

作业 6

郭一隆

June 8, 2018

1 KNN 算法

1.1 问题描述

- 问题背景：有两组二维数据，在空间中的样本分布如图1所示。对一个新的样本点，请尝试用 KNN 算法判断它的所属组别。
- 提供数据：trainData.txt
 - 格式：200 行，每行为一个样本点的二维坐标及其组别 (1 或 2)，即 $x, y, label$
 - 读取方式：Matlab 下可以直接使用 `load` 进行载入，得到一个 200x3 的矩阵；其他语言也不难读取
- 具体要求：
 - 测试样本点集为 $\{(x, y) | x = -6:0.1:6, y = -6:0.1:6\}$
 - 尝试采用不同的 k 值（如 1, 3, 5），观察结果的变化并进行分析

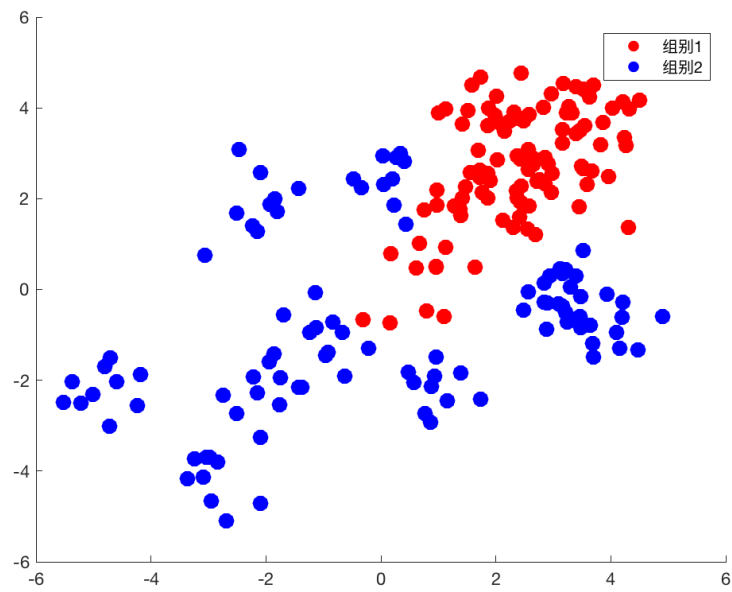


Figure 1: 训练样本分布

1.2 问题求解

采用欧式距离作为样本点之间的距离标准，分别实验 $K = 1, 3, 5, 10$ 情况下的 KNN 分类结果如下

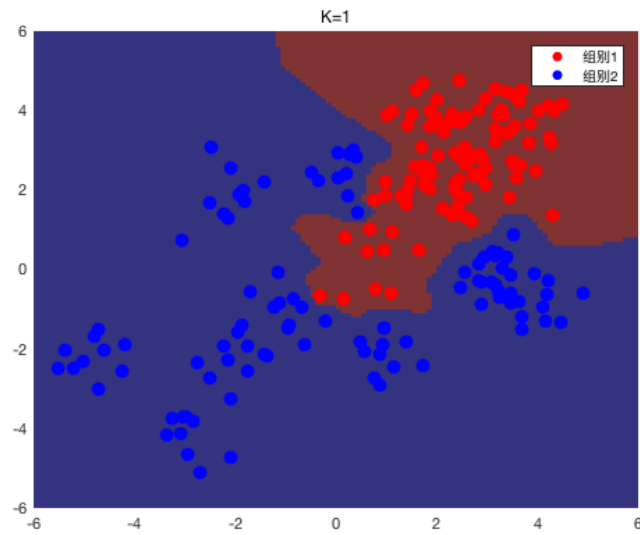


Figure 2: K=1 结果

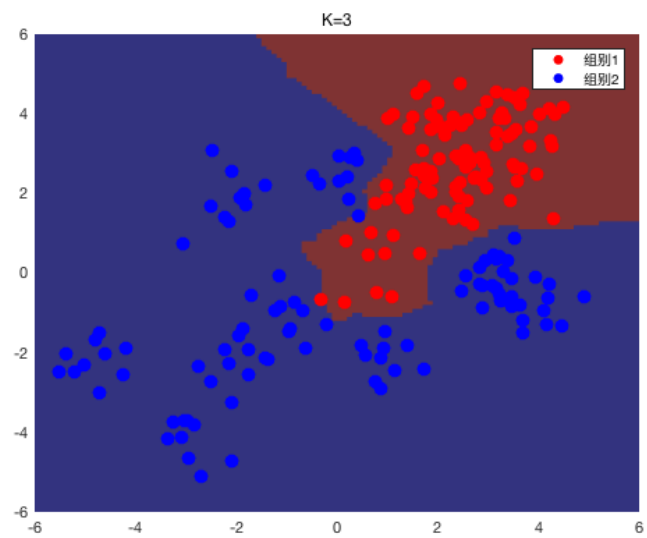


Figure 3: K=3 结果

