

## 4.机器学习笔记 (Machine Learning Tips)

- 4.1指标(Metrics)
- 4.2分类器(Classification)
- 4.3回归(Regression)
- 4.4模型选择(Model Selection)
- 4.5算法诊断(Diagnostics)

# 4.机器学习笔记 (Machine Learning Tips)

## 4.1指标(Metrics)

给定一组数据点 $\{x^{(2)}, \dots, x^{(m)}\}$ , 其中每个 $x^{(i)}$ 拥有 $n$ 个特性, 以及相关的一组输出 $\{y^{(1)}, \dots, y^{(m)}\}$ , 我们需要对分类器的预测效果进行评估。

## 4.2分类器(Classification)

在二进制分类上下文中, 以下是用于评估模型性能的重要指标。

- 混淆矩阵 (Confusion Matrix)

评估一个模型的性能时, 混淆矩阵是用于总结分类模型的预测成效的一种 $N \times N$ 表格。定义如下:

|              |   | Predicted class                              |   |
|--------------|---|--|---|
|              |   | +  | -   |
| Actual class | + | <b>TP</b><br>True Positives                  | <b>FN</b><br>False Negatives<br>Type II error |
|              | - | <b>FP</b><br>False Positives<br>Type I error | <b>TN</b><br>True Negatives                   |

- 主要指标(Main Metrics)

以下指标通常用于评估分类模型的性能:

| 指标                          | 公式                          | 相关解释            |
|-----------------------------|-----------------------------|-----------------|
| 准确率(Accuracy)               | $\frac{TP+TN}{TP+TN+FP+FN}$ | 模型的整体性能         |
| 精确率(Precision)              | $\frac{TP}{TP+FP}$          | 阳性预测的准确性        |
| 召回率/灵敏度(Recall Sensitivity) | $\frac{TP}{TP+FN}$          | 实际阳性样本的覆盖率      |
| 特异度(Specificity)            | $\frac{TN}{TN+FP}$          | 实际阴性样本的覆盖率      |
| F1分数(F1 score)              | $\frac{2TP}{2TP+FP+FN}$     | 对类别不均衡问题有效的混合指标 |

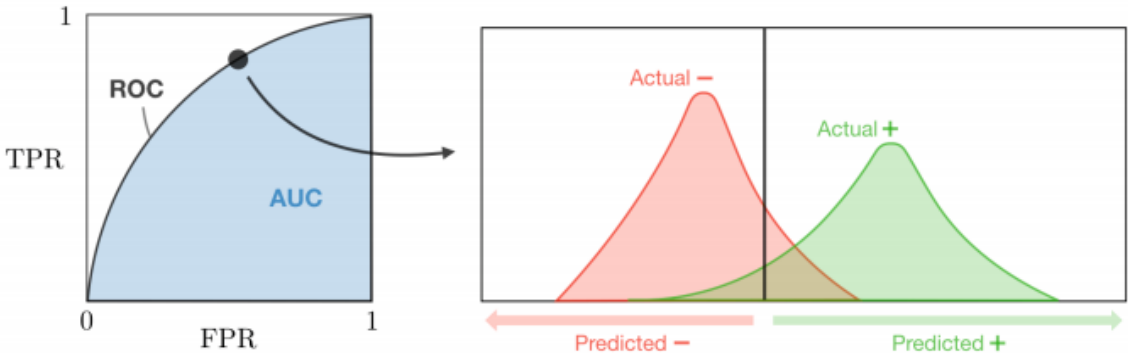
- ROC

接收者操作特征曲线，也提到了ROC，通过改变阈值得到真阳性率(TPR)和伪阳性率(FPR)的曲线。这些指标的汇总如下表：

| 指标        | 公式                 | 等价对应    |
|-----------|--------------------|---------|
| 真阳性率(TPR) | $\frac{TP}{TP+FN}$ | 召回率、灵敏度 |
| 假阳性率(FPR) | $\frac{FP}{TN+FP}$ | 特异度     |

- AUC

接收者操作特征曲线下的面积，也称AUC或AUROC，是ROC下方的面积，如下图所示：



### 4.3回归(Regression)

- 基础指标(Basic Metrics)

