线性回归

①例子：<a> 股价涨跌幅回归 <b> 房价预测 $y = a_1x_1 + a_2x_2 + c$.

②原理：用极大似然解释最小二乘

预估目标建模：$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$ ⓐ 样本i的似然函数

预估目标 模型参数 特征 误差值

样本i: $\varepsilon^{(i)}$

<b> 误差值i的概率密度

$$p(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}}$$

<a> 误差值 $\{\varepsilon^{(1)}, \varepsilon^{(2)}, \dots \varepsilon^{(m)}\}$ 服从某均值为0 方差为σ的高斯分布. ← 根据大数定理（概率论㉓㉔）

!※ 假定M样本独立，则 $\{\varepsilon^{(i)}\}$ 独立分布（不会相互影响）

!※ 假定M样本同分布，比如是因为某些线性特征未考虑

※ 均值 $\sum \varepsilon^{(i)}/M = 0$，如果不为0 可通过平移使其为0

<c> 样本i似然函数值的概率密度

$$p(y^{(i)}|x^{(i)},\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}}$$

$y^{(i)}$（样本标签）

误差值的高斯分布以0为中心，似然函数（的预估值）以 $\theta^T x^{(i)}$ 为中心，方差都是 σ^2.

<d> 所有样本的似然函数.

$$L(\theta) = \prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}}$$

对数似然（方便计算，乘法变加法）

$$\underline{\ln L(\theta)} = \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}} = \underbrace{m \ln \frac{1}{\sqrt{2\pi}\sigma}}_{\text{固定值}} - \underbrace{\frac{1}{\sigma^2}\cdot\frac{1}{2}}_{\text{固定值}} \underbrace{\sum_{i=1}^{m} (y^{(i)} - \theta^T(x^{(i)}))^2}_{\text{希望最小}}$$

希望最大（为什么希望 $\ln L(\theta)$ 最大；　）

ⓐ 见 LR&Softmax ⑬，可能同理.

似然最大 <=> 相对熵（y 与 预估y）最小

另外一种直观理解

ⓑ 高斯分布中心点的概率密度最大，而在这个中心点预估值与样本y值相同，误差为0

预估目标 ☆ $J(\theta) = \sum_{i=1}^{m} (y^{(i)} - \theta^T(x^{(i)}))^2$

③假设的内涵性，简化性，发散性

内涵性：根据常理应该是正确的，往往是正确的但不一定总是正确

简化性：只是接近真实，往往做了简化（如词袋模型）

发散性：在某个简化假设下推导出的结论，假设不成立时有时也能用

④ 最小二乘求解:

$$\theta^* = \arg\min_{\theta} J(\theta) = \frac{1}{2}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 = \frac{1}{2}(X\theta - Y)^T(X\theta - Y)$$

$$= \frac{1}{2}(\theta^T X^T - Y^T)(X\theta - Y) = \frac{1}{2}(\theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta + Y^T Y)$$

求偏导:

$$\frac{\partial J(\theta)}{\partial(\theta)} = \frac{1}{2}\cdot\frac{\partial(\theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta + Y^T Y)}{\partial(\theta)}$$

$$= \frac{1}{2}\Big(\underbrace{2X^TX\theta}_{(线代㉑)} \ \underbrace{-X^TY}_{(线代⑳)} \ \underbrace{-(Y^TX)^T}_{(线代⑳)} + 0\Big) \quad 与\theta无关,消去$$

$$= X^TX\theta - X^TY$$

令 $X^TX\theta - X^TY = 0$ 求驻点，然而只有 X^TX 可逆，才有解.

为防止 X^TX 不可逆，同时防止过拟合，在主对角线增加扰动因子 λI

$$(X^TX + \lambda I)\theta - X^TY = 0$$

X^TX 半正定：线代⑯
一定有 $z^T(\lambda I)z > 0$, $(X^TX+\lambda I)$正定

一定可逆

$\boxed{阻止\theta_i变得过大}$ $\qquad \theta = (X^TX + \lambda I)^{-1}X^TY$

⑤ 正则项与$\boxed{防止过拟合}$. 与④使用的 λI 不同. 实际应用中，使用下面三种之一

- L1-norm (Lasso): $J(\vec{\theta}) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}|\theta_j|$
- L2-norm (Ridge): $J(\vec{\theta}) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2$ // 本质为假定θ服从高斯分布
- Elastic Net: $J(\vec{\theta}) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\left(\rho\cdot\sum_{j=1}^{n}|\theta_j| + (1-\rho)\sum_{j=1}^{n}\theta_j^2\right)$