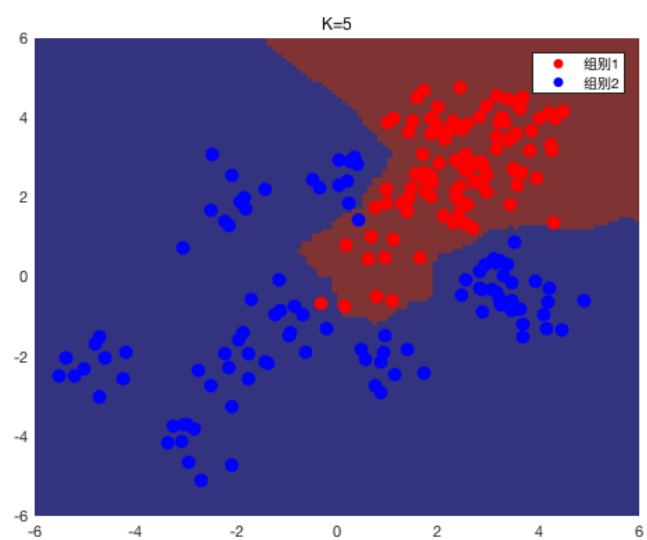


Figure 4: K=5 结果

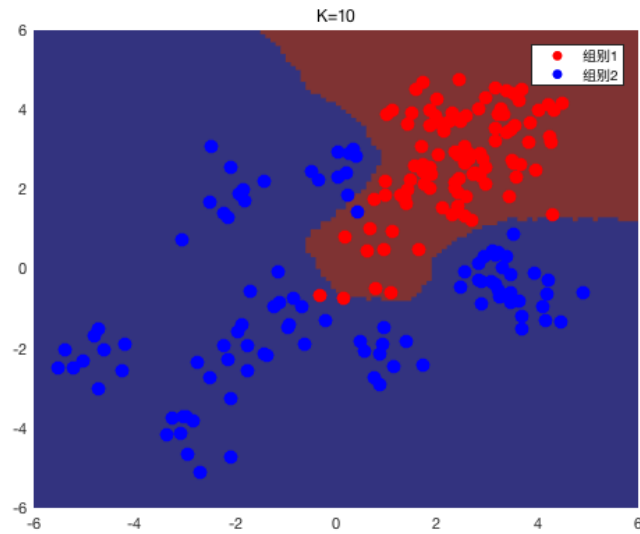


Figure 5: K=10 结果

1.3 总结

- K 的取值会影响分界面;
- K 过小容易产生过拟合;
- 较大的 K 会使分界面趋于平缓;
- KNN 在分界面处对于密度差异较大的两类样本不能有效地区分。

2 FISHER 判别准则

2.1 问题描述

对下列两种情况，求采用 Fisher 判别准则时的投影向量 \mathbf{w} 和分类界面，并作图。

$$\begin{cases} \omega_1 = \{(2, 0), (2, 2), (2, 4), (3, 3)\} \\ \omega_2 = \{(0, 3), (-2, 2), (-1, -1), (1, -2), (3, -1)\} \end{cases}$$

$$\begin{cases} \omega_1 = \{(1, 1), (2, 0), (2, 1), (0, 2), (1, 3)\} \\ \omega_2 = \{(-1, 2), (0, 0), (-1, 0), (-1, -1), (0, -2)\} \end{cases}$$

