日期:　　　/

线性模型

一. 线性回归

1. 回归

① 定义: 给定一组数据点 $(x_1, y_1) \cdots, (x_n, y_n)$, 根据这些数据点研究 x 和 y 之间关系的分析方法就是回归.

② 线性回归:

(1) 利用数理统计中回归分析, 来确定 2 种或 2 种以上变量间相互依赖的定量关系的一种统计分析方法.

(2) 以一个线性函数 $f(x) = w^T x + b$ 描述两者关系。

③ 逻辑回归: 以逻辑函数描述

2. 线性模型表示形式

① 一般形式: $f(x) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b$.

$x = [x_1, x_2, \ldots, x_d]^T$ 是由属性描述的样本, 其中 x_i 是 x 在第 i 个属性上的取值.

② 向量形式:

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

其中 $w = (w_1; w_2; \cdots; w_d) = [w_1, w_2, \cdots, w_d]^T$

3. 线性模型优点..

日期: /

① 形式简单，易于建模。

② 可解释性

③ 非线性模型的基础：引入层级结构和高维映射。

一个例子

- 综合考虑色泽、根蒂和敲声来判断西瓜好不好
- 其中根蒂的系数最大，表明根蒂最要紧；而敲声的系数比色泽大，说明敲声比色泽更重要

$$f_{\text{好瓜}}(\boldsymbol{x}) = 0.2 \cdot x_{\text{色泽}} + 0.5 \cdot x_{\text{根蒂}} + 0.3 \cdot x_{\text{敲声}} + 1$$

4. 数据。

① 给定 $D=\{(x_1,y_1),(x_2,y_2),\dots,(x_m,y_m)\}$.

$x_i=(x_{i1};x_{i2}\cdots;x_{id})$ $y_i\in R$.

② 离散属性处理。

- 有序关系：连续化为连续值
- 无序关系：有 k 个属性值，则转为 k 维向量

5. 单一属性的线性回归流程

① 模型表示 $f(x_i)=wx_i+b$

② 代价函数 (loss function) 均方误差.

③ 均方误差最小化模型求解方法：最小二乘法.

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

日期: /

$$= \underset{(w,b)}{\arg\min} \frac{1}{m} \sum_{i=1}^{m} (wx_i + b - y_i)^2$$

$$= \underset{(w,b)}{\arg\min} \sum_{i=1}^{m} (wx_i + b - y_i)^2$$

最小化均方误差

$$E_{(w,b)} = \sum_{i=1}^{m} (y_i - wx_i - b)^2.$$

分别对 w 和 b 求偏导可得

$$\begin{cases} \dfrac{\partial E_{(w,b)}}{\partial w} = 2\left(w \sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} (y_i - b) x_i \right) = 0 \\ \dfrac{\partial E_{(w,b)}}{\partial b} = 2\left(mb - \sum_{i=1}^{m} (y_i - wx_i) \right) = 0 \end{cases}$$

得:

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

$$b = \frac{1}{m} \sum_{i=1}^{m} (y_i - wx_i)$$

，其中 $\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$

6. 多元线性回归

① 数据集.

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$$

$$\begin{aligned} \boldsymbol{x}_i &= (x_{i1}; x_{i2}; \ldots; x_{id}) \quad (d \text{ 为属性维数}) \\ &= (x_{i1}, x_{i2}, \cdots, x_{id})^{\mathrm{T}} \\ &= [x_{i1}, x_{i2}, \cdots, x_{id}]^{\mathrm{T}} \end{aligned}$$

$$y_i \in \mathbb{R} \quad i = 1, 2, \cdots, m$$

② 目标: $f(x_i) = w^T x_i + b$ 使得 $f(x_i)$ 逼近 y_i.

其中: w, b 为向量形式: $\hat{w} = (w;b) = [w, b]^T$

$$\mathbf{X} = \begin{array}{c} \overbrace{\qquad\qquad\qquad}^{\boldsymbol{x}_1} \\ \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix} \end{array} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} & 1 \\ \boldsymbol{x}_2^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} & 1 \end{pmatrix} = \begin{bmatrix} \hat{\boldsymbol{x}}_1^{\mathrm{T}} \\ \hat{\boldsymbol{x}}_2^{\mathrm{T}} \\ \vdots \\ \hat{\boldsymbol{x}}_m^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{m\times(d+1)}$$

$$\hat{\boldsymbol{x}}_i = [x_{i1}, x_{i2}, \cdots, x_{id}, 1]^{\mathrm{T}} \qquad \boldsymbol{y} = (y_1; y_2; \ldots; y_m)$$

③ 求解: 最小二乘模型.

