COMP 551 Lecture 3 - Model evaluation 1 (M1.2)

Junji Duan

2024/1/15

Today's Outline

- Objectives
  - Evaluating generalization performance
  - Confusion table
  - Receiver Operator Characteristic (ROC)

Evaluating generalization performance

0.1 泛化误差（或泛化准确性)[Generalization error]

- 我们真正关心的是模型在新数据上的表现, 即我们希望了解模型对未见数据的泛化能力。
- 我们假设训练数据和未见数据来自同一分布。我们通常假设存在某种分布 $p(x, y) = p(y \mid x)p(x)$, 这样我们的训练数据就是从这个分布中独立抽样得来的, 即 $x^{(n)} \sim p_x$ 和 $y^{(n)} \sim p_{y|x}$, 这些数据是独立同分布的 (i.i.d.)。
- 我们也假设未见数据同样是来自这个分布的样本。

泛化误差是我们模型 $f : x \mapsto y$ 在此分布下的预期误差:

$$\mathrm{Err}(f) = \mathbb{E}_{x,y\sim p}[\ell(f(x), y)].$$

这里 ℓ 是某种损失函数, 比如分类误差 $\ell(y, \hat{y}) = \mathbb{I}(y \neq \hat{y})$ 或在回归中常用的平方损失 $\ell(y, \hat{y}) = (y - \hat{y})^2$

0.2 测试集 [Test set]

- 不幸的是，我们无法访问真实的数据分布，我们只有从分布中抽样得到的样本。
- 我们可以通过将数据集的一部分保留作为测试集来估计泛化误差, 这部分数据在学习或选择模型时不被使用。
- 这部分数据集被称为测试集, 我们用 $\mathcal{D}_{\text{train}}$ 和 $\mathcal{D}_{\text{test}}$ 来表示我们原始数据集 $\mathcal{D}$ 的这种划分。

测试误差是:

$$\widehat{\mathrm{Err}}(f) = \mathbb{E}_{x,y\sim\mathcal{D}_{\text{test}}}\left[\ell(f(x), y)\right] = \frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{x,y\in\mathcal{D}_{\text{test}}} \ell(f(x), y).$$

其中 $|\mathcal{D}_{\text{test}}|$ 是测试集 $\mathcal{D}_{\text{test}}$ 的基数（即测试样本的数量)。

0.3 预测准确率

$$\text{Accuracy } = \frac{\text{Correct predictions}}{\text{Total number of predictions}}$$

Confusion table

0.4 True/false positives and negatives

- 真正例 (True Positive, TP) : 这是分类器正确预测的正例数量。例如, 预测一个人患有癌症, 并且这个人确实患有癌症。
- 假正例 (False Positive, FP) : 分类器错误预测的正例数量。例如, 预测一个人患有癌症, 但这个人实际上是健康的。
- 真负例 (True Negative, TN) : 分类器正确预测的负例数量。例如, 预测一个人健康, 并且这个人确实健康。
- 假负例 (False Negative, FN) : 分类器错误预测的负例数量。例如, 预测一个人健康, 但这个人实际上患有癌症。

0.5 True and false positive rates

- 真正率（True Positive Rate, TPR），也称为灵敏度（sensitivity），是正确识别的正例占所有实际正例的比例。计算公式为:

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

- 假正率（False Positive Rate, FPR），也称为1-特异性（1-specificity），是错误标记为正例的负例占所有实际负例的比例。计算公式为:

$$FPR = \frac{FP}{FP + TN}$$

Receiver Operator Characteristic (ROC)

0.6 分类模型和概率预测 (Classification model)

在机器学习的二分类问题中（例如判断病人是否患有癌症），模型常常会输出一个概率值而不是直接的决策结果。这个概率值表示了给定输入数据后，模型预测样本属于某一类别的置信度。公式

$$p\left(y^{(*)} = 1 \mid \mathbf{x}^{(*)}\right) = \frac{1}{K} \sum_{n \in \mathcal{N}_K\left(\mathbf{x}^{(*)}, \mathcal{D}\right)} \mathbb{I}\left(y^{(n)} = 1\right)$$

描述的是在给定输入 $\mathbf{x}^{(*)}$ 的情况下, 样本属于类别 1（例如癌症）的概率。这里利用了 K 个最近邻样本的信息, 通过计算这些邻近样本中属于类别 1 的比例来估计概率。

0.7 分类阈值 (Classification threshold)

在实际应用中, 我们需要根据这个概率值决定如何分类, 这通常通过设置一个阈值（默认通常是 0.5）来完成。如果预测概率大于这个阈值, 我们判断结果为正类（例如判断为癌症）, 否则为负类（非癌症）。调整阈值可以改变模型的敏感性和特异性。

0.8 ROC 曲线和 AUC 值

为了评估模型在不同阈值下的表现，我们可以绘制 ROC 曲线。ROC 曲线是通过在不同的分类阈值下计算真正率 (TPR) 和假正率 (FPR) 来得到的。真正率 (也叫敏感性) 是模型正确识别出的正样本比例, 假正率是错误标记为正样本的负样本比例。

- 真正率 $(TPR) = TP/(TP + FN)$
- 假正率 (FPR) = FP/(FP + TN)

ROC(Receiver Operating Characteristic) 曲线将 FPR 作为横坐标, TPR 作为纵坐标。AUC（Area Under the Curve-曲线下面积）是评估模型整体性能的一个重要指标, AUC 值越高, 模型的性能通常认为越好。

0.9 模型效果的极端情况

1. 对于一个“哑”模型 (Dummy model), 它对所有输入都预测概率为 0.5 :

- 这种模型无法区分任何样本, 其 ROC 曲线会是一条从 $(0, 0)$ 到 $(1, 1)$ 的对角线。AUC 为 0.5 , 表示没有预测能力。

2. 对于一个完美的模型 (Perfect model):

- 这个模型能够完美区分所有样本, 其 ROC 曲线会在左上角形成一个完美的“”形（先垂直上升到 $TPR = 1$, 然后水平移动到 $FPR = 1$ AUC 为 1 , 表示完美预测。

3. 对于与完美模型完全相反的模型:

- 这个模型的预测结果总是错误的, 其 ROC 曲线将沿着 $y = 0$ 先从左向右, 然后在 $FPR = 1$ 处向上到 $(1, 1)$ 。这样的 AUC 接近 0 , 是一个非常糟糕的模型。