Machine Learning Homework 5

Bodong Zhang

u0949206

1 Warm up: Margins

1 Because it is XOR function, so

When x1=-1, x2=-1, (x1=-1, x1x2=1) output=-1

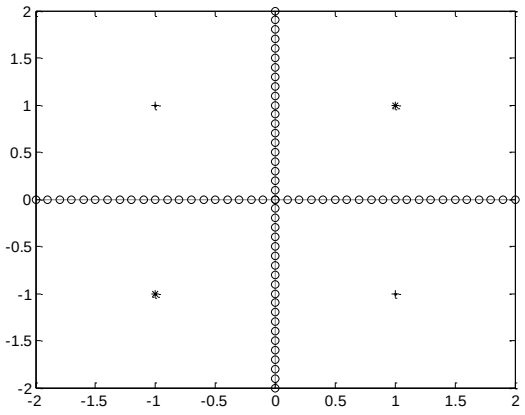
When x1=-1, x2=1, (x1=-1, x1x2=-1) output=1

When x1=1, x2=-1, (x1=1, x1x2=-1) output=1

When x1=1, x2=1, (x1=1, x1x2=1) output=-1

So the separating line is x1x2=0, if x1x2>0, output=-1, else output=1. The maximal margin is 1.

The figure showing separating line is below.

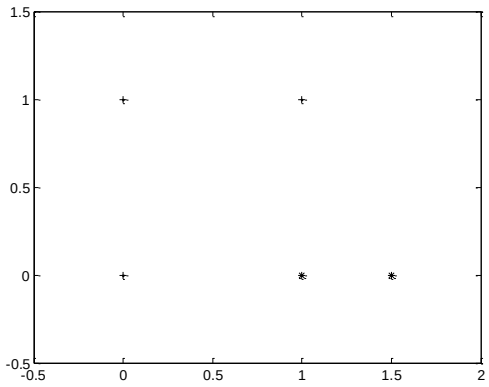


'*' means negative output, '+' means positive output. The two lines means the separating line, points in the first quadrant and third quadrant would be predicted as negative, points in the second and fourth quadrant would be predicted as positive.

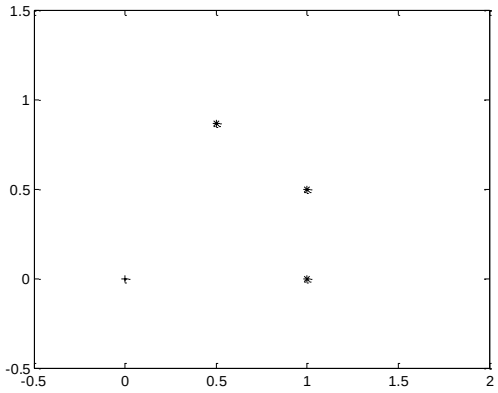
2 (a)For D1, the separating line with maximum margin is v2-v1+0.5, if v2-v1+0.5>0, predict positive, if v2-v1+0.5<0, predict negative. The maximum possible margin for D1 is $0.5/\sqrt{2} = \sqrt{2}/4$

For D2, maximum margin is $(1, 0)\cdot n/2=(1,0)\cdot(\sqrt{3}/2, 1/2)/2=\sqrt{3}/4$

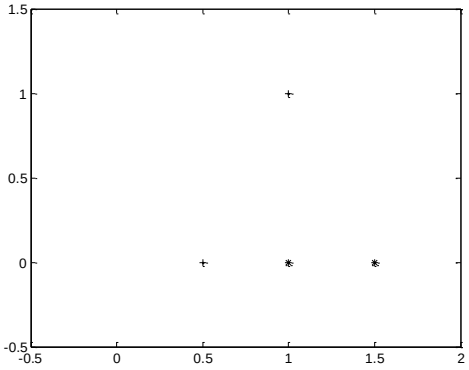
For D3, maximum margin is $(0.5, 0)\cdot n/2=(0.5,0)\cdot(2/\sqrt{5}, -1/\sqrt{5})/2=\sqrt{5}/10$



D1



D2



D3

(b) For D1, R=3/2, y=$\sqrt{2}$/4, so the mistake bound is 18.

For D2, R= $\sqrt{5}$/2, y=$\sqrt{3}$/4, so the mistake bound is 20/3

For D3, R=3/2,y=$\sqrt{5}$/10, so the mistake bound is 45.

D3 has the greatest mistake bound.

(c) Based on the mistake bound of each, the dataset D2 is “easiest” to learn, D1 is second, D3 is the worst.

