第二章 线性模型

Castor Ye

1 基本形式

给定由 d 个属性描述的示例 $x = (x_1; x_2; \cdots; x_d)$，其中 x_i 是 x 在第 i 个属性上的取值，线性模型（linear model）试图学得一个通过属性的线性组合来进行预测的函数，即：

$$f(x) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b$$

一般用向量形式写成：

$$f(x) = w^T x + b$$

其中 $w = (w_1; w_2; \cdots; w_d)$，$w$ 和 b 学得之后，模型就得以确定。

2 线性回归

给定数据集 $D = \{(x_1, y_2), (x_2, y_2), \cdots, (x_m, y_m)\}, x_i = (x_{i1}; x_{i2}; \cdots; x_{id}), y_i \in \mathbb{R}$。"线性回归"（linear regression）试图学得一个线性模型以尽可能准确地预测实值输出标记。例如：根据历年城市 GDP 预测未来 GDP，或者根据历年天气数据预测今年农作物收成等。

对于连续值的属性，一般都可以直接或经过预处理（归一化等）后被学习器所用。但对于离散值，我们可以做以下处理：

i. 若属性间存在"序"（order）关系，可通过连续化将其转化为连续值。例如：身高属性分为"高""中""矮"，可转化为数值：$\{1, 0.5, 0\}$。

ii. 若属性间不存在"序"（order）关系，则通常将其转化为向量的形式。例如：性别属性分为"男""女"，可转化为二维向量：$(1, 0), (0, 1)$。

线性回归试图学得：

$$f(x_i) = wx_i + b,\text{ 使得} f(x_i) \simeq y_i$$

为了确定 w 和 b，我们可以引入均方误差：

$$\begin{aligned}(w^*, b^*) &= \underset{(w,b)}{\arg\min} \sum_{i=1}^{m} \left(f(x_i) - y_i\right)^2 \\ &= \underset{(w,b)}{\arg\min} \sum_{i=1}^{m} \left(y_i - wx_i - b\right)^2\end{aligned}$$

均方误差有非常好的几何意义，它对应了常用的“欧几里得距离（两点直线距离）”（Euclidean distance）。基于均方误差最小化来进行模型求解的方法称为“最小二乘法”（least square method）。在线性回归中，最小二乘法就是试图找到一条直线，使所有样本到直线上的欧氏距离之和最小。

求解 w 和 b 使 $E_{(w,b)} = \sum\limits_{i=1}^{m}(y_i - wx_i - b)^2$ 最小化的过程，称为线性回归模型的最小二乘“参数估计”（parameter estimation）。我们可将 $E_{(w,b)}$ 分别对 w 和 b 求导，得到：

$$\frac{\partial E_{(w,b)}}{\partial w} = 2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}(y_i - b)x_i\right)$$

$$\frac{\partial E_{(w,b)}}{\partial b} = 2\left(mb - \sum_{i=1}^{m}(y_i - wx_i)\right)$$

令上面两式为零可得到 w 和 b 最优解的闭式（closed-form）解：

$$w = \frac{\sum\limits_{i=1}^{m} y_i\left(x_i - \bar{x}\right)}{\sum\limits_{i=1}^{m} x_i^2 - \frac{1}{m}\left(\sum\limits_{i=1}^{m} x_i\right)^2} \quad b = \frac{1}{m}\sum_{i=1}^{m}(y_i - wx_i)$$

其中 $\bar{x} = \frac{1}{m}\sum\limits_{i=1}^{m} x_i$，为 x 的均值。

注意：

i. 这里 $E_{(w,b)}$ 是关于 w 和 b 的凸函数，当它关于 w 和 b 的导数均为零时，得到 w 和 b 的最优解。

ii. 对区间 $[a, b]$ 上定义的函数 f，若该函数对区间中任意两点 x_1, x_2 均有 $f(\frac{x_1 + x_2}{2}) \leq \frac{f(x_1) + f(x_2)}{2}$，则称 f 为区间 $[a, b]$ 上的凸函数。

