Lecture Notes on Machine Learning

February 25, 2014

First, we will introduce how to build a classifier on multiple classes, we use the so-called strategy "one v.s. rest". Suppose in a data set, we have multiple classes: C_1, C_2, C_3, C_4, based on the method we use, we can build four binary classifiers L_i, which is defined as:

$$\begin{cases} L_i(x) > 0 \Longrightarrow x \ in \ class \ C_i, \\ L_i(x) < 0 \Longrightarrow x \ in \ class \ M_i. \end{cases}$$

where $M_i = C_1 \cup ... \cup C_{i-1} \cup C_{i+1} \cup ... \cup C_4$. Then we define the scores on class C_i by $\max\{L_i(x), 0\}$, compare all the scores on different classes, then x is predicted to be in the class with maximum score.

In previous lecture, we introduced the Gaussian kernel, and in this lecture, we will introduce a new kernel – polynomial kernel. It is defined as:

$$k(x, y) = (1+ < x, y >)^r$$

where $x, y \in \mathbb{R}^p$, and r is an integer and usually selected as $r \leq 5$.

To show it is a kernel, we need to show the matrix M with entries $M_{i,j} = k(x_i, x_j)$ is positive definite. To prove this, we need one important propostion:

Propostion 1. *If $A(x, y)$,$B(x, y)$ are kernels, then*

$$\begin{cases} k(x, y) = A(x, y) + B(x, y) \\ k(x, y) = A(x, y)B(x, y) \end{cases}$$

are also kernel.

Here, it is not difficult to find the matrix defined by functions $k_1(x, y) = 1$ and $k_2(x, y) =< x, y >$ are positive definite. Therefore, $k(x, y) = (1+ < x, y >)^r$ is a kernel by applying above propostion.

Next, we will take a quick review on SVM (support vector machine). Consider the Euclidean model of SVM build on feature space $\mathbb{R}^p$, where

$$x_i(observation) \Longrightarrow y_i(label) \begin{cases} +1 \ in \ 1st \ class \\ -1 \ in \ 2nd \ class \end{cases}$$

We want to get a seperator function:

$$g(x) =< u, x > +b$$

where we will have the prediction on x:

$$\begin{cases} g(x) > 0 \Longrightarrow class\ 1 \\ g(x) < 0 \Longrightarrow class\ 2 \end{cases}$$

However, when we build SVM on training set $\{x_i\}$, it is required that:

$$\begin{cases} g(x_i) \geq 1 \Longrightarrow class\ 1 \\ g(x_i) \leq -1 \Longrightarrow class\ 2 \end{cases}$$

which is equivalent to

$$y_i g(x_i) - 1 \geq 0$$

Define the size of error by

$$\xi_i = \max(0, 1 - y_i g(x_i))$$

then, in order to find u, b, we need to solve the following minimization problem:

$$\min\{\frac{1}{2}||u||^2 + c\sum \xi_i\}$$

with constraints

$$\begin{cases} \xi_i \geq 0 \\ \xi_i - (1 - y_i g(x_i)) \geq 0 \end{cases}$$

To solve this problem, we need to introduce Lagrange multipliers $\alpha_i, \beta_i \geq 0$, and use KKT algorithm, we will get a saddle point problem:

$$L(U) = \frac{1}{2}||u||^2 + c\sum \xi_i - \sum \alpha_i(\xi_i - (1 - y_i g(x_i))) - \sum \beta_i \xi_i$$

take partial derivatives, which are equal to 0:

$$\begin{aligned} \frac{\partial L}{\partial u} = 0 &\Longrightarrow u = \sum \alpha_i y_i x_i \\ \frac{\partial L}{\partial b} = 0 &\Longrightarrow \sum \alpha_i y_i = 0 \\ \frac{\partial L}{\partial \xi_i} = 0 &\Longrightarrow \alpha_i + \beta_i = c \end{aligned}$$

Let $G(\alpha, \beta) = \min L(U)$, then we need to solve the max problem:

$$\max_{0 \leq \alpha_i \leq c} G(\alpha, \beta)$$

To solve this max problem, we need to compare all the values on all the corners of (α, β). Since u is the linear combination of x_i with $\alpha_i > 0$, we call all of these x_i to be support vectors.

We need to check the performance on both training and test sets, in order to have good generalization capacity, we don't want to see significant difference between these two accuracies. For SVM, this condition will be satisfied if

$$aver(\frac{\#\ of\ support\ vectors}{N})$$

is small, where N is the total number of observations in training set.