机器学习基础 | 分类模型评估指标

原创 AhongPlus dataxon 2019-06-09

收录于话题
#机器学习

4个

在处理机器学习的分类问题中，我们需要评估分类结果的好坏以选择或者优化模型，本文总结二分类任务中常用的评估指标。

对于多分类任务的评估指标，可以参考：

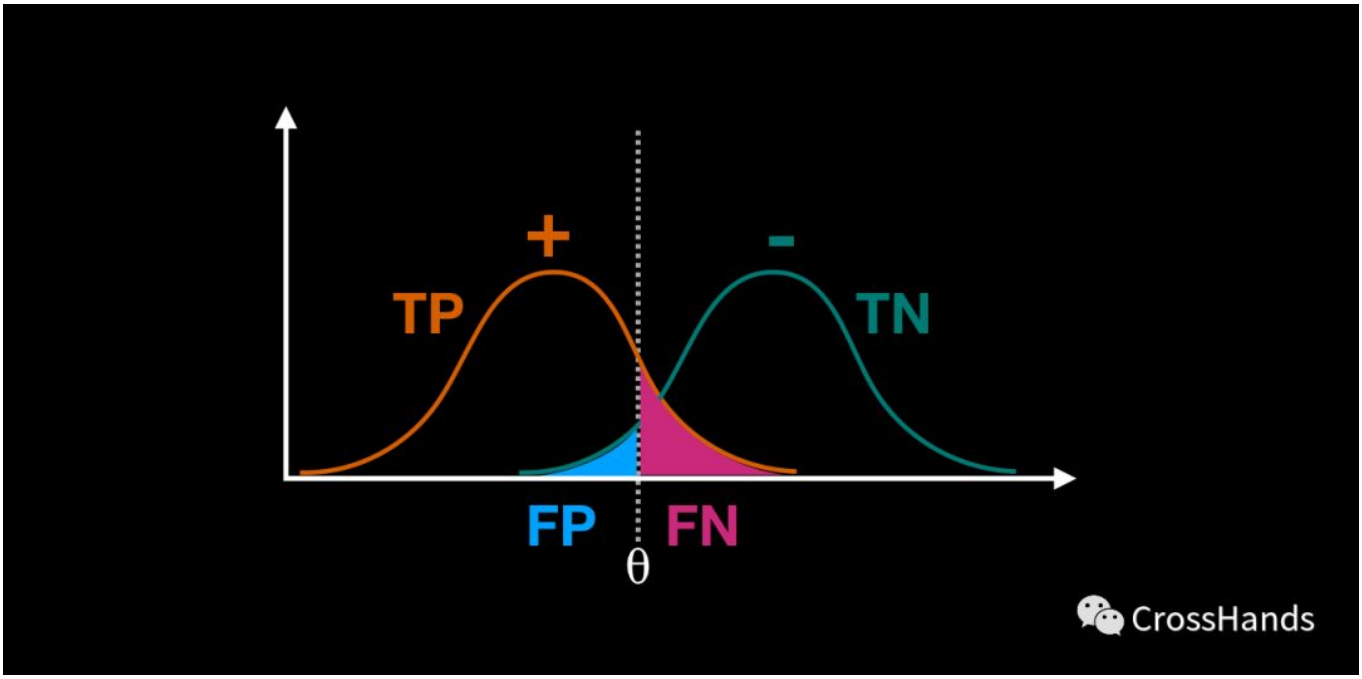
https://turi.com/learn/userguide/evaluation/classification.html#accuracy

先从我们最熟知的混淆矩阵(confusion matrix)说起。

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	True positive	False positive, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	False negative, Type II error	True negative	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

图源：https://en.wikipedia.org/wiki/Confusion_matrix

鉴于混淆矩阵看着比较抽象，可以参考下图



常用的评估指标可以分为3类：

1. 成对指标，包括正确率(精度)&错误率，Precision&Reall，TPR(Sentitivity)&TNR(Specificity)等;
2. 综合指标，包括F-Score，MCC，BCR等；
3. 图形指标，包括ROC以及延伸得到的Gini、AUC、Lift\Gain曲线、代价曲线等；

1 成对指标

1.1 错误率和正确率

错误率定义为分类错误的样本数占样本总数的比例

$$Err = \frac{FP + FN}{N_{sample}}$$

正确率(精度)定义为分类正确的样本数占总数的比例

$$Acc = 1 - \frac{FP + FN}{N_{sample}} = \frac{TP + TN}{N_{sample}}$$

注意：N_{sample}表示样本总数(即confusion matrix中TP, FP, TN, FN之和)。

1.2 Precision、Recall

Precision(准确率、查准率)，即判断为正例的样本中有多大比例是真的正例。

$$P = \frac{TP}{TP + FP}$$

Recall(召回率、查全率)，即正例样本中有多大比例的正例被发现(判定为正例)，该指标也称为True Positive Rate(TPR)、Sensitivity.

$$R = \frac{TP}{TP + FN}$$

考虑到FP和FN的关系是I类错误和II类错误的关系，会此消彼长，故Precision和Recall也有这种关系。

对比多个模型的表现时，可以用P-R图(横轴为Recall，纵轴为Precision).

