数学建模

谈欣

2023 年 9 月 4 日

目录

1 数据预处理常用公式

1.1 特征缩放

“StandardScaler”和“MinMaxScaler”是常用于数据预处理的特征缩放方法，用于将特征数据进行标准化或归一化，以便更好地适应机器学习算法。

1.1.1 StandardScaler（标准化）

标准化通过将特征的值按照均值（mean）和标准差（standard deviation）进行缩放，使得特征的均值为 0，标准差为 1。对于每个特征 X_i，标准化后的特征 X_i' 计算如下：

$$X_i' = \frac{X_i - \mu}{\sigma} \tag{1}$$

其中，X_i 是原始特征的值，(μ 是特征 X_i 的均值，σ 是特征 X_i 的标准差。

1.1.2 MinMaxScaler（归一化）

归一化将特征的值缩放到一个指定的范围，通常是 [0, 1]。对于每个特征 X_i，归一化后的特征 X_i' 计算如下：

$$X_i' = \frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)} \tag{2}$$

其中：X_i 是原始特征的值。$\min(X_i)$ 是特征 X_i 的最小值。$\max(X_i)$ 是特征 X_i 的最大值。

通过归一化，所有特征的值都缩放到了 [0, 1] 的范围内，这有助于处理不同特征值范围差异较大的情况，使得模型更容易收敛和训练。

2 相关性分析

2.1 三大主流相关系数

以下是皮尔逊相关系数（Pearson correlation coefficient）、斯皮尔曼相关系数（Spearman rank correlation coefficient）和肯德尔相关系数（Kendall tau rank correlation coefficient）的数学公式：

这些相关系数用于不同类型的数据和关系度量，可以帮助你了解变量之间的关系或相关性。

2.1.1 皮尔逊相关系数

皮尔逊相关系数（Pearson correlation coefficient）用于衡量两个连续变量之间的线性关系。它的数学公式如下：

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{3}$$

其中，n 是数据点的数量，X_i 和 Y_i 分别是两个变量的第 i 个观测值，$\bar{X}$ 和 $\bar{Y}$ 分别是两个变量的均值。

2.1.2 斯皮尔曼相关系数

斯皮尔曼相关系数（Spearman rank correlation coefficient）用于衡量两个变量之间的单调关系，不要求变量是连续的，而是基于它们的秩次（排名）。其数学公式如下：

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{4}$$

其中，n 是数据点的数量，d_i 是两个变量在排序后的秩次之差。

2.1.3 肯德尔相关系数

肯德尔相关系数（Kendall tau rank correlation coefficient）也用于衡量两个变量之间的单调关系，基于它们的秩次。其数学公式如下：

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j) \tag{5}$$

其中，n 是数据点的数量，X_i 和 Y_i 是两个变量的第 i 个观测值，$\text{sign}(X_i - X_j)$ 表示 X_i 与 X_j 的秩次差的符号。

3 K-最近邻数学模型

4 决策树数学模型

4.1 基础决策树

4.2 随机森林

随机森林是一种集成学习方法，由多棵决策树组成。其公式可以分为两部分：随机森林的生成和随机森林的预测。随机森林的生成过程如下：

1. 从原始数据集中采样出 n 个样本作为训练集，采用 bootstrap 技术进行采样，即每次从原始数据集中随机抽取一个样本并将其放回，重复 n 次得到大小为 n 的采样集。

2. 从所有特征中随机选择 m 个特征，其中 $m \ll d$，d 为原始特征的总数。

3. 利用上述采样得到的数据集和特征集构建一棵决策树，具体建树过程可以使用 ID3、C4.5 或 CART 等决策树算法。

4. 重复步骤前三步 T 次，得到 T 棵决策树，这些决策树构成了随机森林。

随机森林的预测步骤：

1. 对于每个测试样本，对随机森林中的每棵决策树进行预测，得到预测结果。

2. 对 T 个预测结果进行投票，将得票最多的类别作为随机森林的最终预测结果。

随机森林的预测公式可以表示为：

$$\widehat{y} = \arg\max_y \sum_{i=1}^{T} I\left(\widehat{y}_i = y\right). \tag{6}$$

