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

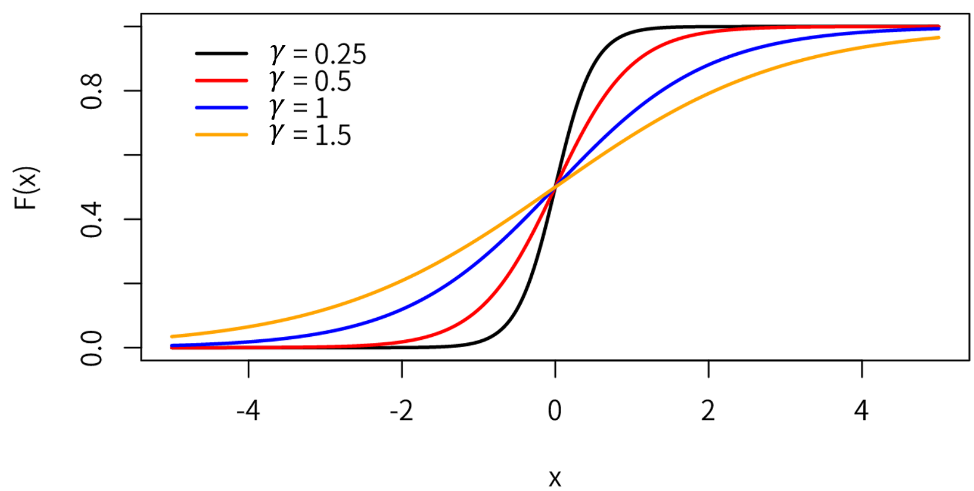

Figure 1: $F(x)$ Curve with Different γ

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 $P(Y=0|X=x)$?

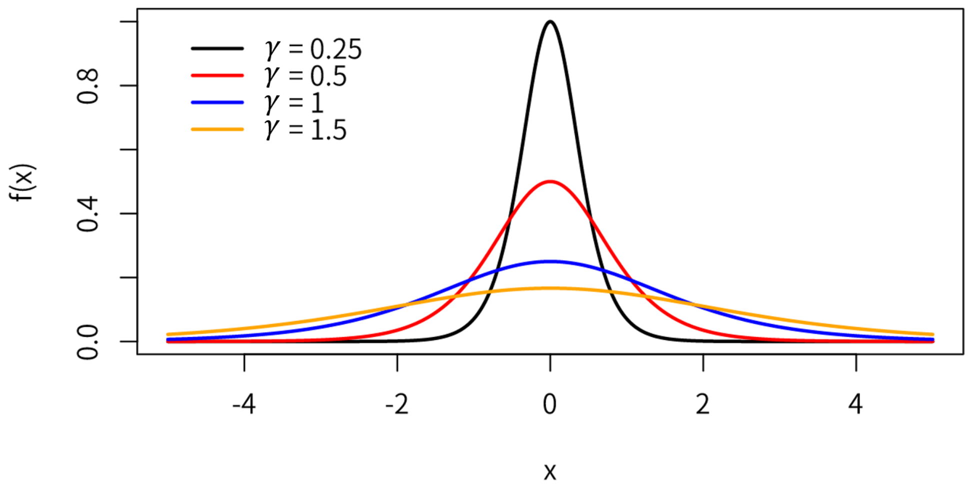

Figure 2: $f(x)$ Curve with Different γ

Log-odds:

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

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

2. Model Estimation

Suppose for the ith subject we observe x_i and y_i. Let $p(x_i;\beta) = P(Y=1|X=x_i)$. Maximum likelihood estimation:

$$\ell(\beta) = \sum_{i=1}^{N}\Big\{y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta))\Big\}$$

Q: derive the blue part by yourself.

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

$$\frac{\partial \ell(\beta)}{\partial \beta} = ??? = 0$$

Optimization: Newton-Raphson algorithm

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

where

$$\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^\top} = ??? \tag{2.1}$$

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

$$\frac{\partial \ell(\beta)}{\partial\beta} = ???$$
$$\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^\top} = ???$$

Then the Newton step is

$$\beta^{new} = \beta^{old} + ???$$

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

Question*: Write Newton-Raphson algorithm to estimate logistic regression by y-ourself.