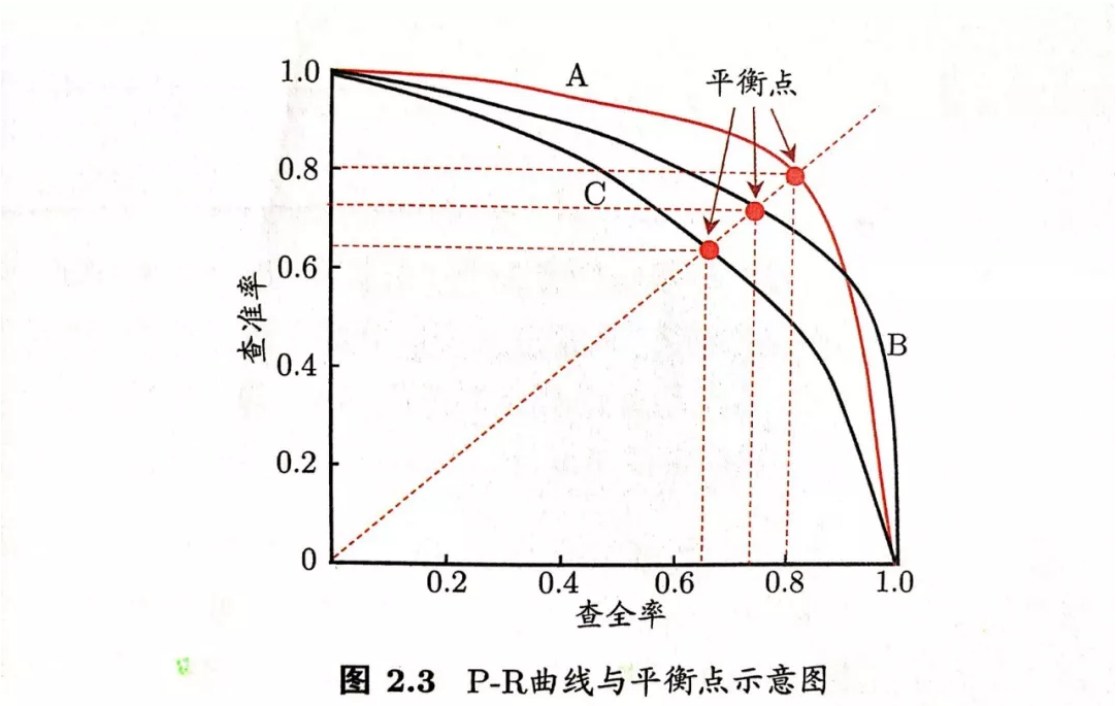


图 2.3 P-R曲线与平衡点示意图

截图来自《机器学习》周志华，更多信息可以参考本书2.3节

1.3 TPR(Sensitivity)、TNR(Specificity)

TPR(True Positive Rate)，正例样本中被正确判定为正例的样本数比例，该指标也称为Sensitivity(敏感度)。

$$TPR = Sensitivity = \frac{TP}{TP + FN}$$

TNR(True False Rate)，指负例样本中被正确判定为负例的样本数比例,该指标也称为Specificity(特异度)。

$$TNR = Specificity = \frac{TN}{TN + FP}$$

2 综合指标

2.1 F-Score

假设我们要判断人群中的好人(正例)和坏人(负例)，如果我们的关注点是“不能冤枉好人”，那么就要尽可能把好人识别出来(判断为好人的标准趋于宽松，坏人也可能被识别为好人)，此时的Precision会趋于更小，Recall会趋于更大；当我们关注的是“不能放过坏人”(比如风控业务中，“坏”客户造成的业务损失很大)，此时判断好人的标准更加严格，更多的“真”好人会被纳入“嫌疑对象”(判定为负例)，此时的Precision会趋于更大，但是Recall会降低；如果我们“尽可能既不能冤枉好人，又不能放过坏人”，那么就需要

在Precision和Recall中取得平衡，此时可以看F1-Score上的表现(不过对于正负例样本不均衡的情况下，F1-Score表现并不好)。

F1-Score是Precision和Recall的调和平均值，即

$$\frac{1}{F_1} = \frac{1}{2} \cdot (\frac{1}{P} + \frac{1}{R})$$

由此可推导得到

$$\begin{aligned} F_1 &= \frac{2 \times P \times R}{P + R} \\ &= \frac{2TP}{2TP + FP + FN} \end{aligned}$$

更一般地，某些场景下关注Precision和Recall的权重不同

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

当$\beta > 1$时，Recall的权重更大，$\beta < 1$时Precision的权重更大。

2.2 Matthews Correlaton Coefficient

简称MCC(马修斯相关系数，Brian W. Matthews, 1975)，

更多参考：https://en.wikipedia.org/wiki/Matthews_correlation_coefficient

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FN)(TN + FP)}}$$

从公式中可以看出MCC和2*2列联表的卡方检验很相近。MCC的值域为[-1,1].

MCC的好处是:

- 正例和负例的不同划分对于最终结果没有影响

  ① TP = 0, FP = 0; TN = 5, FN = 95.

  ② TP = 95, FP = 5; TN = 0, FN = 0.

