Data Mining and Machine Learning Notes
Lecture date: 02-05-2015
Prepared by : Kamalpreet Kaur and Sho Nakagome
Instructor: Robert Azencott
Draft 2

1 From last lecture

We learned how to do linear discrimination between 2 classes using Gaussian assumptions, using the maximum likelihood principle. Cases described by:
v = vector of attributes, $v \in \mathbb{R}^k$.
N examples v(1)....v(N), y(1)....y(N) are given and constitute the training data set. Each of the y(j) has a value equal to +1 or -1.

1.1 Result of linear classification

We obtained:
w = vector of coefficients

$$\begin{pmatrix} w_1 \\ . \\ . \\ w_k \end{pmatrix}; w \in \mathbb{R}^{\mathbb{k}}$$

$b \in \mathbb{R}$

Separator: Sep(x) is defined as a separator which separates 2 classes in $\mathbb{R}^{\mathbb{k}}$. If the value of Sep(x) is equal to +1, x belongs to $Class_1$. If Sep(x) is equal to -1, x belongs to $Class_{-1}$. Here, x is a new case.

$$\begin{pmatrix} x_1 \\ . \\ . \\ x_k \end{pmatrix}; x \in \mathbb{R}^k$$

As a result, Sep(x) could be written as follows
•$Sep(x) =< w, x >_{\mathbb{R}^k} +b$

If written explicitly,
•$Sep(x) = w_1x_1 + w_2x_2 + ... + w_kx_k + b$

The equation Sep(x)=0 defines a straight line in 2-dimension, hyperplane in 3 or more dimension. It is called separator because the hyperplane $(\mathbf{x}|\mathbf{sep}(\mathbf{x}) = \mathbf{0})$ separates $\mathbb{R}^k$ into two half spaces.

How to compute w and b is the key question. There are multiple methods, and the one discussed last week was maximum likelihood method.

Let us see the concept of Sep(x) in Hilbert Space, H.

2 Linear discriminations using a positive definite kernel

Consider cases:

$v(1)...v(N) \in H$ (Hilbert Space of infinite dimension)

and each vector has its own class tag y(1)...y(N). Each y(j) takes a value equal to +1 or -1 so that each case v(j) belongs to one of the 2 classes. We want to separate $Class_1$ and $Class_{-1}$,for which we will consider Sep(x),
$Sep(x) =< w, x >_{\mathbb{R}^k} +b$
where x is a new case,$x \in \mathbb{R}^k$
Decision Rule:
•$sep(x) > 0 \Rightarrow x \in Class_1$

•$sep(x) < 0 \Rightarrow x \in Class_{-1}$

To have linear function on H,we need the form
$Sep(x) =< w, x >_H +b, b \in \mathbb{R}, w \in H$

Select any orthonormal basis $e_1, e_2, ..., e_m, ...$ in H, then

• $\| (e_m) \| = 1$

• $< e_m, e_j >_H = 0 for j \neq m$

•$x = x_1e_1 + x_2e_2 + ... + x_me_m + ...$ where x is the new case with k attributes.

•$w = w_1e_1 + w_2e_2 + ... + w_me_m + ...$, w is the vector of coefficients.

Also,the following series converge
• $\sum_{(i=1)}^{\infty} x_i^2 < \infty$

• $\sum_{(i=1)}^{\infty} w_i^2 < \infty$

• $< w, x >_H = \sum_{(j=1)}^{\infty} w_j x_j$

The goal would be to find a good w and b to classify the training set. Let us assume vectors $v(1)...v(N)$ are not in H but in $\mathbb{R}^k$. We know that each vector v(j) has a class tag y(j). Assume that we have a kernel, we select a kernel K(u,v) which is "positive definite" and defined for $u, v \in \mathbb{R}^k$. Therefore,the following two properties hold:
1) Symmetric
$K(u, v) = K(v, u)$

2)The following Matrix M is positive semi-definite
$M = M_{ij} = K(v(i), v(j))$

2.1 Case I:Gaussian Kernel

We take Gaussian kernel

$K(u, v) = exp[-\frac{(\|(u-v)\|_{\mathbb{R}^k})^2}{\sigma^2}]$

where $\sigma > 0$ and fixed.
According to Aronszajn theorem :
H is a space of functions on $\mathbb{R}^k$ and ϕ_v is a function defined on $\mathbb{R}^k$ by the formula
$\phi_v(u) = K(u, v) \forall u \in \mathbb{R}^k$
H is essentially the set of all linear combinations of these functions.

$\psi = \lambda_1\phi_{z_1} + \lambda_2\phi_{z_2} + ... + \lambda_r\phi_{z_r}$

where $\psi \in H$ and $z_1...z_r$ are arbitrary vectors in $\mathbb{R}^k$, and $\lambda_1....\lambda_r$ are arbitrary numbers in $\mathbb{R}$. For example, if we want to compute $\psi(u)$

$\psi(u) = \sum_{(j=1)}^{r} \lambda_j \phi_{z_j} = \sum_{(j=1)}^{r} \lambda_j K(u, z_j)$

In this Hilbert space,the scalar product is defined by:
$< \phi_v, \phi_z >= K(v, z)$

Example

If we have
$< \lambda_1\phi_{z_1} + \lambda_2\phi_{z_2} + \lambda_3\phi_{z_3}, \mu_1\phi_{x_1} + \mu_2\phi_{x_2} >_H$
where $z_1, z_2, z_3 \in \mathbb{R}^k, x_1, x_2 \in \mathbb{R}^k$ and $\lambda_1, ..\lambda_3, \mu_1, \mu_2 \in \mathbb{R}$

