第七章 判别分析算法

判别分析算法最早由Fisher在1936年提出，是一种经典而常用的机器学习方法，本质上也是一种线性算法，常用来做特征提取、数据降维和任务分类，可用于二分类或多分类问题，在人脸识别或检测等领域发挥重要作用。根据每一种分类的协方差矩阵是否相同，判别分析算法分为线性判别分析和二次判别分析，其中线性判别分析假定每一种分类的协方差矩阵相同，而在样本示例集数据量较大、或者观测类别较多时，等协方差矩阵的假设会被拒绝，就需要用到二次判别分析。

目录

CONTENTS

PART 01

判别分析算法的基本原理

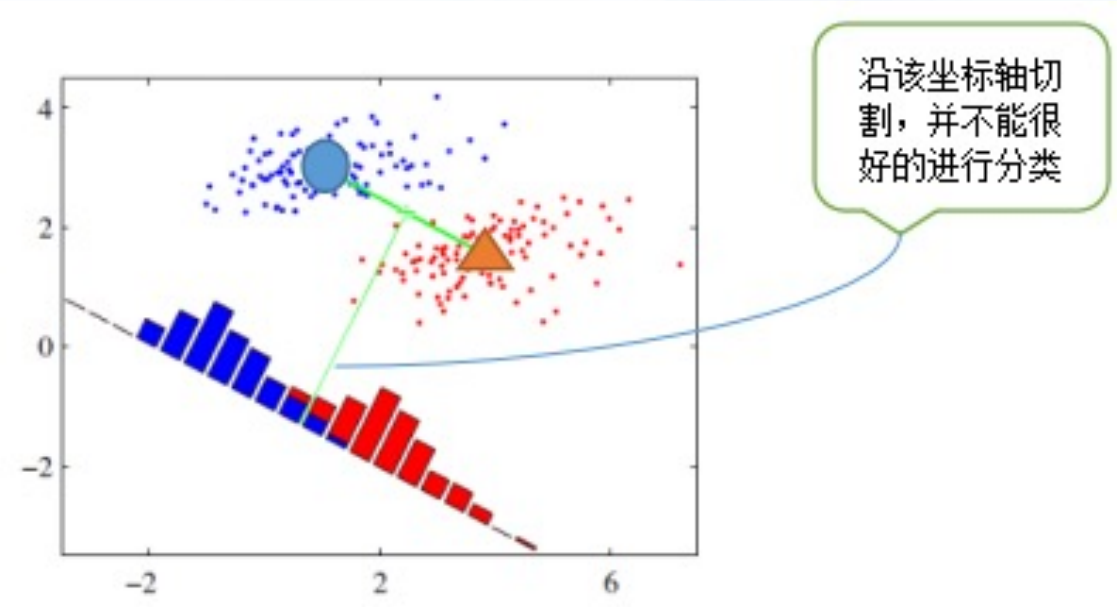
线性判别分析的基本原理

线性判别分析算法使用贝叶斯规则来确定示例属于哪一类的后验概率，该算法假设每个类别中的观测值均来自多元正态分布，并且预测变量的协方差在响应变量Y的所有k个水平上都是相同的，或者说假定不同分组样本的协方差矩阵近似相等。

线性判别分析算法的基本思想是“类间大、类内小”，实现过程是：首先将样本示例全集分为训练样本和测试样本，针对训练样本，设法找到一条直线，将所有样本示例投影到这条直线上，使得相同分类的样本示例在该直线上的投影尽可能落在一起，而不同分类的样本示例在该直线上的投影尽可能远离，或者一言以蔽之，就是使得同类之间的差异性尽可能小，不同类之间的差异性尽可能大；然后针对测试样本，将其投影到已经找到的直线上，根据具体投影点的落地位置来判定样本示例的类别。

