线性模型

给定包含d个属性的样本$\mathbf{x}=(x_1;x_2;\dots x_d)$，线性模型学习一个【通过各个属性的线性组合来进行预测】的函数，即

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

学习到$\mathbf{w}$和b之后，即可确定一个线性模型。线性模型的优势在于参数$\mathbf{w}$直观地表达了样本的各个属性对最终预测结果的重要性（正负，大小），因此线性模型具有很好的可解释性。

1. 单变量线性回归

线性回归模型试图根据数据集，学习一个【预测实数值输出】的线性模型。先考虑样本中只有一个属性，即$d=1$的情况，用(x_i,y_i)表示一个样本，则线性回归试图学习参数w和b，使得

$$f(x_i)=wx_i+b$$

尽可能与y_i接近。衡量接近的方法是计算$f(x)$与y的均方误差（MSE），即：

$$\sum_{i=1}^{N}(y_i-f(x_i))^2$$

令MSE最小化，可得：

$$\begin{aligned}\frac{\partial\sum_{i=1}^{N}(y_i-wx_i-b)^2}{\partial w}=&\frac{\partial\sum_{i=1}^{N}(w^2x_i^2-2wx_i(y_i-b)+(y_i-b)^2)}{\partial w}\\=&\frac{\partial w^2\sum_{i=1}^{N}x_i^2}{\partial w}-2\sum_{i=1}^{N}x_i(y_i-b)\\=&2w\sum_{i=1}^{N}x_i^2-2\sum_{i=1}^{N}x_iy_i+2b\sum_{i=1}^{N}x_i\end{aligned}$$

$$\begin{aligned}\frac{\partial\sum_{i=1}^{N}(y_i-wx_i-b)^2}{\partial b}=&\frac{\partial\sum_{i=1}^{N}(w^2x_i^2-2wx_i(y_i-b)+(y_i-b)^2)}{\partial b}\\=&\frac{\partial 2b\sum_{i=1}^{N}wx_i-2b\sum_{i=1}^{N}y_i+Nb^2}{\partial b}\\=&2Nb+2\sum_{i=1}^{N}wx_i-2\sum_{i=1}^{N}y_i\end{aligned}$$

令对b的导数为0，可得：

$$b^*=\frac{1}{N}\sum_{i=1}^{N}(y_i-wx_i)=\bar{y}-w\bar{x}$$

将结果代入对w的导数，并令其为0，可得：

$$w\sum_{i=1}^{N}x_i^2-\sum_{i=1}^{N}x_iy_i+\bar{y}\sum_{i=1}^{N}x_i-w\bar{x}\sum_{i=1}^{N}x_i=0$$

求解可得：

$$
\begin{aligned}
w^*(\sum_{i=1}^{N}x_i^2-\bar{x}\sum_{i=1}^{N}x_i)&=\sum_{i=1}^{N}x_iy_i-\bar{y}\sum_{i=1}^{N}x_i\\
w^*&=\frac{\sum_{i=1}^{N}x_iy_i-\bar{y}\sum_{i=1}^{N}x_i}{\sum_{i=1}^{N}x_i^2-\bar{x}\sum_{i=1}^{N}x_i}\\
&=\frac{\sum_{i=1}^{N}x_iy_i-N\bar{x}\bar{y}}{\sum_{i=1}^{N}x_i^2-2N\bar{x}^2+N\bar{x}^2}\\
&=\frac{\sum_{i=1}^{N}x_iy_i-\sum_{i=1}^{N}x_i\bar{y}-\sum_{i=1}^{N}\bar{x}y_i+\sum_{i=1}^{N}\bar{x}\bar{y}}{\sum_{i=1}^{N}x_i^2-2\bar{x}\sum_{i=1}^{N}x_i+\sum_{i=1}^{N}\bar{x}^2}\\
&=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2}
\end{aligned}
$$

