Contents

K近邻算法的实践

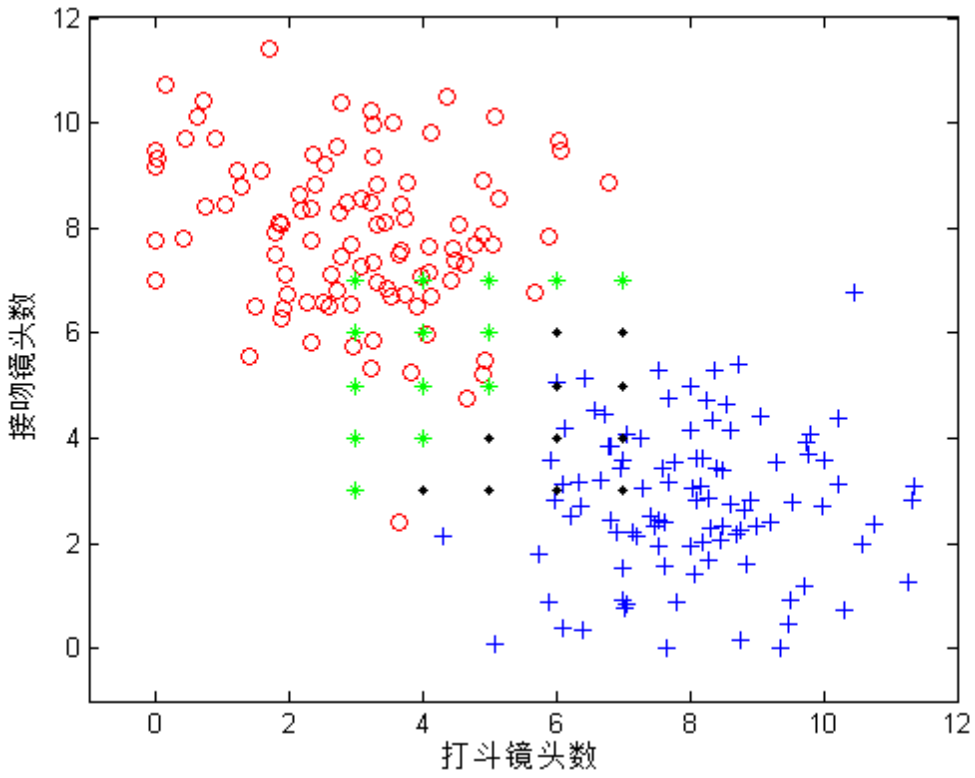
山东理工大学 数学院 周世祥 文本分类，聚类分析，预测分析，模式识别，图像处理等领域。

1. 初始化距离值为最大值，便于在搜索过程中迭代掉；
2. 计算待分类样本和每个训练样本的距离dist；
3. 得到目前k个最邻近样本中的最大距离maxdist；
4. 如果dist小于maxdist，则将该训练样本作为k最邻近样本；
5. 重复步骤（２），（３），（４），直到未知样本和所有训练样本的距离都算完；
6. 统计k邻近样本中每个类标号出现的次数；
7. 选择出现频率最大的类标号作为未知样本的类标号。

```
% 假设有一个具体应用为区分某一电影为动作片还是武侠片，首先，需要建立已知标签的样本，通过人工统计电影
% 中打斗镜头和接吻镜头数，并对相应的电影进行标签标注。之后，如果有一部未看过的电影，通过机器计算的方式
% 判断其为动作片还是爱情片。
```

```
clear all;
close all;
clc;
%%利用高斯分布，生成动作片数据和标签
aver1=[8 3];  %均值
covar1=[2 0;0 2.5];  %2维数据的协方差
data1=mvnrnd(aver1,covar1,100);   %产生高斯分布数据
for i=1:100    %令高斯分布产生数据中的负数为0
    for j=1:2   %因为打斗镜头数和接吻镜头数不能为负数
        if data1(i,j)<0
            data1(i,j)=0;
        end
    end
end
label1=ones(100,1);  %将该类数据的标签定义为1
plot(data1(:,1),data1(:,2),'+');  %用+绘制出数据
axis([-1 12 -1 12]); %设定两坐标轴范围
xlabel('打斗镜头数'); %标记横轴为打斗镜头数
ylabel('接吻镜头数'); %标记纵轴为接吻镜头数
hold on;
%%利用高斯分布，生成爱情片数据和标签
aver2=[3 8];
covar2=[2 0;0 2.5];
data2=mvnrnd(aver2,covar2,100); %产生高斯分布数据
for i=1:100    %另高斯分布产生数据中的负数为0
    for j=1:2  %因为打斗镜头数和接吻镜头数不能为负数
        if data2(i,j)<0
            data2(i,j)=0;
        end
    end
end
plot(data2(:,1),data2(:,2),'ro');  %用o绘制出数据
label2=label1+1; %将该类数据的标签定义为2
data=[data1;data2];
```

```
label=[label1;label2];
K=11;   %两个类，一般K取奇数有利于测试数据属于那个类
%测试数据，KNN算法看这个数属于哪个类，测试数据共计25个
%打斗镜头数遍历3-7，接吻镜头书也遍历3-7
for movenum=3:1:7
    for kissnum=3:1:7
        test_data=[movenum kissnum];  %测试数据，为5X5矩阵
        %%下面开始KNN算法，显然这里是11NN。
        %求测试数据和类中每个数据的距离，欧式距离（或马氏距离）
        distance=zeros(200,1);
        for i=1:200
            distance(i)=sqrt((test_data(1)-data(i,1)).^2+(test_data(2)-data(i,2)).^2);
        end
        %选择排序法，只找出最小的前K个数据,对数据和标号都进行排序
        for i=1:K
            ma=distance(i);
            for j=i+1:200
                if distance(j)<ma
                    ma=distance(j);
                    label_ma=label(j);
                    tmp=j;
                end
            end
            distance(tmp)=distance(i);  %排数据
            distance(i)=ma;
            label(tmp)=label(i);        %排标签
            label(i)=label_ma;
        end
        cls1=0; %统计类1中距离测试数据最近的个数
        for i=1:K
            if label(i)==1
                cls1=cls1+1;
            end
        end
        cls2=K-cls1;    %类2中距离测试数据最近的个数
        if cls1>cls2
            plot(movenum,kissnum, 'k.'); %属于类1（动作片）的数据画小黑点
        else
            plot(movenum,kissnum, 'g*'); %属于类2（爱情片）的数据画绿色*
        end
        label=[label1;label2]; %更新label标签排序
    end
end
```