其中，$\widehat{y}$ 表示随机森林的预测结果，$\widehat{y}_i$ 表示第 i 棵决策树的预测结果，T 表示随机森林中的决策树数量，y 表示所有可能的类别，$I(\cdot)$ 表示指示函数，当条件成立时取值为 1，否则取值为 0。

4.3 XGBoost

4.4 LightGBM

LightGBM 是一种基于决策树的集成学习算法，是 GBDT 算法的改进版。相对于传统的 GBDT 算法，LightGBM 具有更快的训练速度、更低的内存消耗以及更高的准确率。可以将 LightGBM 的优化用公式表达，如下式：

$$LightGBM = XGBoost + Histogram + GOSS + EFB. \tag{7}$$

LightGBM 的核心算法可以用如下的公式表示：

$$Obj(\theta) = \sum_{i}^{n} l(y_i, \widehat{y}_i) + \sum_{k} \Omega(f_k). \tag{8}$$

其中，$\Omega(f_k)$ 表示正则化项，$\sum\limits_{k}$ 表示所有树的叶子节点，f_k 表示第 k 棵树。LightGBM 的目标是最小化 $Obj(\theta)$，即同时优化模型的预测精度和模型复杂度，θ 表示模型参数，$l\left(y_i, \widehat{y}_i\right)$ 表示预测值 $\widehat{y}_i$ 与真实值 y_i 之间的损失。

5 模型评价指标

5.1 分类模型评价指标

5.1.1 AUC 值计算

当计算 AUC（曲线下面积）时，首先需计算真正例率（True Positive Rate，TPR）与假正例率（False Positive Rate，FPR）：

$$TPR = \frac{TP}{TP + FN} \tag{9}$$

$$FPR = \frac{FP}{FP + TN} \tag{10}$$

其中，TP（真正例数）表示模型正确地预测为正例的样本数；FN（假负例数）表示模型错误地将正例标记为负例的样本数；FP（假正例数）表示模型错误地将负例标记为正例的样本数；TN（真负例数）表示模型正确地预测为负例的样本数。

然后根据 FPR 和 TPR 绘制 ROC 曲线——ROC 曲线是一条以 FPR 为横轴，TPR 为纵轴的曲线，通常由多个点组成。ROC 曲线表示了在不同分类阈值下，模型的性能变化。AUC 即 ROC 曲线下面积的大小，可以使用积分来计算，公式如下：

$$AUC = \int_0^1 \mathrm{TPR}(FPR)\, dFPR \tag{11}$$

其中，TPR(FPR) 是 ROC 曲线上的真正例率（TPR）作为假正例率（FPR）的函数。

5.1.2 support, recall, f1-score

支持率（Support）等于某个类别的真正例（True Positives，TP）和假负例（False Negatives，FN）的总和，也可以理解为在数据集中属于该类别的所有样本数量。计算公式如下：

$$Support = TP + FN \tag{12}$$

其中，TP（True Positives）是模型正确预测为正例的样本数量；FN（False Negatives）是模型将实际正例错误预测为负例的样本数量。

Recall 是衡量模型找出所有正例（真正例）的能力，也称为敏感性（Sensitivity）或真正例率（True Positive Rate，TPR）。

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

F1-Score 是综合考虑了精确度（Precision）和召回率（Recall）的指标，用于平衡这两者之间的关系。它是精确度和召回率的调和平均值。

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$F1 - Score = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{15}$$

分类模型的 Recall、F1-Score 和 Support 是用来评估模型性能的重要指标，通常用于二分类或多分类任务的性能评估。

5.2 回归模型评价指标

以下是均方误差（Mean Squared Error，MSE）、平均绝对误差（Mean Absolute Error，MAE）和 R 平方分数（R-squared score）的数学公式：

5.2.1 均方误差 (MSE)

均方误差是一种用于度量回归模型性能的指标。它计算了模型预测值与实际观测值之间的平方误差的平均值。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{16}$$

