

线性回归模型：

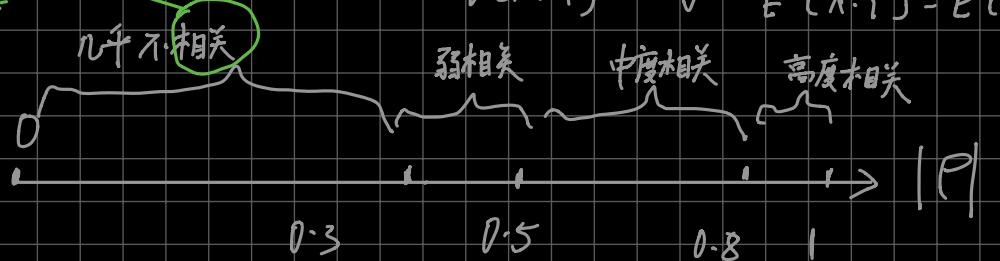
相关性分析：

$$\rho = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}}$$

相关系数

$$\sum (X_i - \bar{X})(Y_i - \bar{Y})$$

不代表没有关系，只是
说非线性相关。



$$D(X) = E(X^2) - E^2(X)$$

$$D(X \cdot Y) = E(X^2 \cdot Y^2) - E^2(X \cdot Y)$$

$$\text{cov}(X, Y) = E(X \cdot Y) - E(X) \cdot E(Y)$$

一元线性回归模型：

$$y = ax + b + \epsilon$$

定义：使得拟合线能够精地捕捉到每一个点，即使误差项 ϵ 的平方 ϵ^2 最小，找到该拟合线（即求参数 a, b ）

直接目的：求解 a, b

目标函数：

$$\text{How: } J(a, b) = \sum_{i=1}^n \epsilon^2 = \sum_{i=1}^n (y_i - [a + b x_i])^2$$

样本点到
拟合线距离

\Rightarrow 求 $J(a, b)$ 取最小值 a, b 的值

\Rightarrow 对 a, b 求偏导，并令偏导数均为 0 即可.

$$\left\{ \begin{array}{l} \frac{\partial J}{\partial a} = 0 \\ \frac{\partial J}{\partial b} = 0 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} a = \bar{y} - b \bar{x} \\ b = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2} \end{array} \right.$$

推导：

$$\text{I. } J(a, b) = \sum_{i=1}^n \frac{1}{n} = \sum_{i=1}^n (y_i - (a + b x_i))^2$$

$$= \sum_{i=1}^n (y_i^2 + a^2 + b^2 x_i^2 + 2ab x_i - 2y_i a - 2y_i b x_i)$$

$$\text{II. } \left\{ \begin{array}{l} \frac{\partial J}{\partial a} = \sum_{i=1}^n (2a + 2b x_i - 2y_i) = 0 \quad \textcircled{1} \\ \frac{\partial J}{\partial b} = \sum_{i=1}^n (2b x_i^2 + 2a x_i - 2y_i x_i) = 0 \quad \textcircled{2} \end{array} \right.$$

$$\text{III. } \Rightarrow \left\{ \begin{array}{l} 2na + 2b \sum_{i=1}^n x_i - 2 \sum_{i=1}^n y_i = 0 \\ 2b \sum_{i=1}^n x_i^2 + 2a \sum_{i=1}^n x_i - 2 \sum_{i=1}^n x_i y_i = 0 \end{array} \right.$$

$$\text{IV. } \Rightarrow \left\{ \begin{array}{l} 2a + 2b \cdot \bar{x} - 2\bar{y} = 0 \quad \textcircled{1} \\ 2b \cdot \frac{\sum_{i=1}^n x_i^2}{n} + 2a \cdot \bar{x} - 2 \cdot \frac{\sum_{i=1}^n x_i y_i}{n} = 0 \quad \textcircled{2} \end{array} \right.$$

$$\left\{ \begin{array}{l} \textcircled{1} \Rightarrow a = \bar{y} - b \bar{x} \\ \textcircled{2} \Rightarrow b = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2} \end{array} \right.$$

$$\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}$$

多元线性回归模型:

$$Y = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{Bmatrix}$$

一元: $y = a + b x + \varepsilon$

多元: $\underbrace{Y}_{n \times p} = \underbrace{X \cdot \beta}_{n \times p} + \underbrace{a}_{n \times 1} + \underbrace{\varepsilon}_{n \times p}$

$$X = \begin{Bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{Bmatrix}_{n \times p}$$

目标函数:

平方项的和可转化为矩阵乘法

$$J(\beta) = \sum \varepsilon^2 = \sum (Y - X\beta)^2$$

$$= (Y - X\beta)' (Y - X\beta)$$

$$= (Y' - \beta' X')(Y - X\beta)$$

$$= Y' \cdot Y - Y' \cdot X\beta - \beta' X' Y + \beta' X' X \cdot \beta$$

求导: $\frac{\partial J}{\partial \beta} = 0 \Rightarrow \beta = \underbrace{(X' \cdot X)^{-1}}_{\text{偏回归系数}} \cdot X' \cdot Y$

