数据探索性分析与数据预处理

分析报告

赵文天 2120171105

一． 数据集与数据预处理

1. 数据集

本次选用的数据集为 Titanic: Machine Learning from Disaster 数据集，分为训练集和测试集两部分，测试集不包含标签。训练集和测试集中包含的属性如下：

属性	定义
pclass	票等级
sex	性别
Age	年龄
sibsp	登船的兄弟姐妹或配偶数量
parch	登船的父母或子女数量
ticket	票编号
fare	费用
cabin	船舱号
embarked	登船地点

2. 数据预处理

为了方便对样本进行分类以及聚类，选取了一些属性并进行了预处理，作为分类和聚类使用的特征。选取的标称属性包括 pclass，sex，embarked 三个属性，数值属性包括 Age, sipsp, parch, fare 四个属性。

由于所有标称属性可能的取值个数都较小，这里用特征向量中的多个维度表示标称属性的一个可能取值，如 embarked 属性有三种可能取值，该属性取值为'S', 'C', 'Q'时，对应的向量分别为[1, 0, 0]，[0, 1, 0]， [0, 0, 1]。数值属性用特征向量中的一个维度表示。同时，对数值属性进行了归一化，使得其取值在(0, 1)区间内。最终，得到一个样本对应的特征。

对于缺失的标称属性，用该属性中取值最多的属性进行填补；对于缺失的数值属性，用该属性的均值进行填补。

二． 分类方法以及可视化

这里选取了两种方法对数据进行分类，分别是神经网络与支持向量机。分类器的输入是样本对应的特征，输出是对该样本进行预测的标签，即幸存或没有幸存。通过输入有标签的数据对分类器进行训练，然后对没有标签的数据的标签进行预测。

神经网络方法：在使用神经网络进行分类时，使用 2 个全连接层的神经网络。使用 relu 激活函数，训练时的 dropout 概率为 0.3。2 个全连层后接 softmax 激活函数，网络的输出为 2 个单元，分别表示没有幸存和幸存的概率。训练时，使用 adam optimizer，学习率为 1e-5。神经网络使用 tensorflow 实现。

支持向量机方法：支持向量机（SVM）通过寻找一个分割超平面对数据进行分类，在训练过程中，我们希望超平面最近的样本（即支持向量）距离超平面的距离尽可能远。经典的 SVM 可以处理线性可分的情况。对于线性不可分的情况，可以使用核函数将样本映射到高维空间，使得在原始空间中线性不可分的样本在高维空间中变得线性可分。这里使用 sklearn 中的 svm 实现。

为了验证分类结果的有效性，我们在训练集上分别对两个分类器进行了训练，并在测试集上进行预测，提交分类结果。使用支持向量机与神经网络的分类器分别取得了 77.99%与 77.51%左右的正确率，证明两种分类方法在这个二分类问题上都是有效的。

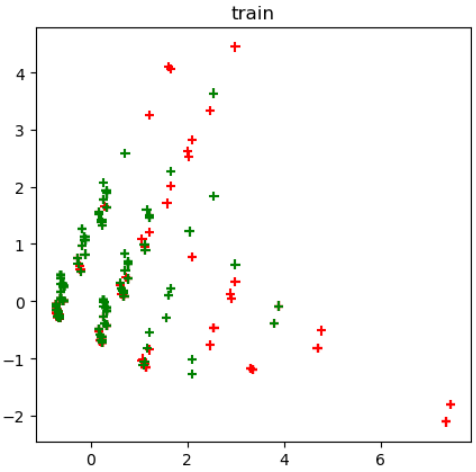
神经网络方法：

Submission and Description	Public Score	Use for Final Score
submission_nn.csv a few seconds ago by Wentian add submission details	0.77990	☐