  这两种条件下(正例和负例的划分不一样)得到的F1-Score差异非常大(①中为0，②中为0.97)，从这里还可以看出F1-Score在正负例样本比例差异不一致的情况下会高估模型的分类效果。
- 综合考虑了正负例样本比例不一致的情况

  TP = 90, FP = 4; TN = 1, FN = 5. 这种条件下得到的分类正确率(Acc)为0.91，F1-Score为0.95，MCC得到的值为0.135. 例如风控业务中“坏”用户占整体用户的比例很小，如果看正确率或者F1-Score那就入坑了，此时MCC可能更合适。

2.3 Balanced Classification Rate

简称BCR，BCR为正例样本和负例样本中各自预测正确率的均值。

$$BCR = \frac{1}{2}(TPR + TNR)$$

与BCR对应的是BER(Balanced Error Rate)，也称之为Half Total Error Rate(HTER).

$$BER = 1 - BCR$$

同MCC一样，正负例的标签选择对BCR的计算没有影响，而且一定程度上也能克服正负例样本不均衡导致的评估指标虚高。

3 图形指标

3.1 ROC、AUC

在分类模型中对样本归属类别的判断通常不是直接得到0或者1，而是一个连续的值区间(比如Logistic回归得到的预测值落在概率区间[0,1])，然后通过划定阈值来判断正例或者负例(比如概率>=0.5判定为正例)。

如果要看分类模型在不同决策阈值下的表现如何，则可以借助ROC曲线(Receiver Operating Characteristic Curve，受试者操作特征曲线)。

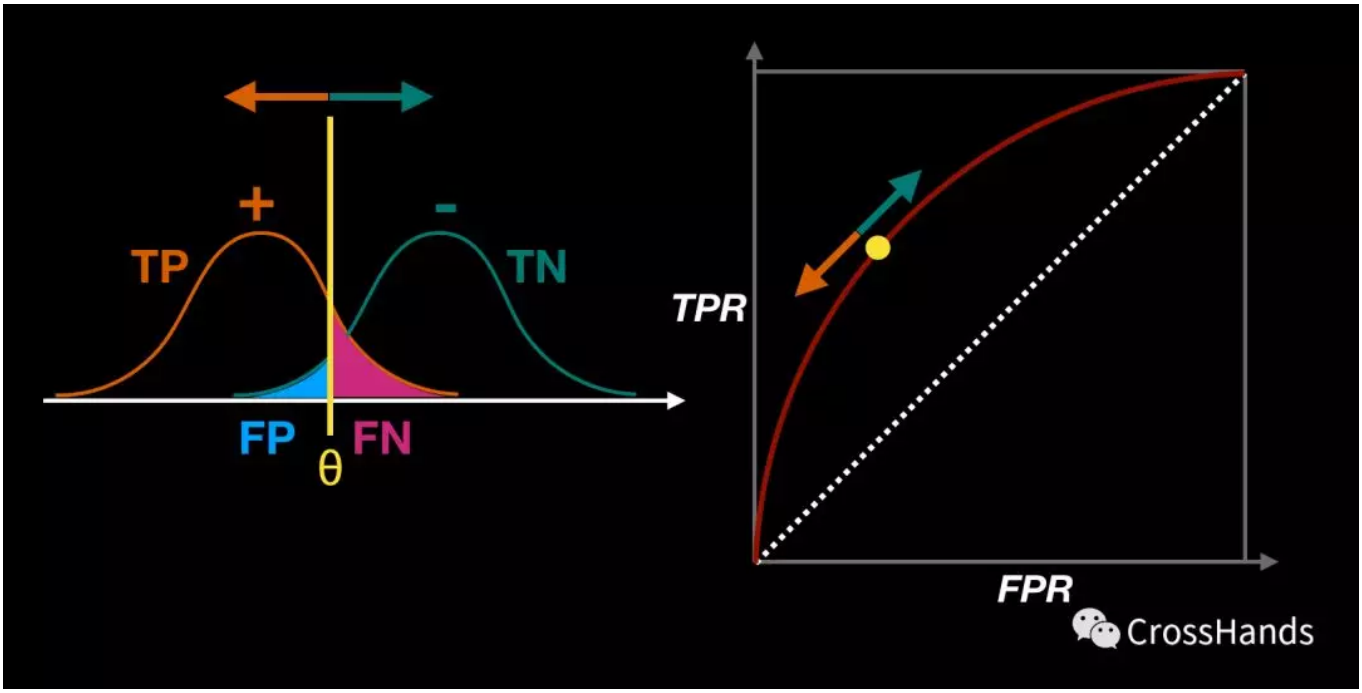
ROC曲线中：

- 横轴(x轴)是False Positive Rate(FPR)，就是负例样本中被错误判定为正例的样本比例，$FPR = \frac{FP}{FP+TN} = \frac{FP}{N}$，FPR=1-Specificity.
- 纵轴(y轴)是True Positive Rate(TPR，等价于Sensitivity)，即正例样本中被正确判定为正例的样本数比例，$TPR = \frac{TP}{TP+FN} = \frac{TP}{P}$.

