

# CLASSIFICATION: LOGISTIC REGRESSION

## 1. Logistic Regression Model

Logistic distribution. Suppose  $X$  is a continuous random variable following logistic distribution. Then the distribution function and the density function take the forms as

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

$$f(x) = \frac{\exp\{-(x - \mu)/\gamma\}}{\gamma\{1 + \exp(-(x - \mu)/\gamma)\}^2},$$

where  $\mu$  is a position parameter and  $\gamma > 0$  is a shape parameter.

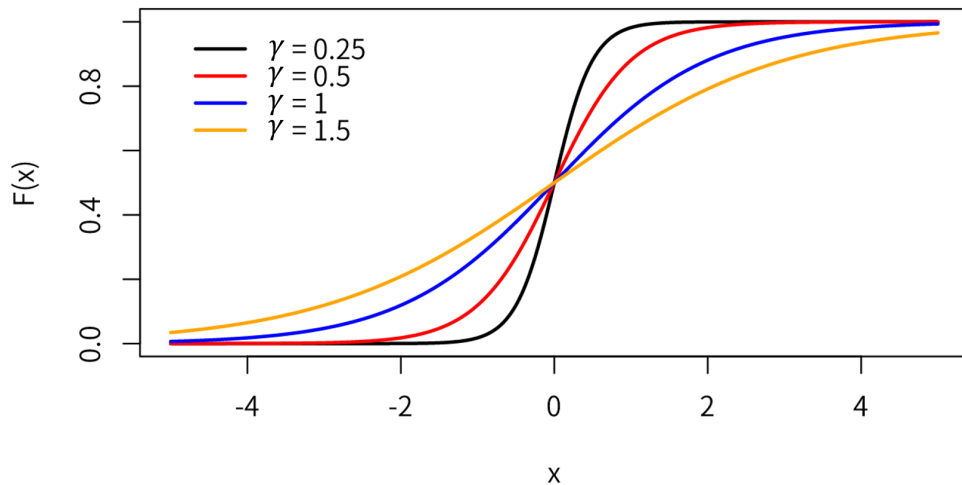


Figure 1:  $F(x)$  Curve with Different  $\gamma$

Output (class label):  $Y$ .

For binary responses ( $Y \in \{0, 1\}$ ):

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

Q: Could you give  $P(Y = 0|X = x)$ ?

Comment: The form of  $P(Y = 0|X = x)$  is similar to the c.d.f of logistic distribution.

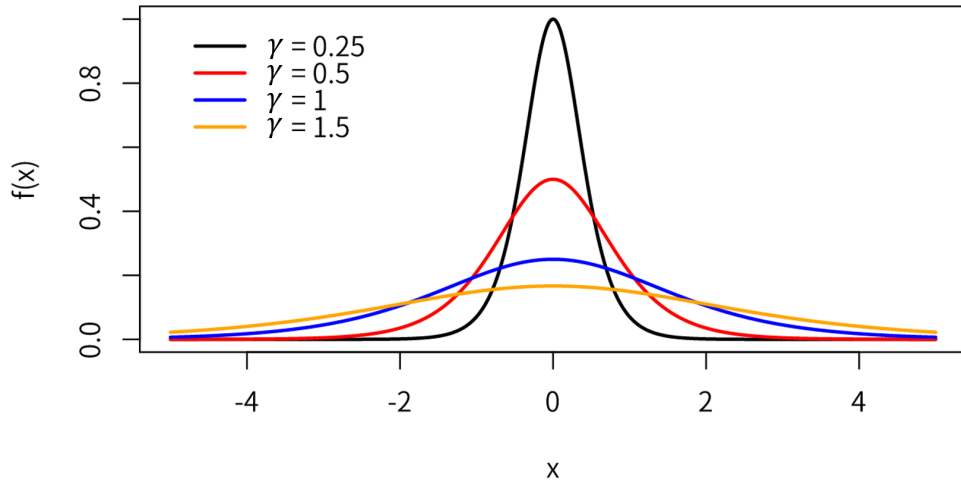


Figure 2:  $f(x)$  Curve with Different  $\gamma$

Log-odds:

$$\log \left( \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} \right) = x^\top \beta. \quad (1.1)$$

Particularly, if  $x^\top \beta \rightarrow +\infty$ , then  $P(Y = 1|X = x) \rightarrow 1$ ; if  $x^\top \beta \rightarrow -\infty$ , then  $P(Y = 1|X = x) \rightarrow 0$ .