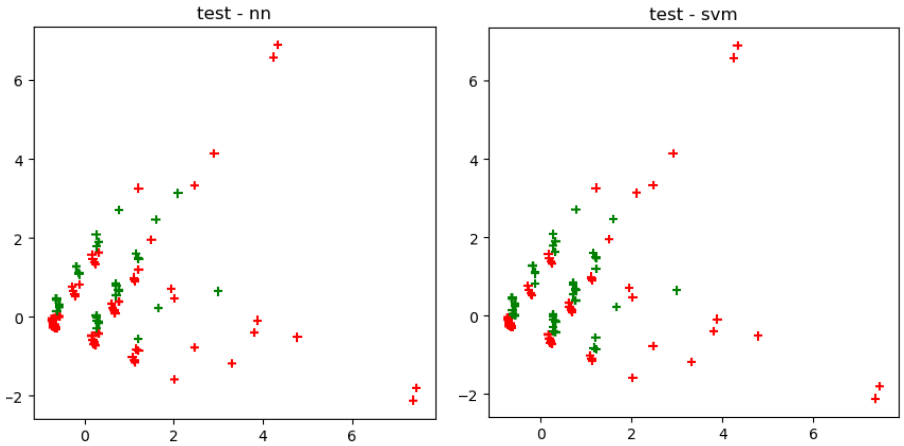
支持向量机方法：

submission_svm.csv a few seconds ago by Wentian add submission details	0.77511	☐

在对分类结果进行可视化之前，我们先用 PCA 方法对对训练数据进行降维并对降维后的数据绘制散点图，观察训练数据的分布：

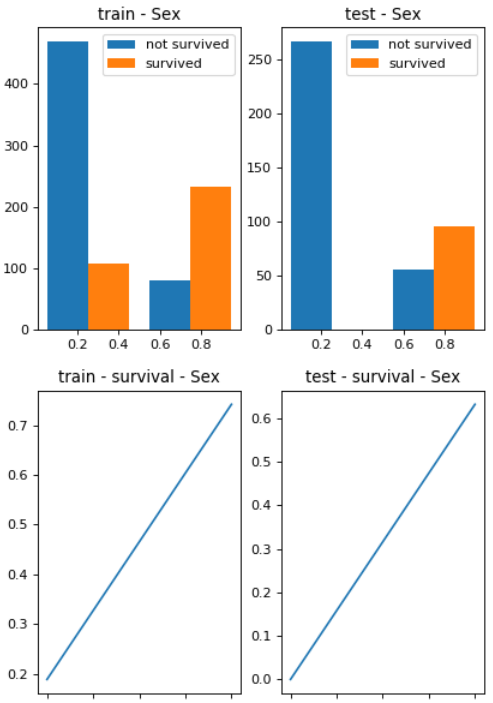


其中红色和绿色的点分别表示标签为没有幸存和幸存的样本。对神经网络的分类结果与 SVM 的分类结果分别进行可视化：

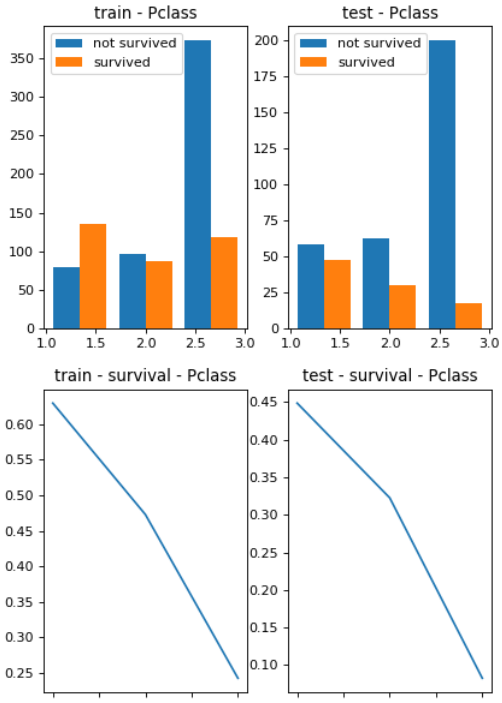


可见两种方法对测试集的划分较为相似，这是因为两种方法都使用了相同的监督信息进行训练，且原始数据的分布较为简单，训练后得到的分类器对数据进行了相似的分类。

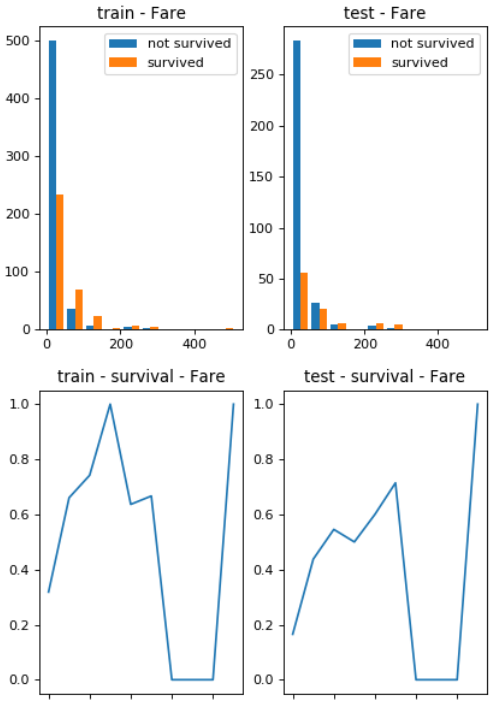
对训练和测试数据中的一些属性作直方图，进行分析：



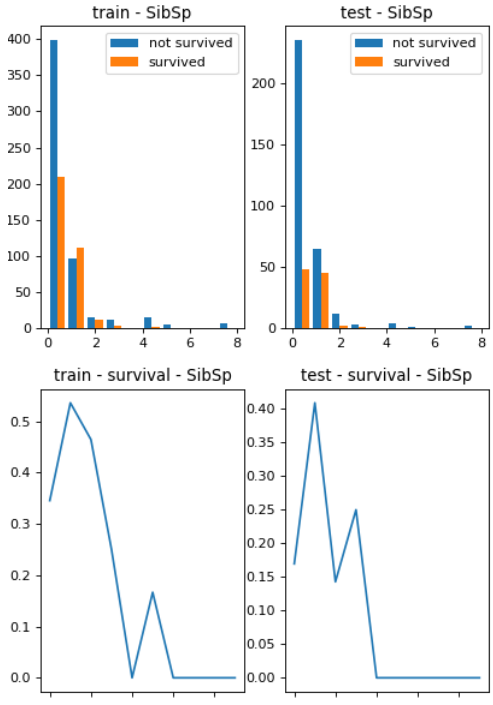
(a)



(b)



(c)



(d)