iii. 对实数集上的函数，可以通过求二阶导数来判别：若二阶导数在区间上非负，则称为凸函数；若二阶导在区间上恒大于零，则为严格凸函数。

更一般的情形是如本节开头的数据集 D，样本由 d 个属性描述，此时我们试图学得：

$$f(x_i)=w^Tx_i+b_i,\quad \text{使得} f(x_i)\simeq y_i$$

这称为“多元线性回归”（multivariate linear regression）。

类似的，可利用最小二乘法来对 w 和 b 进行估计。为便于讨论，我们把 w 和 b 吸收入向量形式 $\hat{w}=(w;b)$，相应的，把数据集 D 表示为一个 $m\times(d+1)$ 大小的矩阵 X，其中每行对应于一个示例，该行前 d 个元素对应于示例的 d 个属性值，最后一个元素恒置为 1，即：

$$\hat{w}=(w;b)=\begin{bmatrix} w_1 & w_2 & \cdots & w_d & b\end{bmatrix}^T$$

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1\\ x_{21} & x_{22} & \cdots & x_{2d} & 1\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ x_{m1} & x_{m2} & \cdots & x_{md} & 1\end{bmatrix}=\begin{bmatrix} x_1^T & 1\\ x_2^T & 1\\ \vdots & \vdots\\ x_m^T & 1\end{bmatrix}$$

$$X\cdot\hat{w}=\begin{bmatrix} w_1x_{11}+w_2x_{12}+\cdots+w_dx_{1d}+b\\ w_1x_{21}+w_2x_{22}+\cdots+w_dx_{2d}+b\\ \cdots\\ w_1x_{m1}+w_2x_{m2}+\cdots+w_dx_{md}+b\end{bmatrix}=\begin{bmatrix} f\left(x_1\right)\\ f\left(x_2\right)\\ \cdots\\ f\left(x_m\right)\end{bmatrix}$$

再把标记也写成向量形式 $y=(y_1;y_2;\cdots;y_m)$，则有：

$$\hat{w}^*=\underset{\hat{w}}{\arg\min}\left(y-X\hat{w}\right)^T\left(y-X\hat{w}\right)$$

令 $E_{\hat{w}}=\left(y-X\hat{w}\right)^T\left(y-X\hat{w}\right)$，对 $\hat{w}$ 求导得到：

$$\frac{\partial E_{\hat{w}}}{\partial\hat{w}}=2X^T(X\hat{w}-y)$$

令上式为零可得 $\hat{w}$ 最优解的闭式解，但由于涉及矩阵逆的计算，比单变量情形要复杂一些，下面做一个简单讨论：

当 X^TX 为满秩矩阵（full-rank matrix）或正定矩阵（positive definite matrix）时，即矩阵行列式不为零时，令上式为零得到：

$$\hat{w}^*=(X^TX)^{-1}X^Ty$$

其中 $(X^TX)^{-1}$ 是矩阵 (X^TX) 的逆矩阵。令 $\hat{x}_i=(x_i,1)$，则最终学得的多元线性回归模型为：

$$f(\hat{x}_i)=\hat{x}_i^T(X^TX)^{-1}X^Ty$$

对于非满秩矩阵，我们不进行深入。

另一方面，有时候原始的线性回归并不能满足需求。例如：y 并不是线性变化，而是指数变化。此时我们可以采用线性模型来逼近 y 的衍生物，例如 $\ln y$，如下图所示。这就是“对数线性回归”（log-linear regression），它实际上是在试图让 $e^{w^T x+b}$ 逼近 y。

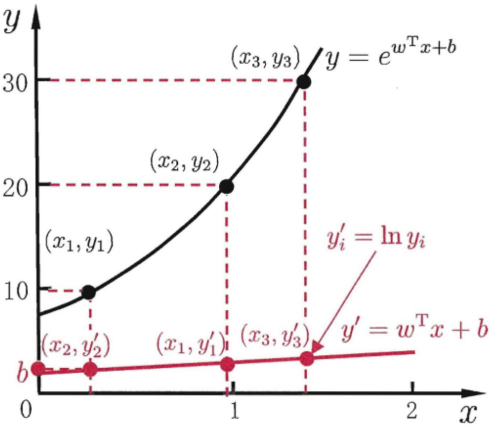

图 1: 对数线性回归示意图