Lecture Notes (02/12/2015)
Math 6397, Spring 2015
Instructor: Robert Azencott
Prepared by: Pei Yang (1113771)
Draft 1

1 Brief review of maximum margin linear discriminant (MMLD)

Suppose we have a training set consisting of N vectors

$$X(1), \cdots, X(N)$$

in $\mathbf{R}^k$. And for these vectors, the classification results are

$$y(1), \cdots, y(N)$$

with

$$y(j) = \begin{cases} +1, & \text{if } X(j) \text{ belongs to Class 1;} \\ -1, & \text{if } X(j) \text{ belongs to Class 2.} \end{cases}$$

Now for a new case $X \in \mathbf{R}^k$, if $< w, X >_{\mathbf{R}^k} + b \geq 1$, we say that X belongs to Class 1; and if $< w, X >_{\mathbf{R}^k} + b \leq -1$, we say that X belongs to Class 2.

Then the maximum margin linear discriminant method is to find $w \in \mathbf{R}^k, b \in \mathbf{R}$ that maximize the distance between the two hyper-planes $< w, X >_{\mathbf{R}^k} + b = 1$ and $< w, X >_{\mathbf{R}^k} + b = -1$, which is $\frac{2}{||w||}$. Apparently, this is equivalent to minimizing $||w||^2$.
Remark Notice that if the classification for $X(j)$ is correct, then it always holds that

$$y(j)(< w, X(j) >_{\mathbf{R}^k} + b) \geq 1 \tag{1.1}$$

no matter which class $X(j)$ belongs to.

2 MMLD with slack variables

We are not able to have correct classifications for all the $X(j), j = 1, 2, \cdots, N$. Suppose we make a mistake when we evaluate $X(j)$, that is to day, (1.1) is violated. Introducing the **slack variable** $\xi_j \geq 0$ (positive and small) such that

$$y(j)(< w, X(j) >_{\mathbf{R}^k} + b) = 1 - \xi_j.$$

It's reasonable that $\Sigma_{j=1}^N \xi_j$ should also be minimized. Combine this with what is discussed in the previous section, we get the following linear minimisation problem

Minimisation Problem for MMLD

$$\begin{cases} \underset{w\in\mathbf{R}^k, b\in\mathbf{R}}{min}\{\frac{1}{2}||w||^2 + C\Sigma_{j=1}^N \xi_j\} \\ \xi_j \in \mathbf{R}, j = 1, 2, \cdots N \\ \text{with constraints} \\ \xi_j \geq 0 \\ \xi_j = 1 - y(j)(< w, X(j) >_{\mathbf{R}^k} + b) \end{cases} \tag{2.2}$$

where C is a fixed constant.
Remark 1 This minimisation problem is linear with respect to w, b and $\xi_1, \xi_2, \cdots, \xi_N$.
Remark 2 There are totally $N + k + 1$ unknowns and $2N$ constraints.

3 Lagrangian multipliers method

We will solve the minimisation problem by Lagrangian multipliers method.

3.1 Review on the method

Suppose we have a smooth function $F : \mathbf{R}^n \to \mathbf{R}$. We want to find the minimum of $F(z), z \in \mathbf{R}^n$ where z is the vector of unknowns.
1. If there are no constraints, then one needs to compute the gradient of F:

$$(\frac{\partial F}{\partial z_1}, \cdots, \frac{\partial F}{\partial z_n})$$

where $(z_1, \cdots, z_n) = z$ and set them to be 0:

$$\frac{\partial F}{\partial z_1} = 0, \cdots, \frac{\partial F}{\partial z_n} = 0$$

and solve for $z_1, \cdots, z_n$.
2. If there are n constraints:

$$g_1(z) \geq 0, \cdots, g_n(z) \geq 0$$

then one needs to consider the **Lagrangian**

$$L(z, \alpha) = F(z) - \Sigma_{j=1}^n \alpha_j g_j(z)$$

where $\alpha = (\alpha_1, \cdots, \alpha_n)(\alpha_j \geq 0, j = 1, \cdots, n)$and seek the saddle point of $L(z, \alpha)$, that is, to find z_0, α_0 such that

$$L(z_0, \alpha_0) = \underset{\alpha}{max}(\underset{z}{min}\{L(z, \alpha)\}).$$

Remark If F is a convex function of z, then for fixed α, one only needs to consider equations

$$\frac{\partial L}{\partial z_1} = \cdots = \frac{\partial L}{\partial z_n} = 0,$$

without worrying about the second order derivatives.