图中，上方分别是训练数据和测试数据中，某个属性的分布直方图；下方分布为训练数据和测试数据中，存活率（被标记或分类为存活的样本数与样本总数之比）的分布。图(a) (b) (c) (d) 分别表示性别，票等级，费用以及兄弟姐妹数量四个属性的分布。对于性别属性，在作图时男性被标记为 0，女性被标记为 1。可见在训练和测试数据中女性的存活率都显著高于男性。对于票等级（Pclass）属

性，该属性的值越接近 3，存活率越低。根据数据字典，1 为最高等级，3 为最低等级；这说明票等级越低，存活的概率越低。

对于费用属性，训练集和测试集的存活率分布相似。费用在 0-100 之间时，存活率随费用的增加而提高；费用在 100-300 之间，存活率下降；费用在 300-400 之间存活率几乎为 0；费用大于 400 时存活率接近 1。对存活率呈现这种分布的分析如下：费用在 100 至 300 之间的人数与其他区间相比较少，费用在 300 至 400 之间的人数很少，因此统计出的存活率也较低。费用在 0-100 之间的人数较多，且费用趋近于最大值时存活率接近 1，这说明在这些区间内，乘客花费的费用越高，存活的可能性越大。将其与票等级属性（票等级越高，存活率越高）共同比较可以发现，乘客越富有，其生存的概率越大。