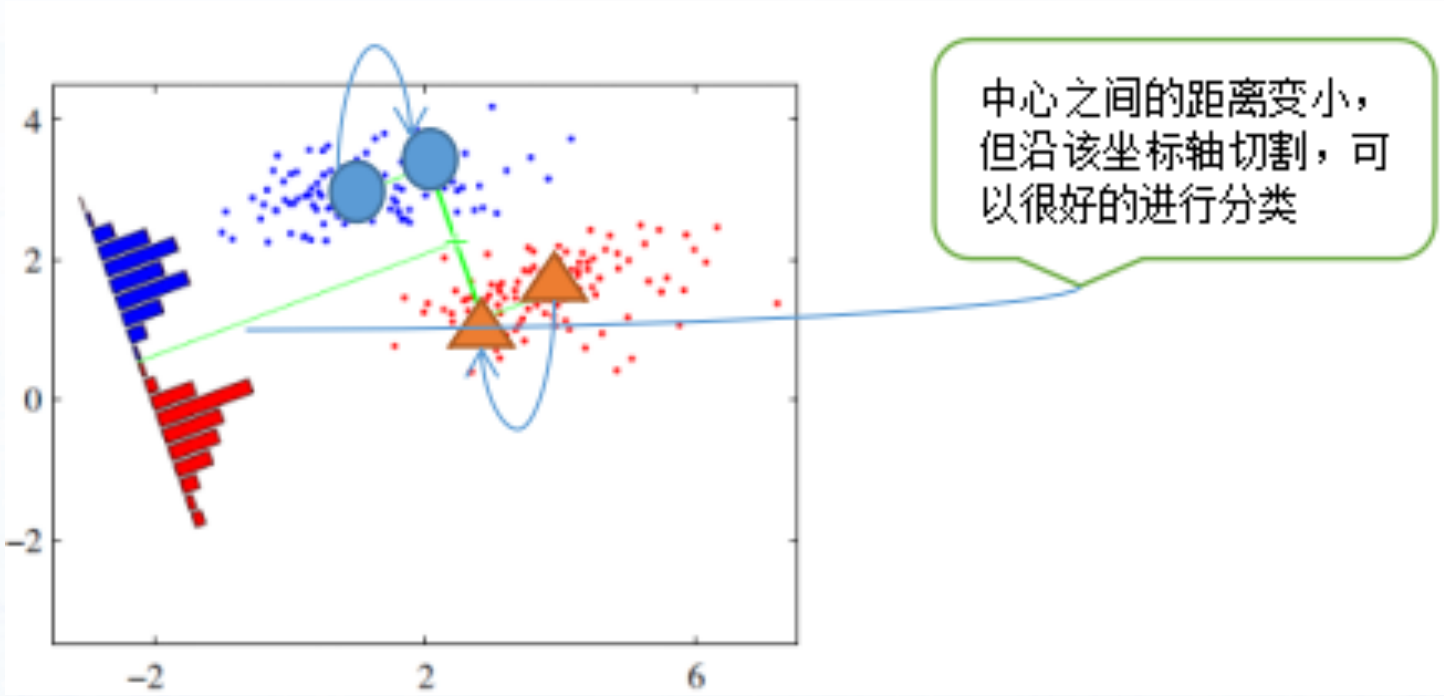
线性判别分析的基本原理

作为一种有监督的机器学习方法，线性判别分析在分类方面具有独特的优势，相对于主成分分析PCA算法（将在后面详细介绍）这种无监督学习方法，线性判别分析充分利用了数据内部的原始分类信息。主成分分析PCA算法通过寻找k个向量，将数据投影到这k个向量展开的线性子空间上，是最大化两类投影中心距离准则下得到的分类结果，该算法将数据整体映射到了最方便表示这组数据的坐标轴上，或者说实现了投影误差最小化。但是由于PCA算法将整组数据整映射时没有利用数据原始分类信息，分类效果并不理想。



线性判别分析的基本原理

线性判别分析LDA算法如图所示，两组输入映射到了另一个坐标轴上，可以看出这是一种样本示例区分性更高的一种投影方式，虽然在增加了分类信息之后，两类中心之间的距离在投影之后有所减小，但投影后的样本点在每一类中分布得更为集中了，或者说每类内部的方差比上图中更小，从而两类样本的可区分度提高了。



线性判别分析的算法过程

从算法的角度，线性判别分析 LDA 的实现过程是：

1、分别计算每一类的均值向量，均值向量之间的差异用来衡量类间距离。假定样本示例全集中共有两类，则有：

$$N = N_1 + N_2$$

$$\mu_{p1} = \frac{1}{N_1}\sum_{i=1}^{N_1} p_i$$

$$\mu_{p2} = \frac{1}{N_2}\sum_{i=1}^{N_2} p_i$$

2、分别计算每一类的协方差矩阵，协方差之和用来衡量类内距离。

$$\sigma_{p1} = \frac{1}{N_1}\sum_{i=1}^{N_1} (p_i - \mu p_1)(p_i - \mu p_1)^T$$

$$\sigma_{p2} = \frac{1}{N_2}\sum_{i=1}^{N_2} (p_i - \mu p_2)(p_i - \mu p_2)^T$$

线性判别分析的算法过程

3、基于前两步结果，可以写出代价函数。

$$\text{cost}(w) = \frac{(w^T \mu p_1 - w^T \mu p_2)^2}{w^T \sigma_{p1} w + w^T \sigma_{p2} w}$$

代价函数中的分子是两个类别均值向量大小之差的平方，分子的值越大，代表类间差异性越大；函数中的分母为两个类别样本点的协方差之和，分母的值越大，代表类内差异越大。

4、求解权重系数 w 使得代价函数最优化

在前面我们可以看到，代价函数是关于权重 w 的函数，所谓“类间大、类内小”的目标，就是要求出使代价函数最大时的权重 w。或者说满足以下公式：

$$w = \underset{w}{\text{argmax}} \left(\frac{(w^T \mu p_1 - w^T \mu p_2)^2}{w^T \sigma_{p1} w + w^T \sigma_{p2} w} \right)$$

