**机器学习模型的选择与优化**

**\*\*\*\*\*\*\*\*\*\*\*\* 机器学习通用的规范实践训练流程 \*\*\*\*\*\*\*\*\*\*\*\***

**数据加载及展示**

**数据预处理**

**模型建立及训练**

**模型预测**

**结果展示及模型表现评估**

**\*\*\*\*\*\*\*\*\*\*\*\* 数据载入 \*\*\*\*\*\*\*\*\*\*\*\***

* **数据质量决定模型表现得上限，因此建模前对数据进行五检查：**
  1. **样本代表性：采集数据得方法是否合理，采集到得数据是否有代表性**
  2. **标签统一化：对于样本结果，要确保每个样本都遵循一样的标签规则**
  3. **数据合理性：样本中的异常数据点是否合理、如何处理**
  4. **数据重要性：数据属性的意义，是否为无关数据**
  5. **属性差异性：不同属性数据的数量级差异性如何**

**\*\*\*\*\*\*\*\*\*\*\*\* 数据预处理 \*\*\*\*\*\*\*\*\*\*\*\***

* **数据预处理的方法和好处：**
  1. **根据实际场景扩充或减少样本——数据质量提升，提高模型表现**
  2. **对不合理标签数据进行预处理——帮助模型学习到正确信息**
  3. **删除不重要的属性数据、数据降维——降低噪声影响，减少过拟合，节约运算时间**
  4. **对数据进行归一化或标准化——平衡数据影响，加快数据收敛**
  5. **过滤掉异常数据——降低噪声影响，提高鲁棒性**

**\*\*\*\*\*\*\*\*\*\*\*\* 模型建立 \*\*\*\*\*\*\*\*\*\*\*\***

* **模型建立的三大核心问题：**
  1. **选用什么算法？**
  2. **核心结构、参数如何设置？**
  3. **模型表现不好怎么办**
* **算法选择与核心结构、参数设置：**
  1. **逻辑回归—— 边界函数一阶线性、二阶、高阶多项式等**
  2. **KNN —— 核心参数K值得选择等**
  3. **决策树 —— 树分支逻辑（ID3、C4.5、CART）、最小分支样本数等**
  4. **朴素贝叶斯—— GaussianNB、ComplementNB、CategoricalNB、 MultinomialNB、BernoulliNB等**
  5. **神经网络 —— 几层、每层神经元数、激活函数等**
  6. **其他算法**
* **同时进行建立多个模型，进行多模型比对，对比模型表现**
* **模型表现：模型表现不好，则需要从前往后找问题，判断数据是否有问题、算法选择是否合适、核心结构与参数是否合理等。**

**\*\*\*\*\*\*\*\*\*\*\* 模型表现评估指标 \*\*\*\*\*\*\*\*\*\*\*\***

* **基于训练数据和预测数据的模型预测效果**
* **基于混淆矩阵计算的评估指标**

**\*\*\*\*\*\*\*\*\*\*\*\* 模型预测表现效果 \*\*\*\*\*\*\*\*\*\*\*\***

**通过机器学习训练，最终模型可能有三种表现效果：**

1. **欠拟合：训练数据和预测数据效果都不好。**
2. **好模型：训练数据和预测数据效果都很好。**
3. **过拟合：训练数据效果很好（甚至比好模型结果还好），预测数据效果不好。**

**\*\*\*\*\*\*\*\*\*\*\*\* 模型表现欠拟合解决方法 \*\*\*\*\*\*\*\*\*\*\*\***

* **欠拟合的解决方法（不仅限于以下几种）：**
  1. **选择其他模型（比如采用非线性模型）**
  2. **增加模型复杂度**
  3. **增加数据样本（比如加入进特征组合、高次特征，来增大假设空间）**
  4. **采集新的维度数据**
  5. **如果已正则化，尝试减少正则化程度λ**