注：上面公式中的N、P是指负例样本和正例样本各自的样本数量。

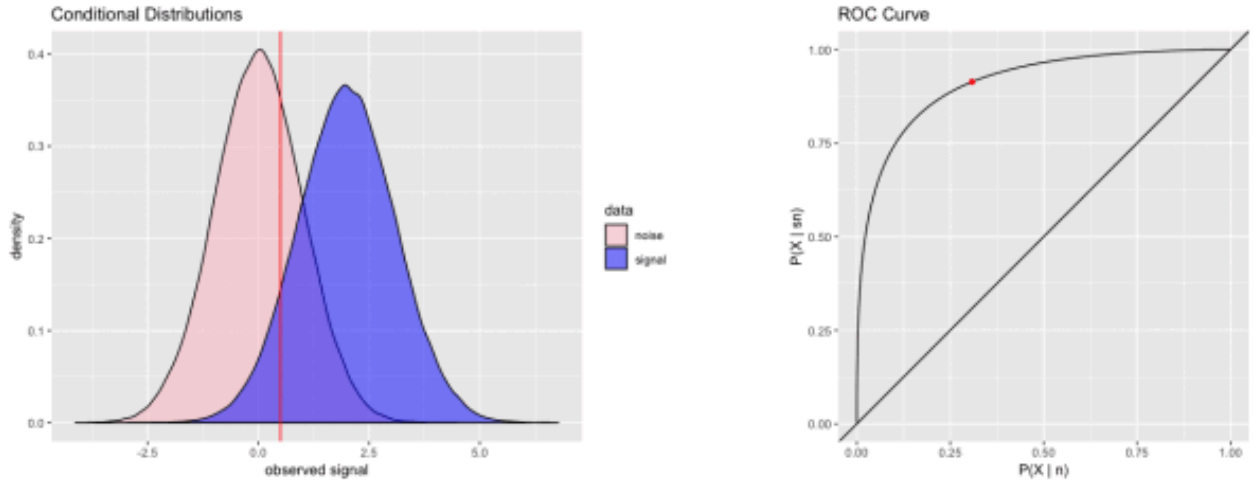
将样本按预测为正例的概率从高到低进行排序后，依次计算每个概率值作为判定阈值对应的TPR和FPR，再将排序后的每个数据点的TPR和FDR值描点到坐标系中，就得到ROC曲线。

如下图示，我们可以看到，将决策阈值往正例方向移动时，对应的TPR和FDR都会下降(FDR和TPR是正相关的关系，所以作ROC曲线图将样本按TPR从小到大排序时，FDR也是从小到大的顺序)。



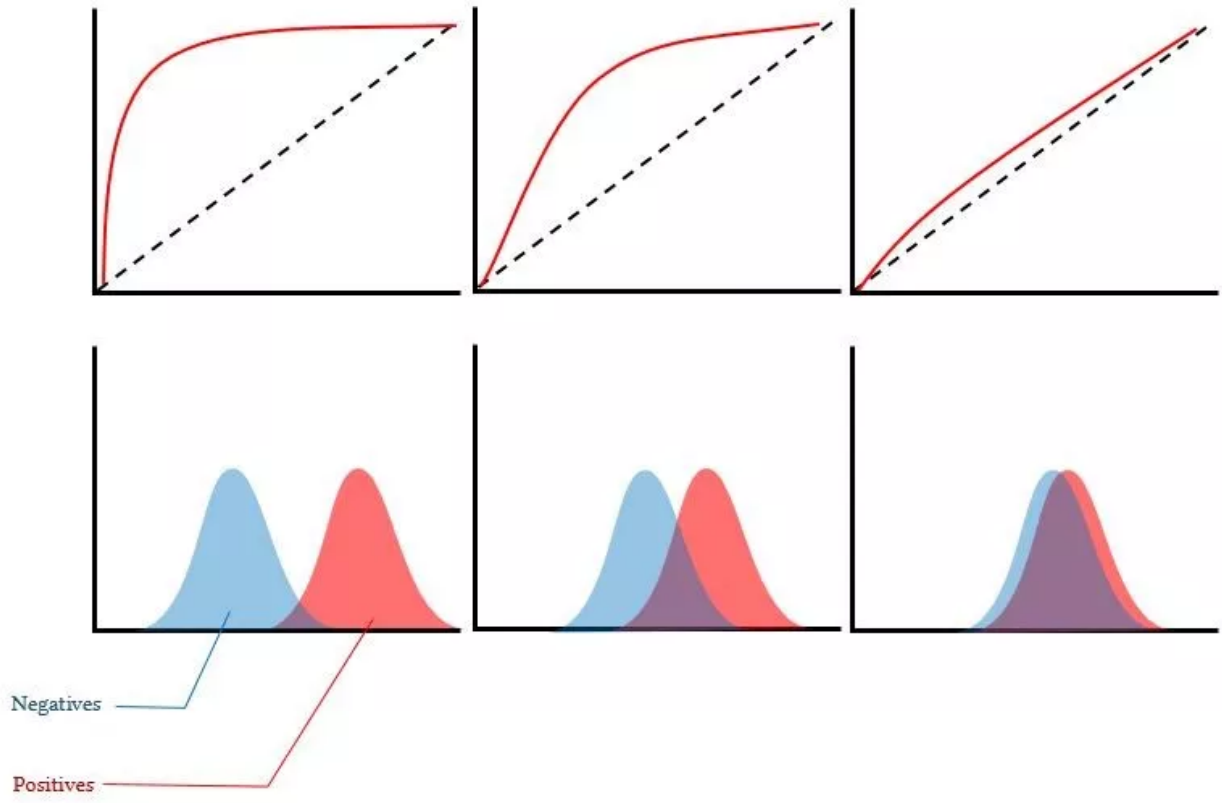
注：ROC曲线图中左下角到右上角的虚线表示“随机操作”下的值(作为参考线)

ROC中决策阈值变化的动态展示如下：



来源：https://rviews.rstudio.com/2019/01/17/roc-curves/

ROC曲线中曲线相对于随机线的最高点，表示正例和负例分布的分离程度(一般来说分离程度越大，ROC曲线的“小山包”隆起越明显)，“小山包”的面积(ROC曲线和随机线围住的面积)就是Gini指数，如下图所示：



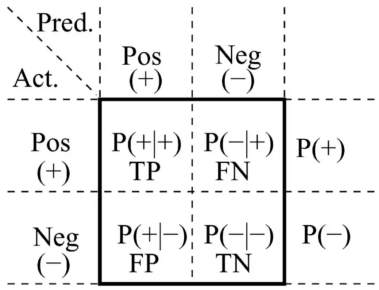
来源：https://derangedphysiology.com/main/cicm-primary-exam/required-reading/research-methods-and-statistics/Chapter%203.0.5/receiver-operating-characteristic-roc-curve

如果模型A的ROC曲线完全包裹模型B的ROC曲线，则表明模型A是优于模型B的；两个模型的ROC曲线发生交叉时，则可以通过ROC曲线下的面积(Area under the ROC curve，简称AUC)来进行比较，AUC取值范围为[0,1].

