15 | 从回归到分类：联系函数与降维

2018-07-07 王天一

讲述：王天一
时长 19:11 大小 8.79M

线性模型最初被用来解决回归问题（regression），可在实际应用中，更加普遍的是分类问题（classification）。要用线性模型解决分类问题的话，就需要将线性模型原始的连续输出转换成不同的类别。

在分类问题中，一种特殊的情况是类别非黑即白，只有两种，这样的问题就是二分类问题，它可以看成是多分类问题的一个特例，也是今天讨论的对象。

将回归结果转化为分类结果，其实就是将属性的线性组合转化成分类的标准，具体的操作方式有两种：一种是**直接用阈值区分回归结果**，根据回归值与阈值的关系直接输出样本类别的

标签；另一种是**用似然度区分回归结果**，根据回归值和似然性的关系输出样本属于某个类别的概率。

这两类输出可以分别被视为**硬输出**和**软输出**，它们代表了解决分类问题不同的思路。

硬输出是对数据的分类边界进行建模。实现硬输出的函数，也就是将输入数据映射为输出类别的函数叫作**判别函数**（discriminant）。判别函数可以将数据空间划分成若干个决策区域，每个区域对应一个输出的类别。不同判别区域之间的分界叫作**决策边界**（decision boundary），对应着判别函数取得某个常数时所对应的图形。用线性模型解决分类问题，就意味着得到的决策边界具有线性的形状。

最简单的判别函数就是未经任何变换的线性回归模型 $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$，它将回归值大于某个阈值（可以通过调整截距 b 设置为 0）的样本判定为正例，小于阈值的样本则判定为负例。

在处理多分类任务时，判别函数对每个类别都计算出一组系数 $\mathbf{w}_k$ 和 b_k，并选择使 $y_k(\mathbf{x})$ 最大的 k 作为输出类别。如果分类的边界较为复杂，还可以通过基函数的扩展或者核技巧来突破线性的限制，相关的内容会在后面的文章中涉及。

今天我们先来看看**基于软输出的分类方法。**软输出利用的是似然度，需要建立关于数据的概率密度的模型，常见的具体做法是对线性回归的结果施加某种变换，其数学表达式可以写成

$$y(\mathbf{x}) = g^{-1}(\mathbf{w}^T\mathbf{x} + b)$$

这里的 $g(\cdot)$ 被称为**联系函数**（link function），其反函数 $f(\cdot) = g^{-1}$ 则被称为**激活函数**（activation function）。**正是联系函数架起了线性模型从回归到分类的桥梁。**

最典型的软输出分类模型就是逻辑回归。在“人工智能基础课”中我曾介绍过，逻辑回归（logistic regression）是基于概率的分类算法，估计的是样本归属于某个类别的后验概率，那么根据贝叶斯定理，二分类问题中的后验概率就可以写成

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_1)p(C_1) + p(\mathbf{x}|C_2)p(C_2)}$$

对这个表达式做个简单的变量代换，就可以得到

$$p(C_1|\mathbf{x}) = \frac{1}{1+\exp(-a)} = \sigma(a)$$

这里的 $\sigma(\cdot)$ 表示对数几率函数（logistic function），也就是逻辑回归的联系函数，这个非线性的联系函数可以将任意输入映射到 [0, 1] 之间。对数几率函数的自变量 a 可以改写成

$$a = \ln \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)} = \ln \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} + \ln \frac{p(C_1)}{p(C_2)} = \mathbf{w}^T\mathbf{x} + b$$

逻辑回归并不能直接给出参数 $\mathbf{w}$ 的解析解，因此需要结合最优化的方法使用。确定参数最常用的方式是使用最大似然估计（maximum likelihood estimation），找到如训练数据匹配度最高的一组参数。

在二分类问题中，若假设当 $\mathbf{x}$ 属于类 C_1 时，输出的分类结果 r 为 1，属于类 C_2 时，输出的分类结果 r 为 0，那么每个单独的分类结果都满足参数为 $\sigma(\mathbf{x})$ 的两点分布，所有结果构成的向量 $\mathbf{r}$ 就会满足二项分布，这时的似然概率就可以写成分类结果的连乘

$$p(\mathbf{w}, b|\mathbf{x}) = \prod_i \sigma(\mathbf{x_i})^{r_i}[1-\sigma(\mathbf{x_i})]^{(1-r_i)}$$

