线性模型

二维线性回归

$f(x_i) = wx_i + b$

(xi, yi)

均方误差

$$E(w,b) = \sum_{i=1}^{m}(f(x_i) - y_i)^2$$

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

$$= \sum_{i=1}^{m}\left((y_i - wxi)^2 - 2b(y_i - wx_i) + b^2\right)$$

$$(w^*, b^*) = arg_{w,b}min(E(w,b))$$

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

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

易证明$E(w,b)$是凸函数, 所以当上面两式都为0时，$E(w,b)$取到最小值
解上面两式得:

$$w = \frac{\sum y_i(x_i - \frac{1}{m}\sum x_j)}{\sum x_i^2 - \frac{1}{m}(\sum x_i)^2}$$

$$b = \frac{1}{m}\sum(y_i - wx_i)$$

多元线性回归

当变量的预测与d个属性有关系时的回归过程叫做多元线性回归:

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

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{bmatrix}$$

$$\hat{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \\ b \end{bmatrix}$$

考虑m个样本，可以有如下矩阵:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix}$$

简记为:

$$X = \begin{pmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \vdots \\ x_m^T & 1 \end{pmatrix}$$

所以根据$f(x_i)$函数得到的预测值为: $X\hat{w}$
那么损失函数为$(y - X\hat{w})^T(y - X\hat{w})$
得到的最佳解应为:

$$\hat{w}^* = arg_{\hat{w}} min(y - X\hat{w})^T(y - X\hat{w})$$

对损失函数展开

$$Los(\hat{w}) = y^Ty - \hat{w}^TX^Ty - y^TX\hat{w} + \hat{w}X^TX\hat{w}$$

参考矩阵求导公式:

$$\frac{\partial Los}{\partial \hat{w}} = -2X^Ty + 2X^TX\hat{w} \tag{1}$$

令上式为0, 若X^TX为满秩矩阵,则存在逆矩阵

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

若将(1)式为0视为一个方程组,则w的系数矩阵$A = X^TX$在$(d+1) > m$时一定为不满秩的矩阵(其中d为维度数, m为样本数)
证明如下: X是$m \times (d+1)$的矩阵,所以当A满秩时其秩应为$(d+1)$

$$Rank(X^TX) \le Rank(X) = m < (d+1)$$

所以不满秩,使得方程具有多个解, 选择哪一个$\hat{w}^*$取决于算法, 可以选择正则化方法。

广义线性模型

考虑单调可微函数$g(*)$, 令

$$y=g^{-1}(w^Tx+b)$$

用线性模型实现非线性效果, g叫做联系函数

对数几率回归

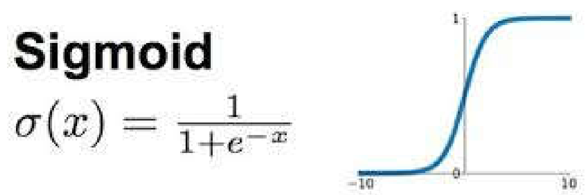
以上分析均适用于连续值的预测, 对于分类问题可以利用阶跃函数映射到0，1域上

设z为预测值

$$y=\begin{cases}0, & z<0\\ 0.5, & z=0\\ 1, & z>0\end{cases}$$

由于其不连续, 我们希望能找到一个连续的函数来逼近于阶跃函数, 这样的函数可以是对数几率函数:

$$y=\frac{1}{1+e^{-z}}$$



上式可以变形为:

$$\ln\frac{y}{1-y}=z=w^Tx+b$$

直观解释:
y只能在$(0,1)$之间取值，那么y就解释为样本x作为正例的可能性，$1-y$就解释为样本x作为反例的可能性

回归分类方法

回归方法实际上是一种分类方法，优点在于无需事先假设数据分布, 通过对数几率回归可以得到概率, 对数几率函数是任意阶可导的凸函数，具有良好的数学性质。
当给定x时易知:

$$\text{预测为}y=1\text{的概率为}:P(y=1|x)=\frac{e^{w^tx+b}}{1+e^{w^tx+b}}$$

$$\text{预测为}y=0\text{的概率为}: P(y=0|x)=\frac{1}{1+e^{w^t x+b}}$$

所以要调整参数使得$\sum_{i=1}^{m}\ln p(y_i|xi;w,b)$越大越好，意为xi的预测值与其标签相符合的概率越大越好
令:$\beta=(w;b),\hat{x}=(x;1)$再令: $p_1(\hat{x};\beta)=p(y=1|\hat{x};\beta)$ 意思是在 β 参数下预测值为1的概率; 相似的令: $p_0(\hat{x};\beta)=p(y=0|\hat{x};\beta)=1-p_1(\hat{x};\beta)$