更多关于ROC曲线的资料：

- ROC曲线的直观展示 *http://www.navan.name/roc/*
- *https://www.dataschool.io/roc-curves-and-auc-explained/*
- *https://derangedphysiology.com/main/cicm-primary-exam/required-reading/research-methods-and-statistics/Chapter%203.0.5/receiver-operating-characteristic-roc-curve*
- *https://en.wikipedia.org/wiki/Receiver_operating_characteristic*

3.2 代价曲线(Cost Curve)



来源：Cost curves: An improved method for visualizing classifier performance

前面提到的指标都有一个前提，那就是正例或者负例预测错误的代价是一样的(FP,FN)。

定义实际为正例预测为负例的损失为$C(-|+)$，实际为负例预测为正例的损失为$C(+|-)$.

代价曲线(Cost Curve)中：

- 横轴是正例概率代价(Probability Cost(+)，简记为$PC(+)$，其值域为[0,1])，与之对应的是是负例概率代价$PC(-) = 1 - PC(+)$，设$p(+)$为样本为正例的概率，样本为负例的概率为$p(-) = 1 - p(+)$. $PC(+) = \frac{p(+) * C(-|+)}{p(+) * C(-|+) + p(-) * C(+|-)}$
- 纵轴是归一化代价(Normalized Expected Cost)

$$\begin{aligned} C_{norm} &= \frac{FNR * p(+) * C(-|+) + FPR * p(-) * C(+|-)}{p(+) * C(-|+) + p(-) * C(+|-)} \\ &= FNR * PC(+) + FPR * PC(-) \\ &= PC(+) * (FNR - FPR) + FPR \end{aligned}$$

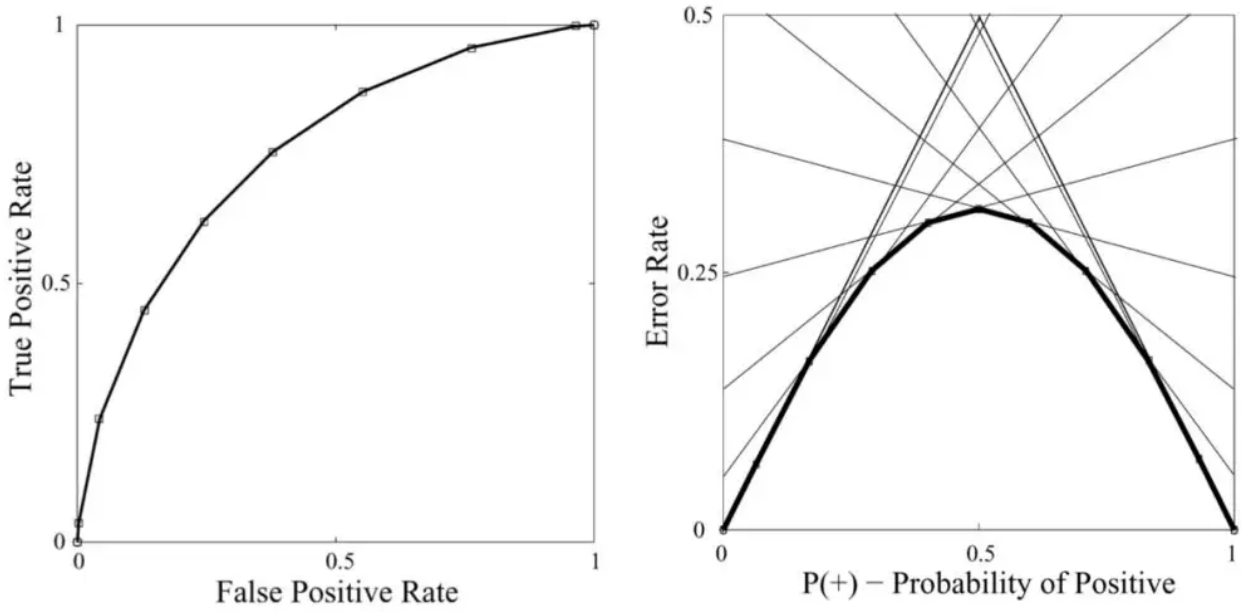


Fig. 4 (a) Ten ROC points and their ROC convex hull — (b) Corresponding set of cost lines and their lower envelope

绘制代价曲线时，ROC曲线上每个点的坐标(TPR,FPR)映射到代价曲线上就是一条*左起于(0,FPR)到右侧(1,1-TPR)*的线段，所有线段绘制好后包裹而成的"小山丘"的面积就是期望的总体代价。