对似然概率求对数并求解最大值，就可以得到最优的参数了。

和逻辑回归相似的另一种分类模型是线性判别分析，它不仅要估计数据的概率密度，还应用了降维的思想。在前面的两篇文章中，我和你分享了对数据进行线性降维和非线性降维的方法。

其实降维不光是数据预处理的一种手段，它还可以用来执行分类任务——本质上讲，分类问题就是将高维的数据投影到一维的类别标签上。

维度的下降会导致信息的损失，从而使数据在标签维度上产生重叠。属于相同类别的数据重叠在一起并不是严重的问题，但类别不同的数据的重叠就会增加分类问题的错误率，因此**好的分类算法既要让相同类别的数据足够接近，又要让不同类别的数据足够远离**。基于这一原则进行分类的方法就是线性判别分析。

用于二分类的**线性判别分析**由著名的统计学家罗纳德·费舍尔于 1936 年提出，按人类的年龄计算已是耄耋之年。归根结底，线性判别分析也是从概率出发，假设不同类别的数据来源于均值不同而方差相同的正态分布，通过判定数据归属于不同正态的可能性来确定类别。

但在设计线性判别分析时，费舍尔利用了一种不同的思路。在计算二分类问题的决策边界时，线性判别分析首先要计算两个类别中数据的均值，以此作为特征来区分不同的类别，让不同类别的数据足够远离就是让两个均值在决策边界上的投影之间的距离足够大。

但仅是均值远离还不够。数据在不同维度上的分布不同会导致有些方向的方差较大，而有些方向的方差较小。如果仅仅考虑均值而忽略了方差，就可能导致判决边界落在波动较大的方向上，由此产生的长尾效应容易使不同类别的数据互相重叠，从而影响分类的精度。因此在投影时，还要让相同类别的数据尽可能集中分布，以避免混叠的出现。

假定训练数据分属两个类别 C_1 和 C_2，每个类别中数据的均值用向量 $\mathbf{m}_1$ 和 $\mathbf{m}_2$ 表示，那么这两个均值在超平面 $\mathbf{y} = \mathbf{w}^T\mathbf{x} + b$ 上的投影就等于

$$m_k = \mathbf{w}^T\mathbf{m}_k (k = 1, 2)$$

降维后两个类各自的方差可以表示为

$$s_k = \sum_{n \in C_k} (y_n - m_k)^2 (k = 1, 2)$$

要同时保证类间距最大和类内方差最小，可以通过最大化下面的目标函数来实现

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

其中待求解的参数 $\mathbf{w}$ 需要满足归一化条件 $||\mathbf{w}||_2^2 = 1$，而这并不会对 $\mathbf{w}$ 的方向造成影响。将线性回归模型代入 $J(\mathbf{w})$ 的表达式，可以将它改写成

