线性分类模型（一）：线性判别模型分析

原创 石头 机器学习算法那些事 2018-10-31

前言

前几篇文章介绍了线性回归算法，线性分类模型分为判别式模型和生成式模型，本文首先简单复习了与算法相关的数学基础知识，然后分析各线性判别式分类算法，如最小平方法，Fisher线性判别法和感知器法，最后总结全文。

目录

相关数学知识回顾

1、直线方程和平面方程

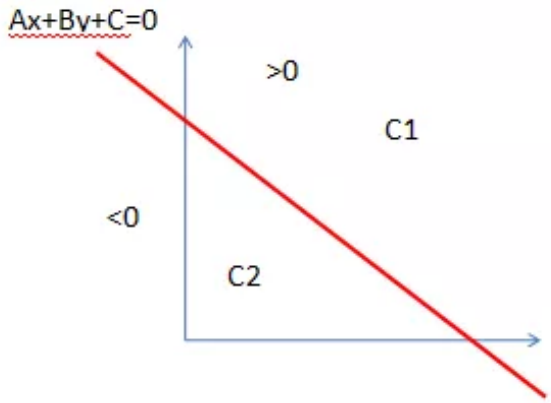
直线方程$l: Ax + By + C = 0$，点P的坐标(x_0, y_0)

(1)A落在直线l上方时，$Ax_0 + By_0 + C > 0$

(2)A落在直线l下方时，$Ax_0 + By_0 + C < 0$

(3)A落在直线l时，$Ax_0 + By_0 + C = 0$

拓展到分类思想：直线l为分类决策方程，坐标点落在直线l上方时，则分类为C1；坐标点落在直线l下方时，则分类为C2（如下图）。



平面方程类似，在这里不展开。

2、点到直线和点到平面的距离

点到直线的距离：

直线l方程：$Ax+By+b=0$，若某一点A的坐标为(x_0,y_0)

求直线l法向量和点A到直线l的距离d；

解：直线l的方向向量$\vec{c}=(-B,A)$,法向量与方向向量垂直相交

$\therefore$法向量$\vec{m}=(A,-B)$

点A到直线l的距离$\mathrm{d}=\left|\frac{Ax_0+By_0+b}{\sqrt{A^2+B^2}}\right|$

点到平面的距离

平面方程$Ax+By+Cz+D=0$，点P的坐标(x_0,y_0,z_0)

求平面的法向量和点P到平面的距离d

如下图：

法向量$\vec{n}=(A,B,C)^T$

$$d=\left|\frac{Ax_0+By_0+Cz_0+D}{\sqrt{A^2+B^2+C^2}}\right|$$

等价于下面两向量的内积：

$\vec{w}=(D,x_0,y_0,z_0)^T$，法向量$\vec{n}=(1,A,B,C)^T$

$$d=\frac{\vec{w}^T*\vec{n}}{\sqrt{A^2+B^2+C^2}}$$

$\because \sqrt{A^2+B^2+C^2}$为常数

$\therefore d\propto \vec{w}^T*\vec{n}$

因此点到直线或点到面的距离可以用向量内积表示

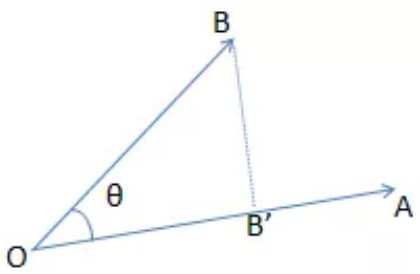
拓展到分类思想：平面方程为决策方程，正确分类的情况下，当点P到决策方程的距离越大，则分类模型越好；错误分类的情况下，点P到决策方程的距离作为损失函数，损失函数最小化过程即是模型参数最优化过程。

