Data Mining and Machine Learning Notes
7th Lecture delivered on 02-10-2015
Prepared by : Kamalpreet Kaur and Sho Nakagome
Instructor: Robert Azencott
Draft 1

February 22, 2015

1 Recommendations:Coding of data[Project 1]

Suppose there are N cases. Each case has a vector of attribute

$$\begin{pmatrix} v_1 \\ . \\ . \\ v_k \end{pmatrix}; v \in \mathbb{R}^{\mathbb{k}}$$

★ Note that this vector should not include class attribute.

1.0.1 Normalization of vector v

As we have N cases, there are N vectors namely v(1),v(2),....v(N). Each have k attributes. Mean of these vectors is given by

$\overline{v} = \frac{v(1)+.....+v(N)}{N}$

Now subtract this mean value from each v(j),j=1,...,N to get $\underline{CenteredData}$; $u(j)$ v(j) - $\overline{v}$ = u(j)

To normalize now centered data, we will calculate variance and standard deviation of each attribute of v(j),j=1....N For example lets take one attribute , say 2nd attribute.
$\therefore$ We have $u_2(1)u_2(2)u_2(3).....u_2(N)$ for N cases

Variance = $\sigma_2(j)^2 = \frac{\sum_{(j=1)}^{N} u_2(j)^2}{N}$

Standard Deviation = $\sqrt{\frac{\sum_{(j=1)}^{N} u_2(j)^2}{N}}$
In this way ,find variance and standard deviation for all k attributes.

$$\begin{pmatrix} \sigma_1 \\ . \\ . \\ \sigma_k \end{pmatrix}$$

Now calculate $\frac{u_i}{\sigma_i} = w_i$

$$\begin{pmatrix} w_1 \\ . \\ . \\ w_k \end{pmatrix}$$

is $\underline{normalizedvectorofattribute}$

1.0.2 Coding of Categorical Attribute

In case of 2 categories, use 0/1 or +1/-1. In more than 2 categories, use $\underline{impartialcoding}$(binary) instead of single numeric variable (2,1,0,-1,-2) because in case of numbers, points are either too far or too close. But in case of impartial coding, all points are somewhat at same distance. For example

, in case of 5 categories use long vectors

$$\begin{aligned} A &= 000 \\ B &= 001 \\ C &= 010 \\ D &= 100 \\ E &= 110 \end{aligned}$$

2 Main Topic : Maximium Margin (MM) Technique for Linear Classification

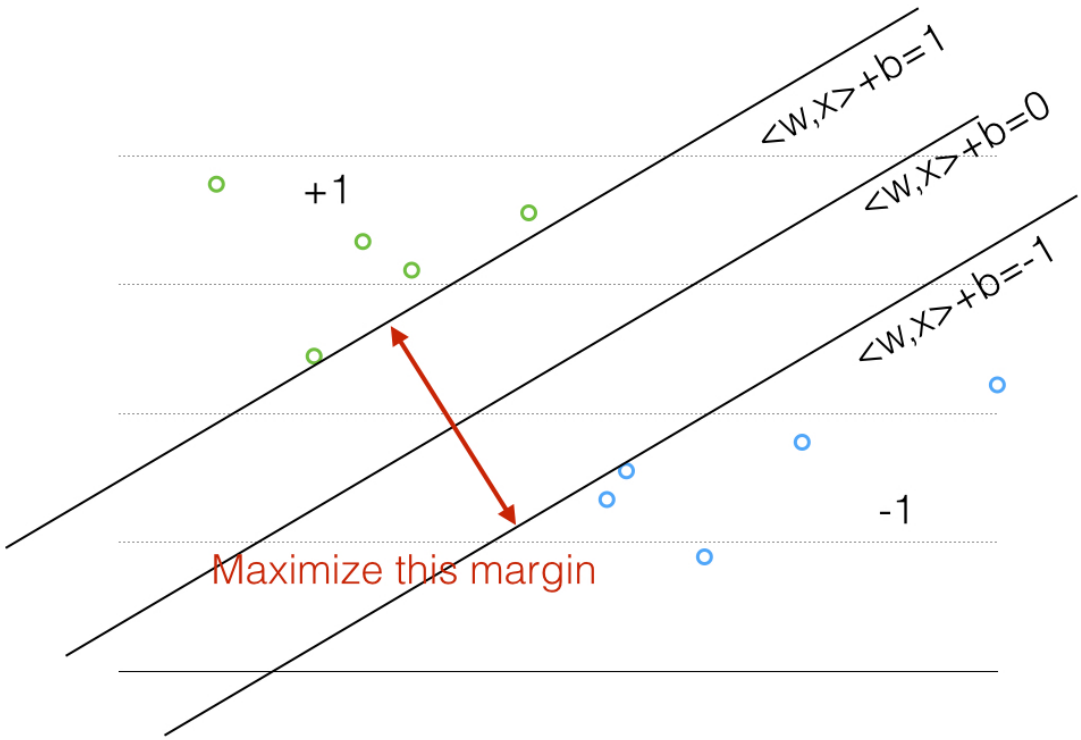
• Studied by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963 - Theory of Machine Learning
For Linear classification in $\mathbb{R}^{\mathbb{k}}$,one of the key aspect about Maximum Margin Technique is that it is more robust.