2 Kernels

1(a) Assume $K1(x,z)=(\phi_1(x), \phi_2(x), \phi_3(x), \phi_4(x)\ldots \phi_n(x))*(\phi_1(z), \phi_2(z), \phi_3(z), \phi_4(z)\ldots \phi_n(z))^T$

$K2(x,z)=(\alpha_1(x), \alpha_2(x), \alpha_3(x), \alpha_4(x)\ldots \alpha_m(x))*(\alpha_1(z), \alpha_2(z), \alpha_3(z), \alpha_4(z)\ldots \alpha_m(z))^T$

So $K(x,z)=K1(x,z)K2(x,z)=$

($\phi_1(x)\alpha_1(x)$, …, $\phi_1(x)\alpha_m(x)$, $\phi_2(x)\alpha_1(x)$, …, $\phi_2(x)\alpha_m(x)$,…, $\phi_n(x)\alpha_1(x)$, …, $\phi_n(x)\alpha_m(x)$)*

($\phi_1(z)\alpha_1(z)$, …, $\phi_1(z)\alpha_m(z)$, $\phi_2(z)\alpha_1(z)$, …, $\phi_2(z)\alpha_m(z)$,…, $\phi_n(z)\alpha_1(z)$, …, $\phi_n(z)\alpha_m(z)$)

So K is a kernel.

1(b) First we need to show that $K(x,z)=\alpha K1(x,z)+\beta K2(x,z)$ is a valid kernel if K1 and K2 are valid kernels.

Assume $K1(x,z)=(\phi_1(x), \phi_2(x), \phi_3(x), \phi_4(x)\ldots \phi_n(x))*(\phi_1(z), \phi_2(z), \phi_3(z), \phi_4(z)\ldots \phi_n(z))^T$

$K2(x,z)=(\varepsilon_1(x), \varepsilon_2(x), \varepsilon_3(x), \varepsilon_4(x)\ldots \varepsilon_m(x))*(\varepsilon_1(z), \varepsilon_2(z), \varepsilon_3(z), \varepsilon_4(z)\ldots \varepsilon_m(z))^T$

(Without losing generality, assume m<=n)

So $K(x,z)= \alpha K1(x,z)+\beta K2(x,z)=(\sqrt{\alpha}\phi_1(x), \sqrt{\beta}\varepsilon_1(x), \sqrt{\alpha}\phi_2(x), \sqrt{\beta}\varepsilon_2(x),\ldots, \sqrt{\alpha}\phi_m(x), \sqrt{\beta}\varepsilon_m(x),\ldots , \sqrt{\alpha}\phi_n(x))*$

$(\sqrt{\alpha}\phi_1(z), \sqrt{\beta}\varepsilon_1(z), \sqrt{\alpha}\phi_2(z),\sqrt{\beta}\varepsilon_2(z),\ldots, \sqrt{\alpha}\phi_m(z), \sqrt{\beta}\varepsilon_m(z),\ldots , \sqrt{\alpha}\phi_n(z))$.

Also K1(x,z)K2(x,z) is valid kernels, any polynomial with positive coefficients P(K1(x,z)) can be formed by these two transforms. So a polynomial over a kernel with positive coefficients is a kernel.

2. x^Tz is a valid kernel function, based on 1(a), $(x^Tz)^2$ is a valid kernel function. $\exp(-||x-z||^2)$ is Gaussian kernel, then based on 1(a) and 1(b), we know this is a valid kernel function.

3. $\exp(-\frac{||x-z||^2}{2\sigma^2})=\exp(-\frac{||x||^2}{2\sigma^2})\exp(-\frac{||z||^2}{2\sigma^2})\exp(\frac{x\cdot z}{\sigma^2})$, so by theory of constructing kernel by multiplying a function f to both x and z, we just need to prove $\exp(\frac{x\cdot z}{\sigma^2})$ is a kernel function. $x\cdot z=x^Tz$, so x^Tz is kernel function, also $\exp(x)=\sum_{n=0}^{\infty}\frac{x^n}{n!}$, so $\exp(\frac{x\cdot z}{\sigma^2})$ is infinite polynomial of x^Tz, so it is also a kernel function. So $\exp(-\frac{||x-z||^2}{2\sigma^2})$ is kernel function.

3.1 Support Vector Machine

(The data is randomly shuffled, so each time the result may be different!)