对于兄弟姐妹数量（sibsp）属性，存活率都呈现随兄弟姐妹数量的增加而减少的趋势，说明同行的兄弟姐妹数量越少，存活率越高。

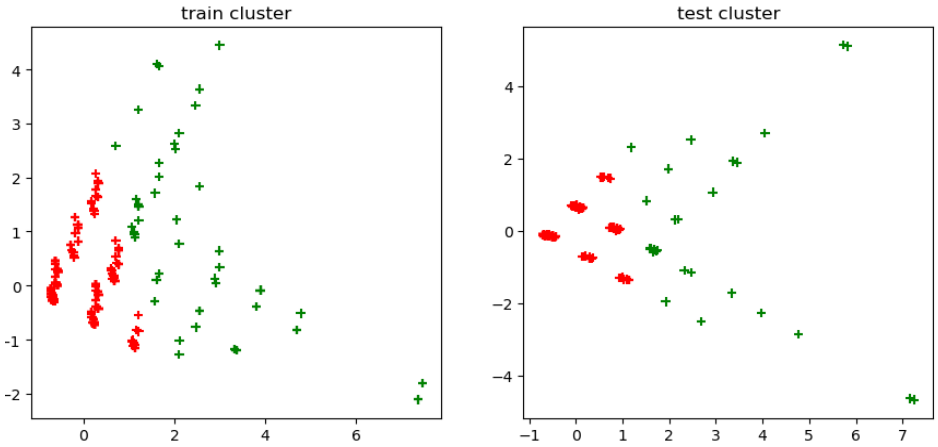
三． 聚类方法以及可视化

在这一部分，采用了两种不同方法进行聚类。忽略训练数据的标签，将所有样本的特征进行无监督的聚类。

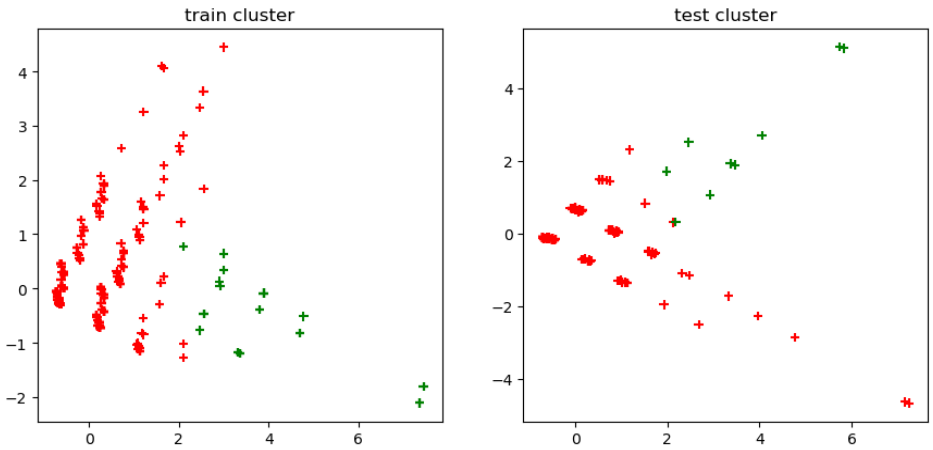
K-means：迭代地确定样本所属的类别。首先选取 k 个初始点作为聚类中心，每次迭代时将所有样本分配到最近的点所属的类别。之后对每一个类别，以其中所有样本的均值作为新的聚类中心。

层次聚类：通过计算不同样本之间的相似性创建有层次的聚类树。这里采用合并的方法进行聚类。初始状态为每个样本属于一类，每次迭代计算每两个类之间的距离，并合并最近的两个类别，直到所有样本归为一个类别。

将训练集和测试集的样本用两种方法进行聚类，并对聚类结果进行可视化：



使用 k-means 进行聚类的结果



使用层次聚类进行聚类的结果

在上图中，红色和绿色的散点分别表示聚类得到的两个类别。可见，两种聚类方法都倾向于将相似的样本归为同一类别，与之前的分类方法得到的结果比较，发现分类方法得到的两个类别的样本中存在距离很近的样本，且两个类别的分布存在重叠，与聚类方法的结果有显著差异。