$$J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}$$

这里有这么几个概念。

类间协方差矩阵（between-class covariance matrix）

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$$

类内协方差矩阵（within-class covariance matrix）

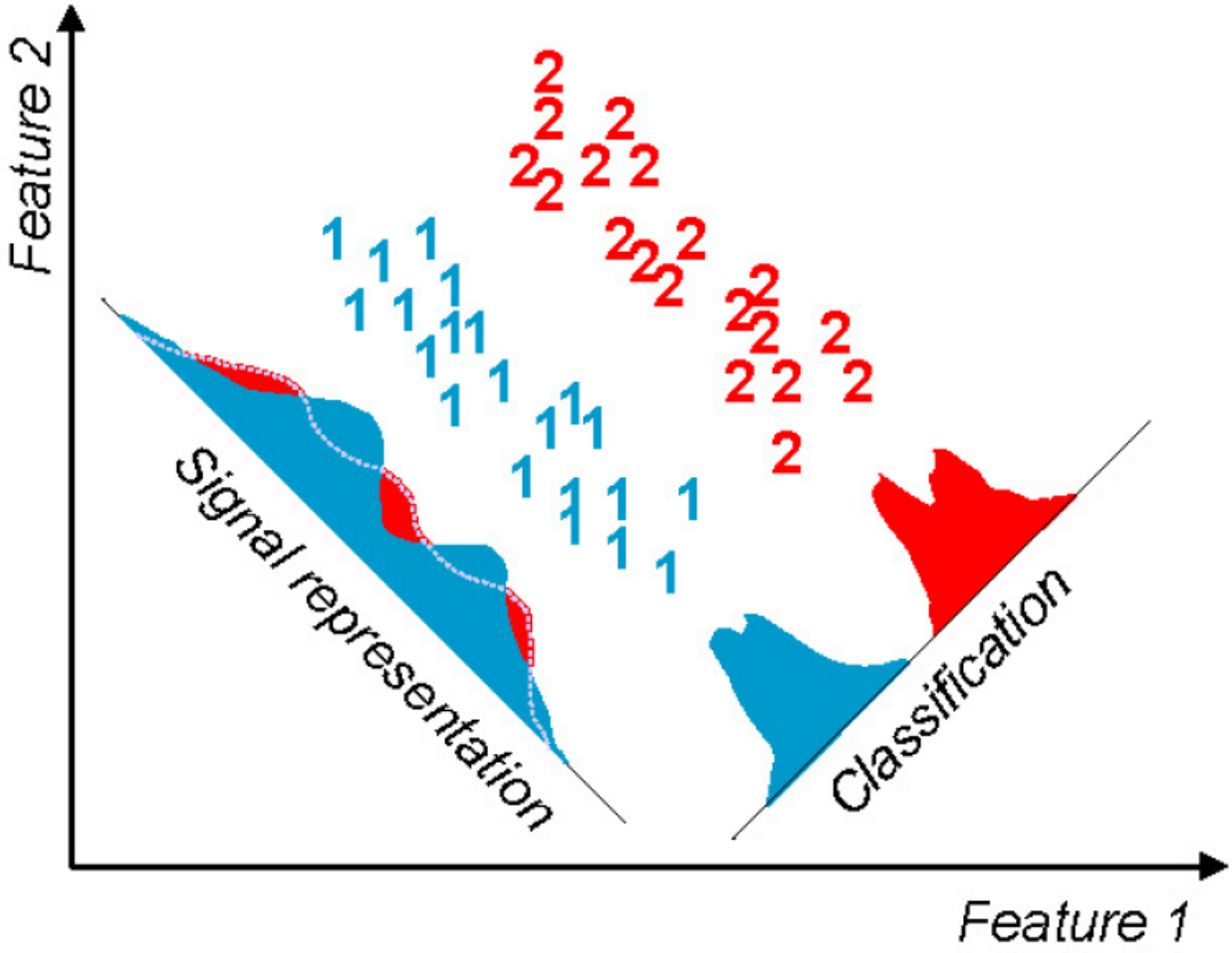
$$\mathbf{S}_W = \sum_{n \in C_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{n \in C_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T$$

两者之商的学名叫作**广义瑞利商**（generalized Rayleigh quotient）。可以求出，使广义瑞利商最大化的解析解为 $\mathbf{w} = \mathbf{S}_{\mathbf{w}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$。

单从形式上看，线性判别分析和主成分分析同属于降维技术，有不少相似的地方，但两者却有本质上的区别。主成分分析的目的是保留不确定性，是通过选择方差最大的主成分来实现信息损失最小的数据低维度重构，整个过程是无监督的。

相比之下， 线性判别分析在降维时要利用数据的类别，因而属于监督学习的范畴，学习的目的则是消除不确定性。消除的不确定性就是类间方差，这部分信息被提取到了类别标签中。分类之后数据的方差越小，意味着降维后剩余的类内不确定性就越小。