1 Based on experiments

The training accuracy is 0.975

The testing accuracy is 0.973

2

C= 2.0 gamma0= 0.1

The average training accuracy is 0.5366666666666667

The average cross validation accuracy is 0.5366666666666666

C= 2.0 gamma0= 0.01

The average training accuracy is 0.5366666666666667

The average cross validation accuracy is 0.5366666666666666

C= 2.0 gamma0= 0.001

The average training accuracy is 0.5379166666666666

The average cross validation accuracy is 0.4883333333333333

C= 1.0 gamma0= 0.1

The average training accuracy is 0.5954166666666667

The average cross validation accuracy is 0.5066666666666666

C= 1.0 gamma0= 0.01

The average training accuracy is 0.5945833333333332

The average cross validation accuracy is 0.5083333333333333

C= 1.0 gamma0= 0.001

The average training accuracy is 0.5966666666666667

The average cross validation accuracy is 0.5216666666666666

C= 0.5 gamma0= 0.1

The average training accuracy is 0.6345833333333333

The average cross validation accuracy is 0.5566666666666666

C= 0.5 gamma0= 0.01

The average training accuracy is 0.6279166666666667

The average cross validation accuracy is 0.5166666666666666

C= 0.5 gamma0= 0.001

The average training accuracy is 0.6166666666666666

The average cross validation accuracy is 0.5216666666666667

C= 0.25 gamma0= 0.1

The average training accuracy is 0.5770833333333334

The average cross validation accuracy is 0.5633333333333332

C= 0.25 gamma0= 0.01

The average training accuracy is 0.6183333333333334

The average cross validation accuracy is 0.5349999999999999

C= 0.25 gamma0= 0.001

The average training accuracy is 0.5845833333333333

The average cross validation accuracy is 0.5349999999999999

C= 0.125 gamma0= 0.1

The average training accuracy is 0.5833333333333333

The average cross validation accuracy is 0.5533333333333333

C= 0.125 gamma0= 0.01

The average training accuracy is 0.5974999999999999

The average cross validation accuracy is 0.5783333333333334

C= 0.125 gamma0= 0.001

The average training accuracy is 0.5620833333333334

The average cross validation accuracy is 0.5333333333333333

C= 0.0625 gamma0= 0.1

The average training accuracy is 0.5395833333333334

The average cross validation accuracy is 0.5233333333333333

C= 0.0625 gamma0= 0.01

The average training accuracy is 0.6033333333333333

The average cross validation accuracy is 0.5433333333333333

C= 0.0625 gamma0= 0.001

The average training accuracy is 0.5891666666666667

The average cross validation accuracy is 0.5516666666666666

The best C that has highest accuracy in cross validation is 0.125

The best gamma that has highest accuracy in cross validation is 0.01

Use the best C and gamma to train whole training set again:

Training complete, the first 10 dimension of w is (first dimension is bias)

5.189021940271396E-6 -0.015038996796650663 0.16198496802100926 -0.006195306468811616 0.01126603683254884 0.07150166236553021 0.00445280332466967 -0.13178110749795935 0.002631435197541526 0.01565469302365862

Training accuracy=0.6316666666666667

The testing accuracy is 0.585

3 In " handwriting data" set,

In training,p=0.9680851063829787

In training,r=0.9963503649635036

In training,F1=0.9820143884892086

In testing,p=0.9637599093997735

In testing,r=0.9964871194379391

In testing,F1=0.9798503166378815

In "madelon" set

In training,p=0.6101694915254238

In training,r=0.5901639344262295

In training,F1=0.6000000000000001

In testing,p=0.5771812080536913

In testing,r=0.5685950413223141

In testing,F1=0.57285595337219

3.2 Ensemble of Decision Trees

1. Each time randomly choose 8 features for each tree and do experiment.

The 0th tree Accuracy= 0.712
The 1th tree Accuracy= 0.686
The 2th tree Accuracy= 0.906
The 3th tree Accuracy= 0.648
The 4th tree Accuracy= 0.675
After training is complete, the w is (first dimension is bias)

-0.012738853503184686 0.1747042766151047 0.12738853503184702 0.5277525022747954
0.058234758871701604 0.13575710326619575
The training accuracy is 0.906
The testing accuracy is 0.8988195615514334

2.(a)In this experiment, the feature values are continuous. But we also need to use decision tree to make decision. So for each feature, we first get minimum and maximum value in training set, and divide values between min and max into 5 groups and categorize as A, B, C, D, E based on value.(If a number in test set is below minimum in training set, it would be E, if a number in test set is above maximum in training set, it would be A) For C and y_0, we can use the value that has the best accuracy that we get by cross-validation before.

When N=10, the accuracy on training set is 0.5925, the accuracy on test set is 0.565.

When N=30, the accuracy on training set is 0.6085, the accuracy on test set is 0.5267.

When N=100, the accuracy on training set is 0.6225, the accuracy on test set is 0.61.

2(b) Choose N=100.
In training set, the accuracy is 0.61, p=0.609, r=0.614, F1=0.611,
In test set, the accuracy is 0.5967, p=0.606, r=0.61, F1=0.608