## 2. Model Estimation

Suppose for the  $i$ th subject we observe  $x_i$  and  $y_i$ . Let  $p(x_i; \beta) = P(Y = 1|X = x_i)$ .

Maximum likelihood estimation:

$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^N \left\{ y_i \log p(x_i; \beta) + (1 - y_i) \log(1 - p(x_i; \beta)) \right\} \\ &= \sum_i \left\{ y_i x_i^\top \beta - \log(1 + \exp(x_i^\top \beta)) \right\} \end{aligned}$$

Q: derive the blue part by yourself.

To maximize the log-likelihood, we set its derivatives to zero. The score equations are

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^N x_i \{y_i - p(x_i; \beta)\} = 0.$$

Optimization: Newton-Raphson algorithm

$$\beta^{new} = \beta^{old} - \left( \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^\top} \right)^{-1} \frac{\partial \ell(\beta)}{\partial \beta}.$$

where

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^\top} = - \sum_i x_i x_i^\top p(x_i; \beta) \{1 - p(x_i; \beta)\}. \quad (2.1)$$

Define  $\mathbf{W}$  as a  $N \times N$  diagonal matrix of weights with the  $i$ th diagonal element  $p(x_i; \beta^{old})(1 - p(x_i; \beta^{old}))$  and  $\mathbf{p} = (p(x_1; \beta), \dots, p(x_N; \beta))^\top$ . Then we have,

$$\begin{aligned} \frac{\partial \ell(\beta)}{\partial \beta} &= \mathbf{X}^\top (\mathbf{y} - \mathbf{p}), \\ \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^\top} &= -\mathbf{X}^\top \mathbf{W} \mathbf{X}. \end{aligned}$$

Then the Newton step is

$$\begin{aligned} \beta^{new} &= \beta^{old} + (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{y} - \mathbf{p}), \\ &= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \{ \mathbf{X} \beta^{old} + \mathbf{W}^{-1} (\mathbf{y} - \mathbf{p}) \}, \\ &= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{z}. \end{aligned}$$

This algorithm is then referred to as *iteratively reweighted least squares*.

Other algorithms can be found in the 附录 A & B 《统计机器学习》。

Comment:

- (1)  $\hat{\beta}$  converge in distribution to  $N(\beta, (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1})$ . The inference can be done.
- (2) Likelihood Ratio Test:

$$\begin{aligned} LR &= -2 \max_{\beta_0} \ell(\beta_0, \beta_1 = 0) + 2 \max_{\beta_0, \beta_1} \ell(\beta_0, \beta_1) \\ &= DEV_0 - DEV_1 \end{aligned}$$

LR asymptotically ( $N$  is large enough) follows Chi-square distribution with degree of freedoms  $p_0$ , where  $p_0$  is number of parameters in  $\beta_1$ .

### 3. Multi-nominal Logistic Regression Model

If  $Y \in \{1, \dots, K\}$ , then the multi-nominal logistic regression model takes the form,

$$P(Y = k|X = x) = \frac{\exp(\beta_k^\top x)}{1 + \sum_{k=1}^{K-1} \exp(\beta_k^\top x)}, \quad k = 1, 2, \dots, K - 1. \quad (3.1)$$

### 4. Model Evaluation

		Predicted class		
		yes	no	Total
Actual class	yes	TP	FN	P
	no	FP	TN	N
Total		P'	N'	P + N

Figure 3: Classification evaluation: Confusion matrix.

Measure	Formula
accuracy, recognition rate	$\frac{TP + TN}{P + N}$
error rate, misclassification rate	$\frac{FP + FN}{P + N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP + FP}$
$F$ , $F_1$ , $F$ -score, harmonic mean of precision and recall	$\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$
$F_\beta$ , where $\beta$ is a non-negative real number	$\frac{(1 + \beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$

Figure 4: Evaluation measures.

