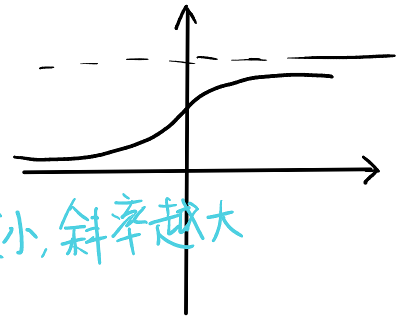
Logistic Regression 逻辑回归

1. Logistic Distribution

$$F(x) = \frac{1}{1+\exp(-(x-\mu)/r)}$$



r越小，斜率越大

· Output Y (class label)

Binary response $Y \in \{0,1\}$

· $$P(Y=1|X) = \frac{\exp(x^T\beta)}{1+\exp(x^T\beta)}$$

$$P(Y=0|X) = \frac{1}{1+\exp(x^T\beta)}$$

$\Phi(x^T\beta)$ 正态分布 cdf

if $x^T\beta \to +\infty$ $P(Y=1|X) \to 1$

$\to -\infty$ $P(Y=1|X) \to 0$

2. Model Estimation

$$L(\beta) = \prod_{i=1}^{N} p(x_i,\beta)^{y_i} \cdot \{1-p(x_i;\beta)\}^{1-y_i}, \quad p(x_i,\beta) = \frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)}$$

$$l(\beta) = \sum_{i=1}^{N} (y_i x_i^T\beta - \log(1+\exp(x_i^T b)))$$

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{N} y_i x_i - \frac{\exp(x_i^T b)}{1+\exp(x_i^T b)} \cdot x_i \qquad E\left\{\frac{\partial l}{\partial \beta}\Big|_{\beta=\beta_0}\right\} = \sum_{i=1}^{n} x_i E\{y_i - p(x_i,\beta)\} = 0$$

$$= \sum_{i=1}^{N} x_i (y_i - p(x_i,\beta))$$

· Optimization Newton-Raphson algorithm

$$\beta^{new} = \beta^{old} - \Big(\underbrace{\frac{\partial^2 l}{\partial\beta\partial\beta^T}}_{\text{二阶信息}}\Big)^{-1} \frac{\partial l(\beta)}{\partial \beta} \qquad \frac{\partial^2 l}{\partial\beta\partial\beta^T} = -\sum x_i x_i^T p(x_i,\beta)\{1-p(x_i,\beta)\}$$

Let $\sum x_i x_i^T = X^TX$, $W = \mathrm{diag}\{w_1, \cdots, w_n\}$ $w_i = p(x_i,\beta)(1-P(x_i,\beta))$

$$\Rightarrow \frac{\partial^2 l}{\partial\beta\partial\beta^T} = X^TWX$$

3. Multi-nomial Logistic Regression model

$$Y=\{1,2,\cdots,K\} \quad P(Y=k\mid X=x)=\frac{\exp(\beta_k^T x)}{1+\sum_{i=1}^{k}\exp(\beta_i^T x)}$$

4. Model Evaluation

1. 总体衡量

(1) Accuracy（精度） $\frac{TP+TN}{P+N}$

(2) Error Rate（错分率） $\frac{FP+FN}{P+N}$

		Predicted Yes	Predicted No
Actual	Yes	TP	FN
	No	FP	TN

2. 查准率与查全率

(1) Precision（查准率）$=\frac{TP}{TP+FP}$

(2) Recall（查全率）$=\frac{TP}{TP+FN}$

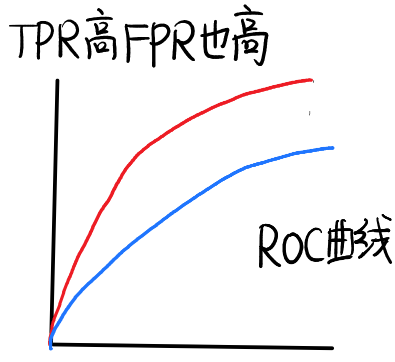
(3) F_1度量（precision和recall的调和平均）$=\frac{2\times P\times R}{P+R}$

(4) F_β度量：$F_\beta=\frac{(1+\beta^2)\times P\times R}{\beta^2 P+R}$

3. ROC 曲线 & AUC

评估方法不依赖于阈值

原理：如果一个分类器能尽量多地将正例排在负例之前，则认为它有较好的分类能力



ROC（Receiver Operation Characteristic 曲线）

横轴 假正例率 False positive rate $FPR=\frac{FP}{TN+FP}$

纵轴 真正例率 True positive rate $TPR=\frac{TP}{TP+FN}$

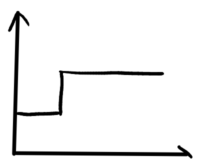
AUC（Area Under Curve） ROC曲线下的面积

有限样本下绘制的ROC曲线

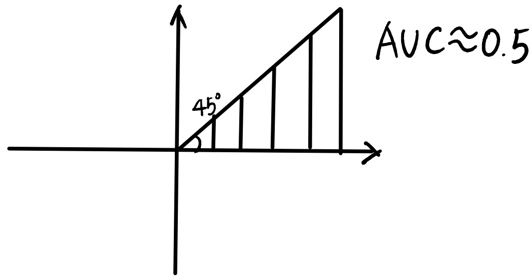
(1) 将阈值设成最大，此时TPR=FPR=0，此时无预测的正例

(2) 将样本的预测值从高到低排序，将阈值依次设成预测值，即将每个样本标成一个点

(3) 设前一个点 (x,y) $\begin{cases}\text{当前预测真正例} & (x,\ y+\frac{1}{m^+})\\ \text{当前预测假正例} & (x+\frac{1}{m^-},\ y)\end{cases}$



(4) 假如一个分类器随机预测正例负例



AUC的取值只与排序有关

排序"损失"(rank loss)

假设有 m^+ 正例和 m^- 负例，令 D^+ 与 D^- 分别代表正例与反例的集合

$$\ell_{rank}=\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left[I\left(\underbrace{f(x^+)}_{\text{模型预测值}}<f(x^-)\right)+\frac{1}{2}I\left(f(x^+)=f(x^-)\right)\right]$$

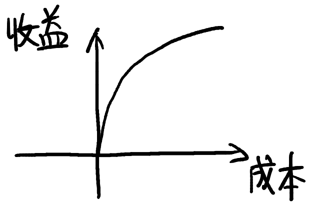
若正例预测值小于负例，则记一个罚分；若相等，则记0.5罚分

$AUC=1-\ell_{rank}$

4. 成本-收益线

成本的质量(覆盖率) $\frac{TP+FP}{P+N}$ (预测的正值比例)

收益的质量(捕获率) Recall (查全率)



5. 多次采样

由于五折交叉验证存在随机性，一般重复K次实验取度量的平均值

6. 类别不均衡

(1) 设置域值 $\frac{p_i}{1-p_i}>\frac{m^+}{m^-}$ (不再是 $\frac{p_i}{1-p_i}\geq 1$)

(2) 过采样(Over-sampling) SMOTE算法

(3) 欠采样(Under-sampling) Easy-Ensemble算法

· 广义线性模型

(1) 指数分布簇：$f(y|\theta,\psi)=\exp\left\{\frac{yb(\theta)-c(\theta)}{a(\psi)}+d(y,\psi)\right\}$

θ 典型参数：与 y 的均值 μ 有关

ψ 刻度参数：与 y 的方差 σ^2 有关