Assignment 3 EECS545

Jiazhao Li

2018/2/22

1 Naive Bayes spam filter

1.1 Error

When we classify there is 1/800 mistake,and Error rate : 0.00125

1.2 Prove

$$w_c = \left[log(\frac{\theta_{c11}}{1-\theta_{c11}}), log(\frac{\theta_{c21}}{1-\theta_{c21}}), ..., log(\frac{\theta_{cD1}}{1-\theta_{cD1}}) \right]^T$$
$$w_{\bar{c}} = \left[log(\frac{\theta_{\bar{c}10}}{1-\theta_{\bar{c}10}}), log(\frac{\theta_{\bar{c}20}}{1-\theta_{\bar{c}20}}), ..., log(\frac{\theta_{\bar{c}D0}}{1-\theta_{\bar{c}D0}}) \right]^T$$

It is binary classification $1 - \theta_{cd1} = \theta_{cd0}$,and$1 - \theta_{\bar{c}d1} = \theta_{\bar{c}d0}$
the metric for classification is $w^T \Phi(x) \geqslant 0 \Rightarrow (w_c - w_{\bar{c}})^T \Phi(x) \geqslant 0$
$\Rightarrow w_c^T \Phi(x) \geqslant w_{\bar{c}}^T \Phi(x)$

Notice that $\Phi(x) = [x_1, x_2, ...x_D, 1]^T$, where x_d can only be 1 or 0.
$w_c^T \Phi(x) = \sum_{d=1}^D log(\frac{\theta_{cd1}}{1-\theta_{cd1}})^{1(x_d=1)} + log(\pi_c) + \sum_{d=1}^D log(1 - \phi_{cd1})$

Substitution $1 - \theta_{cd1} = \theta_{cd0}$ and transform sum to times through log.
$w_c^T \Phi(x) = log(\pi_c * \prod_{d=1}^D \frac{\theta_{cd1}}{\theta_{cd0}})^{1(x_d=1)} * \prod_{d=1}^D (\theta_{cd0})$
$= log(\pi_c * \prod_{d=1}^D (\theta_{cd1})^{1(x_d=1)} * \prod_{d=1}^D (\theta_{cd0})^{1(x_d=0)}$
$= log(P(y = c)P(x = x^{new}|y = c)) = log(P(y = c|x = x_{new}))$,the posterior for label c.

The same format for $w_{\bar{c}}$,we can get posterior for label $\bar{c}$
$w_{\bar{c}}^T \Phi(x) = log(P(y = \bar{c})P(x = x^n ew|y = \bar{c})) = log(P(y = \bar{c}|x = x_{new}))$
In fact we compare $P(y = c|x = x_{new}) and (y = \bar{c}|x = x_{new})$
Or we can present $P(y = c|x = x_{new}) - P(y = \bar{c}|x = x_{new}) \geqslant 0$,as:
if $(w_c - w_{\bar{c}})^T \Phi(x) \geqslant 0$,we assign y= 1,Otherwise,y=0

2 Valid Kernel

2.1 properties

(i).$\forall b \subset R^n, a \geqslant 0, b^T \kappa(x,x')b = b^T a\kappa_1(x,x')b \geqslant 0 \Rightarrow b^T \kappa_1(x,x')b \geqslant 0$
$\Rightarrow \kappa isPSD$
κ_1 is valid kernel $\Rightarrow a\kappa_1$ is still symmetric matrix.
Hence, κ is a valid kernel

(ii).$\forall b \subset R^n, b^T \kappa(x,x')b = b^T \kappa_1(x,x')b + b^T \kappa_2(x,x')b \geqslant 0 \Rightarrow \kappa isPSD$
κ_1, κ_2 are valid kernels $\Rightarrow \kappa$ is still symmetric matrix.
Hence, κ is a valid kernel

(iii)c.let $\kappa_1(x,x') = \Phi_1(x)^T \Phi_1(x'), where \Phi_1(x) = [a_1(x), a_2(x), ...a_n(x)]$
$\kappa_2(x,x') = \Phi_2(x)^T \Phi_2(x'), where \Phi_2(x) = [b_1(x), b_2(x), ...b_m(x)]$

$\kappa = \kappa_1\kappa_2 = \sum_{n=1}^{N} a_n(x)a_n(x') \sum_{m=1}^{M} b_m(x)b_m(x') = \sum_{n=1}^{N} \sum_{m=1}^{M} a_n(x)b_m(x)a_n(x')b_m(x')$

$= \sum_{n=1}^{N} \sum_{m=1}^{M} c_{mn}(x)c_{mn}(x') = \Phi(x)^T \Phi(x'), where \Phi(x) = [c_1(x), c_2(x), ...c_{mn}(x)]$
Hence, κ is a valid kernel

(iv).f is one function map R^s to R,we know that Φ is exactly the map function on
R^s. We can take $\Phi(x) = f(x)$, so$\kappa(x,x') = \Phi(x)^T \Phi(x')$
Hence, κ is a valid kernel

(v).$b \subset R^n, b^T \kappa_1^d(x,x')b = b^T (U^T \Sigma U)^d b = b^T U^T \Sigma^d U b \geqslant 0$
since $k_1(x,x') = b^T U^T \Sigma U b \geqslant 0, \Sigma isPSD \Rightarrow \Sigma^d isPSD$
Hence, κ is a valid kernel

(vi). $\kappa(x,x') = p(\kappa(x,x'))$,from (v) we have proved that the power of valid kernel is still a valid kernel. from(i) we can conclude that coefficients will not affect the kernel.p() function is just form one linear combination of power of valid kernel, so it will still be one valid kernel.