Lasso 具有特征选择能力，原理：

希望求 $J(\theta)$ 极值，同时希望 $\lambda\sum_{j=1}^{n}|\theta_j|$ 比较小，即 $\sum_{j=1}^{n}|\theta_j| < r$，转化为约束条件求极值(拉格朗日)问题，这个约束条件使得 $|\theta_j|$ 不能太大，当特征数(阶数)增多时，一些与优化目标关系较弱特征对应的 $|\theta_j|$ 会非常接近零

⑥ 超参数 λ 的选择：用 N-Fold 交叉验证做实验，看哪个 λ 取值下训练出的模型 MSE(均方误差)最小.

⑦伪逆矩阵：前面线性回归使用λI扰动因子保证公式有解，如果不用λI来解线性方程组 $A\vec{x}=\vec{B}$，另一个可选方法是使用伪逆矩阵.

$$\begin{cases} A\text{可逆时：} \vec{x} = A^{-1}\vec{B} \\ A\text{不可逆：} \vec{x} = A^{+}\vec{B} = (A^TA)^{-1}A^TB \end{cases}$$

其中 $A^+=(A^TA)^{-1}A^T$ 叫做A的伪逆

A可逆时：$A^+ = (A^TA)^{-1}A^T = A^{-1}(A^T)^{-1}A^T = A^{-1}$

如何得到 A^+：ⓐ奇异值分解 $A_{m\times n} = U\Sigma V^T$ ⓑ直接令 $A^+ = V\Sigma^{-1}U^T$

⑧ L1-Norm梯度下降

⑨批量梯度下降：把所有样本梯度加到一起然后下降.（梯度一定下降）

⑩随机梯度下降：拿到一个样本下降一次（梯度震荡下降）

优势：ⓐ快 ⓑ不用等到所有样本，适合在线学习 ⓒα步长较小时，利用噪声样本跳出"洼地"

⑪ Mini-batch梯度下降：若干样本下降一轮（折衷方法，实际中使用较多）

⑫下降步长α：

<a> 固定步长

<b> 自适应α：参考斜率大小，参考一定步长时loss是上升还是下降

<c> 回溯线性搜索

⑬样本与预估目标

预估目标与样本特征不呈线性关系时也没关系（与θ呈线性关系就行）

事先对特征做变换，例如 $x \Rightarrow x^2$，不再仅使用一次函数，用二次，三次函数来拟合 $y=\theta_{11}x_1+\theta_{12}x_1^2+\theta_{13}x_1^3+\theta_{21}x_2+\theta_{22}x_2^2+\cdots$

但次数过高的缺点是过拟合.

⑭过拟合解决：

<a> 正则项 <b> 删减无关特征，例如PCA（降维），one-hot编码，手造出一些组合特征，基于业务逻辑

⑮ 超参数判定系数

方法1：均方误差(MSE) $\sum_{i=1}^{m}(\underbrace{h_\theta(x^{(i)}, x^{(i)})}_{\text{似然函数}} - y^{(i)})^2$ 其实就是残差平方和

=RSS RSS

预估的 $\hat{y}^{(i)}$ 样本真实的 $y^{(i)}$

方法2：判定系数(R^2)

残差平方和(Residual Sum of Square)：预估偏差.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$$

样本均方差和值(Total Sum of Square)：体现样本本身的发散程度

原理：方法1残差平方和太大可能是因为样本本身震荡造成的

所以用TSS（m*样本均方差）来作为基准.

R2值越大，说明训练的越准确，越小说明误差越大，正常位于0，1，

为负值时说明误差极大

方法3：判定系数(ESS)，又叫回归平方和.

$$ESS = \sum_{i=1}^{m}(\hat{y}_i - \bar{y})^2$$

预测值 样本均值

~~RSS X X RSS~~

TSS ≥ ESS + RSS 仅当无偏估计时成立.

样本均方差 预测值 残差平方和.

