模式识别导论上机题5-核函数

薛犇 1500012752

1. 程序实现说明

本次实验采用Matlab作为编程语言，使用的版本为2016b。

在实验的一开始，利用importdata函数读取hw5_data.txt中的数据，保存在向量 raw_data中。

```
raw_data = importdata('hw5_data.txt');
```

从中分离出样本x与标签y，并按照8:2的比例设置training与test集。

```
% process data
x = raw_data(:, 1:3);
y = raw_data(:, 4);

x_1 = x(y==0,:);
x_2 = x(y==1,:);
y_1 = y(y==0,:);
y_2 = y(y==1,:);

x_1_train = x_1(1:40,:);
x_2_train = x_2(1:40,:);
y_1_train = y_1(1:40,:);
y_2_train = y_2(1:40,:);
x_1_test = x_1(41:50, :);
x_2_test = x_2(41:50, :);
y_1_test = y_1(41:50,:);
y_2_test = y_2(41:50,:);
x_train = cat(1, x_1_train, x_2_train);
y_train = cat(1, y_1_train, y_2_train);
x_test = cat(1, x_1_test, x_2_test);
y_test = cat(1, y_1_test, y_2_test);
```

随后，选取核函数，本次采用rbf核作为核函数。定义如下：

$$K(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$$

在matlab中的实现如下：

```
% rbf kernel
function y = rbf(x1, x2)
    sigma = 1;
    gamma = 1 / (2 * sigma^2);
    y = exp(-gamma * sum((x1-x2).*(x1-x2)));
end
```

然后设置超参数的值，本次需要设置fisher算法中的 t 的大小，与rbf核的参数 σ 的大小。本次设置如下：

```
% hyper params
sigma = 1;
t = 1;
```

随后开始计算fisher方法中的一些参数的值，由于fisher最终的对偶表达式中，α的定义如下：

$$\alpha = \frac{d}{\lambda}(N + tK)^{-1}\Gamma$$

所以，实际上只要求出

$$N, K, \Gamma$$

其中

$$N = N_1 + N_2$$

$$N_1 = [\sum_{x \in w_1}(K(x_i, x) - \frac{1}{n_1}\sum_{x \in w_1}K(x_i, x))(K(x_j, x) - \frac{1}{n_1}\sum_{x \in w_1}K(x_j, x))]_{n \times n}$$

$$N_2 = [\sum_{x \in w_2}(K(x_i, x) - \frac{1}{n_2}\sum_{x \in w_2}K(x_i, x))(K(x_j, x) - \frac{1}{n_2}\sum_{x \in w_2}K(x_j, x))]_{n \times n}$$

$$K = [K(x_i, x_j)]_{n \times n}$$

$$\Gamma = (\frac{1}{n_1}\sum_{x \in w_1}K(x_1, x) - \frac{1}{n_2}\sum_{x \in w_2}K(x_1, x), ..., \frac{1}{n_1}\sum_{x \in w_1}K(x_n, x) - \frac{1}{n_2}\sum_{x \in w_2}K(x_n, x))$$

在matlab中的实现如下：

```
% K
for i=1:n
for j=1:n
K(i,j) = rbf(x_train(i,:),x_train(j,:));
end
end

% N
for i=1:n
for j=1:n
```

```
tmp_i = K(i,1:n1);
tmp_j = K(j,1:n1);
tmp_i = tmp_i - mean(tmp_i);
tmp_j = tmp_j - mean(tmp_j);
N1(i,j)=sum(tmp_i.*tmp_j);

tmp_i = K(i,n1+1:n);
tmp_j = K(j,n1+1:n);
tmp_i = tmp_i - mean(tmp_i);
tmp_j = tmp_j - mean(tmp_j);
N2(i,j)=sum(tmp_i.*tmp_j);
end
end

N = N1 + N2;

% Gamma
for i=1:n
tmp_1 = K(i,1:n1);
tmp_2 = K(i,n1+1:n);
Gamma(i) = mean(tmp_1)-mean(tmp_2);
end
```

以上可以算出权重，再根据偏移量b的定义：

$$b = -\frac{1}{n}\sum_{i=1}^{n}\sum j = 1^{n}\alpha_i K(x_i, x_j)$$

可以算出b。

之后就是**Test**的部分，利用公式如下：

$$f(x) = w^T\phi(x) + b$$

$$= \sum_{i=1}^{n}\alpha_i K(x_i, x) + b$$

就可以算出每个test样本的预测值pred：

```
% Testing
pred = zeros(np,1);

for i=1:np
for j=1:n
pred(i) = pred(i) + a(j)*rbf(x_train(j,:),x_test(i,:));
end
end

pred = pred + b

pred(pred > 0) = 0; % belongs to class 1
pred(pred < 0) = 1; % belongs to class 2
```

```
res = zeros(np,1);
res(pred==y_test) = 1;

% final accuracy
accuracy = sum(res)/np
```

2. 实验结果分析

实验结果如下：

```
pred =

    0.0199
    0.0417
    0.0484
    0.0550
    0.0592
    0.0597
    0.0331
    0.0182
    0.0679
    0.0586
    0.0015
   -0.0466
   -0.0556
   -0.0524
   -0.0447
   -0.0063
   -0.0033
   -0.0511
   -0.0178
   -0.0323


accuracy =

    0.9500
```

可以看出，在小样本集中，分类结果尚能满意。

而当调整超参数如下时：

```
t = 1
sigma = 2
```

结果如下：

```
pred =

    0.0778
```

```
    0.1475
    0.1487
    0.1336
    0.1521
    0.1515
    0.1099
    0.0397
    0.1545
    0.1472
   -0.0329
   -0.1415
   -0.1672
   -0.1460
   -0.1355
   -0.1067
   -0.0726
   -0.1311
   -0.1153
   -0.0974


accuracy =

     1
```

可以达到完全分开。