2.2 Gaussian Kernel

$\kappa(x,y) = exp(-\frac{|x|_2^2}{2\sigma^2})exp(-\frac{|y|_2^2}{2\sigma^2})exp(-\frac{xy}{\sigma^2})$
from (iv) $\Rightarrow \kappa(x,y) = f(x)f(y)exp(-\frac{xy}{\sigma^2})$,we need to prove third term $exp(-\frac{xy}{\sigma^2})$is one valid kernel.
Using Taylor Theorem,$exp(-\frac{xy}{\sigma^2}) = \sum_{n=0}^{N} \frac{(xy)^n}{n!\sigma^{2n}}$,which is the linear combinations of power of (xy),we have known (xy) is the most common kernel.Hence, GK is a valid kernel.
Then we can express this valid kernel

$$\kappa(x,y) = f(x)f(y)(1 + \frac{2xy}{2\sigma^2} + \frac{(\frac{2xy}{2\sigma^2})^2}{2!} + ... + \frac{(\frac{2xy}{2\sigma^2})^n}{n!})$$

$$= f(x)f(y)(1\times 1+\sqrt{\frac{1}{\sigma^2}}x\sqrt{\frac{1}{\sigma^2}}y+\sqrt{\frac{(\frac{1}{\sigma^2})^2}{2!}}x^2\sqrt{\frac{(\frac{1}{\sigma^2})^2}{2!}}y^2+...+\sqrt{\frac{(\frac{1}{\sigma^2})^n}{n!}}x^n\sqrt{\frac{(\frac{1}{\sigma^2})^n}{n!}}y^n)$$

$= \Phi(x)\Phi(y)$,where$\Phi(x) = exp(-\frac{x^2}{2\sigma^2})[1, \sqrt{\frac{1}{\sigma^2}}x, \sqrt{\frac{1}{\sigma^4 2!}}x^2 ... \sqrt{\frac{1}{\sigma^{2n} n!}}x^n]^T$,when n infinite, this is infinite dimension kernel.

3 Kernel Perceptron

The algorithm in Kernel Perceptron: As conclusions in lecture,in the normal perceptron $y(x) = sgn(w^T x)$ we know w can be presented with linear combinations of $\Phi(x)$,then we get $w = \sum_i \alpha_i y_i \Phi(x_i)$.
This is algorithm:

Initialization the parameters $\alpha = [0, 0...0] \subset R^n$
Iteration for 30 times to make converge:
  for each sample data(x_j, y_j):
    Make the prediction:$y_{pre} = sgn(\sum_i^n \alpha_i y_i \kappa(x_i, x_j))$
    if $y_{pre} \neq y_j$:
      updata the parameters $a[j] = a[j] + 1$
    else:
      continue

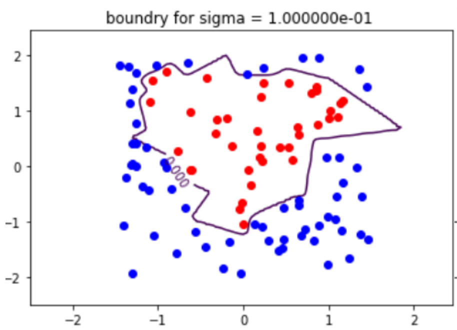

Figure 1: sigma = 1

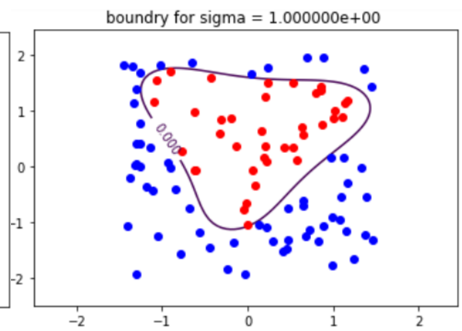

Figure 2: sigma = 0.1

4 LDA,QDA

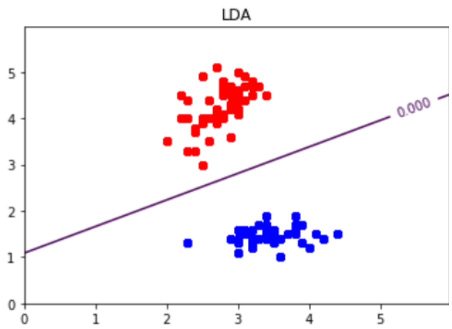

Figure 3: LDA

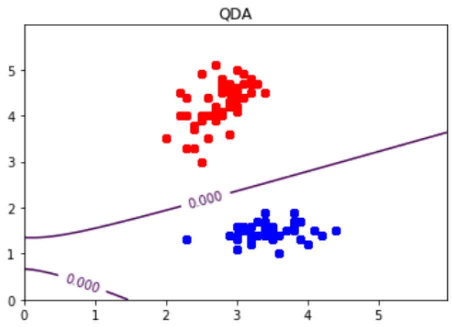

Figure 4: QDA