在实际应用中，通常可以先使用主成分分析进行特征提取，再利用线性判别分析做训练。这就相当于先把数据的信息集中在某些特征上，再利用不同的类别把这些信息提取出来。



主成分分析（左）与线性判别分析（右）的对比

图片来自 https://zybuluo.com/anboqing/note/117518

将线性模型扩展到分类问题中时，线性判别分析和逻辑回归作为两种具有代表性的模型，都是基于概率生成线性的分类边界，因此有必要做一比较。

线性判别分析就像个傲娇的老师，只愿意指导天赋异禀的学生，这体现在它对数据的要求上：第一，每个类别的数据必须服从潜在的多元正态分布；第二，每个类别的数据必须具有相同或者相近的协方差矩阵；第三，数据的属性之间不能存在较强的共线性，计算出的协方差矩阵应为满秩矩阵；第四，数据中尽可能不存在异常点。

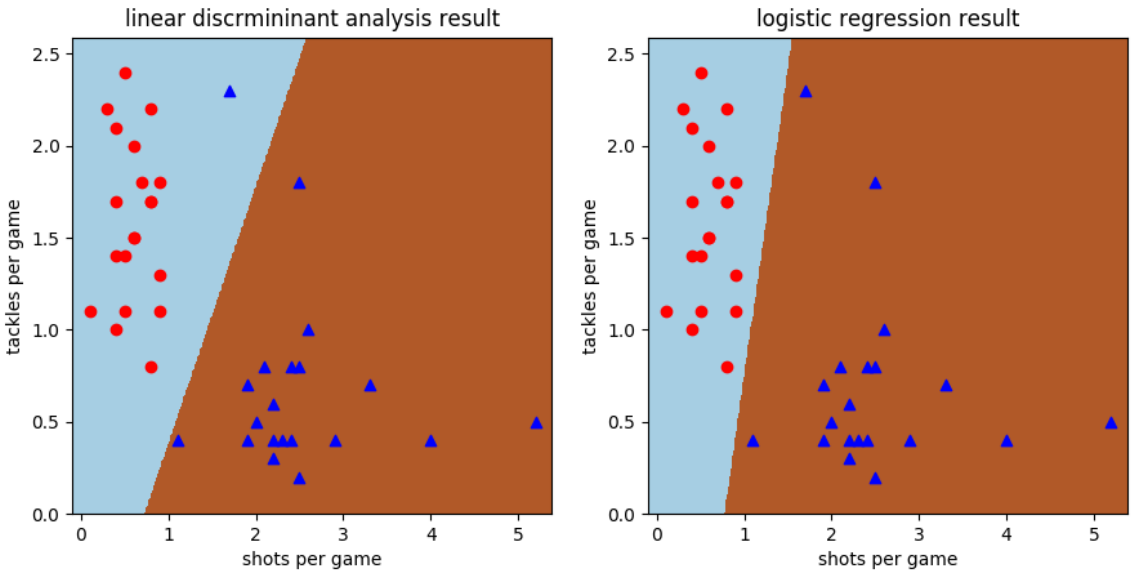
虽然在实际问题中，一定程度上放宽这些条件并不会对线性判别分析的性能产生太大的影响，但这些条件还是严重地限制了方法的应用，使找到一个能解决的问题比解决这个问题更加困难。

相比之下，逻辑回归就没有那么多讲究了，这个老师不管学生好坏都能因材施教。它无需对数据分布做出任何先验假设（两点分布是二分类问题必然的结果），对数据的协方差矩阵和共线性也没有特殊的要求。即使当数据集中存在一些异常点，逻辑回归也能完成精确地分类。整体来说，线性判别分析只能在所有条件都满足时发挥出最佳的性能，在任何其他的场景下都要略逊逻辑回归一筹。

虽然实现的方式有所不同，但本篇所介绍的两种解决分类问题的方法在思想上是一致的，那就是**根据数据的概率密度来实现分类**。这两种基于似然度（likelihood-based）的模型在执行分类任务时不是以每个输入样本为单位，而是以每个输出类别为单位，将每个类别的数据看作不同的整体，并寻找它们之间的分野。这样看来，数据和人一样，也要面临站队的问题啊！

