lec04

Classifier Evaluation

评估 classifier/model/system

- Absolute goodness: 训练好的模型在外部环境按照预期运行，但是在我们部署之前无法得知结果
- Relative goodness: 我们有一个小的具有代表性的测试集；我们比较分类器在此测试数据集（黄金标准）上产生的输出，并测量它与数据集中标签的相似程度

Gold Standard

一个能够评估我们目的的数据集，被称为**测试集 (test data)**，其中的每个测试实例都有其正确的标签注释。存在许多措施来比较训练好的分类器的预测标签和测试数据集中的实际/目标标签。**测试集永远不能加入训练**

混淆矩阵 Confusion Matrix

	Actual Yes (+)	Actual No (-)
Predicted Yes (+)	True Positive (TP)	False Positive (FP)
Predicted No (-)	False Negative (FN)	True Negative (TN)

举例子来说

- 我们预测病人患有癌症，但是更进一步的检查显示病人并没有癌症：FP
- 我们预测病人没有患癌，但是病人死于癌症：FN

实际上，FP 与 FN 在真实世界中的重要性相差很大

Evaluation Measures

Accuracy 准度

被正确分类的对象占比称为**准度**

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{\text{TP} + \text{TN}}{N(\text{样本总和})}$$

Precision 精度

预测为 True 并且实际上也为 True 的对象占比称为**精度**

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall 召回率

实际上为 True 的对象被正确分类的占比称为**召回率**

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Recall vs. Precision

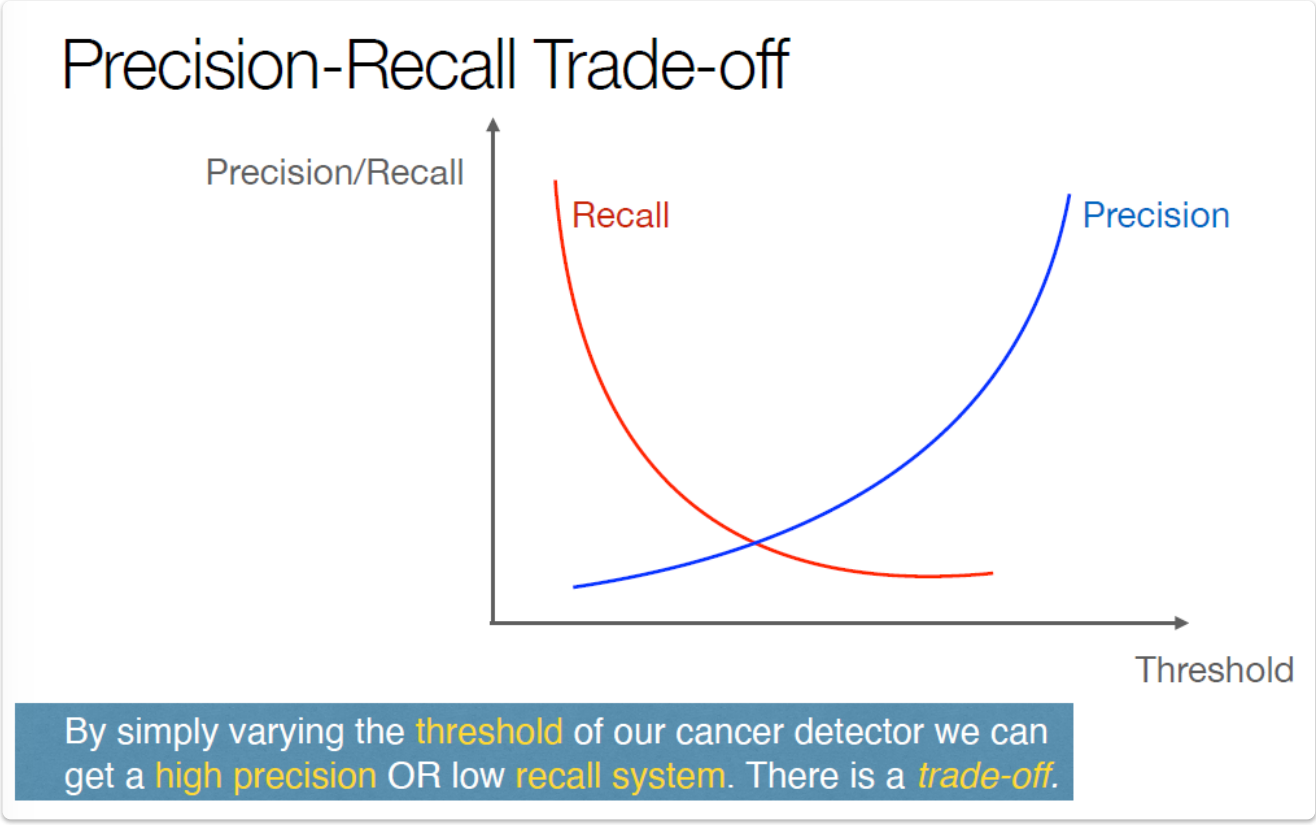
癌症检测

对于癌症检测来说，我们希望尽可能多地检测出癌症
即，在实际上患癌的人群中，让预测患癌的比率更大，因此我们希望有高 Recall

产品推荐

对于产品推荐来说，我们希望把产品推荐给尽可能多的潜在用户
即，在预测会购买的人群中，让实际上购买的人群的占比更大，因此我们希望有高 Precision

Precision-Recall Trade-off



F-score

F-score 是 precision 和 recall 的调和平均数，给了较小的数字更大的权重

$$\text{F-score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Multiple Classes

	A	B	C
A			
B			
C			

对于这个混淆矩阵，纵向是预测，横向是实际
对于 class A
precision：

$$\frac{\text{预测为A实际上也为A的数量}}{\text{预测为}A\text{的数量}}$$

recall：

$$\frac{\text{预测为}A\text{实际上也为}A\text{的数量}}{\text{实际为}A\text{的数量}}$$

F-score：

$$\text{F-score}_{\text{A}} = \frac{2 \times \text{Precision}_{\text{A}} \times \text{Recall}_{\text{A}}}{\text{Precision}_{\text{A}} + \text{Recall}_{\text{A}}}$$

Macro F-score：

$$\frac{1}{C} \sum_{i=1}^{C} \text{F-score}_i$$

其中，C 是 class 的数量，i 是第 i 个 class