$$(\boldsymbol{w}^*, b^*) = \underset{\boldsymbol{w},b}{\arg\min} \sum_{i=1}^{m} (f(\boldsymbol{x}_i) - y_i)^2 = \underset{\boldsymbol{w},b}{\arg\min} \sum_{i=1}^{m} \left[(\boldsymbol{w}^T \boldsymbol{x}_i + b) - y_i\right]^2$$

$$\Rightarrow \quad \hat{\boldsymbol{w}}^* = \underset{\hat{\boldsymbol{w}}}{\arg\min} \sum_{i=1}^{m} (\hat{\boldsymbol{w}}^{\mathrm{T}} \hat{\boldsymbol{x}}_i - y_i)^2 = \underset{\hat{\boldsymbol{w}}}{\arg\min} \sum_{i=1}^{m} (\hat{\boldsymbol{x}}_i^{\mathrm{T}} \hat{\boldsymbol{w}} - y_i)^2$$

$$= \underset{\hat{\boldsymbol{w}}}{\arg\min} (\mathbf{X}\hat{\boldsymbol{w}} - \boldsymbol{y})^{\mathrm{T}} (\mathbf{X}\hat{\boldsymbol{w}} - \boldsymbol{y})$$

$$= \underset{\hat{w}}{\arg\min} \|X\hat{w} - y\|_2^2$$

$$= \arg\min \hat{w}^T X^T X \hat{w} - 2\hat{w}^T X^T y + y^T y$$

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

$$\Rightarrow \frac{\partial E_{\hat{w}}}{\partial w} = 2X^T (X\hat{w} - y) = 0$$

□ 若 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 是满秩矩阵或正定矩阵(矩阵可逆), 则

$$\hat{\boldsymbol{w}}^* = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}} \boldsymbol{y}$$

其中 $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$ 是 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的逆矩阵, 线性回归模型为

$$f(\hat{\boldsymbol{x}}_i) = \hat{\boldsymbol{x}}_i^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}} \boldsymbol{y}$$

日期: /

7. 对数线性回归

① 定义：输出标记的对数为线性模型逼近的目标

② 应用场合：样本对应标记 y 不是线性变化，而是指数变化

③ 表示形式：$\ln y = w^T x + b$

8. 广义线性模型

① 表示形式：$y = g^{-1}(w^T x + b)$

$g(\cdot)$ 称联系函数，单调可微函数

例 $g(\cdot) = \ln(\cdot)$ $\quad \ln y = w^T x + b$

二、二分类任务：

对于二分类任务，输出标记为 $y \in \{0,1\}$，即我们更倾向于选择介于0和1之间的概率，而线性回归的预测结果 $z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$ 一般是一个连续值，因此，我们需要将 $z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$ 做某种变换，将其转换到输出介于0和1之间的值(类似于概率范围)。

$$\left\{\begin{array}{l} \text{线性回归模型实际预测值 } z = w^T x + b \\ y \in \{0,1\} \end{array}\right\} z \text{ 与 } y \text{ 联系}$$

1. 最理想函数：单位阶跃函数

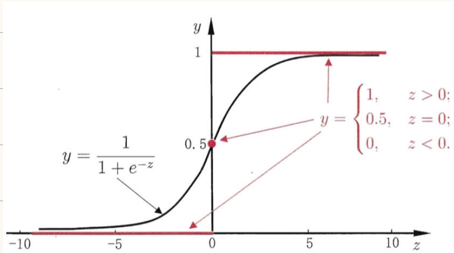
$$y = \begin{cases} 0, & z < 0 \\ 0.5, & z = 0 \\ 1, & z > 0 \end{cases}$$

日期:　　　/

缺点：不连续，$x=0$ 不可导。

2. 对数几率函数 logistic function (sigmoid)：

① 表示形式　$y=\frac{1}{1+e^{-z}}$ ⇒ 将 z 变为 $(0,1)$ 的 y



对数几率函数定义域 $(-\infty,+\infty)$，值域 $(0,1)$

② 几率(odds)定义：指该事件发生的概率与该事件不发生概率的比值。

$$odds=\frac{p}{1-p}$$ （p 为事件发生概率）

- $odds>1$，事件发生概率大
- $odds<1$，事件发生概率小
- 变化范围为 $[0,+\infty)$

③ 对数几率（log odds 或 logits）：$\log\frac{p}{1-p}$

- $\log\frac{p}{1-p}>0$，事件发生概率大
- $\log\frac{p}{1-p}<0$，事件发生概率小

3. 对数几率回归

日期:　　/

$$y = g^{-1}(w^T x + b) \rightarrow g^{-1} = \frac{1}{1+e^{-z}}$$

① 表现形式: $\ln\frac{y}{1-y} = w^T x + b$ 或 $y = \frac{1}{1+e^{-(w^T x+b)}}$

其中 y 表示样本 x 作为正例的可能性.