$= \lambda_1\mu_1 < \phi_{z_1}, \phi_{x_1} > +\lambda_1\mu_2 < \phi_{z_1}, \phi_{x_2} > +\lambda_2\mu_1 < \phi_{z_2}, \phi_{x_1} > +\lambda_2\mu_2 < \phi_{z_2}, \phi_{x_2} > +\lambda_3\mu_1 < \phi_{z_3}, \phi_{x_1} > +\lambda_3\mu_2 < \phi_{z_3}, \phi_{x_2} >$

$= \lambda_1\mu_1 K(z_1, x_1)+\lambda_1\mu_2 K(z_1, x_2)+\lambda_2\mu_1 K(z_2, x_1)+\lambda_2\mu_2 K(z_2, x_2)+\lambda_3\mu_1 K(z_3, x_1)+ \lambda_3\mu_2 K(z_3, x_2)$

$= \sum_i \sum_j \lambda_i\mu_j K(z_i, x_j)$

Now, we are going to construct the separator Sep(x) where x is new case and x is in $\mathbb{R}^k$. We want to construct the separator in H but x is not in H. Therefore at first, we consider

$x \rightsquigarrow \mathcal{X}$,where $\mathcal{X}$ is equal to $\phi(x) \in H$

$\phi_x(u) = K(u, x)$
then,

$Sep(\mathcal{X}) =< w, \mathcal{X} >_H +b$
where w is a vector in H,$\mathcal{X} \in H$ and b is a number.

2.2 Computation of 'w'

It can be proved that the best 'w' must be a linear combination of the $\phi_{v(j)}$.

$\Rightarrow w = \sum_{j=1}^{N} \alpha_j \phi_{v(j)}$
where $\phi_{v(j)}$ are functions constructed from the training data.

2.3 How do we compute $< w, \mathcal{X} >_H$?

$< w, \mathcal{X} >_H = < \sum_{(j=1)}^{N} \alpha_j \phi_{v(j)}, \phi_x >_H$
where $v(j) \in \mathbb{R}^{\mathbb{k}}, \alpha_j$ is computed by SVM program on basis of given data
.
$= \sum_{(j=1)}^{N} \alpha_j < \phi_{v(j)}, \phi_x >_H$

$= \sum_{(j=1)}^{N} \alpha_j K(v(j), x)$

$= \sum_{(j=1)}^{N} \alpha_j K(x, v(j))$; (by symmetry)

$K(x, v(j)) = exp[-\frac{\|(x-v(j))\|^2}{\sigma^2}]$(Gaussian Kernel)
which can be computed easily

$\therefore Sep(x) = b + \sum_{(j=1)}^{N} \alpha_j K(v(j), x)$

is a linear combination of exponentials but nonlinear function of x.
We are looking at sign of the separator. First of all, if you know alpha and b, the computation is elementary. Once you have the value of the separator, you look at the sign, if the sign is positive, then you are going to classify x into $Class_1$ and vice versa. Classification is based on the separator, which is a non-linear function of x. However, viewed in the Hilbert space, it is a linear function of X.
$Sep(x) = 0$ is called a hypersurface

3 Advantages of this approach:

1st Advantage:
Linear function in the hilbert space H are complicated non-linear function of the real data which are in $\mathbb{R}^{\mathbb{k}}$. However, doing the computation to find the best linear classifier in H is a much simpler problem than computing the non-linear function.

2nd advantage:
This kind of linear combination of exponentials can actually approximate any function of x. Even for extremely complicated x, separating function can be approximated by linear combination of the kernel functions.

4 Case II: Polynomial kernel

Now we will take polynomial kernel, $K(u,v) = (1+ < u, v >_{\mathbb{R}}^{k})^{r}$, k is large. Generally r is small: 2 to 4. If it is bigger than 5, use exponentials.For instance let us consider the case r = 2. H is the set of all linear combinations of the functions ϕ_v.
$\phi_v(u) = (1+ < u, v >)^2 = K(u,v) = (1 + u_1v_1 + u_2v_2 + ... + u_kv_k)^2$

H = space of all polynomial of degree 2

$\psi_v(u)$ = Polynomial of degree 2 in u

Generic Polynomial of degree 2 is given by

$\psi'(u) = a + \sum_{(j=1)}^{k} \gamma_j u_j + \sum_{(i=1)}^{k} \sum_{(j=1)}^{k} \epsilon_{ij} u_i u_j$

where γ_j, ϵ_{ij} are arbitrary coefficients.

$\therefore Sep(\mathcal{X}) =< w, x >_H + b = \psi(x)$ = polynomial of degree 2 (in this case) where x is a new case, x $\in \mathbb{R}^k$

$\phi_x(u) = K(u,x) = (1+ < u, x >)^2$

For optimal "w" : $w = \sum_{(j=1)}^{N} \alpha_j \phi_{v(j)}$
where v(j) are the examples from the training set in $\mathbb{R}^k$
$< w, \phi_x >_H =< \sum_{(j=1)}^{N} \alpha_j \phi_{v(j)}, \phi_x >_H$

$= \sum_{(j=1)}^{N} \alpha_j < \phi_{v(j)}, \phi_x >_H$

$= \sum_{(j=1)}^{N} \alpha_j K(v(j), x)$

$= \sum_{(j=1)}^{N} \alpha_j (1+ < v(j), x >_{\mathbb{R}^k})^2$

Goal

Find w,b such that Separator has "maximum margin" in H.
So in Hilbert Space H, each $\mathcal{X}(1)...\mathcal{X}(N)$ has a class tag $y(1)...y(N)$ respectively. Hyperplane $< w, \mathcal{X} > + b = 0$ separates the space into 2 half spaces.