5、根据权重系数判断新样本分类

求出权重系数后，前面提到，类间距离使用均值向量之间的差异来衡量，所以针对新样本或者测试样本，新样本点距离哪一个类别的均值向量更近，那么新样本就被预测分配到哪个类别，数学形式为：

$$k = \underset{k}{\text{argmin}} |(w^T x - w^T \mu_k|$$

二次判别分析的基本原理

二次判别分析QDA（Quadratic Discriminant Analysis）是与线性判别分析类似的另外一种线性判别分析算法，二者区别在于线性判别分析假设每一种分类的协方差矩阵相同，类别之间的判别边界是条直线，而二次判别分析假设每一种分类的协方差矩阵不同，类别之间的判别边界不是直线，所以，与LDA相比，QDA通常更加灵活。

从方差-偏差的角度来看，二次判别分析QDA和线性判别分析LDA也是一种典型的方差-偏差取舍选择，LDA因为假设每一种分类的协方差矩阵相同，所以相对方差更低；而QDA因为假设每一种分类的协方差矩阵不同，所以方差会更高，但由于更加灵活，相对偏差更低。从应用场景来看，如果样本示例全集容量比较小，对协方差矩阵很难估计准确时，采取LDA方法相对更合适；而如果样本示例全集容量比较大，或者合理预期类间协方差矩阵差异比较大时，则采取QDA方法相对更合适。

二次判别分析的基本原理

二次判别分析 QDA 的数学形式为：

$$
\begin{aligned}
h(x) &= \underset{k}{\operatorname{argmax}} P(k \mid x) \\
&= \underset{k}{\operatorname{argmax}} \ln P(k \mid x) \\
&= \underset{k}{\operatorname{argmax}} \ln f_k(x) + \ln P(k) \\
&= \underset{k}{\operatorname{argmax}} \ln \left(\frac{e^{-\frac{(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}{2}}}{|\Sigma_k|^{\frac{1}{2}} (2\pi)^{\frac{p}{2}}} \right) + \ln P(k) \\
&= \underset{k}{\operatorname{argmax}} -\frac{1}{2} (x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k) - \ln \left(|\Sigma_k|^{\frac{1}{2}} (2\pi)^{\frac{p}{2}} \right) + \ln P(k) \\
&= \underset{k}{\operatorname{argmax}} -\frac{1}{2} (x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k) - \frac{1}{2} \ln (|\Sigma_k|) + \ln P(k)
\end{aligned}
$$

其中二次判别函数为：

$$
-\frac{1}{2} (x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k) - \frac{1}{2} \ln (|\Sigma_k|) + \ln P(k)
$$

针对特定样本示例，二次判别算法就是找到最优决策规则，使得将样本示例分别到类别 k，能够使得其二次判别函数最大化。

PART 02

数据准备

数据准备

本节我们以“数据7.1”为例进行讲解。“数据7.1”记录的是某商业银行在山东地区的部分支行经营数据（虚拟数据，不涉及商业秘密），案例背景是该商业银行正在推动支行开展转型，实现所有支行的做大做强。数据文件中的变量包括这些商业银行全部支行的转型情况（V1）、存款规模（V2）、EVA（V3）、中间业务收入（V4）、员工人数（V5）。转型情况（V1）又分为三个类别：“0”表示“未转型网点”；“1”表示“一般网点”；“2”表示“精品网点”。

下面以转型情况（V1）为响应变量，以存款规模（V2）、EVA（V3）、中间业务收入（V4）、员工人数（V5）为特征变量，构建判别分析算法模型，包括线性判别分析和二次判别分析。

载入分析所需要的模块和函数

在进行分析之前，我们首先载入分析所需要的模块和函数，读取数据集并进行观察。

示例

参阅教材内容

线性判别分析降维优势展示

在前面，我们提到线性判别分析不仅可以用来进行任务分类，还可以进行降维处理，由于其依据使用贝叶斯规则，充分利用了既有分类信息，所以在降维时可以很好的保存样本特征和类别的信息关联。下面我们进行演示。

一、绘制三维数据的分布图

二、使用PCA进行降维

三、使用LDA进行降维

示例

参阅教材内容

数据读取及观察

大家首先需要将本书提供的数据文件放入安装python的默认路径位置，并从相应位置进行读取。

示例

参阅教材内容

PART 03

特征变量相关性分析

特征变量相关性分析

示例

参阅教材内容

PART 04

使用样本全集开展线性判别分析

模型估计及性能分析

示例

参阅教材内容

运用两个特征变量绘制LDA决策边界图

示例

参阅教材内容

PART 05

使用分割样本开展线性判别分析

使用分割样本开展线性判别分析

示例

参阅教材内容

PART 06

使用分割样本开展二次判别分析

模型估计

示例

参阅教材内容

运用两个特征变量绘制QDA决策边界图

示例

参阅教材内容

PART 07

习　题

习题

一、使用“数据6.1”数据文件（详情已在第6章中介绍），以收入档次（V1）为响应变量，以工作年限（V2）、绩效考核得分（V3）和违规操作积分（V4）为特征变量，构建判别分析算法模型。

1、载入分析所需要的库和模块

2、数据读取及观察。

3、特征变量相关性分析。

4、使用样本示例全集开展线性判别分析。

（1）模型估计及性能分析；

（2）运用两个特征变量绘制LDA决策边界图。

5、使用分割样本开展线性判别分析。

感谢聆听

THANKS