② 本质: 是一种分类学习方法

③ 优点:

- 无需事先假设数据分布
- 可得到"类别"的近似概率预测
- 可直接应用现有数值优化算法求取最优解

④. 概率解释:

y 表示样本 x 为正例可能性.(后验概率)

$$\begin{cases} y = p(y=1|x) \\ 1-y = p(y=0|x) \end{cases} \Rightarrow \ln\frac{p(y=1|x)}{p(y=0|x)} = w^T x + b$$

$$\Rightarrow \begin{cases} p(y=1|x) = \frac{e^{w^T x+b}}{1+e^{w^T x+b}} = h_\beta(x) \\ p(y=0|x) = \frac{1}{1+e^{w^T x+b}} = 1 - h_\beta(x) \end{cases}$$

其中 $\beta = (w;b) = [w,b]^T$

⑤ 极大似然法:

对于 $p(x|\theta)$

(1) 若 x 为变量, θ 已知, 则 $p(x|\theta)$ 为概率函数[概率分布参数为 θ 时, 不同样本 x 出现概率].

日期: /

2、当x已知，θ变量时，p(x|θ)为似然函数[即概率分布的参数取不同值时，某个样本x出现的概率]，在已有观测样本时，寻找数据分布的最佳参数

3、求解对数几率回归：构造似然函数+(牛顿法)

□ $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m$ 为独立同分布的采样，定义似然函数 L 为混合密度函数 (m个样本同时出现)：

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= p(y_1|\boldsymbol{x}_1;\boldsymbol{\beta}) \times p(y_2|\boldsymbol{x}_2;\boldsymbol{\beta}) \times \cdots \times p(y_m|\boldsymbol{x}_m;\boldsymbol{\beta}) \\
&= \prod_{i=1}^{m} p(y_i|\boldsymbol{x}_i;\boldsymbol{\beta}) = \prod_{i=1}^{m} (h_{\boldsymbol{\beta}}(\boldsymbol{x}_i))^{y_i} (1 - h_{\boldsymbol{\beta}}(\boldsymbol{x}_i))^{1-y_i}
\end{aligned}
$$

思想：找到一组参数使所有观测样本联合概率最大化.

$p(y_i|x_i;\beta)$：在β条件下，x_i被预测正确的概率

正确分类概率 $p(y_i|\boldsymbol{x}_i;\boldsymbol{\beta}) = \begin{cases} p(y=1|\boldsymbol{x}_i;\boldsymbol{\beta}), & \text{if } y_i = 1, \\ p(y=0|\boldsymbol{x}_i;\boldsymbol{\beta}) = 1 - p(y=1|\boldsymbol{x}_i;\boldsymbol{\beta}) & \text{if } y_i = 0, \end{cases}$

$$= (p(y=1|\boldsymbol{x}_i;\boldsymbol{\beta}))^{y_i} \times (1 - p(y=1|\boldsymbol{x}_i;\boldsymbol{\beta}))^{1-y_i},$$

$$
\begin{aligned}
\ln p(y_i|\boldsymbol{x}_i;\boldsymbol{\beta}) &= \begin{cases} \ln p(y=1|\boldsymbol{x}_i;\boldsymbol{\beta}), & \text{if } y_i = 1, \\ \ln p(y=0|\boldsymbol{x}_i;\boldsymbol{\beta}) = \ln[1 - p(y=1|\boldsymbol{x}_i;\boldsymbol{\beta})] & \text{if } y_i = 0, \end{cases} \\
&= y_i \ln p(y=1|\boldsymbol{x}_i;\boldsymbol{\beta}) + (1-y_i) \ln p(y=0|\boldsymbol{x}_i;\boldsymbol{\beta})
\end{aligned}
$$

其中 $\boldsymbol{\beta} = (\boldsymbol{w}_i; b)$.

对数似然函数

$$\ell(\boldsymbol{w}, b) = \ln L(\beta) = \sum_{i=1}^{m} \ln p(y_i|\boldsymbol{x}_i; \boldsymbol{w}, b)$$

概率中分号表示分号后的 $\boldsymbol{w}, b$ 是待估**参数**，它们是确定的，只是当前未知。它们不是随机变量。

日期: /

$$\ln L(\boldsymbol{\beta})=\sum_{i=1}^{m}\left[y_i \ln h_{\boldsymbol{\beta}}(\boldsymbol{x}_i)+(1-y_i)\ln(1-h_{\boldsymbol{\beta}}(\boldsymbol{x}_i)\right]$$

$$\max_{\boldsymbol{\beta}} \ln L(\boldsymbol{\beta}) \Leftrightarrow \min_{\boldsymbol{\beta}} -\ln L(\boldsymbol{\beta})$$ （最大化转化为求最小化）

$$\begin{aligned}\ell(\boldsymbol{\beta})&=-\ln L(\boldsymbol{\beta})\\&=\sum_{i=1}^{m}\left[y_i \ln h_{\boldsymbol{\beta}}(\boldsymbol{x}_i)+(1-y_i)\ln(1-h_{\boldsymbol{\beta}}(\boldsymbol{x}_i)\right]\\&=\sum_{i=1}^{m}\left[-y_i\boldsymbol{\beta}^T\hat{\boldsymbol{x}}_i+\ln\left(1+e^{\boldsymbol{\beta}^T\hat{\boldsymbol{x}}_i}\right)\right]\end{aligned}$$