3、向量内积的数学意义

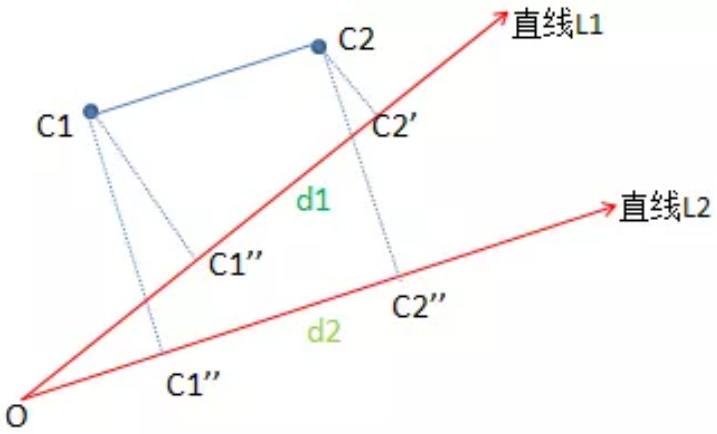
向量A与向量B的内积定义为：

$$A*B=|A||B|\cos\theta,\quad \theta \text{ 是向量}A\text{和向量}B\text{的夹角}$$

几何意义：向量A与向量B的内积等于向量A在向量B的投影与向量B的乘积，当向量B是单位向量时，则等于向量A在单位向量方向的投影，单位向量类似于基函数或者可以理解成坐标轴,即向量A在向量B的投影可理解成向量A在向量B方向的坐标，如下图，B'是B在OA坐标轴方向的投影。



拓展到分类思想：C1与C2属于不同的类，给定一条决策性直线l，当C1与C2在直线L2的投影间距越大，则分类效果越好。增加不同类间的距离可以作为模型参数优化的方向。
如下图，C1和C2的在直线L2的投影距离|C1"C2"|大于|C1'C2'|，因此决策方程直线L2优于直线L1.



4、梯度下降法

梯度的定义如下：

$$gradf(x_0, x_1, \ldots, x_n) = \left(\frac{\partial f}{\partial x_0}, \ldots, \frac{\partial f}{\partial x_j}, \ldots, \frac{\partial f}{\partial x_n}\right)$$

函数f(x0,x1,...,xn)在梯度方向是函数值变化（增加或减少）最快的方向（本文只给出结论，后续文章会有详细的说明）。

拓展到分类思想：损失函数最小化过程即是模型参数最优化过程，损失函数最小化可通过梯度下降法来实现，当迭代到一定程度，损失函数收敛，则迭代结束，参数w即是要求的最优参数。
流程图如下：

假设损失函数$E(w)$，设置初始化参数$w_1 = w_0$和步长 α
求模型的最优参数w
(1) 计算损失函数的梯度，梯度表达式如下：

$$\frac{\partial E(w)}{\partial w}$$

(2)更新参数w，

$$w'_1 = w_1 - \frac{\partial E(w)}{\partial w} * \alpha$$

(3) 重新计算损失函数$E'(w)$，
(4)若$E(w) - E'(w)$小于阈值，则w参数是最优参数；反之进入步骤(1)

判别式模型和生成性模型

我们常把分类问题分成两个阶段：推断阶段和决策阶段，对于输入变量x，分类标记为Ck。推断阶段和决策阶段具体表示为：

推断阶段：估计P(x,Ck)的联合概率分布，对P(x,Ck)归一化，求得后验概率P(Ck|x)。
决策阶段：对于新输入的x，可根据后验概率P(Ck|x)得到分类结果。

判别式模型和生成性模型的区别

判别式模型：简单的学习一个函数，将输入x直接映射为决策，称该函数为判别式函数。
生成式模型：推断阶段确定后验概率分布，决策阶段输出分类结果，生成式模型包含两个阶段。

本文介绍判别式线性分类模型的三种算法。

最小平方法

最小平方法与最小二乘法的算法思想类似，K类判别函数由K个方程决定，

训练集 $\{\vec{x_n}, \vec{t_n}\}, n = 1,2,\cdots N$ ，K类判别函数为 $y_k(\vec{x})$ ，k=1,2,...,K

参数矩阵为 $\vec{W}$ ，目标矩阵 $\vec{T}$ 。

令损失函数 $E(\vec{W})$

$$E(\vec{W}) = \frac{1}{2} Tr\{(\vec{X}\vec{W} - \vec{T})^T (\vec{X}\vec{W} - \vec{T})\}$$