Generate $X = (1, X_1, X_2)$, where $X_j \sim N(0, I_N)$.

Set true parameter $\beta = (0.5, 1.2, -1)^\top$.

Set $N = 200, 500, 800, 1000$.

Estimate β using NR algorithm for $R = 200$ times. For each j, draw $(\widehat{\beta}_j^{(r)} - \beta_j)$ in boxplot for $N = 200, 500, 800, 1000$. Submit your code (with detailed comments) + report your plot & findings in pdf.

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

Comment:

(1) $\widehat{\beta}$ 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 β_1.

3. Multi-nominal Logistic Regression Model

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

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

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.

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

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	$\frac{2 \times precision \times recall}{precision + recall}$
F_β, where β is a non-negative real number	$\frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$

Figure 4: Evaluation measures.

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：

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

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_β度量：

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

F_β是加权调和平均:

$$\frac{1}{F_\beta} = \frac{1}{1+\beta^2}\Big(\frac{1}{P} + \frac{\beta^2}{R}\Big) \tag{4.2}$$

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、船舰），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。

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：

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

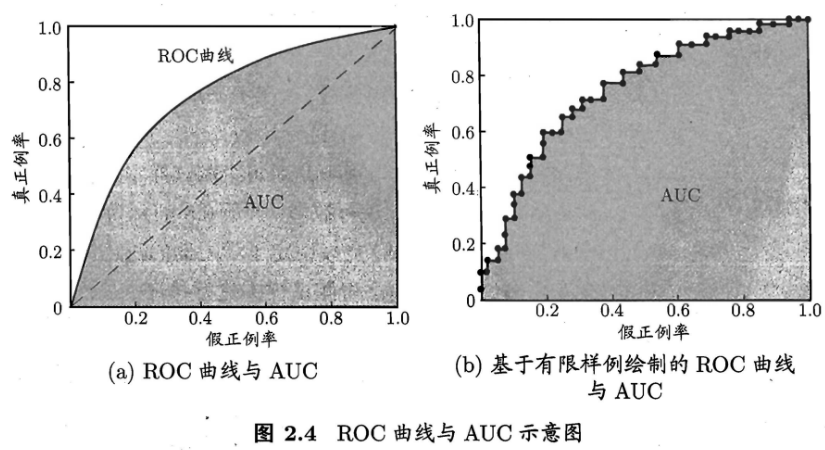

Figure 5: ROC和AUC.

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。

$$AUC = 1 - \ell_{rank}. \tag{4.4}$$

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。

5. 类别不均衡

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。

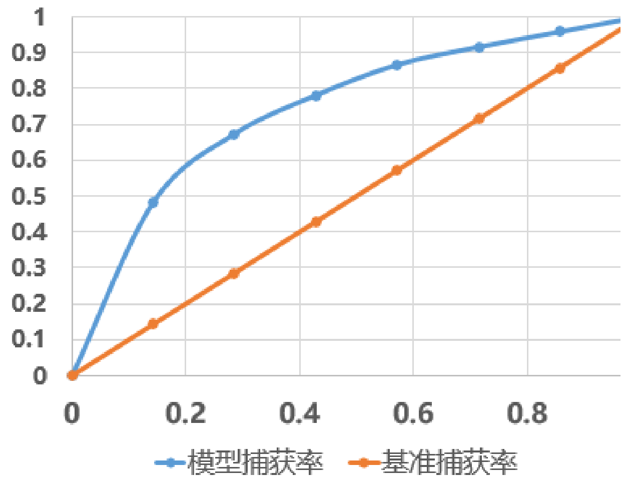

Figure 6: 成本收益曲线.

6. 广义线性模型

1. 指数分布族

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

θ: 典型参数, 与y的均值μ有关

ψ : 刻度参数，与方差有关

举例：正态分布

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

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μ

(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：$\log(\mu) = x^\top\beta$.