动手推导

- 求解

$$\boldsymbol{\beta}^*=\arg\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$$

- 牛顿法第 *t*+1 轮迭代解的更新公式

$$\boldsymbol{\beta}^{t+1}=\boldsymbol{\beta}^t-\left(\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\right)^{-1}\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}$$

其中关于 β 的一阶、二阶导数分别为

$$\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=-\sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\left(y_i-p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})\right)$$

动手推导

$$\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}=\sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\hat{\boldsymbol{x}}_i^{\mathrm{T}}p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})\left(1-p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})\right)$$

动手推导

对数几率回归-梯度下降法求解

- 关于参数 β 的更新公式

$$\boldsymbol{\beta}^{t+1}=\boldsymbol{\beta}^t-\alpha\frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta}$$

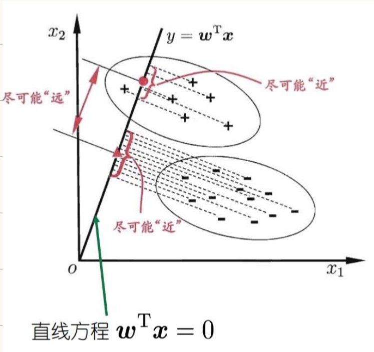
其中 α 是学习率. 关于 β 的一阶偏导数为

$$\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=-\sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\left(y_i-p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})\right)$$

日期: /

4. 线性判别分析 LDA

① 基本思想：给定训练集，设法将样例投影到一条适当选择的直线上，使得同类样例的投影点尽可能接近，异类样例投影点中心尽可能远离【类内方差小，类间方差大】



目标：

② 在保留尽可能多的类区分信息时进行降维

③ 寻找最佳投影向量 w，X到 w 上投影为

$$y = w^T X$$

④ 目标函数

数据：m个d维样本 $x_1, \cdots, x_m$，属于示例集合 X_i 中两个不同类别 w_0, w_1 即

$$\begin{cases} X_0 = \{x_i \mid w(x_i) = w_0\} & |X_0| = n_1 \\ X_1 = \{x_i \mid w(x_i) = w_1\} & |X_1| = n_2 \end{cases}$$

日期: /

线性组合成 $(x_1, x_2, \cdots, x_m)$ 产生 m 个投影结果 $(y_1, \cdots, y_m)$ 相应属于集合 Y_0, Y_1，即 $y_i = w^T x_i$

投影前样本均值 | 投影后样本均值

$\mu_i = \frac{1}{n_i} \sum_{x \in X_i} x \quad i=0,1$

$\tilde{\mu}_i = \frac{1}{n_i} \sum_{y \in Y_i} y = \frac{1}{n_i} \sum_{x \in X_i} w^T x = w^T \mu_i$

均值投影

定义：(2个指标) 投影后样本均值之差，投影后各类的类内散度（类似于方差）

投影后二点的样本均值之差 $|\tilde{\mu}_0 - \tilde{\mu}_1| = |w^T(\mu_0 - \mu_1)|$

投影后第 i 类类内散度：$\tilde{s}_i^2 = \sum_{y \in Y_i} (y - \tilde{\mu}_i)^2$

Fisher 线性可分准则要求

$$\max \quad J(w) = \frac{|\tilde{\mu}_0 - \tilde{\mu}_1|^2}{\tilde{s}_0^2 + \tilde{s}_1^2}$$

→ 尽可能大

→ 尽可能小

定义**类间散度矩阵** $\boldsymbol{S}_b = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T$，可证明其为对称半正定的. 投影后两类样本均值之差展开为

$$|\tilde{\mu}_0 - \tilde{\mu}_1|^2 = \boldsymbol{w}^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}.$$

投影后第 i 内的类内散度为:

$$\begin{aligned}\tilde{s}_i^2 &= \sum_{y \in Y_i} (y - \tilde{\mu}_i)^2 \\ &= \sum_{y \in Y_i} (\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu}_i)^2 \\ &= \sum_{\boldsymbol{x} \in X_i} \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{\Sigma}_i\boldsymbol{w}.\end{aligned}$$

$\tilde{s}_0^2 + \tilde{s}_1^2 = w^T S_w w$

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