Tr表示求矩阵的迹（这里就不给出详细解释了，若有问题请微我）

最小化误差平方和，即 $\frac{\partial E(\vec{W})}{\partial \vec{W}} = 0$

即可求得最优参数矩阵$\vec{W}$,$\vec{W}$的每一列表示决策函数的最优参数$\vec{w}$

求得最优参数w后，输入变量x所属K类的判别方法如下：

K类判别函数由K个决策方程决定，假设输入$\overrightarrow{x_0}=(x_1,x_2,x_3)^T$

决策方程：$y_k(\vec{x})=\overrightarrow{w_k}^T\vec{x}+w_0$

分类思想：计算点到K个决策方程的距离，选取距离最大对应的决策方程即是输入向量所属类别。

分类算法：由点到平面的距离（可看第一节）可知：

$y_k(\overrightarrow{x_0})=\overrightarrow{w_k}^T\vec{x}+w_0$就是输入到输出的距离

选取最大距离对应的决策方程即是分类结果，$k=(y_k(\overrightarrow{x_0}))_{\max}$

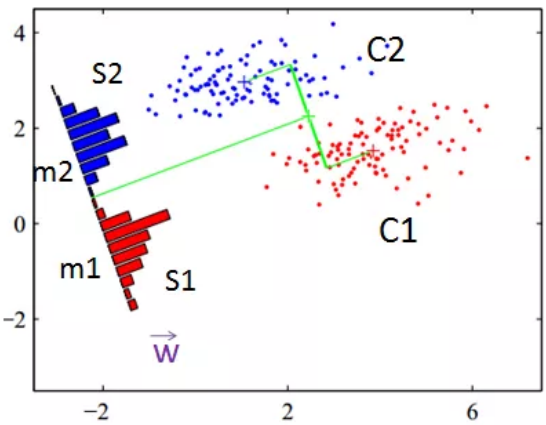
Fisher线性判别函数

第一节讲到，若两个类在同一个决策方程的投影距离相隔越大，则该决策方程越好。再深入一点，相同类投影到决策方程的方差越小，则该决策方程越好，方差代表类投影到决策方程的聚集程度。**这就是Fisher线性判别法参数优化思想。**

参数优化思想：同类样本投影到决策方程的方差最小，不同类样本投影到决策方程的均值间隔最大。用表达式J(w)表示，J（w）越大越好。

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

如下图：



其中，m1，m2分别表示不同类在决策方程的投影均值；S1,S2分别表示不同类投影到决策方程的方差。

$$令\frac{\partial J(\mathrm{w})}{\partial w}=0,\ 即可求得最优参数\mathrm{w}$$

求得最优参数w后，输入变量x所属类的判别方法如下：

设置一阈值y_0，若$\vec{w}^T\vec{x}\geq y_0$,则分类结果为$C2$，反之为$C1$。

感知器算法

感知器算法的目的是找到能够准确分离正负样本训练数据集的超平面。
超平面定义：

$$\vec{w}^T * \vec{x} + b = 0$$

其中，参数$\vec{w}$和b是超平面参数

感知器学习策略：

对训练数据集某一样本点(x,y)，若wx+b>0，则y=1；若wx+b<0，则y=-1；
即感知机模型为：

$$y = sign(wx + b)$$

因此，对于误分类的数据(xi,yi)来说：

$$-y_i(w \cdot x_i + b) > 0$$

成立．因为当 $w \cdot x_i + b > 0$ 时，$y_i = -1$，而当 $w \cdot x_i + b < 0$ 时，$y_i = +1$．因此，误分类点 x_i 到超平面 S 的距离是

$$-\frac{1}{\|w\|} y_i(w \cdot x_i + b)$$

因此，**感知器学习策略是最小化误分类点到平面S的距离**，不考虑分母项。
假设训练数据集有M个误分类点，损失函数为：

$$L(w,b) = -\sum_{x_i \in M} y_i(w \cdot x_i + b)$$

运用梯度下降算法最小化损失函数L(w,b)。

$$\frac{\partial L(w,b)}{\partial w} = -y_i x_i$$
$$\frac{\partial L(w,b)}{\partial w} = -y_i$$

设学习率η，感知器学习策略步骤：

（1）、选取处置w0，b0；
（2）、选取训练集(xi,yi)；
（3）、如果yi(w*xi+b)<=0，则更新权值参数w,b：

$$w \leftarrow w + \eta y_i x_i$$

$$b \leftarrow b + \eta y_i$$

（4）、转至（2），直至训练数据集没有误分类点，得到超平面最优参数w，b。

感知机学习算法由于采用不同的初值或选取不同的误分类点，参数解可能不同（例题可参考《统计学习方法》）。

因此，对某一输入点，若感知机模型大于0，则分类为1；反之分类为-1。

总结

本文介绍了线性判别分类的三种方法，第一种判别方法是根据点到判别函数的距离来分类，，第二种方法是根据输入样本在判别函数的投影距离进行分类，第三种方法则采用感知机模型进行分类。

参考

Christopher M.Bishop <<Pattern Reconition and Machine Learning>>

李航 《统计学习方法》

推荐阅读文章

深入理解线性回归算法（三）：浅谈贝叶斯线性回归

深入理解线性回归算法（二）：正则项的详细分析

深入理解线性回归算法（一）

线性回归：不能忽视的三个问题

浅谈频率学派和贝叶斯学派

浅谈先验分布和后验分布

-END-


长按二维码关注
机器学习算法那些事
微信：beautifulife244
砥砺前行 不忘初心