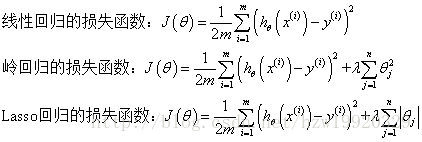
**\*\*\*\*\*\*\*\*\*\*\*\* 模型表现过拟合解决方法 \*\*\*\*\*\*\*\*\*\*\*\***

* **过拟合的原因：**
  1. **使用了过于复杂的模型结构（比如高阶决策边界）**
  2. **训练数据不足，有限的训练数据（训练样本只有总体样本中的小部分，不具备代表性）**
  3. **样本里的噪音数据干扰过大，模型学习到了噪音信息（使用过多与结果不相关的属性数据）**
* **过拟合的解决方法（不仅限于以下几种）：**
  1. **简化模型结构（降低模型复杂度，能达到好的效果情况下尽可能选择简单的模型）**
  2. **数据增强（按照一定的规则扩充样本数据）**
  3. **数据预处理，保留主成分信息（数据降维）**
  4. **增加正则化项——回归任务中实现**

**\*\*\*\*\*\*\*\*\*\*\*\* 回归任务中的增加正则项 \*\*\*\*\*\*\*\*\*\*\*\***

* **增加正则项定义：在机器学习过程中，模型求解的核心目标就是最小化损失函数，增加正则项是指在损失函数中添加一个额外项λ，实现对求解参数的数值约束，防止模型过拟合。**
* **回归任务中增加正则项的实现方式：**

1. **岭回归：回归模型引入Ridge回归正则项**
2. **Lasso回归：回归模型引入Lasso回归正则项**

****

**额外项λ称为正则化参数，若λ选取过大，会把所有参数θ均最小化，造成欠拟合；若λ选取过小，会导致对过拟合问题解决不当，因此要选取合适的λ。**

**\*\*\*\*\*\*\*\*\*\*\*\* 混淆矩阵 \*\*\*\*\*\*\*\*\*\*\*\***

**混淆矩阵（Confusion Matrix）：又称为误差矩阵，用于统计各类别样本预测正确与错误的数量，从而帮助用户全面的评估模型表现。**

|  |  |  |
| --- | --- | --- |
|  | Predicted ---- 0 | Predicted ---- 1 |
| Actual ---- 0 | **TN** | **FP** |
| Actual ----1 | **FN** | **TP** |

**True Positives (TP)：预测准确，预测为正样本的数量（实际为1，预测为1）**

**True Negatives (TN)：预测准确，预测为负样本的数量（实际为0，预测为0）**

**False Positives (FP)：预测错误，预测为正样本的数量（实际为0，预测为1）**

**False Negatives (FN)：预测错误，预测为负样本的数量（实际为1，预测为0）**

**混淆矩阵的优点：**

1. **分类任务中，相比单一的准确率指标，混淆矩阵提供了更全面的模型评估信息（TP\TN\FP\FN）**
2. **基于混淆矩阵，我们可以计算出多样的模型表现衡量指标，从而实现模型的综合评估**

**\*\*\*\*\*\*\*\*\*\*\*\* 基于混淆矩阵计算的评估指标 \*\*\*\*\*\*\*\*\*\*\*\***

* **准确率（Accuracy）：总样本中预测正确的比例。**
  + **Accuracy = ( TP + TN ) / ( TP + TN +FP +FN )**
* **错误率（Misclassification Rate）：总样本中预测错误的比例。**
  + **Misclassification Rate = ( FP +FN ) / ( TP + TN +FP +FN )**
* **召回率（Recall）：正样本中，预测正确的比例。**
  + **Recall = TP / ( TP + FN )**
* **特异度（Specificity）：负样本中，预测正确的比例。**
  + **Specificity = TN / ( TN +FP )**
* **精确度（Precision）：预测结果为正的样本中，预测正确的比例。**
  + **Precision = TP / ( TP + FP )**
* **F1分数（F1 Score）：综合Precision和Recall的指标。**
  + **F1 Score = 2 \* Precision \* Recall / ( Precision + Recall )**

**不同的衡量指标在不同的应用场景中的重要性是不同的，因此在要根据应用场景的实际情况选择使用哪些评估指标。**