(定值) v.s (预测值 v.s 样本真实值)

样本均值

in fold

va

bias

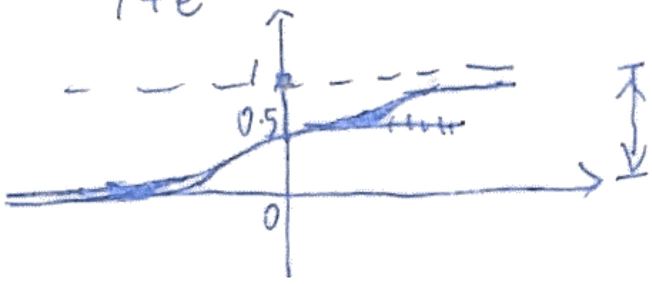
Logistic回归/Softmax回归

①Logistic回归用于2分类；Softmax回归用于N分类：两者都是从信息熵推导出来的分类器；不推荐用线性回归做分类，如PPT中的例子

②LR建模（$h_\theta(x)$）：不像线性回归直接用θx来对预测回归建模，而是把θx放在sigmoid函数中。

sigmoid函数：$g(z)=\frac{1}{1+e^{z}}$

特性ⓐ



值域在(0,1)

特性ⓑ $g'(z) = \cdots = g(z)\cdot(1-g(z))$

把$z=\theta x$代入到$g(z)=\frac{1}{1+e^{z}}$中，得到建模用的 函数.

模型参数 特征

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

③ LR参数估计（即LR参数求解）

假定 $P(y=1|x;\theta)=h_\theta(x)$ ⓐ $P(y=0|x;\theta)=1-h_\theta(x)$ ⓑ 推导出

则用一个公式统一表示 $P(y|x;\theta)=(h_\theta(x))^{y}(1-h_\theta(x))^{1-y}$ 或 ★也与二项分布一致

扩展到多个样本，得到似然函数

$$L(\vec{\theta})=P(\vec{Y}|X;\theta)=\prod_{i=1}^{m}p(y^{(i)}|x^{(i)};\theta)=\prod_{i=1}^{m}(h_\theta(x^{(i)}))^{y^{(i)}}(1-h_\theta(x^{(i)}))^{1-y^{(i)}}$$

求对数，得到对数似然函数

$$l(\vec{\theta})=\ln(L(\vec{\theta}))=\sum_{i=1}^{m}\left(y^{(i)}\ln \underbrace{\boxed{h_\theta(x^{(i)})}}_{g(\theta^T x)}+(1-y^{(i)})\ln(1-\underbrace{\boxed{h_\theta(x^{(i)})}}_{g(\theta^T x)})\right)$$

关于θ_j求偏导，

$$\frac{\partial l(\vec{\theta})}{\partial \theta_j}=\cdots=\sum_{i=1}^{m}(y^{(i)}-g(\theta^T x^{(i)}))\,x_j^{(i)}$$

偏导即为梯度方向，延梯度方向更新θ_j（梯度上升，让似然函数取最大值）

$$\theta_j := \theta_j + \underbrace{\alpha}_{学习率}(y^{(i)}-h_\theta(x^{(i)}))\,x_j^{(i)}$$

理解

期望 ★似然函数取最大值 $\Rightarrow h_\theta(x^{(i)})^{y^{(i)}}(1-h_\theta(x^{(i)}))^{1-y^{(i)}}$：$y^{(i)}$取值(0,1)决定$h_\theta(x^{(i)})$该是0,还是1，$h_\theta$与y一致才能取得最大似然

④ 为什么LR预估值可看作概率

发生概率 $p = p(y=1|\vec{x};\theta) = h_\theta(\vec{x}) = g(\theta^T\vec{x}) = \frac{1}{1+e^{-\theta^T x}}$

不发生概率 $1-p = p(y=0|\vec{x};\theta) = 1-h_\theta(\vec{x}) = 1-g(\theta^T x) = \frac{e^{-\theta^T x}}{1+e^{-\theta^T x}}$

用对数概率公式来表示 (logit函数)

代入

$$\mathrm{logit}(p) = \log\frac{p}{1-p} = \log\frac{h_\theta(x)}{1-h_\theta(x)} = \log\frac{1}{e^{-\theta^T x}} = \theta^T x$$

线性函数

到这里可以看出，作者是想设计一个函数 $g(z)$，使得 $\log\frac{p}{1-p}$ 是线性的

即 $\log\frac{p}{1-p} = \theta^T x$，用这个等式可推出 $\begin{cases} g(z) = \frac{1}{1+e^{z}} \\ z = \theta x \end{cases}$

原因见本节⑬

⑤ LR损失函数：用负的对数似然函数作为损失函数

(似然函数取最大值时，损失函数取最小值)

对比线性回归，把MSE (均方误差) 作为损失函数.