$$p(y_i|xi;w,b)=[p_1(\hat{x};\beta)]^{y_i}[p_0(\hat{x};\beta)]^{1-y_i}$$

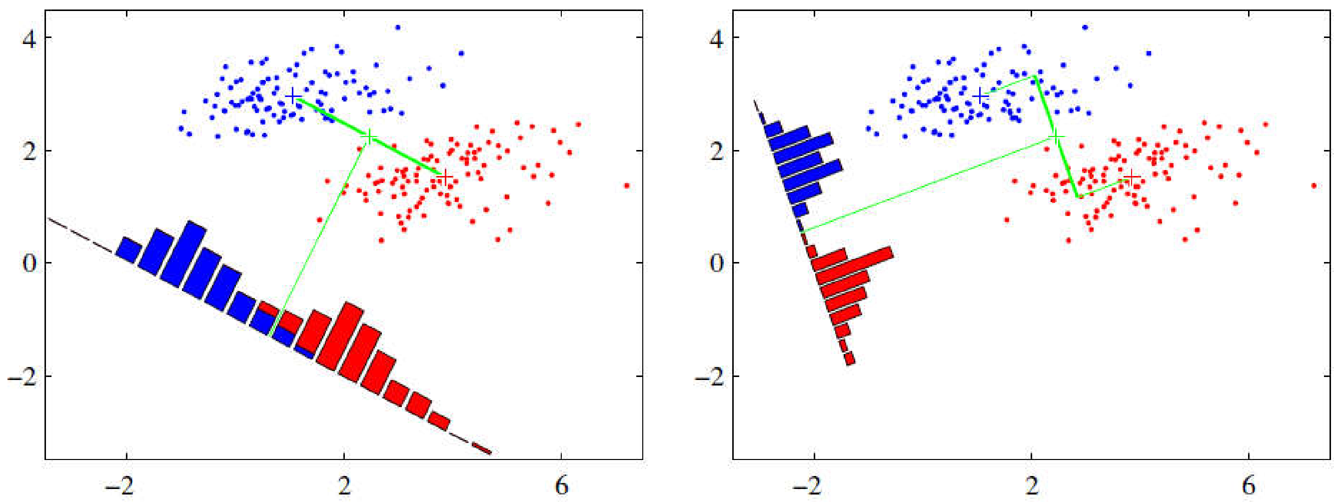
代入$\sum_{i=1}^{m}\ln p(y_i|xi;w,b)$得到

$$\ell(\beta)=\sum_{i=1}^{m}(y_i\beta^T\hat{x}_i-\ln(1+e^{\beta^T\hat{x}_i}))$$

可以用梯度下降或者牛顿法迭代优化参数

线性判别分析(LDA)

- 思想:将两个类别的样本投影到一条直线上, 如下图:



当有新的待分类数据到达时, 同样投影到直线完成分类
- 数学描述:
设:X_1,X_2 为两个标签数据的集合, μ_1,μ_2 是他们的均值,Σ_1,Σ_2 是协方差矩阵 w^T为投影向量, 则完成映射后的均值方差分别为: $w^T\mu_0,w^T\mu_1,w^T\Sigma_0 w,w^T\Sigma_1 w$
我们希望投影函数使得均值尽量远离, 类内部的方差尽量小，所以得到最大化目标:

$$J=\frac{||w^T\mu_0-w^T\mu_1,||^2}{w^T\Sigma_0 w+w^T\Sigma_1 w}=\frac{w^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T w}{w^T(\Sigma_0+\Sigma_1)w}$$

令$S_b=(\mu_0-\mu_1)(\mu_0-\mu_1)^T,S_w=\Sigma_0+\Sigma_1,$

$$J=\frac{w^T S_b w}{w^T S_w w}$$

注意到此时J是一个齐次形式, 所以最终的解与w的长度无关, 只与w的方向有关, 所以不妨令分母w^TS_ww = 1, 只需要求分子的最大值即可:

$$目标函数: \max_w = w^T S_b w$$

$$约束条件: \text{s.t.} \ \ w^T S_w w = 1$$

使用**Lagrange**乘数法:
转化为求函数 $L(w) = \max_w = w^T S_b w - \lambda(w^T S_w w - 1)$的最大值

$$\frac{\partial L}{\partial w} = (S_b^T + S_b)w - \lambda(S_w^T + S_w)w$$

由于S_b和S_w 具有斜对角的对称性, 所以上式可以等价地写作$S_b w = \lambda S_w w$

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

对上面矩阵进行维度分析, 后面两项相乘得到的是一个标量, 所以$S_b w$与$\mu 0 - \mu 1$方向相同, 又因为w不限制大小, 所以可以令$S_b w = \lambda(\mu 0 - \mu 1)$ 所以$w = S_w^{-1}(\mu 0 - \mu 1)$, 而$S_w^{-1}$通常可以通过SVD方法来得到

- 推广: 在多分类任务中LDA算法常用类似手段实现数据降维

多分类学习

考虑N个类别$C_1, C_2, \cdots C_N$, 分类的基本思想: 将多分类化为若干个二分类, 拆分的基本策略: 一对一(OvO), 一对其余(OvR), 多对多(MvM)

- OvO: 将数据类别两两配对共产生$N(N-1)/2$个二分类器, 将每一个二分类器的结果投票产生最终结果
- OvR: 将每一个类看成正类, 将其他的类看成反类, 取置信概率最大的类
- MvM: 一部分作为正类, 另一部分作为反类
  ECOC技术: 使用M个分类器每个分类器对N个类别做不同的划分, 得到M不同的编码(分类器f_i化为正类第i的分量为1, 否则为-1), 对数据进行M次二分类后得到预测的编码(分类器f_i化为正类第i的分量为1, 否则为-1), 比较海明距离或者欧式距离对数据N分类
  PS: 引入编码技术是使得分类的结果具有一定的容错性

类别不平衡问题

上面介绍的方法只适用于正例和反例大体相等的情况, 对于类别数据量不平衡的问题有如下解决办法:

- 欠采样: 对多出的数据去除, 这样会损失重要信息, 可以分别取出形成多数据集
- 过采样: 用插值办法增加数据少的一边的数据
- 阈值移动: $\frac{y'}{1-y'} = \frac{y}{1-y} \times \frac{m^-}{m^+}$来修正
  对阈值移动的理解:
  1. 前面的分类器假定正例和反例一样多其阈值为0.5, 即y>0.5可判断为正例, 当正例和反例不一样多时, 阈值会发生移动
  2. 从模型的角度讲, 生成的模型更加依赖于数据多的类, 因此要加以修正