对于回归模型，以下指标通常用于评估模型的性能：

| 总平方和   | 回归平方和   | 残差平方和   |
|--|---|---|
| $SS_{\text{tot}} = \sum_{i=1}^n (y_i - \bar{y})^2$ | $SS_{\text{reg}} = \sum_{i=1}^n (f(x_i) - \bar{y})^2$ | $SS_{\text{res}} = \sum_{i=1}^n (y_i - f(x_i))^2$ |

- 决定系数(Coefficient of determination)

决定系数通常使用 $R^2$ 或 $r^2$ 表示,提供了一种测量方法来评估观察到的结果与模型之间的拟合程度，定义如下：

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

- 主要指标(Main Metrics)

以下指标通常用于评估回归模型的性能，考虑到变量n的数量：

| Mallow's Cp   | 赤池信息量 (AIC)          | 贝叶斯信息量(BIC)               | 可调拟合优度(Adjusted $R^2$ )          |
|---|----------------------|---------------------------|----------------------------------|
| $\frac{SS_{\text{res}} + 2(n+1)}{\hat{\sigma}^2_m}$ | $2[(n+2) - \log(L)]$ | $\log(m)(n+2) - 2\log(L)$ | $1 - \frac{(1-R^2)(m-1)}{m-n-1}$ |

当L是可能性时， $\widehat{\sigma}^2$ 是每个响应相关的方差估计。

## 4.4模型选择(Model Selection)

- 相关词汇(Vocabulary)

在选择模型时，我们将数据集的三个不同部分区分如下：

| 训练集                    | 验证集  | 测试集                 |
|------------------------|--|---------------------|
| - 模型训练<br>- 通常占数据集的80% | - 模型性能评估<br>- 通常占数据集的20%<br>- 也被称为 hold-out 或开发集 | - 模型给与预测<br>- 不可见数据 |

一旦模型被确定，它将在整个数据集上进行训练，并在不可见的测试集上进行测试。如下图所示：

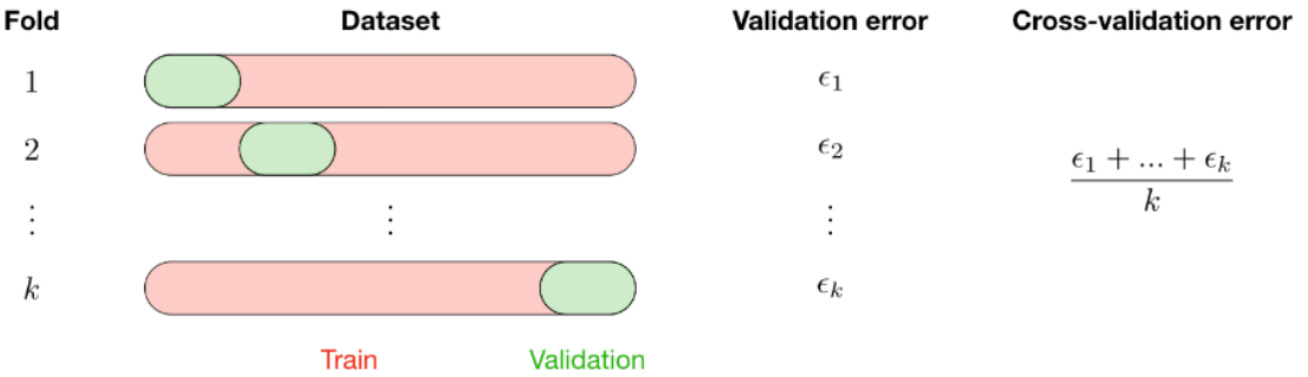


- 交叉验证(Cross-validation)

交叉验证也称为CV，是一种用于那些不太依赖初始训练集的模型的方法，下表总结了不同的类型：

| k-fold                               | 留p验证(Leave-p-out)                                      |
|--------------------------------------|--|
| - 对k-1份样本进行训练，留下一份进行评估<br>- 一般k取5和10 | - 对n-p份样本进行训练，剩下p份进行评估<br>- 当p=1时称为留一验证(leave-one-out) |

最常用的方法是k-fold交叉验证，将训练数据分割成k份，其中一份被保留作为验证模型的数据，其他K-1份用来训练。交叉验证重复K次，每份验证一次，平均K次的误差得到一个估测值，并将其命名为交叉验证误差(cross-validation error)。



- 正则化(Regularization)