为了让损失函数好看些，用 $y=1$ 表示正例，$y=-1$ 表示负例，则

似然函数改写：$L(\theta) = \prod_{i=1}^{m} p_i^{y_i}(1-p_i)^{1-y_i}$ 改为 $L(\theta) = \prod_{i=1}^{m} p_i^{(y_i+1)/2}(1-p_i)^{-(y_i-1)/2}$

对数似然改写：$l(\theta) = \sum_{i=1}^{m} \ln\left(p_i^{y_i}\cdot(1-p_i)^{1-y_i}\right)$ 改为 $l(\theta) = \sum_{i=1}^{m} \ln\left(p_i^{(y_i+1)/2}\cdot(1-p_i)^{-(y_i-1)/2}\right)$

损失函数

$$loss(y,\hat{y}_i) = -l(\theta) = -\sum_{i=1}^{m}\left[\frac{1}{2}(y_i+1)\ln p_i - \frac{1}{2}(y_i-1)\ln(1-p_i)\right]$$

$$= \sum_{i=1}^{m}\left[\frac{1}{2}(y_i+1)\ln\frac{1}{p_i} - \frac{1}{2}(y_i-1)\ln\frac{1}{1-p_i}\right] \qquad p_i = \frac{1}{1+e^{-\theta^T x}}$$

$$= \sum_{i=1}^{m}\left[\frac{1}{2}(y_i+1)\ln(1+e^{-\theta^T x}) - \frac{1}{2}(y_i-1)\ln(1+e^{+\theta^T x})\right] \qquad 1-p_i = \frac{e^{-\theta^T x}}{1+e^{-\theta^T x}}$$

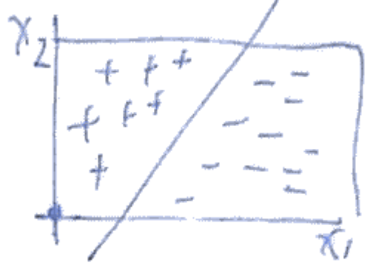
$$= \begin{cases} \sum_{i=1}^{m}\left[\ln(1+e^{-\theta^T x})\right] & y_i = 1 \\ \sum_{i=1}^{m}\left[\ln(1+e^{\theta^T x})\right] & y_i = -1 \end{cases}$$

备注：也有用MSE做损失函数，不是不可以，只不过能用LR自带损失函数的，尽量用LR的

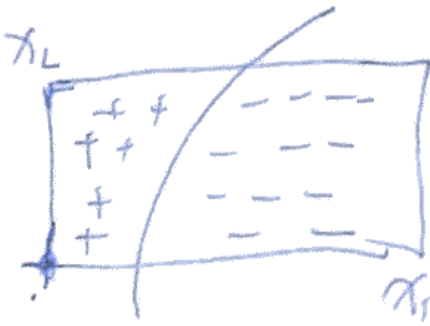
$$\Rightarrow loss(y,\hat{y}_i) = \sum_{i=1}^{m}\left[\ln(1+e^{-y_i\cdot\theta^T x})\right]$$

⑥ 梯度上升：沿似然函数正梯度上升，本质上与沿损失函数负梯度下降相同 → 见⑬：还等价于 $\hat{y}_{(i)}$ 与 $y_{(i)}$ 相对熵最小

⑦ 维度提升：



提升前：$z(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$ 分类界面



提升后：$z(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1 x_2 + \theta_5 x_2^2$ 分类界面

⑧ 线性回归，LR ⟷ 广义线性模型 ⟺ 指数族

⑨ Softmax 回归：K分类，第k类的参数为 $\vec{\theta}_k$，组成二维矩阵 $\theta_{k\times n}$

概率密度：$p(c=k|x;\theta) = \dfrac{e^{\theta_k^T x}}{\sum_{i=1}^{K} e^{\theta_i^T x}}$ 详细见⑩

※ soft-max 并不是直接用 $\dfrac{P_1}{P_1+\cdots+P_k}$ 来建模（找最大值），而是用 $\dfrac{e^{P_1}}{e^{P_1}+\cdots+e^{P_k}}$ 来近似

$\boxed{\text{假设样本间独立，}}$ $\boxed{\text{概率就可以相乘，得到似然函数}}$

似然函数：$L(\theta) = \prod_{i=1}^{m}\prod_{k=1}^{K} p(c=k|x^{(i)};\theta)^{y_k^{(i)}} = \prod_{i=1}^{m}\prod_{k=1}^{K}\left[\dfrac{e^{\theta_k^T\cdot x}}{\sum_{i=1}^{K} e^{\theta_i^T x}}\right]^{y_k^{(i)}}$