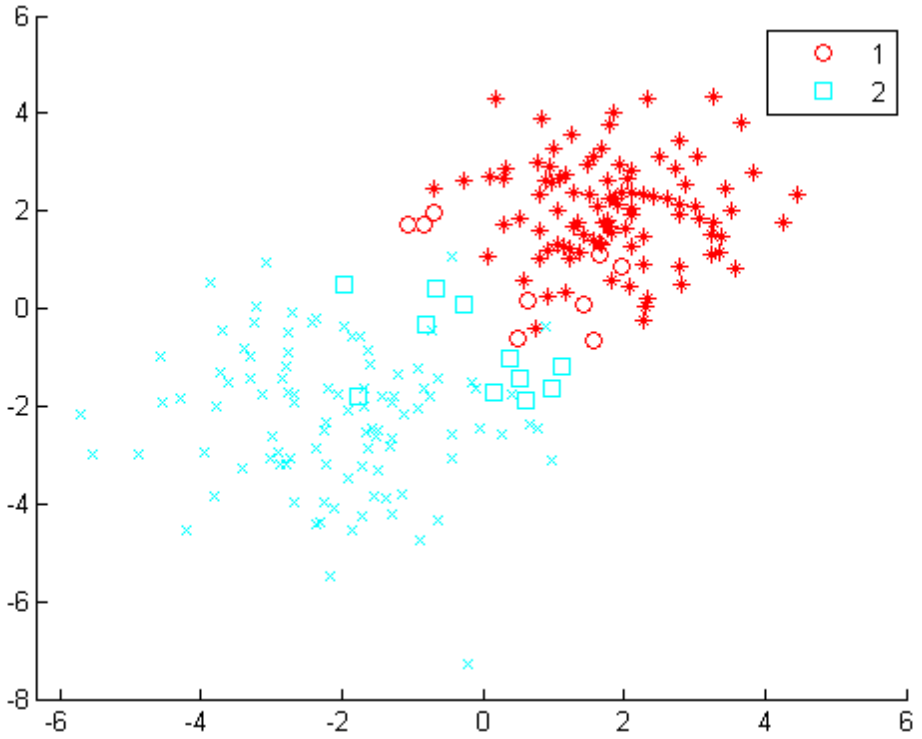
Matlab 自带KNN算法函数knnclassify实现

help knnclassify

语法：

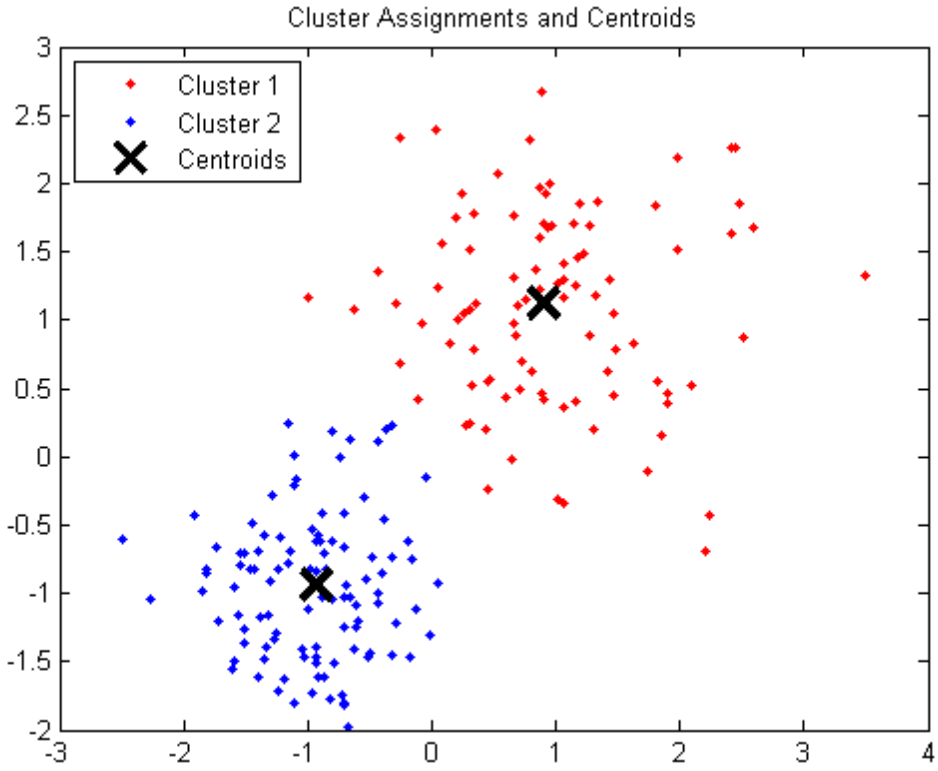
- Class = knnclassify(Sample, Training, Group)
- Class = knnclassify(Sample, Training, Group, k)
- Class = knnclassify(Sample, Training, Group, k, distance)
- Class = knnclassify(Sample, Training, Group, k, distance, rule)
- 'euclidean' — Euclidean distance (default)
- 'cityblock' — Sum of absolute differences
- 'cosine' — One minus the cosine of the included angle between points (treated as vectors)
- 'correlation' — One minus the sample correlation between points (treated as sequences of values)
- 'hamming' — Percentage of bits that differ (suitable only for binary data)

```
clc
close all;
clear
%生成200个样本数据
training = [mvnrnd([2  2],eye(2), 100); mvnrnd([-2 -2], 2*eye(2), 100)];
%mvnrnd([2  2],eye(2),100)表示随机生成多元正态分布100X2矩阵，每一列以2，2为均值，eye(2)为协方差
%200个样本数据前100标记为标签1，后100个标记为标签2
group = [ones(100,1); 2*ones(100,1)];
%绘制出离散的样本数据点