To apply the Lagrangian multipliers method to our minimisation problem, we compute the Lagrangian as follows

$$L(z,\alpha,\beta)=\frac{1}{2}||w||^2+C\Sigma_{j=1}^N\xi_j-\Sigma_{j=1}^N\alpha_j\xi_j-\Sigma_{j=1}^N\beta_j(\xi_j-1+y(j)(<w,X(j)>_{\mathbf{R}^k}+b))$$

where $\alpha=(\alpha_1,\cdots,\alpha_N),\beta=(\beta_1,\cdots,\beta_N),z=(w,b,\xi_1,\cdots,\xi_N)\in\mathbf{R}^{N+k+1}$.
We have

$$grad_wL=w-\Sigma_{j=1}^N\beta_jy(j)X(j)=\mathbf{0}$$

which gives

$$w=\Sigma_{j=1}^N\beta_jy(j)X(j). \tag{3.3}$$

And

$$\frac{\partial L}{\partial b}=-\Sigma_{j=1}^N\beta_jy(j)=0 \tag{3.4}$$

Also

$$\frac{\partial L}{\partial \xi_j}=C-\alpha_j-\beta_j=0$$

gives

$$\alpha_j+\beta_j=C, i=1,\cdots,N. \tag{3.5}$$

To get β_j, we plug (3.3),(3.4),(3.5) into L and get

$$\begin{aligned}L(z,\alpha,\beta)&=\frac{1}{2}||\Sigma_{j=1}^N\beta_jy(j)X(j)||^2+\Sigma_{j=1}^N\beta_j-\Sigma_{j=1}^N\beta_jy(j)<\Sigma_{i=1}^N\beta_iy(i)X(i),X(j)>\\&=-\frac{1}{2}\Sigma_i\Sigma_j\beta_i\beta_jy(i)y(j)<X(i),X(j)>+\Sigma_{j=1}^N\beta_j.\end{aligned} \tag{3.6}$$

We denote the above formula with $LL(\beta)$. Then it is sufficient to maximize $LL(\beta)$:

$$\max_{C\geq\beta_1,\cdots,\beta_N\geq 0}\{LL(\beta)\} \tag{3.7}$$

$$\text{with constraints } \Sigma_{j=1}^N\beta_jy(j)=0. \tag{3.8}$$

Remark The above problem is explicitly solvable if N is very small (like N=2,3, etc.).

After finding β, we can get α by (3.5) and get w by (3.3). To get b, we need the Karush-Kuhn-Tucker (KKT) condition. By KKT, the constraints in the original minimisation problem should hold if $0<\beta_j<C$ (hence $0<\alpha_i<C$), that is

$$\xi_j=0,\ \text{and}\ \xi_j=1-y(j)(<w,X(j)>_{\mathbf{R}^k}+b)$$

which gives $y(j)(<w,X(j)>_{\mathbf{R}^k}+b)=1$, thus $<w,X(j)>_{\mathbf{R}^k}+b=y(j)$, so

$$b=y(j)-<w,X(j)>_{\mathbf{R}^k}. \tag{3.9}$$

1. If $0 < \beta_j < C$ (hence $0 < \alpha_i < C$), then by KKT, $\xi_j = 0$, which means that $X(j)$ is correctly classified.

2. If $\beta_j = C$ (hence $\alpha_j = 0$), then $\xi_j > 0$, and this means $X(j)$ is not correctly classified.

Support Vector All the points $X(j)$ with $\beta_j > 0$ are called support vectors. Because $w = \Sigma_{j=1}^{N} \beta_j y(j) X(j)$, w consists of only support vectors.

Remark In the above analysis, the constant C is pre-given. The performance (percentage of correct classifications) of the maximum margin separator depends on C. So in practice, C should be chosen such that the performance $P(C)$ is maximized.
Remark The software package SVM-Light (in C, C++ and Python) is available to do maximum margin linear discriminant.
Remark The same analysis can be done in Hilbert space associated to a kernel K, including linear kernels, polynomial kernels, and Gaussian kernels.