Java

禚文宇
202100800179

日期：October 24, 2022

摘　　要

1　数据集介绍

本次上机作业使用的是 3w 数据集，3w 数据集是第一个公开的记录了油井中罕见的不良真实事件的数据集，可以作为基准数据集，用于开发与实际数据固有困难相关的机器学习技术。

关于该数据集背后的理论的更多信息，可在《石油科学与工程杂志》（Journal of Petroleum Science and Engineering）上发表的论文《油井中罕见不良真实事件的现实和公共数据集》[1] 中找到。

对于数据集中的数据，其部分属性信息如下表所示：

属性	含义
P-PDG	永久井下压力表的压力
P-TPT	压力传感器的数据
T-TPT	温度传感器的数据

表 1: 数据部分属性信息

2　使用的算法介绍

本次上机实验，我尝试了两个分类算法，分别为 K-最近邻（KNN）算法和朴素贝叶斯算法，下面我将简单介绍这两个算法。

2.1　K-最近邻算法（KNN）介绍

K-最近邻算法（KNN）是一种用于分类和回归的非参数统计方法：

1. 在 KNN-分类中，通过 K 个最近邻居中出现次数最多的分类决定了此对象的分类；
2. 在 KNN-回归中，K 的最近邻居的值的平均值将会称为此对象的预测值。

KNN 是一个非常简单的机器学习算法，分为计算距离、取 K 个最近邻居、根据邻居分类三个步骤。

2.1.1 具体过程

1. 计算距离：在 KNN 中，我们通过 Euclid 距离来度量两个对象 $\theta_0 = (x_0, x_1, \cdots, x_n)$ 和 $\theta_1 = (y_0, y_1, \cdots, y_n)$ 之间的距离，具体定义为
$$\mathrm{dis}(\theta_0, \theta_1) = \sum_{i=0}^{n} \sqrt{(x_i - y_i)^2}$$
必须注意到的是，两个对象必须具有相同的数据维度，否则 Euclid 距离将无法计算。
2. 取 K 个最近邻居：本步骤非常容易，即按照 Euclid 距离排序后，取前 K 个互异的数据点即可。
3. 在根据邻居分类：KNN-分类中，我们只需取出这 K 个最近邻居的标签，找出出现次数最多的标签即为返回值。

2.2 朴素贝叶斯算法介绍

朴素贝叶斯算法是一个基于贝叶斯公式的算法：设 $(\Omega, \mathscr{F}, P)$ 是概率空间，$A_1, A_2, \ldots, A_n$ 是样本空间 Ω 的一个分割，则对任意 $B \in \mathscr{F}$，$P(B) > 0$，有
$$P(A_k \mid B) = \frac{P(A_k)P(B \mid A_k)}{\sum_{j=1}^{n} P(A_j)P(B \mid A_j)}$$
我们可以这样理解这个公式：假设某个过程具有 $A_1, A_2, \cdots, A_n$ 这样 n 个可能的前提（原因），而 $P(A_1), P(A_2), \cdots, P(A_n)$ 是人们对这 n 个可能的前提（原因）的可能性大小的一种事前估计，称之为**先验概率**。当这个过程有了一个结果 B 之后，人们会通过条件概率 $P(A_1 \mid B), P(A_2 \mid B), \cdots, P(A_n \mid B)$ 来对这 n 个可能前提的可能性大小做出一个新的认识，因此将这些条件概率称之为**后验概率**，而贝叶斯公式恰好提供了一种计算后验概率的工具。

3 实验过程

3.1 数据清洗

3.2 min-max 标准化

3.3 使用 Java-ML 库中自带的 KNN 算法

3.4 手写 KNN 算法

3.5 使用 Java-ML 库中自带的朴素贝叶斯算法

参考文献

[1] Ricardo Emanuel Vaz Vargas et al. “A realistic and public dataset with rare undesirable real events in oil wells”. In: *Journal of Petroleum Science and Engineering* 181 (2019), p. 106223. ISSN: 0920-4105. DOI: https://doi.org/10.1016/j.petrol.2019.106223. URL: http://www.sciencedirect.com/science/article/pii/S0920410519306357.