在 Scikit-learn 中，线性判别分析在模块 discriminant_analysis 中实现，逻辑回归则在模块 linear_model 中实现。由于逻辑回归需要使用有标签的数据，因而原来的回归数据就不能使用了。

这次使用的数据依然来自于 WhoScored 的英超技术统计：我选取了 17/18 赛季平均评分最高的 20 名中卫和 20 名中锋，他们的首发次数均在 15 次以上。数据的属性包括每个人的场均射门数和场均铲球数两个维度，位置则作为分类标签出现。熟悉足球的朋友肯定明白，中卫的铲球数较多，而中锋的射门数较多，因此这两个指标可以用来作为判断位置的根据。



线性判别分析和逻辑回归在中卫 - 中锋数据集上的决策边界，红色圆点代表中卫，蓝色三角代表中锋

用上面的数据集训练使用不同的线性分类模型，得到的效果也不相同。这个数据集本身是线性可分的，也就是存在将两个类别完全区分开来的线性边界，这条边界也被逻辑回归正确地计算出来。可遗憾的是，线性判别分析并没有找到准确的边界，而是将一个热爱防守的前锋（斯旺西城 18 号乔丹·阿尤，每场的铲球多过射门，这不禁让人想起著名的防守型前锋德克·库伊特）误认成后卫。

直观地从数据分布的图示看，这个被线性判别分析误分类的蓝点和其他蓝点相距较远，反倒是和红点更加接近，怎么看怎么像是个异常点。在计算数据的统计特性时，这个离群索居的样本远离了归属类的均值，也就难怪会被同伴所抛弃。这也印证了前面的说法：线性判别分析需要较强的假设来支持。

今天我和你分享了使用线性模型解决分类问题的方法，其要点如下：

在解决分类问题时，线性模型的回归值可以通过联系函数转化为分类结果；

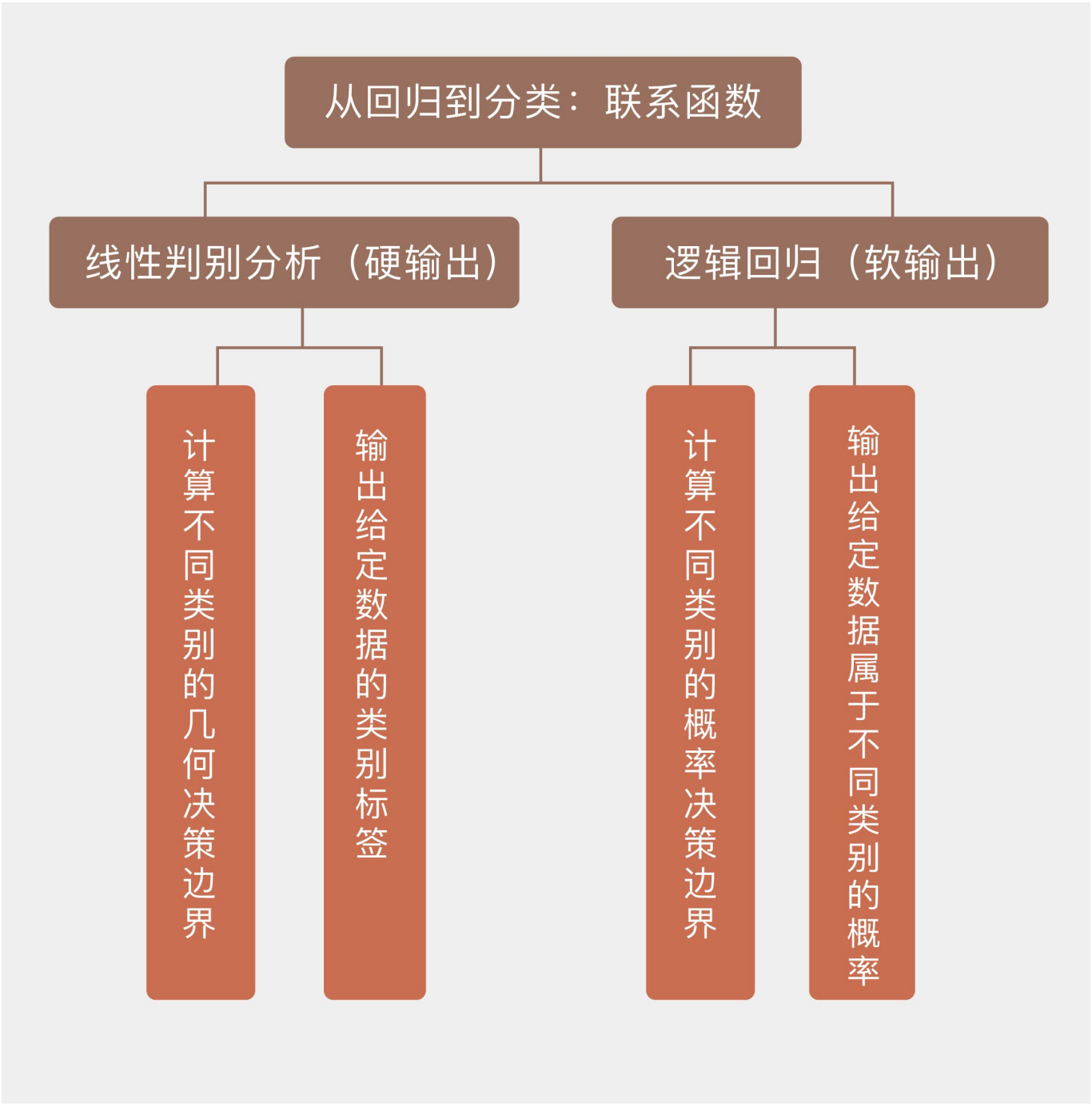
线性判别分析假定数据来自均值不同但方差相同的正态分布，通过最大化类间方差与类内方差的比值计算线性边界；