正则化过程旨在避免模型过度拟合数据，从而处理高方差问题。下表总结了几种常用的正则化技术:

| LASSO   | Ridge   | 弹性网络(Elastic Net)  |
|---|---|--|
| -将系数压缩到0<br>-有利于变量选择  | 使系数变小   | 变量选择和小系数之间的权衡  |
|   |   |  |
| $\min_{\theta} \sum_{i=1}^n \text{loss}(\theta; x_i, y_i) + \lambda \sum_{j=1}^p  \theta_j $ $\lambda \in \mathbb{R}^+$ | $\min_{\theta} \sum_{i=1}^n \text{loss}(\theta; x_i, y_i) + \frac{\lambda}{2} \sum_{j=1}^p \theta_j^2$ $\lambda \in \mathbb{R}^+$ | $\min_{\theta} \sum_{i=1}^n \text{loss}(\theta; x_i, y_i) + \lambda [(1-\alpha) \sum_{j=1}^p  \theta_j  + \frac{\alpha}{2} \sum_{j=1}^p \theta_j^2]$ $\lambda \in \mathbb{R}^+, \alpha \in [0, 1]$ |

- 模型选择(Model Selection)

在训练集上训练模型，在验证集上做评估，然后在验证集上选择最佳性能 的模型，最后在整个训练集上重新训练该模型。

## 4.5算法诊断(Diagnostics)

- 偏差(Bias)

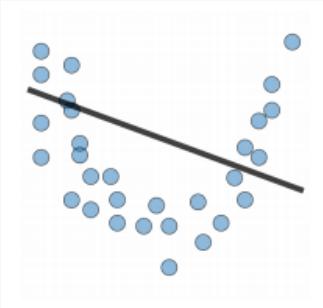
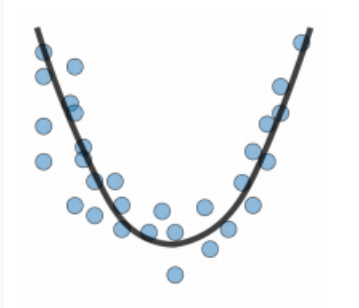
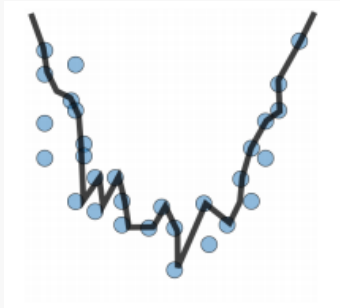
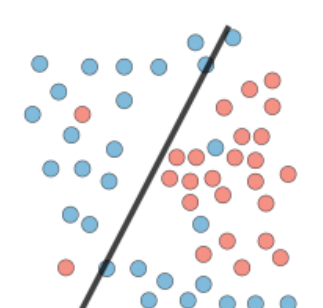
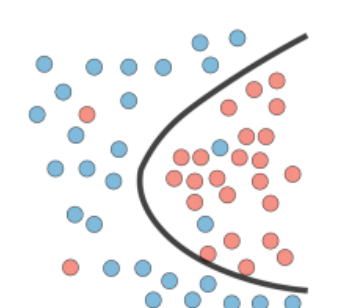
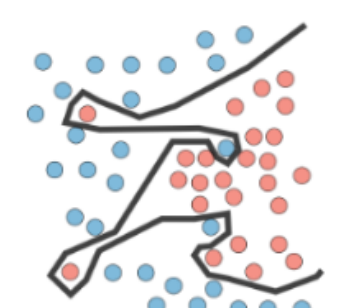



模型的偏差是指我们在给定数据点上模型的样本输出与真实值之间的差值。

- 方差(Variance)

模型的方差是指我们在给定数据点上模型每一次输出结果与模型输出期望之间的误差，即模型的稳定性。

- 偏差/方差权衡(Bias / Variance tradeoff)

模型越简单，偏差越大，模型越复杂，方差越大。

| -    | 欠拟合   | 适合   | 过拟合   |
|------|---|--|---|
| 表现   | -高训练误差<br>-训练误差接近测试误差<br>-高偏差   | -训练误差略低于测试误差   | -低训练误差<br>-训练误差远低于测试误差<br>-高方差  |
| 回归   |    |    |    |
| 分类   |   |   |   |
| 深度学习 |  |  |  |
| 补救措施 | -模型复杂化<br>-添加更多特征<br>-更长时间的训练   |  | -正则化<br>-获取更多的数据  |