gscatter(training(:,1),training(:,2),group,'rc','*x');
hold on;
% 生成待分类样本20个
sample = unifrnd(-2, 2, 20, 2);
%产生一个100X2,这个矩阵中的每个元素为20 到30之间连续均匀分布的随机数
K=3;%KNN算法中K的取值
cK = knnclassify(sample,training,group,K);
gscatter(sample(:,1),sample(:,2),cK,'rc','os');
```



Matlab 自带K均值算法函数kmeans实现

```
clc;
clear;
close all;
X = [randn(100,2)*0.75+ones(100,2);randn(100,2)*0.5-ones(100,2)]; %产生两组随机数据
[idx,C] = kmeans(X,2,'Distance','cityblock','Replicates',5);%利用K均值算法进行分组
plot(X(idx==1,1),X(idx==1,2),'r.','MarkerSize',12) %绘制分组后第一组的数据
hold on
plot(X(idx==2,1),X(idx==2,2),'b.','MarkerSize',12) %绘制分组后第二组的数据
plot(C(:,1),C(:,2),'kx','MarkerSize',15,'LineWidth',3) %绘制第一组和第二组数据的中心点
legend('Cluster 1','Cluster 2','Centroids','Location','NW')
title 'Cluster Assignments and Centroids'
hold off
```



参考文献

1. 冷雨泉等，《机器学习入门到实践：MATLAB实践应用》，清华大学出版社，2019年3月
2. https://baike.baidu.com/item/%E9%82%BB%E8%BF%91%E7%AE%97%E6%B3%95/1151153?fromtitle=Knn&fromid=3479559

Published with MATLAB® R2014a