这是样本中只有一个属性的情况下的线性回归。

2.多变量线性回归

当样本包含多个属性时，此时，我们试图学习到参数$\mathbf{w}$和b，其中$\mathbf{w}$是维度等于属性数目的列向量，b是标量，使得

$$
f(\mathbf{x}_i)=\mathbf{w^T}\mathbf{x}_i+b
$$

与y_i尽可能接近。为简化表述，将b吸收到$\mathbf{w}$中，即$\mathbf{w}\leftarrow(\mathbf{w};b)$，数据集表示为一个行数等于样本数，列数为属性数目+1的矩阵$\mathbf{X}$，label写成向量形式$\mathbf{y}$，则有：

$$
\mathbf{y}=\mathbf{X}\mathbf{w}
$$

均方误差为$(\mathbf{y}-\mathbf{X}\mathbf{w})^\mathbf{T}(\mathbf{y}-\mathbf{X}\mathbf{w})$，对$\mathbf{w}$求偏导数，有：

$$
\begin{aligned}
\frac{\partial(\mathbf{y}-\mathbf{X}\mathbf{w})^\mathbf{T}(\mathbf{y}-\mathbf{X}\mathbf{w})}{\partial\mathbf{w}}&=\frac{\partial((\mathbf{y}-\mathbf{X}\mathbf{w})^\mathbf{T}\mathbf{y}-(\mathbf{y}-\mathbf{X}\mathbf{w})^\mathbf{T}\mathbf{X}\mathbf{w})}{\partial\mathbf{w}}\\
&=\frac{\partial(\mathbf{y^T}\mathbf{y}-\mathbf{w^T}\mathbf{X^T}\mathbf{y}-\mathbf{y^T}\mathbf{X}\mathbf{w}+\mathbf{w^T}\mathbf{X^T}\mathbf{X}\mathbf{w})}{\partial\mathbf{w}}
\end{aligned}
$$

由于$\mathbf{w^T}\mathbf{X^T}\mathbf{y}=(\mathbf{y^T}\mathbf{X}\mathbf{w})^\mathbf{T}$，且二者均为标量，故$\mathbf{w^T}\mathbf{X^T}\mathbf{y}=\mathbf{y^T}\mathbf{X}\mathbf{w}$。上式可变形为：

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

若$\mathbf{X^T}\mathbf{X}$为满秩矩阵或正定矩阵，令上式为0，可得：

$$
\begin{aligned}
\mathbf{X^T}\mathbf{X}\mathbf{w}&=\mathbf{X^T}\mathbf{y}\\
\mathbf{w}&=(\mathbf{X^T}\mathbf{X})^{-1}\mathbf{X^T}\mathbf{y}
\end{aligned}
$$

若$\mathbf{X^T}\mathbf{X}$不满秩，则由多个能使均方误差最小的$\mathbf{w}$值，此时需要引入正则化项，常用的正则化项包括Ridge和Lasso两种。

Ridge正则化在均方误差的基础上增加一个正则化项$\lambda\mathbf{w^T}\mathbf{w}$，此时，误差函数对$\mathbf{w}$的梯度变为:

$$
-\mathbf{2}\mathbf{X^T}\mathbf{y}+\mathbf{2}\mathbf{X^T}\mathbf{X}\mathbf{w}+\mathbf{2}\lambda\mathbf{w}
$$

令梯度为0，可得：

$$\mathbf{w}=(\lambda\mathbf{I}+\mathbf{X^T}\mathbf{X})^{-1}\mathbf{X^T}\mathbf{y}$$

其中，$\mathbf{I}$是形状与$\mathbf{X^T}\mathbf{X}$相同的单位矩阵。

Lasso正则化在均方误差的基础上增加一个正则化项$\lambda\sum_j|w_j|$。当λ足够大时，Lasso正则化会导致向量$\mathbf{w}$中某些维度的取值为0，即产生稀疏的向量。这是因为Lasso正则化约束了向量$\mathbf{w}$中各维度取值的绝对值之和，从PRML146页的图中，可以观察到误差函数的等高线图将与$\mathbf{w}$的取值相交于坐标轴上。由于Lasso正则化加入的正则化项不可导，无法直接求解或使用梯度下降等方法进行优化，一般可以使用坐标轴下降法进行优化，即在每一轮迭代中，在当前点处固定其它坐标轴方向，沿着某一个坐标轴方向进行一维搜索，当所有的坐标轴上都达到收敛时，向量$\mathbf{w}$的取值即为所需结果。

