线性模型与算法

张翼鹏

November 25, 2019

目录

1 基本形式

假设每个样本有k个特征$\boldsymbol{x} = (x_1, x_2, \cdots, x_p)$和一个标签$y$，线性模型试图学习一个由各特征的线性组合组成的函数来找到$\boldsymbol{x}$和y之间的联系并对未知标签的样本进行预测，即

$$f(\boldsymbol{x}) = w_1x_1 + w_2x_2 + \cdots + w_px_p + b$$

向量形式为

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

其中$\boldsymbol{w} = (w_1, w_2, \cdots, w_p)$。$\boldsymbol{w}$和$b$确定后，模型便得以确定。

线性模型形式简单、易于建模，却蕴含着机器学习中一些重要的基本思想。很多功能强大的非线性模型都是在线性模型的基础上改进而得。此外，由于$\boldsymbol{w}$直观地表达了各特征的在模型中的重要程度，因此线性模型拥有很好的可解释性。

2 线性回归

2.1 基础知识

对于数据集$D = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \cdots, (\boldsymbol{x_n}, y_n)\}$，求得一个线性函数使其能较为准确地求得标签值$y_i$。即确定$\boldsymbol{w}$和$b$，建立模型

$$f(\boldsymbol{x_i}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x_i} + b，\text{使得} f(\boldsymbol{x_i}) \approx y_i$$

一般采用最小二乘法进行求解，即均方误差最小化。均方误差具有良好的几何意义，它对应着欧氏距离。最小二乘法即找到一组$\boldsymbol{w}$和b，使得其确定的线性模型满足所有样本点与模型对应点的欧式距离平方和最小。求解过程可以使用梯度下降法或者微积分法（正规方程法）。

对于一元线性回归，即

$$(w^*, b^*) = \arg\min_{(w,b)} \sum_{i=1}^{n} (y_i - wx_i - b)^2$$

解得

$$w^* = \frac{\sum_{i=1}^{n} y_i(x_i - \overline{x})}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} \qquad b^* = \frac{1}{n}\sum_{i=1}^{n}(y_i - wx_i)$$

对于多元线性回归，为方便运算，令$\widehat{\boldsymbol{w}} = (\boldsymbol{w}; b)$；同时给每个样本增加一个特征，值固定为1。将数据集D表示成一个$n * (p+1)$ 阶矩阵$\boldsymbol{X}$和一个n维列向量$\boldsymbol{y}$，即

$$\widehat{\boldsymbol{w}}^* = \arg\min_{\widehat{\boldsymbol{w}}} \sum_{i=1}^{n} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{w}})$$

解得

$$\widehat{\boldsymbol{w}} = (\boldsymbol{X}^{\mathbf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathbf{T}}y$$

2.2 广义线性回归

线性回归的思想同样可以用在很多非线性回归上，例如多项式回归、指数回归。

以平面上的数据集$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$为例，我们想用二次函数拟合这些数据，构造模型

$$f(x) = w_1x^2 + w_2x + w_3$$

我们可以对数据集中所有数据点，计算$x_i^2(i = 1, 2, \cdots, n)$，然后按照线性回归的方法求出$\boldsymbol{w}$。即我们为只有一个特征$x$的数据集额外添加了一个特征$x^2$，将一个一元非线性回归转化成了多元线性回归。

