Lecture Notes 15

Yanshu zhao

March 27, 2012

Contents

1 Leave One Out

Assume the data points are linear separable and iid. And the training and test data are D dimensions, while D is unknown, meaning our problem setting is distribution free.

$$Q(x,a) = \begin{cases} 1 & y \neq \hat{y}(x) \\ 0 & otherwise \end{cases}$$

Where Q is called loss. Risk is defined as expected loss:

$$R(a) = \int Q(x,a)P(x)dx$$

Now we need to know the maximum valve of the expectation of the risk over possible training sets. This means we need to calculate $E(R(a)) \leq ?$; In the past lecture notes, we know $W = \phi = \sum_i \alpha_i X_i Y_i$ where X_i is the ith data point.

$$Y_i = \begin{cases} -1 \\ 1 \end{cases}$$

In order to estimate the risk for the function $Q(z,a)$, we need to use the following statistics: exclude the first vector z_1 from the sequence and obtain the function that minimizes the empirical risk for the remaining $l-1$ elements of the sequence for the given sequence $z_1, z_2, \ldots, z_l$;

Let the function be $Q(z_1, a_{l-1}|z_1)$. In this notation, we indicate that the vector z_1 was excluded from the sequence. We use this excluded vector for computing the value $Q(z_1, a_{l-1}|z_1)$.

Now we excluded the second z_2 from the sequence (while the first vector is retained) and compute the value $Q(z_2, a_{l-1}|z_2)$.

In this manner, we compute the values for all the vectors and calculator the number of errors in the leave on out procedure;

$$L(z_1, z_2....z_l) = \sum_{i=1}^{l} Q(z_i, a_{l-1}|z_i)$$

2 One Way to Compute an Error Bound

We will use L as an estimate for the expectation of the function $Q(z, a_l)$ that minimizes the empirical risk.

First we need to prove $E(L(z_1, ...z_{l+1})/(l+1)) = ER(a_l)$.

The proof consists of the following chain of transformations:

$$EL(z_1, z_2, ...z_{l+1})/(l+1) = \int 1/(l+1) \sum_{i=1}^{l+1} Q(z_i, a_l|z_i)dP(z_1), ..., dP(z_{l+1})$$
$$= \int 1/(l+1) \sum_{i=1}^{l+1} (\int Q(z_i, a_{l|}|z_i)dP(z_i))$$
$$= E(1/(l+1). \sum_{i=1}^{l+1} R(a_l|z_i))$$
$$= ER(a_l)$$

We will introduce the essential support vectors: Essential support vectors are the vectors that, if removed, would result in learning a different SVM. Indeed, if the vector x_i is not an essential support vector, then there exists an expansion of the vector ϕ that defines the optimal hyperplane that doesn't contain the vector x_i.

Since the optimal hyperplane is unique, removing this vector from the training set doesn't change it. Therefore in the leave one out method it will be recognized correctly.

Thus the leave one out method recognizes correctly all the vectors. Therefore the number of $L(z_1, ..., z_{l+1})$ of errors in the leave one out method doesn't exceed K_{l+1}, the number of the essential support vectors; that is,

$$L(z_1, \ldots, z_{l+1}) \leq K_{l+1}$$

Then the equation we will get:

$$ER(a_l) = EL(z_1, z_2, \ldots, z_{l+1})/(l+1) <= EK_{l+1}/(l+1)$$

3 Another Way to Compute and Error Bound

In order to use another way to prove the bound of the number of errors in the leave on out estimator for the optimal hyperplanes, we use the following method.

Suppose we are given the training set:

$$(x_1,y_1), \ldots, (x_{l+1}, y_{l+1})$$

And the maximum of $W(a)$ in the area $a \geq 0$ is achieved at the vector $a^0 = (a_1^0, \ldots, a_l^0)$. Let the vector

$$\phi_0 = \sum_{i=1}^{a} a_i^0 x_i y_i$$

Define the optimal hyperplane passing through the original, where we enumerate the support vectors with $i = 1 : a$

Let us denote by a^p the vector providing the maximum for the functional $W(a)$ under the constraint $a_p = 0$

$$a_i \geq 0 \quad \text{where}(i \neq p)$$

Let the vector $\phi_p = \sum_{i=1}^{a} a_i^p x_i y_i$

The above defines the coefficients of the corresponding separating hyperplane passing through the origin.