3. Logistic回归

Logistic回归是一种用于二分类任务的线性模型。二分类任务中，输出$y\in\{0,1\}$，因此，需要将线性模型$\mathbf{w^T}\mathbf{x}+b$的输出转换成0或1。Logistic回归通过sigmoid函数：

$$\sigma(z)=\frac{1}{1+e^{-z}}$$

实现实数值到(0,1)区间的转换。 即：

$$\mathbf{y}=\frac{1}{1+e^{-\mathbf{w^T}\mathbf{x}+b}}$$

如果令$\mathbf{y}$表示样本$\mathbf{x}$作为正样本的可能性，则有：

$$\mathbf{w^T}\mathbf{x}+b=ln\frac{y}{1-y}$$

即通过线性模型来拟合样本$\mathbf{x}$作为正样本的相对可能性。

求解$\mathbf{w}$和b的取值可以使用最大似然法完成。给定训练数据集$\{x_i,y_i\}$，其对数似然为$\sum lnp(y_i|x_i,\mathbf{w},b)$。如果用$\beta=(\mathbf{w},b),\hat{x}=(\mathbf{x},1),\sigma=\frac{1}{1+e^{-\beta^T\hat{x}}}$，则有

$$\begin{aligned}p(y_i|x_i,\mathbf{w},b)=&\sigma^{y_i}(1-\sigma)^{1-y_i}\\ ln(\sigma^{y_i}(1-\sigma)^{1-y_i})=&y_iln(\sigma)+(1-y_i)ln(1-\sigma)\end{aligned}$$

$$\begin{aligned}NLL=&-\sum ln(\sigma^{y_i}(1-\sigma)^{1-y_i})\\ \frac{\partial NLL}{\partial\beta}=&\sum\{y_i*\frac{1}{\sigma}*\sigma(1-\sigma)\hat{x}_i+(1-y_i)*\frac{1}{1-\sigma}*-\sigma(1-\sigma)\hat{x}_i\}\\ =&-\sum\{y_i(1-\sigma)\hat{x}_i-(1-y_i)\sigma\hat{x}_i\}\\ =&-\sum(y_i\hat{x}_i-\sigma y_i\hat{x}_i-\sigma\hat{x}_i+\sigma y_i\hat{x}_i)\\ =&\sum\{(\sigma-y_i)\hat{x}_i\}\end{aligned}$$

此处NLL指Negative Log-Likelihood。因此，可以据此构造一个逐步更新β的算法，在每一步，均利用上式的梯度更新β的取值。

4. 线性判别分析

线性分类模型可以被看成是一种降维。以二分类为例，假设输入向量$\mathbf{x}$是一个d维向量，分类模型将通过$y = \mathbf{w^T x}$其降维到1维，并在y上设置一个阈值，当y大于这个阈值时判为正类，否则判为负类。将高位向量降维到1为必然会带来一定的信息损失，如在高维空间中可分的两类数据点降维到1维后，有可能出现相互重叠的情况。因此，我们需要选择合适的$\mathbf{w}$值，使得两类数据点被降维后最大程度地可分。设C_1表示第一类数据点，共有N_1个，C_2表示第二类数据点，共有N_2个，两类数据点的均值用$\mathbf{m}_1$和$\mathbf{m}_2$表示。

一种最简单的度量被降维后的数据的可分情况的方法是最大化其均值被降维后的距离，即：

$$m_2 - m_1 = \mathbf{w^T(m_2 - m_1)}$$

由于增大$\mathbf{w}$的取值，可以让上式变得任意大，因此需要将$\mathbf{w}$约束成单位向量。使用拉格朗日乘子法可得$\mathbf{w} \propto \mathbf{(m_2 - m_1)}$。

