continuous variables $\xi_i$ ($i = 1, 2, \ldots, n$) with the following meaning:

$$z_i = \begin{cases} 1 & \text{if observation } x_i \text{ is misclassified outside of the margin} \\ 0 & \text{otherwise} \end{cases}$$
The model reads as follows:

$$\xi_i = \begin{cases} 
1 - y_i(wx_i + b) & \text{if } x_i \text{ falls in the margin} \\
0 & \text{otherwise}
\end{cases}$$

The model reads as follows:

$$\min \frac{1}{2} ||w||^2 + C \left( \sum_{i=1}^{n} \xi_i + 2 \sum_{i=1}^{n} z_i \right)$$  \hspace{1cm} (5)$$

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i - M z_i \quad i = 1, 2, \ldots, n$$  \hspace{1cm} (6)$$

$$z_i \in \{0, 1\} \quad i = 1, 2, \ldots, n$$  \hspace{1cm} (7)$$

$$0 \leq \xi \leq 2 \quad i = 1, 2, \ldots, n$$  \hspace{1cm} (8)$$

where $M$ is a big-M value (large value).

- The SVM with Hard Margin Loss

Introducing a set of binary variables $z_i (i = 1, 2, \ldots, n)$ with the following meaning:

$$z_i = \begin{cases} 
1 & \text{if observation } x_i \text{ is misclassified} \\
0 & \text{otherwise}
\end{cases}$$

The model reads as follows:

$$\min \frac{1}{2} ||w||^2 + C \left( \sum_{i=1}^{n} z_i \right)$$  \hspace{1cm} (9)$$

$$y_i(w \cdot x_i + b) \geq 1 - M z_i \quad i = 1, 2, \ldots, n$$  \hspace{1cm} (10)$$

$$z_i \in \{0, 1\} \quad i = 1, 2, \ldots, n$$  \hspace{1cm} (11)$$

where $M$ is a big-M value (large value).

Case Study – Investigate the performance of the different SVM models

- Consider data sets on the web page http://www.lamsade.dauphine.fr/~furini/). Further information on these real-world training data can be found here: https://archive.ics.uci.edu/ml/datasets.html. Analyse the instances describing their main features, with particular attention to the presence of the outliers. Divide the data sets into the training set and validation set. Compute the classifier using the 3 different SVM models on the training data set and measure its quality in terms of correctly classified points of the validation data set. Determine a good compromise between the different components of the objective function by tuning the parameter $C$.

- Discuss the advantages and the disadvantages of the classifier obtained and propose possible solutions for the principal problems. Investigate how sensitive the hyperplanes are to increasing dimensions of the training sets. Discuss the results as you were in machine learning data center to convince the management to adopt or not the proposed classification method.

- For the Ramp and Hard Margin Loss the presence of binary variable make the models harder from a computational point of view. Determine the maximum size of the models (in terms of number of observations) which can be solved to proven optimality within 10 minutes of CPU time.