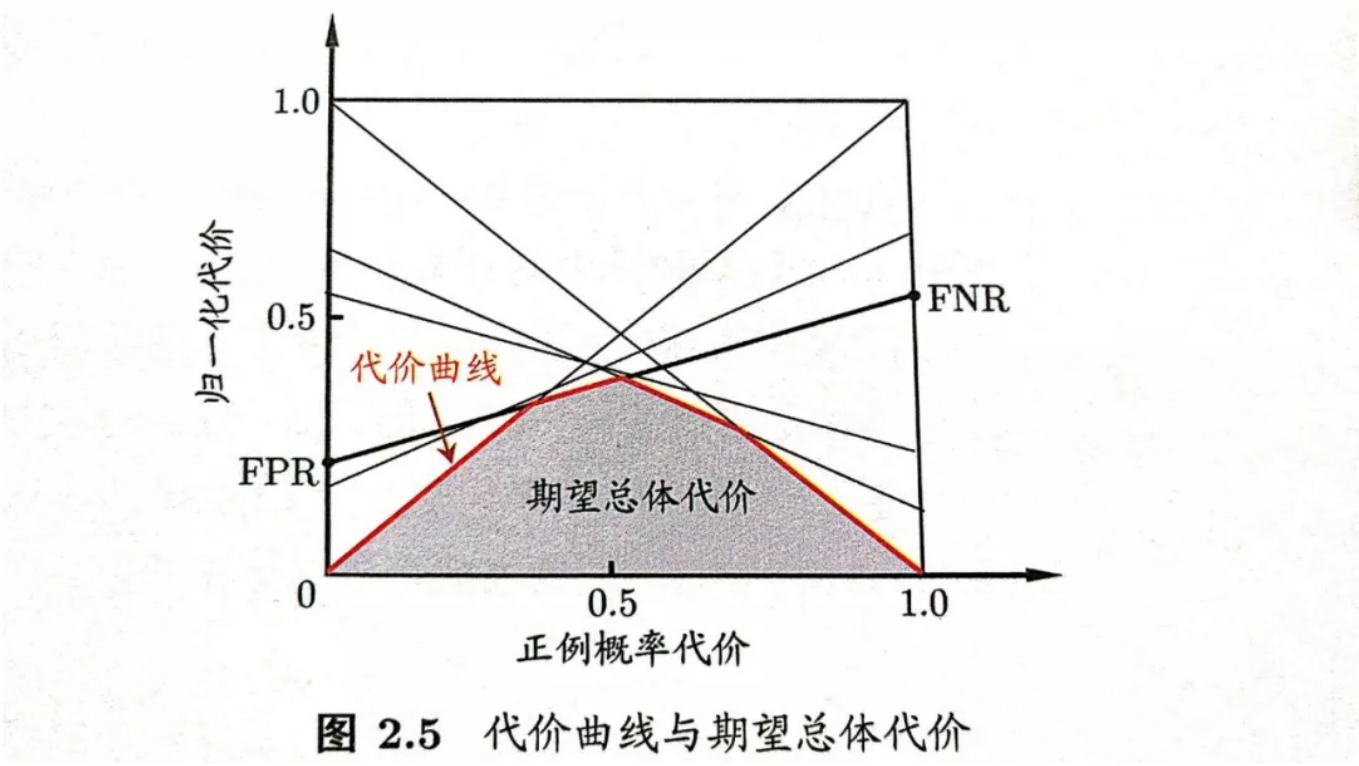


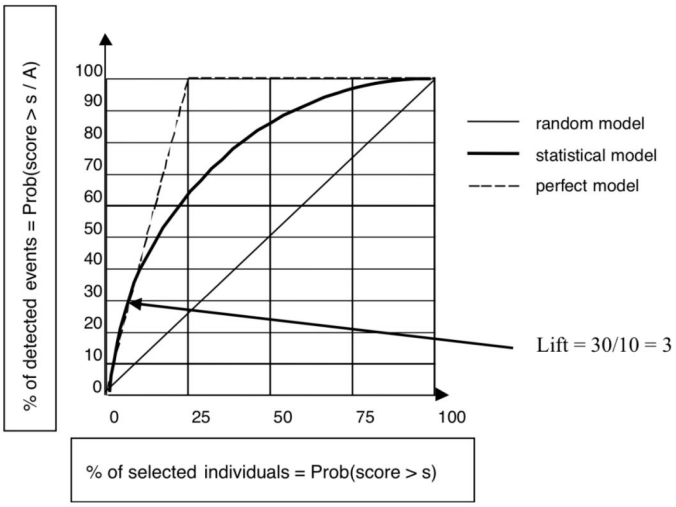
图 2.5 代价曲线与期望总体代价

来源：《机器学习》周志华

更多参考：Cost curves: An improved method for visualizing classifier performance, Chris Drummond & Robert C. Holte, 2006

3.3 Gain/Lift Chart

提升图(Lift Chart，也称为Lift Curve)和收益图(Gain Chart)是从ROC曲线衍生出来的。



截图来自Data Mining and Statistics for Decision Making , Stéphane Tufféry

Lift公式通过贝叶斯推导可以得到

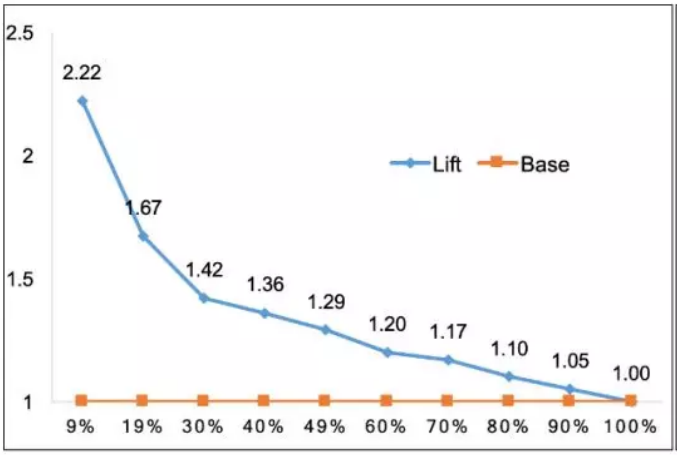
$$
\begin{aligned}
Lift &= \frac{P(A \bigcap B)}{P(A) \times P(B)} \\
&= \frac{P(A|B) \times P(B)}{P(A) \times P(B)} = \frac{P(A|B)}{P(A)} \\
&= \frac{P(B|A) \times P(A)}{P(A) \times P(B)} = \frac{P(A|B)}{P(B)}
\end{aligned}
$$