偏回归系数

岭回归模型：目的 \Rightarrow 解决线性回归的短板。

$$J(\beta) = \sum (y - X\beta)^2 + \lambda \cdot \|\beta\|_2^2$$

引入 L_2 正则项，也为惩罚项。

$$= \sum (y - X\beta)^2 + \lambda \cdot \beta^2$$

欲使 $J(\beta)$ 最小，①求偏导： $\frac{\partial J}{\partial \beta} = 0 \Rightarrow$

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

(此法与解线性回归 β 值相同)

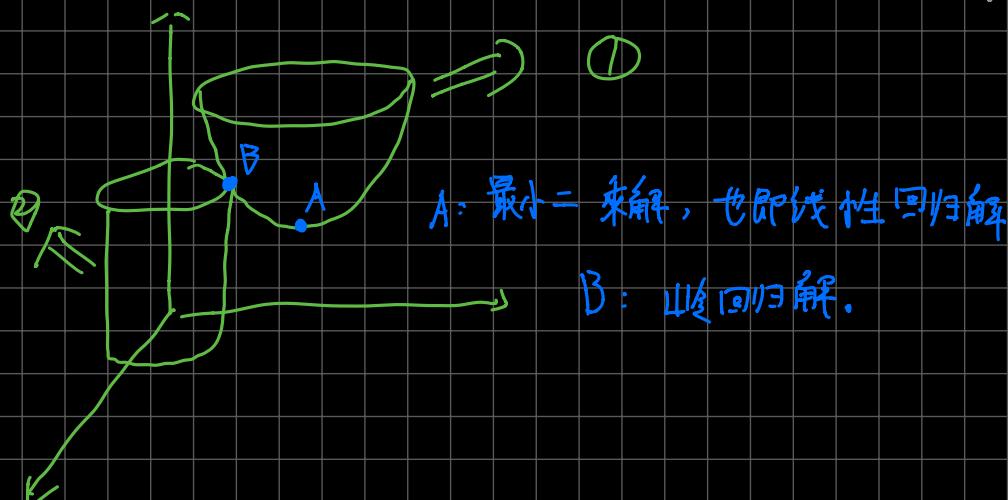
② 凸优化 转化：

目标函数： $J(\beta) = \sum (y - X\beta)^2 + \sum \lambda \cdot \beta^2$

欲求 $\min_{\beta} J(\beta) \xrightarrow{\text{凸优化转化}} \left\{ \begin{array}{l} \text{①} \arg \min \left\{ \sum (y - X\beta)^2 \right\} \\ \text{②} : \sum \beta^2 \leq \gamma \end{array} \right. \Rightarrow$

线性回归
模型求解
方法。
附加约束

几何意义：



Logistic 回归模型:

线性回归预测函数:

$$z = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_p \cdot x_p$$

Logit 变换:

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

令 $h_{\beta}(x) = g(z)$, 则 $h_{\beta}(x)$ 为 Logistic 回归的预测函数.

预测值为 P , 即

$$h_{\beta}(x) = P$$

$$1 - h_{\beta}(x) = 1 - P$$

令 $P(y=1|x; \beta) = P$, 则 $P(y=0|x; \beta) = 1 - P$

y 只有 2 种取值.

优势比/发生比:

$$\text{odds} = \frac{P}{1-P} = \frac{h_{\beta}(x)}{1-h_{\beta}(x)} = e^z$$

② 随机变量表达式:

$$P(y|x; \beta) = h_{\beta}(x)^y \cdot (1 - h_{\beta}(x))^{(1-y)}$$

求解参数 $\beta \Rightarrow$ 极大似然估计法

Step 1: 构造极大似然函数 $L(\beta)$

$$L(\beta) = \prod_{i=1}^n P(y_i|x_i; \beta)$$

$$= \prod_{i=1}^n h_{\beta}(x^{(i)})^{y^{(i)}} (1 - h_{\beta}(x^{(i)}))^{1 - y^{(i)}}$$

Step 2: $L(\beta)$ 对数化:

$$\log(L(\beta)) = \sum_{i=1}^n \log(h_{\beta}(X^{(i)})) \left(y^{(i)} \right) \left(1 - h_{\beta}(X^{(i)}) \right)^{1-y^{(i)}}$$

$$= \sum_{i=1}^n \left[y^{(i)} \cdot \log(h_{\beta}(X^{(i)})) + (1-y^{(i)}) \cdot \log(1-h_{\beta}(X^{(i)})) \right]$$

求 $\log(L(\beta))$ 的 max 值.

Step 3: 构造目标函数 $J(\beta)$

令: $J(\beta) = -\log(L(\beta))$, 即求 $J(\beta)$ 的最小值时的 β , ∵ 即与此前回归求法类似, 但并非直接对 β 求偏导并使 $\frac{\partial J}{\partial \beta} = 0$.