2.2 问题求解

1. 求总类内离散度

$$S_W = \sum_i S_i = \sum_i \sum_{x \in \omega_i} (X - M_i)(X - M_i)^T$$

2. 求得降维线单位向量

$$W_0 = \frac{S_W^{-1}(M_1 - M_2)}{\|S_W^{-1}(M_1 - M_2)\|}$$

3. 将原样本点投影至 W_0 方向

4. 求得降维后样本点的分界面（最小欧式距离判别）

2.3 实验结果

2.3.1 情况 1

$$\begin{cases} \omega_1 = & \{(2, 0), (2, 2), (2, 4), (3, 3)\} \\ \omega_2 = & \{(0, 3), (-2, 2), (-1, -1), (1, -2), (3, -1)\} \end{cases}$$

求得

$$W_0 = (0.9458, 0.3247)$$

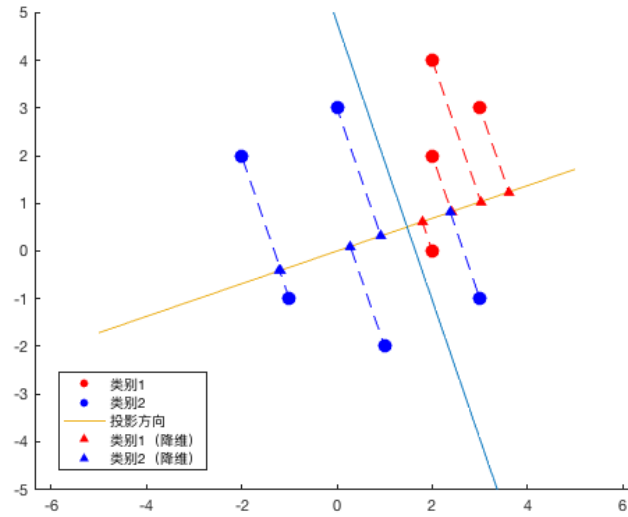


Figure 6: 分界面用蓝线标明

2.3.2 情况 2

$$\begin{cases} \omega_1 = & \{(1, 1), (2, 0), (2, 1), (0, 2), (1, 3)\} \\ \omega_2 = & \{(-1, 2), (0, 0), (-1, 0), (-1, -1), (0, -2)\} \end{cases}$$

求得

$$W_0 = (0.8980, 0.4400)$$

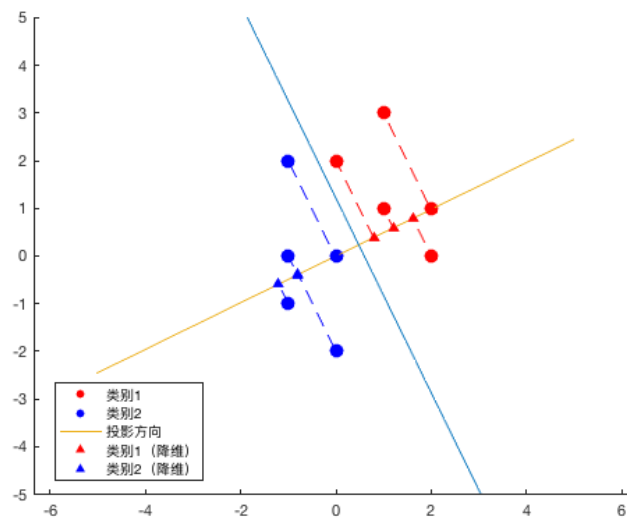


Figure 7: 分界面用蓝线标明

2.4 总结

- Fisher 准则基于不同类别之间的特点（离散度）来寻找投影方向，与 PCA 不同。
- 数据降维后，可以采用简单的线性识别函数，也可以采用 KNN 等较复杂的方法寻找分界面，但要注意避免过拟合问题。

附

- 源码仓库链接：

<https://github.com/Nuulllll/pattern-recognition-assignment-6>

- 运行环境：

MATLAB 版本：9.1.0.441655 (R2016b)

MATLAB 许可证编号：1114839

操作系统：Mac OS X Version: 10.13.4 Build: 17E199

Java 版本：Java 1.7.0_75-b13 with Oracle Corporation Java
HotSpot(TM) 64-Bit Server VM mixed mode

- 运行方法：

直接运行 `run.m` 即可。