Now denote by W_0^p the value of the $W(a)$ for:

$$a_i = a_i^0 \quad where(i \neq p)$$

$$a_p = 0;$$

Consider the vector a^p that maximize the function $W(a)$ under the above constraint. The following obvious inequality is valid:

$$W_0^p \leq W(a^p)$$

On the other hand the following inequality is true:

$$W(a^p) \leq W(a^0)$$

Therefore the inequality:

$$W(a^0) - W(a^p) \leq W(a^0) - W_0^p$$

is valid.

Now let us rewrite the right-hand side of the inequality in the explicit form;

$$W(a^0) - W(a^p) = \sum_{i=1}^{a} a_i^0 - 1/2 * (\phi_0 * \phi_0) - (\sum_{i=1}^{a} a_i^0 - a_p^0 - 1/2 * ((\phi_0 - a_p^0 * y_p * x_p) * (\phi_0 - a_p^0 * y_p * x_p)))$$

$$= a_p^0 - a_p^0 y_p (x_p \phi_0) + 1/2 * (a_p^0)^2 * |x_p|^2$$

Taking into account that x_p is a support vector, we have

$$W(a^0) - W(a^p) = 1/2 * (a_p^0)^2 * |x_p|^2$$

Suppose the optimal hyperplane passing through the origin recognizes the vector x_p incorrectly. This means that the inequality:

$$y_p(x_p * \phi_0) \leq 0$$

is valid. This is possible only if the vector x_p is an essential support vector. Now let us make one step in maximization the function $W(a)$ by fix a_i, $i \neq p$ and change only one parameter $a_p > 0$. We obtain:

$$W(a) = W(a^p) + a_p(1 - y_p(x_p \phi_p)) - 1/2 * (a_p^0)^2 * |x_p|^2$$

From the equality we obtain the best value of a_p:

$$a_p = (1 - y_p(x_p \phi_p)) \div |x_p|^2$$

The increment of the W(a) at this moment equals

$$\Delta W_p = 1/2 * (1 - y_p(x_p \phi_p))^2 \div |x_p|^2$$

Since ΔW_p does not exceed the increment of the function $W(a)$ for the complete maximization, we obtain:

$$W(a^0) - W(a^p) >= \Delta W_p = 1/2 * (1 - y_p(x_p \phi_p))^2 \div |x_p|^2$$

Combining the equations above, finally we get:

$$1/2 * (a_p^0)^2 * |x|^2 >= 1/2 * (1 - y_p(x_p\phi_p))^2 \div |x_p|^2$$

From this equation and use the equation above, we could obtain:

$$a_p^0 >= (1 - y_p(x_p\phi_p))/|x_p|^2 >= 1/|x_p|^2$$

Taking into account that $|x_p| <= D_{l+1}$, we obtain

$$a_p^0 >= 1/D_{l+1}^2$$

Thus if the optimal hyperplane makes the error classifying vector x_p in the leave one out procedure, then the inequality holds. Therefore

$$\sum_{i=1}^{a} a_i^0 >= L_{l+1}/D_{l+1}^2$$

Where $L((x_1, y_1), ..., (x_{l+1}, y_{l+1}))$ is the number of errors in the leave one out procedure on the sample$(x_1, y_1), ..., (x_{l+1}, y_{l+1})$.

Now let us recall the properties of the optimal hyperplane

$$(\phi_0 * \phi_0) = \sum_{i=1}^{a} a_i^0$$

and

$$(\phi_0 * \phi_0) = 1/p_{l+1}^2$$

Combing the above equations, we conclude that the inequality:

$$L_{l+1} \leq D_{l+1}^2/p_{l+1}^2$$

is true with probability 1. Finally we get:

$$ER(a_l) = EL_{l+1}/(l+1) \leq E\frac{D_{l+1}^2/p_{l+1}^2}{l+1}$$

Most of the notes are quoted from the material the professor gives us in the class.