

十一、机器学习系统的设计(Machine Learning System Design)

误差分析 (Error Analysis)

一、核心思想

用证据（数据）而非直觉来指导算法优化。

通过简单快速的实验 → 学习曲线分析 → 误差分析 → 决策下一步优化方向。

二、推荐的构建学习算法流程

1. 快速实现一个简单算法

- 用最基础、能跑通的模型开始（不要过早优化）。
- 在交叉验证集上测试算法。

2. 绘制学习曲线 (Learning Curve)

- 用于判断问题类型：
 - 高偏差 (High Bias) → 模型欠拟合，需要更复杂的模型或更多特征。
 - 高方差 (High Variance) → 模型过拟合，需要更多数据或正则化。
- 根据学习曲线决定下一步优化方向：“增加数据”还是“添加特征”。

3. 误差分析 (Error Analysis)

- 手动检查交叉验证集中的错误样本
- 目的：发现错误分类的系统性规律。
- 步骤：
 1. 按类别对错误样本分组（如在垃圾邮件分类器中：医药品垃圾、仿冒品垃圾、钓鱼邮件等）。
 2. 找出预测误差最大的类别。
 3. 思考：这些类别中是否缺乏某些关键特征？
 4. 从最常见的错误类型开始优化模型。

11.3 类偏斜的误差度量

类偏斜情况表现为我们的训练集中有非常多的同一种类的样本，只有很少或没有其他类的样本。

准确率 (Precision) 和召回率 (Recall) 我们将算法预测的结果分成四种情况：

1. 正确肯定 (True Positive, TP) : 预测为真，实际为真

2. 正确否定 (True Negative, TN) : 预测为假，实际为假

3. 错误肯定 (False Positive, FP) : 预测为真，实际为假

4. 错误否定 (False Negative, FN) : 预测为假，实际为真

则：准确率=TP/(TP+FP)。例，在所有我们预测有恶性肿瘤的病人中，实际上有恶性肿瘤的病人的百分比，越高越好。

召回率=TP/(TP+FN)。例，在所有实际上有恶性肿瘤的病人中，成功预测有恶性肿瘤的病人的百分比，越高越好。

这样，对于我们刚才那个总是预测病人肿瘤为良性的算法，其召回率是0。

		预测值	
		Positive	Negative
实际值	Positive	TP	FN
	Negative	FP	TN

红色为准确率，蓝色为召回率：

11.4 准确率和召回率之间的权衡

Trading off precision and recall

→ Logistic regression: $0 \leq h_\theta(x) \leq 1$

Predict 1 if $h_\theta(x) \geq 0.5$

Predict 0 if $h_\theta(x) < 0.5$

Suppose we want to predict $y = 1$ (cancer)
only if very confident.

$$\rightarrow \text{precision} = \frac{\text{true positives}}{\text{no. of predicted positive}}$$

$$\rightarrow \text{recall} = \frac{\text{true positives}}{\text{no. of actual positive}}$$

准确率 (Precision)=TP/(TP+FP)

例，在所有我们预测有恶性肿瘤的病人中，实际上有恶性肿瘤的病人的百分比，越高越好。

召回率 (Recall)=TP/(TP+FN)

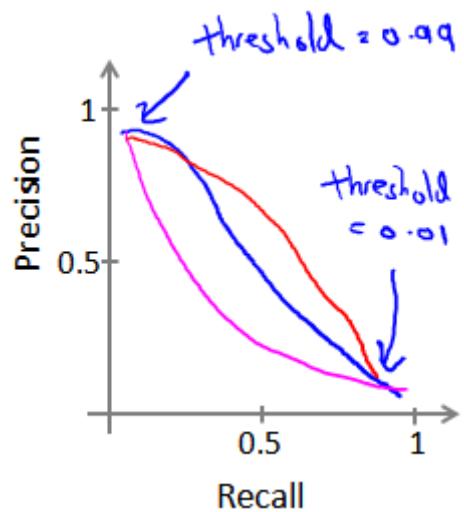
例，在所有实际上有恶性肿瘤的病人中，成功预测有恶性肿瘤的病人的百分比，越高越好。

我们必须设定一个阈值 (threshold)，比如 0.5，当预测概率 > 0.5 时我们认为“恶性”。随着阈值从高 → 低：

- 预测为恶性的人越来越多。
- 真正的恶性 (True Positive) 数也越来越多 (召回率上升)。
- 误判的良性 (False Positive) 也越来越多 (准确率下降)。

如果我们希望提高召回率，尽可能地让所有有可能是恶性肿瘤的病人都得到进一步地检查、诊断，我们可以使用比0.5更小的阈值，如0.3。

我们可以将不同阈值情况下，召回率与准确率的关系绘制成图表，曲线的形状根据数据的不同而不同：



我们希望有一个帮助我们选择这个阀值的方法。一种方法是计算**F1** 值 (**F1 Score**)，其计算公式为：

$$F_1 Score : 2 \frac{PR}{P+R}$$

我们选择使得**F1**值最高的阀值。