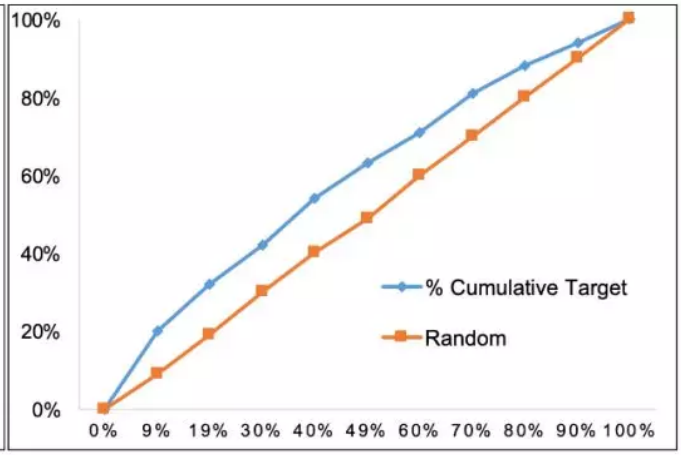
公式中的P(B)可以看做上图中的横轴，也就是每个划分下对应的样本数量占比$P(p > p_\theta)$(p是样本为正例的概率，p_θ是正例概率划分点)，P(A|B)就是每个分组中正例的比例$P(+|p > p_\theta)$。

做lift曲线时，将样本按预测为正例的概率P(+)从大到小排序后，按照P(+)等距划分为N段(取百分位数，一般划分10段)，将每1小段当做一个小组($P(p > p_\theta)$)，然后计算每个分组中正例的比例$P(+|p > p_\theta)$，类似如下的表格

Group	Count	% Count	Count Target=No	Count Target=Yes	% Target=No	% Target=Yes	Cumulative No %	Cumulative Yes %	% Cumulative Overall	KS	% Lift
1	3770	9%	2828	942	8%	20%	8%	20%	9%	12.6%	2.22
2	4181	10%	3630	551	10%	12%	18%	32%	19%	14.5%	1.67
3	4366	11%	3890	476	11%	10%	28%	42%	30%	14.1%	1.42
4	4058	10%	3518	540	10%	12%	38%	54%	40%	16.1%	1.36
5	3825	9%	3404	421	9%	9%	47%	63%	49%	15.9%	1.29
6	4355	11%	3977	378	11%	8%	58%	71%	60%	13.2%	1.20
7	4094	10%	3639	455	10%	10%	68%	81%	70%	13.0%	1.17
8	4270	10%	3949	321	11%	7%	79%	88%	80%	9.1%	1.10
9	3956	10%	3682	274	10%	6%	89%	94%	90%	4.9%	1.05
10	4312	10%	4030	282	11%	6%	100%	100%	100%	0.0%	1.00

来源：http://dni-institute.in/blogs/predictive-model-performance-statistics/





注：表格数据得到的提升图和收益图

更多关于Lift\Gain Curve参考：

- https://en.wikipedia.org/wiki/Lift_(data_mining)
- http://dni-institute.in/blogs/predictive-model-performance-statistics/
- http://www2.cs.uregina.ca/~dbd/cs831/notes/lift_chart/lift_chart.html
- Data Mining and Statistics for Decision Making , Stéphane Tufféry

更多参考指标可以参考如下cheat sheet

Binary classification performances measure cheat sheet

Damien François - v1.1 - 2009 (damien.francois@uclouvain.be)

Confusion matrix for two possible outcomes *p* (positive) and *n* (negative)

		Actual: p	Actual: n	Total
Predicted	p'	**true positive**	**false postive**	P
	n'	**false negative**	**true negative**	N
	total	P'	N'	

Classification accuracy
(TP + TN) / (TP + TN + FP + FN)
Error rate
(FP + FN) / (TP + TN + FP + FN)

Paired criteria

Precision: (or Positive predictive value) proportion of predicted positives which are actual positive
TP / (TP + FP)

Recall: proportion of actual positives which are predicted positive
TP / (TP + FN)

Sensitivity: proportion of actual positives which are predicted positive
TP / (TP + FN)

Specificity: proportion of actual negative which are predicted negative
TN / (TN + FP)

True positive rate: proportion of actual positives which are predicted positive
TP / (TP + FN)

True negative rate: proportion of actual negative which are predicted negative
TN / (TN + FP)

Positive likelihood: likelihood that a predicted positive is an actual positive
sensitivity / (1 - specificity)

Negative likelihood: likelihood that a predicted negative is an actual negative
(1 - sensitivity) / specificity

Combined criteria

BCR: Balanced Classification Rate
½ (TP / (TP + FN) + TN / (TN + FP))
BER: Balanced Error Rate, or **HTER**: Half Total Error Rate: 1 - BCR

F-measure harmonic mean between precision and recall
2 (precision . recall) / (precision + recall)
F_β-measure weighted harmonic mean between precision and recall
$(1+\beta^2)$ TP / $((1+\beta^2)$ TP + β^2 FN + FP)

The harmonic mean between specificity and sensitivity is also often used and sometimes referred to as F-measure.