对数似然：$J(\theta) = \ln L(\theta) = \sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)}\left(\theta_k^T x - \ln\sum_{i=1}^{K} e^{\theta_i^T x}\right)$

或

$y^{(i)} = k$ 时

$y_0^{(i)}=0,\ y_1^{(i)}=0,\ \cdots y_k^{(i)}=1,\ \cdots y_K^{(i)}=0$

one-hot 编码

$\theta_k = \begin{bmatrix}\theta_1^{(k)}\\ \theta_2^{(k)}\\ \vdots\\ \theta_n^{(k)}\end{bmatrix}$

$\theta_i = \begin{bmatrix}\theta_1^{(i)}\\ \theta_2^{(i)}\\ \vdots\\ \theta_n^{(i)}\end{bmatrix}$ $X = \begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{bmatrix}$

用于类别 i 的模型参数。 特征

梯度：$\dfrac{\partial J(\theta)}{\partial \theta_k} = (y_k - p(y_k|X;\theta))\cdot X$

沿着这个梯度对似然概率做梯度上升

⑩ 用 $\dfrac{e^{p_i}}{e^{p_1}+e^{p_2}+\cdots+e^{p_k}}$ 替代 $\dfrac{p_i}{p_1+p_2+\cdots+p_k}$ 做为 Soft-Max 建模时 概率密度的依据

$$3<5 \Longleftrightarrow e^3<e^5 \qquad a=\ln\frac{e^3}{e^3+e^5} \sim \frac{3}{3+5}$$

$$\frac{3}{3+5}<\frac{5}{3+5} \qquad \frac{e^3}{e^3+e^5}<\frac{e^5}{e^3+e^5} \qquad b=\ln\frac{e^5}{e^3+e^5} \sim \frac{5}{3+5}$$

⑪ Soft-max 分类数 $k=2$ 时，会退化成 Logistic 回归.

$k=2$ 时，令 $\theta=\theta_1-\theta_2$.

$$p_1=\frac{e^{\theta_1^T x}}{e^{\theta_1^T x}+e^{\theta_2^T x}}=\frac{1}{1+e^{-(\theta_1-\theta_2)^T x}}=\frac{1}{1+e^{-\theta^T x}}$$

就是 LR 用的 sigmod 函数（令 $z=\theta^T x$）

⑫ 约束条件下求极值问题（见下页）

⑫ 约束条件求极值问题：拉格朗日乘子法

☆ 题目例子：投骰子，6个点等概率 $P_1=P_2=\cdots=P_6$，N次投置的均值为2.71828.

下一次投置出现点5的概率

建模：在给定的约束条件下，使信息熵取最大（~~等~~ 等价于 ~~在~~ 在装箱问题中，不违反约束条件的前提下，装到几个箱子的概率尽可能趋于相同）

→ $\boxed{\text{概率④}}$

目标函数：$H(\vec{p})=-\sum_{i=1}^{6}p_i\ln p_i$ （信息熵，期望它能取到最大值）

约束条件：$\sum_{i=1}^{6}p_i=1$ 即 $(1-\sum_{i=1}^{6}p_i)=0$

$\sum_{i=1}^{6}p_i\cdot i=2.71828$ 即 $(2.71828-\sum_{i=1}^{6}i\cdot p_i)=0$

Lagrange函数：$L(\vec{p},\lambda_1,\lambda_2)=-\sum_{i=1}^{6}p_i\ln p_i+\lambda_1(1-\sum_{i=1}^{6}p_i)+\lambda_2(2.71828-\sum_{i=1}^{6}i\cdot p_i)$

对p_i求偏导，令偏导为0，$\frac{\partial L}{\partial p_i}=-\ln p_i-1-\lambda_1-i\cdot\lambda_2=0$ ⊛

$$\Rightarrow\begin{cases}p_i=e^{-1-\lambda_1-i\lambda_2} & ⓐ\\ \lambda_1=-0.0787 & ⓑ\\ \lambda_2=0.2808 & ⓒ\end{cases}$$

ⓐ：由公式⊛得到

ⓑ,ⓒ：把 $p_i=e^{-1-\lambda_1-i\lambda_2}$ 代入到 $J(\lambda)=(\sum_{i=1}^{6}p_i-1)+(\sum_{i=1}^{6}ip_i-E)^2$

对λ_1，λ_2求偏导，得到梯度，

按梯度下降，求出 ~~除~~ λ_1,λ_2 代入ⓐ解出p_i.

如何理解最大熵模型？：见 →