Machine Learning Homework 2.1

1、 实验目的:

利用贝叶斯公式设计分类器进行分类

2、 试验要求

利用贝叶斯分类器再二维平面中正确分裂单高斯分布生成的点和双高斯分布生成的点

3、 试验环境

Windows10、matlab2016

4、 试验基本原理

(1)、假设 x = {a_1, a_2, ⋯, a_m}是一个待分类的数据，而每个 a 作为 x 的一个特征属性。

(2)、有类别集合 C = {y_1， y_2, ⋯, y_n}.

(3)、计算 P （y_1|x） , P(y_2|x), ⋯, P(y_n|x).

(4)、如果 P(y_k|x) = max{ P(y_1|x） , P(y_2|x), ⋯, P(y_n|x)}。那么就成 x 属于 y_k 类

现在的目的就是为了计算出步骤 3 中每个条件的条件概率:

(5)、由训练集合，统计得到再各类别下各个特征属性的条件概率估计，也就是 P（a_1|y_1)， ⋯， P （a_m|y_1） , ⋯ P （a_1|y_n)， ⋯， P （a_m|y_n）

(6)、如果每个特征属性条件独立，那么根据贝叶斯定理有:

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)}$$

(7)、分母对所有类别为常数，因此只需最大化分子，又由于各特征属性为条件独立，

所以有:

$$P(x|y_i)P(y_i) = P(a_1|y_i)P(a_2|y_i)...P(a_m|y_i)P(y_i) = P(y_i)\prod_{j=1}^{m} P(a_j|y_i)$$

(8)、对于具有多个特征参数的样本，正态分布的概率密度函数为:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

(9)、令 $g_i(x) = P(x|y_i)P(y_i)$ 那么当对于所有 j，j 不等于 i 时，有 $g_i(x) > g_j(x)$，那么可以认定 x 属于 i 类

(10)、 有

$$\begin{aligned} g_i(\mathbf{x}) &= P(y_i)N(\mathbf{x}|\mu_i,\textstyle\sum_i) \\ &= \frac{P(y_i)}{(2\pi)^{\frac{d}{2}}|\sum_i|^{\frac{1}{2}}} exp\{-\tfrac{1}{2}(\mathbf{x}-\mu_i)\textstyle\sum_i^{-1}(\mathbf{x}-\mu_i)^T\} \end{aligned}$$

(11)、对上式右端取对数，并将于样本所属类别无关的项除去，得到化简式子:

$$g_i(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\mu_i)\textstyle\sum_i^{-1}(\mathbf{x}-\mu_i)^T + \ln P(y_i) - \tfrac{1}{2}\ln|\sum_i|$$

5、 试验过程

(1)、产生两类数据，并由均值和协方差公式计算:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T = \frac{1}{N}XX^T - \mu\mu^T$$

(2)、由此计算判别函数 $g_i(x)$

(3)、代码见附件

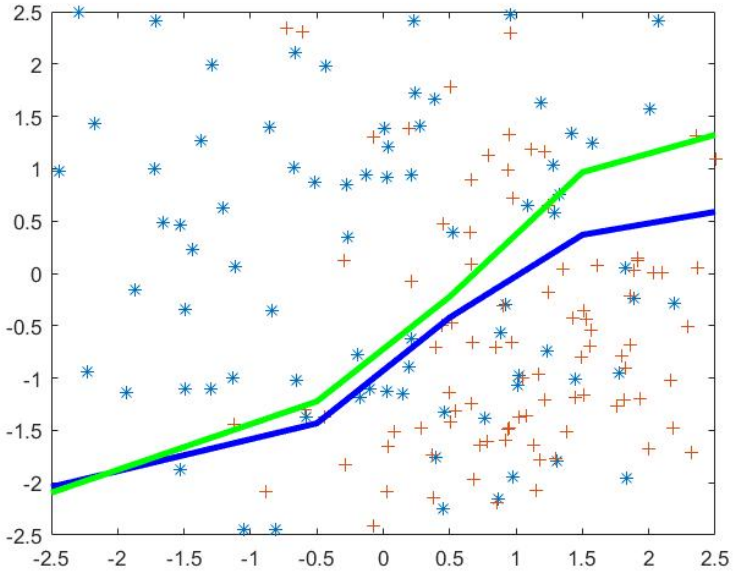
6、 实验结果及其分析

(1) 实验参数设置如下,

```
N = 200; 随机取得的点数
mu1 = [0, 0]; 二维单高斯分布的均值
sigma1 = [3, -0.5; -0.5, 3]; 二维单高斯分布的协方差
r1 = mvnrnd(mu1, sigma1, N);随机生成200个数据点
p=[0.7, 0.3]; 二维高斯混合函数占比
```

```
mu2 = [1 -1; 1 1]; 均值
SIGMA1 = [0.8 0.5; 0.5 0.8]; 协方差
SIGMA2 = [0.9 -0.5; -0.5 0.9]; 协方差
X = cat(3, SIGMA1, SIGMA2);
gm = gmdistribution(mu2, X, p);
[r2, compIdx] = random(gm, N); 随机取200个数据点
```

(2)、实验结果



(3)、实验结果分析

上述实验结果中，绿色的线表示已知两个分布的均值和协方差情况下通过贝叶斯分类得到的最优的分类结果，而蓝色的线表示在给定数据点的情况下通过对数据处理得到二维单高斯分布和二维双高斯分布的均值和协方差，利用估计的均值和协方差通过贝叶斯分类器得到的分类结果，实验发现数据点在 100 左右利用估计的均值和协方差基本可以达到最优的分类标准。