另一方面，我们不仅可以令模型逼近y，还可以使其逼近y的函数。比如我们可以以y的对数作为逼近目标，即

$$\ln y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

该模型被称为“对数线性回归”，也可以写成如下形式：

$$y = e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}$$

该模型形式上依然是线性回归，但实际是在求从输入空间到输出空间的非线性函数映射。

更一般地，考虑单调可微函数$g(\cdot)$，将

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

称为“广义线性模型”，其中函数$g(\cdot)$被称为“联系函数”。

2.3 正则化

在线性回归模型代价函数（均方误差）中加入正则化项（一般分为$L1$正则化和$L2$正则化），可以一定程度避免多重共线性或过拟合等问题。

$L1$正则化项为

$$\lambda \sum_{i=1}^{p} |w_i| \quad \lambda > 0$$

使用$L1$正则化的线性回归模型称为$Lasso$回归。$L1$正则化通过稀疏参数来降低模型的复杂度，可以使部分参数快速收缩为0，具有较快的求解速度。

$L2$正则化项为

$$\lambda \sum_{i=1}^{p} w_i^2 \quad \lambda > 0$$

使用$L2$正则化的线性回归模型称为岭回归。$L2$正则化通过整体缩小参数来防止过拟合，具有较高的准确性和鲁棒性。

3 线性判别分析LDA

LDA是一种有监督的降维方法，目标是使相同类别的数据分布更紧凑，不同类别的数据尽量相互远离，让映射后的样本有最好的分类性能。

以二分类样本为例，首先计算类间散度矩阵S_b：

$$S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T$$

其中μ_0是第0类样本的均值，μ_1是第1类样本的均值。

然后计算类内散列矩阵S_w：

$$S_w = \sum_{x \in X_0} (x - \mu_0)(x - \mu_1)^T + \sum_{x \in X_1} (x - \mu_0)(x - \mu_1)^T$$

其中X_0是第0类样本的集合，X_1是第1类样本的集合。

最后${S_w}^{-1} S_b$的最大特征值所对应的特征向量w即为最佳投影方向（这是降到一维的情况，若要降到k维就取前k个最大特征值对应的特征向量）。

（此段向量加粗表示，之前某些向量未加粗，请注意辨别矢量标量）向量$\boldsymbol{w}$即为我们要找的投影方向（可以表示一条直线），$y = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}$ 即为我们要找的特征映射（注意这个关系式并不是我们要找的直线方程），y表示每个样本点在$\boldsymbol{w}$方向上投影后的点在一维数轴上的坐标，即每个样本点降维后的坐标。

4 logistic回归

思想：利用sigmoid函数，将一般的线性回归模型转换成一种概率函数，进而判断未知样本属于某一类的概率是多少。

步骤：1、目标函数为：

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

目标函数计算的是一个样本在第0类和第1类中属于第1类的概率。

2、使用交叉熵代价函数：

$$J(\theta) = -\frac{1}{n} \sum_{i=1}^{n} [(y^{(i)}) \ln(h_\theta(x^{(i)})) + (1 - y^{(i)}) \ln(1 - h_\theta(x^{(i)}))]$$

3、使用梯度下降法更新参数θ：

$$\theta'_j = \theta_j - \frac{1}{n}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

对于多元分类问题，我们有以下两种方式对上述二元分类模型进行改进。

（1）对于有k个类别的分类问题，首先将第一类看成正样本，其他所有类看成负样本，计算目标函数值计作p_1；然后将第二类看作正样本，其他所有类看作负样本，计算p_2；以此类推，最后若$p_j = \max\limits_{1\le i\le k} p_i$，则将未知样本归为第$j$类。

（2）利用softmax函数。目标函数为：

$$h_w(x) = \frac{1}{\sum_{i=1}^{k} e^{w_i x + b_i}} \begin{bmatrix} e^{w_1\cdot x+b_1} \\ e^{w_2\cdot x+b_2} \\ \vdots \\ e^{w_k\cdot x+b_k} \end{bmatrix}$$

其中k为类别的个数，w_i和b_i为第i个类别对应的权重向量和偏置标量。

我们可以将目标函数求出的第$j(1 \le j \le k)$个分量看成是样本x属于第j类的概率，即：

$$\frac{e^{w_j\cdot x+b_j}}{\sum_{i=1}^{k} e^{w_i\cdot x+b_i}} = P(y_j = 1|x)$$

其中y_j为样本x的期望输出的第j个分量，若该样本属于第j类，则该值为1，否则为0。

使用对数似然代价函数：

$$J(w,b) = -\frac{1}{n}\sum_{i=i}^{n}\sum_{j=1}^{k} y_j^{(i)} \ln\left(\frac{e^{w_j\cdot x^{(i)}+b_j}}{\sum_{l=1}^{k} e^{w_l\cdot x^{(i)}+b_l}}\right)$$

其中$y_j^{(i)}$为第i个样本的期望输出的第j个分量，若该样本属于第j类，则该值为1，否则为0。

利用梯度下降法更新w：

$$w'_r = w_r + \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - P(y_r^{(i)} = 1|x^{(i)}) * x^{(i)})$$

优点：自变量可以是离散型也可以是连续型；借助概率的概念，易于理解；计算代价不高。

缺点：当特征空间很大时，性能不是很好；容易欠拟合，准确度不太高。