## 1. 总体衡量

(1) Accuracy (精度):

$$\frac{TP + TN}{P + N}$$

(2) Error rate (错分率):

$$\frac{FP + FN}{P + N}.$$

## 2. 查准率、查全率与F1

在“抓坏蛋”的分类任务中，我们更关心“预测的坏蛋是否是真的坏蛋”，以及“有多少坏蛋被挑出来了”；而不在乎“是否把好人预测成了好人”。这种情况下，precision（查准率）和recall(查全率)更能代表这类需求的性能度量。它们定义如下：

$$\begin{aligned} precision &= \frac{TP}{TP + FP}, \\ recall &= \frac{TP}{TP + FN}. \end{aligned}$$

查准率高，说明预测为坏蛋的人中坏蛋真的很多；查全率高，说明真实的坏蛋中有很多被模型挑出来了。这两个指标难以两全，例如，如果希望查全率高，则可以把全部人都预测成坏蛋，但这时候查准率往往不高。希望查准率高，则可以把阈值设的高一点，使得预测为坏蛋的人大概率是真的坏蛋。

这里涉及阈值选择，一般阈值越高，则查准率高；阈值越低，查全率高。

F1度量(precision和recall的调和平均):

$$F1 = \frac{2 \times P \times R}{P + R}$$

有时候precision和recall的重要程度不同。比如，在癌症筛查中，可以允许误诊，但是希望能够尽量准确查出癌症，此时，查全率更重要；在一些营销场景中，

由于每一次营销都要付出成本，因此希望查准率更高。 $F_\beta$ 度量：

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R} \quad (4.1)$$

$F_\beta$ 是加权调和平均：

$$\frac{1}{F_\beta} = \frac{1}{1 + \beta^2} \left( \frac{1}{P} + \frac{\beta^2}{R} \right) \quad (4.2)$$

$\beta > 1$ ， recall更重要； $\beta < 1$ ， precision更重要。

### 3. ROC曲线及AUC

之前叙述的方法依赖于阈值（threshold）的设定，因此，阈值设置的好坏往往影响评估度量的差异。这显然是不合理的。

事实上，根据预测概率，我们可以对样本进行排序。直观上，如果一个分类器，能够尽量多的把正样本排序在负样本之前，那么这个分类器具有很好的分类能力。这个排序情况与阈值的设置无关。ROC曲线正是从这个角度出发设计的。

ROC全称是（Receiver Operating Characteristic）[受试者工作特征曲线]。这个名字很怪，跟它历史有关：ROC是由二战中的电子工程师和雷达工程师发明的，用来侦测战场上的敌军载具（飞机、舰船），也就是信号检测理论。

ROC曲线的横轴是“假正例率”（False Positive Rate, FPR），纵轴是“真正例率”（True Positive Rate, TPR）。

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{TN + FP}.$$

一般TPR高的时候FPR也比较高。如果一个分类器的ROC曲线将另一个包裹，则说明后者优于前者。

一般采用采取ROC曲线下的面积AUC（Area Under Curve）来判断分类器性能的优劣。这个面积该如何计算？

**AUC取值只与排序有关**。假设有 $m^+$ 个正例和 $m^-$ 个负例，令 $D^+$ 与 $D^-$ 分别表示正

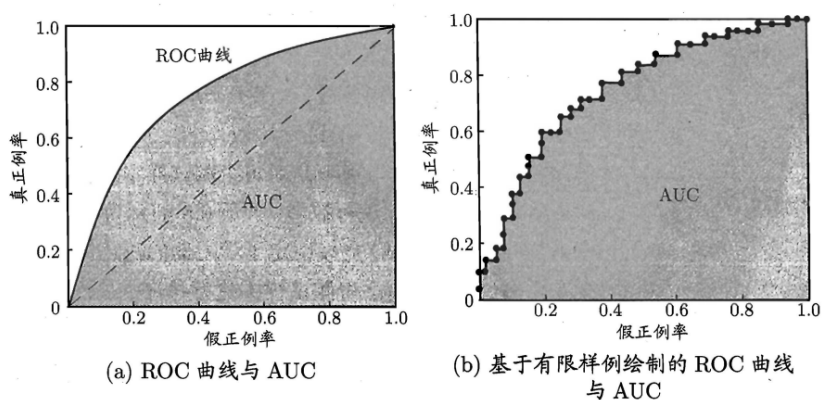


图 2.4 ROC 曲线与 AUC 示意图

随机分类的AUC的期望为0.5

Figure 5: ROC和AUC.

