天善智能
TIANSHAN SOFT
专注于商业智能BI和大数据的垂直社区平台

数据挖掘之模型评估篇

谢佳标（Daniel.Xie）

课程助手|微信：18516242756

数据挖掘模型的优化要遵循有效、适度的原则

- 任何一个数据挖掘模型都是针对一个特定业务需求的，围绕一个具体的业务需求，数据挖掘模型总是可以有办法不断完善、不断提升，即提升精确度、提升转化率等。
- 既然任何一个数据挖掘模型都是针对一个特定业务需求的，那么评价模型是否合格的一个原则性标准就是模型的结论或应用效果是否满足当初的业务需求，即有效的原则。

如何有效地优化模型

- 从业务思路上优化
- 从建模的技术思路上优化
- 从建模的技术技巧上优化

模型效果评价的主要指标体系

- 模型的评价指标和评价体系是建模过程中的一个重要环节，不同类型的项目、不同类型的模型有各自的评价指标和体系。

- 我们将重点介绍目标变量是二元变量（即是与否、1与0）的分类（预测）模型的评价体系和评价指标。

混淆矩阵

- True Positive(TP)：指模型预测为正(1)的，并且实际上也的确是正(1)的观察对象的数量。
- True Negative(TN)：指模型预测为负(0)的，并且实际上也的确是负(0)的观察对象的数量。
- False Positive(FP)：指模型预测为正(1)的，但是实际上是负(0)的观测对象的数量。
- False Negative(FN)：指模型预测为负(0)的，但是实际上是正(1)的观测对象的数量。

		预测的类别	
		1	**0**
实际的类别	**1**	**TP**	**FN**
	0	**FP**	**TN**

评价指标

正确率 $\frac{TP+TN}{TP+FP+TN+FN}$

错误率 $1-\frac{TP+TN}{TP+FP+TN+FN}$

灵敏性/真正率 $\frac{TP}{TP+FN}$

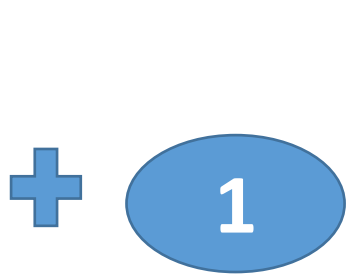
特效性/真负率 $\frac{TN}{TN+FP}$

精度 $\frac{TP}{TP+FP}$

错正率/假正率 $\frac{FP}{\mathrm{FP}+\mathrm{TN}}$

负元正确率 $\frac{TN}{\mathrm{TN}+FN}$

正元错误率 $\frac{TP}{\mathrm{TP}+FP}$



ROC曲线

- ROC曲线是一种有效比较(或对比)两个(或两个以上)二元分类模型(Binary Models)的可视工具，ROC(Receiver Operating Characteristic,接收者运行特征)曲线来源于信号检测理论，它显示了给定模型的灵敏性(Sensitivity)真正率与假正率(False Postive Rate)之间的比较评定。
- 真正率的增加是以假正率的增加为代价的，ROC曲线下面的面积就是比较模型的准确度的指标和依据。面积大的模型对应的模型准确度药膏，也就是要择优应用的模型。面积越接近0.5，对应的模型的准确率就越低。