The goal is to understand Maximum Margin Principles in Hilbert space. However, at first we start from defining MM in $\mathbb{R}^{\mathbb{k}}$.

Let $x(1), x(2), x(3), , x(N) \in \mathbb{R}^{\mathbb{k}}$, (Training sets).
Assume we only have 2 classes +1 and -1
$y(1), y(2), y(3), ...y(N)$ each corresponds to the cases x and each has a classifier value of +1 or -1.
Here, we would like to separate linearly the 2 classes. Therefore, we need to find out linear separator.

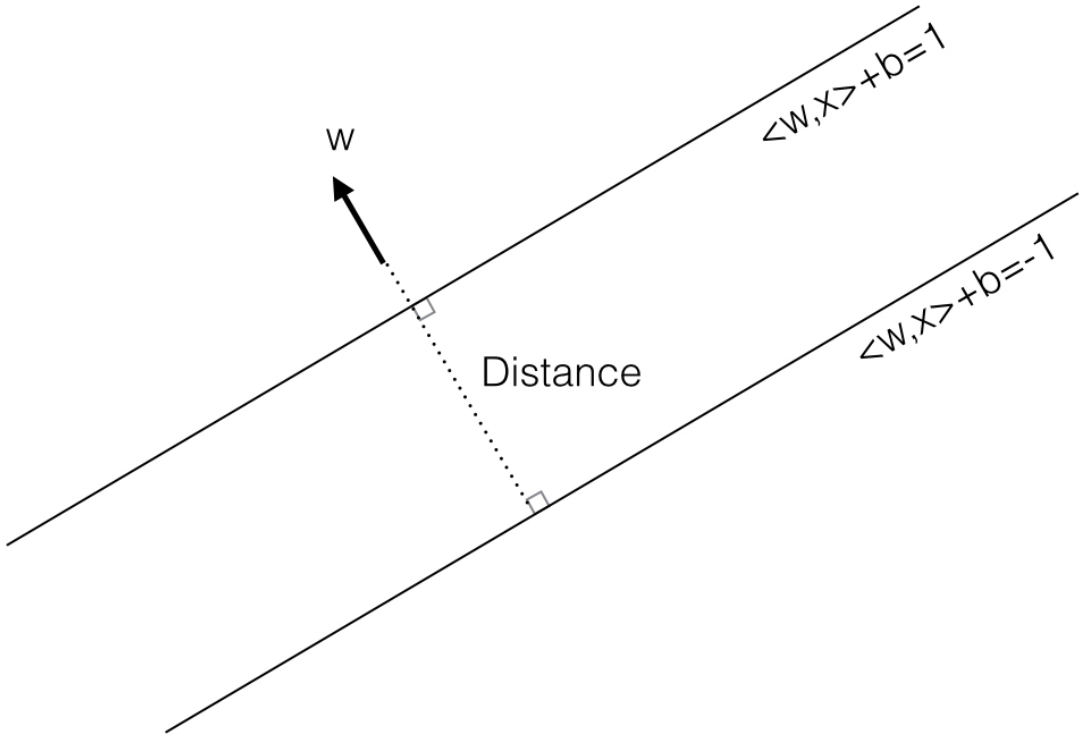
$Sep(x) =< w, x >_{R^k} +b, w \in \mathbb{R}^{\mathbb{k}}, b \in \mathbb{R}$