原因: 未知参数个数 > 方程组个数.

Step 4: 木梯度下降求解参数

$$\beta_j := \beta_j - \alpha \cdot \underbrace{\frac{\partial J(\beta)}{\partial \beta_j}}_{\text{学习率}} \quad (j=1, 2, \dots, p)$$



优势比 odds 的应用 \Rightarrow 比较多系数变化的影响.

$$\frac{\text{发生比1}}{\text{发生比2}} = \frac{\text{odds}(V+1)}{\text{odds}(V)} = \frac{e^{\beta_0} \cdot e^{\beta_1 \cdot G} \cdot e^{\beta_2(V+1)}}{e^{\beta_0} \cdot e^{\beta_1 \cdot G} \cdot e^{\beta_2(V)}} = e^{\beta_2}$$

∴ 当 V 变量增加1时, 发生比增大 e^{β_2} 倍.

评价 Logistic 模型：

I. 混淆矩阵:

预测值			实际值	错误预测
	良性--0	恶性--1		
良性--0	A, True Negative	B, False Negative	A+B, Predict Negtive	
恶性--1	C, False Positive	D, True Positive		C+D, Predict Positive
		A+C, Actual Negative		B+D, Actual Positive

A: 表示正确预测负例的样本个数，用TN表示。

B: 表示预测为负例但实际为正例的个数，用FN表示。

C: 表示预测为正例但实际为负例的个数，用FP表示。

D: 表示正确预测正例的样本个数，用TP表示。

非常关心：准确率：表示正确预测的正负例样本数与所有样本数量的比值，即 $(A+D)/(A+B+C+D)$ 。

迫切需要：正例覆盖率：表示正确预测的正例数在实际正例数中的比例，即 $D/(B+D)$ 。

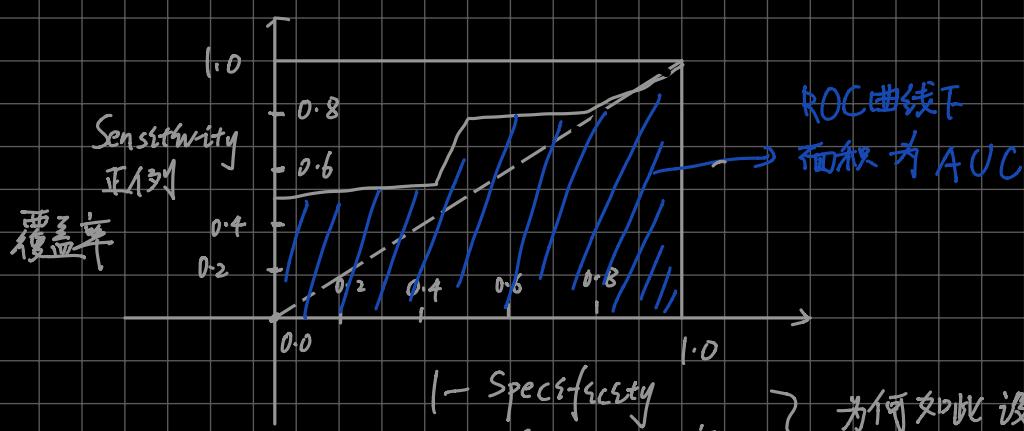
追着的谈为正：负例覆盖率：表示正确预测的负例数在实际负例数中的比例，即 $A/(A+C)$ 。

正例命中率：表示正确预测的正例数在预测正例数中的比例，即 $D/(C+D)$ ，

例如，因此恶性设为

正例。

II. ROC 曲线：



为何如此设置？正例覆盖率↑则
负例覆盖率↓，而我们想让二者同向
变化。

AUC 越大，则模型越好， $AUC > 0.8$ 时模
型即可接受

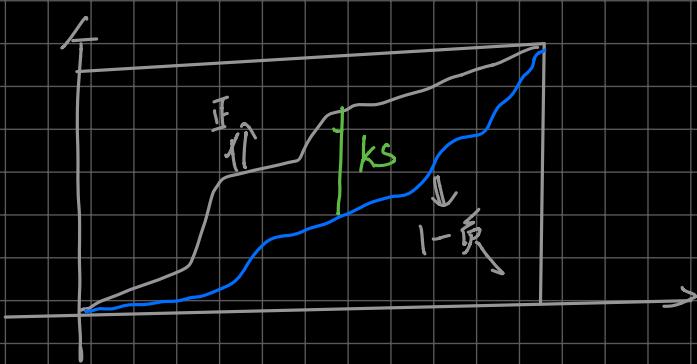
\Rightarrow 因此我们希望正例覆盖率越大，负例覆盖率越大，此时 AUC 很

大，模型好。

III. KS 曲线：

$$KS = \text{Sensitivity} - (1 - \text{Specificity})$$

$$= \underbrace{\text{Sensitivity} + \text{Specificity} - 1}_{\uparrow, KS \uparrow}$$



KS 值越大越好。