其中：y_i 是第 i 个观测值的实际目标值，$\hat{y}_i$ 是模型对第 i 个观测值的预测值。

5.2.2 平均绝对误差 (MAE)

平均绝对误差是另一种用于度量回归模型性能的指标。它计算了模型预测值与实际观测值之间的绝对误差的平均值。

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{17}$$

其中：y_i 是第 i 个观测值的实际目标值，$\hat{y}_i$ 是模型对第 i 个观测值的预测值。

5.2.3 R 方分数 (R-squared score)

R 平方分数是用于评估回归模型拟合度的指标，它表示模型对观测数据方差的解释程度。R 平方分数的取值范围在 0 到 1 之间，越接近 1 表示模型对数据的解释程度越好。

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{18}$$

其中，y_i 是第 i 个观测值的实际目标值，$\hat{y}_i$ 是模型对第 i 个观测值的预测值，$\bar{y}$ 是实际目标值的均值。

5.3 聚类分析评价指标

5.3.1 CH 分数（Calinski and Harabasz score）

Calinski-Harabasz 分数（也称为方差比准则，Variance Ratio Criterion，通常简称 CH 分数）是一种用于评估聚类分析（Cluster Analysis）结果的指标。它可以用来衡量聚类的准确性，但不是唯一的评估指标，还应该与其他指标一起考虑。

Calinski-Harabasz 分数通过比较簇内的数据点之间的方差与簇间的数据点之间的方差来评估聚类的紧密度和分离度。具体来说，CH 分数的计算是通过以下公式完成的：

$$CH = \frac{B}{W} \times \frac{N-k}{k-1} \tag{19}$$

其中，B 是簇间的方差，表示不同簇之间的数据点之间的方差；W 是簇内的方差，表示同一簇内数据点之间的方差；N 是数据点的总数；k 是簇的数量。

CH 分数的值越高，表示簇间方差相对于簇内方差更大，即簇之间更分离，聚类效果更好。因此，较高的 CH 分数通常被认为是较好的聚类结果。

5.3.2 轮廓系数（Silhouette Score）

轮廓系数（Silhouette Score）是一种用于评估聚类质量的指标，它考虑了数据点与其所属簇内的相似度和与最近的邻近簇之间的不相似度。较高的轮廓系数表示数据点更适合于它们所属的簇，而较低的轮廓系数则表示数据点更适合于与其所属簇不同的簇。轮廓系数的计算公式如下：

对于单个数据点 i（注意：该公式适用于每个数据点），首先首先计算数据点 i 与其所属簇内所有其他数据点之间的平均距离（簇内平均距离）：

$$a(i) = \frac{1}{C-1} \sum_{j \neq i} d(i,j) \tag{20}$$

其中，C 是数据点 i 所属簇内的数据点总数（包括数据点 i 自身），$d(i,j)$ 表示数据点 i 与数据点 j 之间的距离，通常可以使用欧氏距离、曼哈顿距离或其他距离度量来计算。

然后计算数据点 i 与最近的不同于其所属簇的簇内所有数据点之间的平均距离（最近簇的簇内平均距离）：

$$b(i) = \min_{k \neq c} \frac{1}{N_k} \sum_{j \in C_k} d(i,j) \tag{21}$$

其中，k 是与数据点 i 所属簇 c 不同的簇的索引，N_k 是簇 k 内的数据点总数，C_k 是簇 k 内的所有数据点的集合，$d(i,j)$ 表示数据点 i 与数据点 j 之间的距离，通常可以使用欧氏距离、曼哈顿距离或其他距离度量来计算。

$b(i)$ 计算了数据点 i 与最近的不同于其所属簇的簇内所有数据点之间的平均距离，用于衡量数据点 i 与其最近邻近簇的数据点的不相似性。借助 $a(i)$ 于 $b(i)$ 计算轮廓系数：

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{22}$$

总体而言，较高的整体轮廓系数表示聚类效果较好，而较低的整体轮廓系数表示聚类效果较差。轮廓系数是一种常用的聚类评估指标，可以帮助选择合适的簇数以及评估聚类算法的性能。