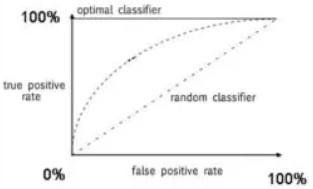
Youden's index: arithmetic mean between sensitivity and specificity
sensitivity - (1 - specificity)

Matthews correlation correlation between the actual and predicted
(TP . TN – FP . FN) / $((TP+FP)\ (TP+FN)\ (TP+FP)\ (TN+FN))^{1/2}$
comprised between -1 and 1

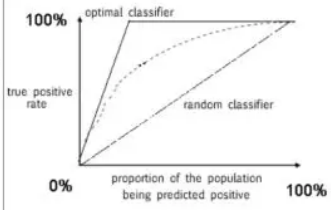
Discriminant power normalised likelihood index
sqrt(3) / π .
(log (sensitivity / (1 – specificity)) + log (specificity / (1 - sensitivity)))
<1 = poor, >3 = good, fair otherwise

Graphical tools

ROC curve receiver operating characteristic curve : 2-D curve parametrized by one parameter of the classification algorithm, e.g. some threshold in the « true postivie rate / false positive rate » space
AUC The area under the ROC is between 0 and 1



(Cumlative) Lift chart plot of the true positive rate as a function of the proportion of the population being predicted positive, controlled by some classifier parameter (e.g. a threshold)



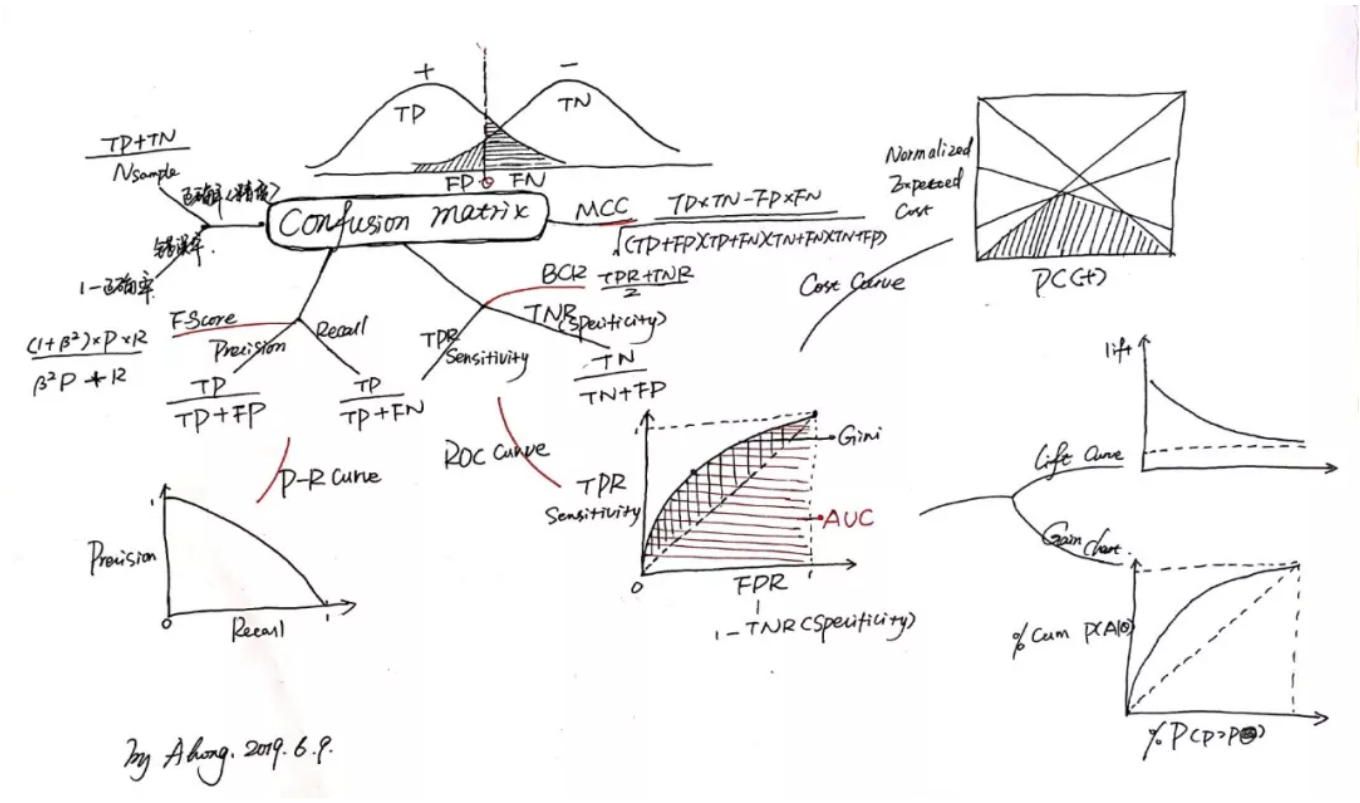
Relationships

sensitivity = recall = true positive rate
specificity = true negative rate
BCR = ½ . (sensitivity + specificity)
BCR = 2 . Youden's index - 1
F-measure = F_1measure
Accuracy = 1 – error rate

References

Sokolova, M. and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45, 4 (Jul. 2009), 427-437.
Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7 (2006) 1–30

来源：http://www.damienfrancois.be/blog/files/modelperfcheatsheet.pdf



参考资料：

- 机器学习，周志华