We define a classification rule:
$x \to +1 if < w, x > +b \geqslant 1$ (Strictly positive)
$x \to -1 if < w, x > +b \leqslant -1$ (Strictly negative)



$Sep(x) =< w, x >_{R^k} +b = 0$, is called separator.
We look at all the separators parallel to hyperplane $< w, x > +b = 0$.
if $x \rightsquigarrow < w, x > +b = +1 \Rightarrow Hyperlane + 1$

if $x \rightsquigarrow < w, x > +b = -1 \Rightarrow Hyperlane - 1$
w is orthogonal to the hyperplane. The distance between two planes is the important thing to evaluate. We need to maximize this distance. The distance between two planes is given by
$d = \frac{2}{\|w\|}$



We want to have hyperplanes that maximize the margin. That is, maximize d or minimize $\| w \|$. Also, at the same time, there should be less mistakes.So lets see what is a good classification.

3 Good classification

If you take for example x(j), then see sign of $< w, x(j) > +b$.
If $< w, x(j) > +b - 1 \geqslant 0 \Rightarrow x(j) \in class + 1$
My estimate is y(j) = +1 if x(j) is in population 1,
$\therefore y(j) * (< w, x(j) > +b - 1) \geq 0$

We consider mistake in classification if $y(j) * (< w, x(j) > +b - 1) < 0$
Situation where x(j) is in population -1,
y(j) = -1 ;

$(< w, x(j) > +b + 1) < 0;$

$\therefore y(j) * (< w, x(j) > +b + 1) > 0$

Using this criteria Count the number of mistakes
$\xi_j = +1$, If there is mistake in classification of x(j)

$\xi_j = 0$,If I there is no mistake in classification of x(j)

$\sum_{(j=0)}^{N} \xi_j$ = No. of mistakes (which we want to minimize).

Therefore, in order for the maximum margin principle to perform best, we need to minimize the following equation:
$\| w \|^2 + c \sum_{(j=0)}^{N} \xi_j$
where $c > 0$ and c is just a fixed parameter.

★The goal is to find w and b in $\mathbb{R}^{\mathbb{k}}$ that minimizes
$\| w \|^2 + c \sum_{(j=0)}^{N} \xi_j$

This can be transformed in a situation where ξ_j is real number greater than 0. These are Slack variables. If y(j) correct classification of x(j)
$y(j) = +1,$

$y(j) * (< w, x(j) > +b - 1) > 0$
Then we define
$y(j) * [< w, x(j) > +b - 1] = -\xi_j$
where $\xi_j > 0$ and small. We defined slack variable like this
$\xi_j = y(j)[1 - < w, x(j) > -b]$
where y(j) and x(j) are fixed. These are also called *constraints*

3.0.3 Conclusion

Minimize w,b,$\xi_1, \xi_2, ..., \xi_N$. $\therefore$ There are 2N linear constraints with N+k+1 unknown variables.
$min \parallel w \parallel^2 + c * \sum_{(j=1)}^{N} \xi_j \Rightarrow$ Quadratic form Q(w,b,$\xi_1, \xi_2, ...\xi_N$)