但是，当数据点的协方差矩阵的对角性很差时，按上述$\mathbf{w}$降维后，两类数据仍可能有明显的重叠（图见PRML188页）。因此，除了最大化均值被降维后的距离，还需要让类内各数据点被降维后的结果尽量接近，从而减少重叠部分。

对于第k类数据，映射后的类内方差可以表示为：

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

其中$y_n = \mathbf{w^T x_n}$。因此，对于二分类问题，我们需要最大化：

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

分子：

$$\begin{aligned}(m_2 - m_1)^2 =& \mathbf{(w^T(m_2 - m_1))^2} \\ =& \mathbf{w^T(m_2 - m_1)(m_2 - m_1)^T w} \\ =& \mathbf{w^T S_B w}\end{aligned}$$

其中，$\mathbf{S_B = (m_2 - m_1)(m_2 - m_1)^T}$

分母：

$$\begin{aligned}s_1^2 + s_2^2 =& \mathbf{\sum(w^T(x - m_1))^2 + \sum(w^T(x - m_2))^2} \\ =& \mathbf{\sum w^T(x - m_1)(x - m_1)^T w + \sum w^T(x - m_2)(x - m_2)^T w} \\ =& \mathbf{w^T S_W w}\end{aligned}$$

其中，$\mathbf{S_w = \sum(x - m_1)(x - m_1)^T + \sum(x - m_2)(x - m_2)^T}$。

令$J(\mathbf{w})$对$\mathbf{w}$的导数为0，可得：

$$\mathbf{(w^T S_B w) S_w w = (w^T S_W w) S_B w}$$

由于$\mathbf{S_B w}$方向始终与$\mathbf{(m_2 - m_1)}$相同，令$\mathbf{S_B w} = \lambda\mathbf{(m_2 - m_1)}$，两边同时乘以$\mathbf{S_W^{-1}}$可得:

$$\mathbf{w \propto S_W^{-1}(m_2 - m_1)}$$

如果类内方差是各向同性的，即$\mathbf{S_W}$为单位矩阵，那么$\mathbf{w}$就正比于两类均值向量之差。

注意，这不能直接求解出分类结果，而只能确定一个将高维样本映射到一维数值的方案，还需要结合其它算法来求出阈值。

对于多分类情况：设有k个分类，对于输入的列向量$\mathbf{x}$，可以使用一个k列的矩阵$\mathbf{W}$计算对应的向量$\mathbf{y}$：

$$\mathbf{y} = \mathbf{W^T x}$$

同二分类情况，用$\mathbf{S_k}$表示第k类的类内协方差，则总类内协方差为：

$$\mathbf{S_W} = \sum \mathbf{S_k}$$

样本总协方差为:

$$\mathbf{S_T} = \sum (\mathbf{x_n} - \mathbf{m})(\mathbf{x_n} - \mathbf{m})^\mathbf{T}$$

其中，$\mathbf{m}$是所有样本的均值向量。总协方差可以分解成类内协方差和类间协方差，即:

$$\mathbf{S_T} = \mathbf{S_W} + \mathbf{S_B}$$

因此有：

$$\mathbf{S_B} = \sum_\mathbf{k} \mathbf{N_k} (\mathbf{m_k} - \mathbf{m})(\mathbf{m_k} - \mathbf{m})^\mathbf{T}$$

其中$\mathbf{N_k}$是第k类的样本数。

我们需要构造一个标量，使得在类间协方差较大，同时类内协方差较小的时候，这个标量的值较大。一种构造方法是：

$$J(\mathbf{w}) = Tr\{(\mathbf{W S_W W^T})^{-1}(\mathbf{W S_B W^T})\}$$

结论表明，此时权重向量可以由矩阵$\mathbf{S_W^{-1} S_B}$的最大的D个特征值对应的特征向量确定。