例、反例集合。定义排序“损失”如下：

$$\ell_{rank} = \frac{1}{m^+m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left( I(f(x^+) < f(x^-)) + \frac{1}{2} I(f(x^+) = f(x^-)) \right) \quad (4.3)$$

理解：若正例的预测值小于反例，则记一个“罚分”，若相等，则记0.5个罚分。

$$AUC = 1 - \ell_{rank}. \quad (4.4)$$

### 3. 成本收益曲线

成本的度量（覆盖率）：

$$\frac{TP + FP}{P + N}.$$

收益的度量（捕获率）：Recall(也就是查全率)

### 4. 多次度量

由于每次抽样时测试集存在随机性，一般重复K次试验后取平均值作为度量。

也可以通过假设检验的方式比较两个分类器的性能优劣。具体见西瓜书2.4节。

## 5. 类别不均衡

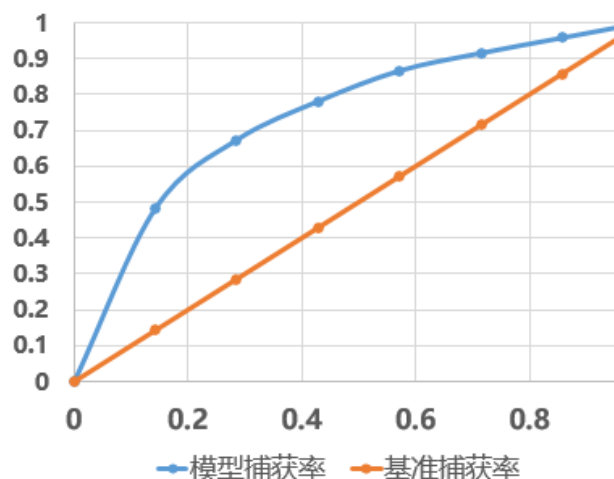


Figure 6: 成本收益曲线.

1. 设置阈值  $\frac{p_i}{1-p_i} > \frac{m^+}{m^-}$ , 则预测为正例 (不再使用1作为cutoff)
2. 过采样(Oversampling): 例如: SMOTE算法, 对正例x进行插值产生新正例
3. 欠采样(Undersampling): 例如: EasyEnsemble算法, 将反例划分为几个子集, 分别学习, 在利用集成学习的方式汇总结果。

## 6. 广义线性模型

1. 指数分布族

$$f(y|\theta, \psi) = \exp \left\{ \frac{yb(\theta) - c(\theta)}{a(\psi)} + d(y, \psi) \right\}.$$

$\theta$ : 典型参数, 与 $y$ 的均值 $\mu$ 有关

$\psi$ : 刻度参数, 与方差有关

举例: 正态分布

$$f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(y - \mu)^2}{2\sigma^2} \right\}$$



泊松分布:

$$f(y|\mu) = \frac{\exp(-\mu)\mu^y}{y!}$$

## 2. 广义线性模型 (Generalized Linear Model)

(1) 因变量 $y$ 的分布为指数族分布, 均值为 $\mu$

(2) 系统成分:  $\eta = x^\top \beta$ .

(3) 链接函数:  $g(\mu) = x^\top \beta$ , 其中 $g(\cdot)$ 为一对一、连续可导的变换。

对于逻辑回归 (二项分布 $n = 1$ ): logit链接函数

$$x^\top \beta = \log \left( \frac{\mu}{1 - \mu} \right) \tag{6.1}$$

对于计数变量: 对数链接函数:  $\log(\mu) = x^\top \beta$ .

其他回归形式: 参考 狗熊会—回归五式。