逻辑回归计算的是不同类别的概率决策边界，输出的是给定数据属于不同类别的后验概率；

基于线性模型的分类方法计算出的决策边界是输入属性的线性函数。

当线性边界不足以完成分类任务时，线性判别分析可以推广为二次判别分析（Quadratic Discriminant Analysis），那么两者之间存在这哪些区别和联系呢？

你可以查阅资料加以了解，并在这里分享你的理解。


从回归到分类：联系函数
线性判别分析（硬输出）
逻辑回归（软输出）
计算不同类别的几何决策边界
输出给定数据的类别标签
计算不同类别的概率决策边界
输出给定数据属于不同类别的概率

上一篇　14 | 非线性降维：流形学习

下一篇　16 | 建模非正态分布：广义线性模型

精选留言 (6)

写留言

林彦

2018-07-09

3

当不同分类样本的协方差矩阵相同时，使用线性判别分析；当不同分类样本的协方差矩阵不同时，则应该使用二次判别分析（Quadratic Discriminant Analysis）。LDA适合均值不同，方差相同的高斯分布，其决策边界是一个平面。QDA适合均值不同，方差也不同的高斯分布。在协方差矩阵相同时，LDA和QDA没有分类结果差异。在不同的协方差矩阵下，LDA和QDA的决策边界存在明显差异。

展开

作者回复: 是的，QDA去掉了对方差相同的限制，这样计算出的似然比，也就是概率密度的比值就不是直线了。

Python
2019-01-22

老师，逻辑回归只适用于带标签的数据的分类任务吗

展开

Python
2019-01-22

x_min,x_max = shots.min() - 0.2,shots.max() +0.2
y_min, y_max = tackles.min() - 0.2, tackles.max() + 0.2

老师为什么要用最小值减去0.2，和最大值加0.2

夏震华(围...
2018-10-08

LDA、QDA： http://www.mamicode.com/info-detail-1819236.html这个比较直观，容易理解

作者回复: 感谢分享

---

paradox
2018-08-11

老师，您好
文中
说LR与LDA是以每个输出类别为单位，将每个类别的数据看作不同的整体，并寻找它们之间的分野。
如何理解呢?

展开

作者回复: 指的是两种模型在分类时利用的都是类别数据整体的统计特性，相比之下，支持向量机使用的支持向量就是每个类别中若干个具有代表性的特例。

鱼大

2018-07-10

干货

展开