**机器学习报告名称：**企业创新性预测

**一、基本信息**

**姓名:** 张宇洋

**学号：**41821219

**二、上机作业内容与结果（包括图形截图）：**

1. 目的：

学习理解各种算法的实现，和现实问题结合起来，学习如何简单的建立模型处理现实中遇到的问题。

1. 原理：

随机森林：

随机森林中有许多的分类树。我们要将一个输入样本进行分类，我们需要将输入样本输入到每棵树中进行分类。打个形象的比喻：森林中召开会议，讨论某个动物到底是老鼠还是松鼠，每棵树都要独立地发表自己对这个问题的看法，也就是每棵树都要投票。该动物到底是老鼠还是松鼠，要依据投票情况来确定，获得票数最多的类别就是森林的分类结果。

核心思想：

将若干个弱分类器的分类结果进行投票选择，从而组成一个强分类器，这就是随机森林bagging的思想（关于bagging的一个有必要提及的问题：bagging的代价是不用单棵决策树来做预测，具体哪个变量起到重要作用变得未知，所以bagging改进了预测准确率但损失了解释性）。

谱聚类：

谱聚类是从图论中演化出来的算法，后来在聚类中得到了广泛的应用。它的主要思想是把所有的数据看做空间中的点，这些点之间可以用边连接起来。距离较远的两个点之间的边权重值较低，而距离较近的两个点之间的边权重值较高，通过对所有数据点组成的图进行切图，让切图后不同的子图间边权重和尽可能的低，而子图内的边权重和尽可能的高，从而达到聚类的目的。

SVM:

在[机器学习](https://zh.wikipedia.org/wiki/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0" \o "机器学习)中，支持向量机（support vector machine，常简称为SVM，又名支持向量网络）是在[分类](https://zh.wikipedia.org/wiki/%E5%88%86%E7%B1%BB%E9%97%AE%E9%A2%98" \o "分类问题)与[回归分析](https://zh.wikipedia.org/wiki/%E8%BF%B4%E6%AD%B8%E5%88%86%E6%9E%90" \o "回归分析)中分析数据的[监督式学习](https://zh.wikipedia.org/wiki/%E7%9B%A3%E7%9D%A3%E5%BC%8F%E5%AD%B8%E7%BF%92" \o "监督式学习)模型与相关的学习[算法](https://zh.wikipedia.org/wiki/%E7%AE%97%E6%B3%95" \o "算法)。给定一组训练实例，每个训练实例被标记为属于两个类别中的一个或另一个，SVM训练算法创建一个将新的实例分配给两个类别之一的模型，使其成为非概率[二元](https://zh.wikipedia.org/w/index.php?title=%E4%BA%8C%E5%85%83%E5%88%86%E7%B1%BB%E5%99%A8&action=edit&redlink=1)[线性分类器](https://zh.wikipedia.org/wiki/%E7%BA%BF%E6%80%A7%E5%88%86%E7%B1%BB%E5%99%A8" \o "线性分类器)。SVM模型是将实例表示为空间中的点，这样映射就使得单独类别的实例被尽可能宽的明显的间隔分开。然后，将新的实例映射到同一空间，并基于它们落在间隔的哪一侧来预测所属类别。

除了进行线性分类之外，SVM还可以使用[核技巧](https://zh.wikipedia.org/w/index.php?title=%E6%A0%B8%E6%8A%80%E5%B7%A7&action=edit&redlink=1)有效地进行非线性分类，将其输入隐式映射到高维特征空间中。

函数间隔与几何间隔：

在分离超平面固定为的时候x+b|表示点x到超平面的相对距离。通过观察和y是否同号，我们判断分类是否正确，这些知识我们在感知机模型里都有讲到。这里我们引入函数间隔的概念，定义函数间隔为：

可以看到，它就是感知机模型里面的误分类点到超平面距离的分子。对于训练集中m个样本点对应的m个函数间隔的最小值，就是整个训练集的函数间隔。

函数间隔并不能正常反应点到超平面的距离，在感知机模型里我们也提到，当分子成比例的增长时，分母也是成倍增长。为了统一度量，我们需要对法向量w加上约束条件，这样我们就得到了几何间隔γ,定义为：

支持向量:

　在感知机模型中，我们可以找到多个可以分类的超平面将数据分开，并且优化时希望所有的点都被准确分类。但是实际上离超平面很远的点已经被正确分类，它对超平面的位置没有影响。我们最关心是那些离超平面很近的点，这些点很容易被误分类。如果我们可以让离超平面比较近的点尽可能的远离超平面，最大化几何间隔，那么我们的分类效果会更好一些。

求解：

SVM的模型是让所有点到超平面的距离大于一定的距离，也就是所有的分类点要在各自类别的支持向量两边。用数学式子表示为：

通过解对偶问题我们可以求出

最终分类超平面为：

1. 内容：

1). 进行数据的修整

2). 对数据进行极差标准化，并且取对数

3). 对数据进行各种方法的聚类并选择最佳参数，选择合适的聚类方法

同时对于相同的企业利用聚类得到的类别均值重新标签。

4). 对模型进行因子分析

5). 对聚类的结果，再选用合适的分类方法以及参数进行分类。

1. 结果：

Bartlett's test of sphericity ：

106084.50882499242 0.0

KMO：

0.5029169271912565

线性分类器：

0.827538726333907

多项式核函数非线性分类器：

0.8244406196213425

高斯核函数非线性分类器：

0.8364888123924269

**三、文件与代码说明：**

<函数说明>：

将代码按过程分成了不同的文件。

<机器学习课程-code>

Fin\_Pretreatment.py  数据的预处理代码，可以对缺失数据利用随机森林进行补全。

Fin\_Cluster.py  代码中比较了各种的聚类方法，最终选择了谱聚类的方法进行聚类，并且重新进行标签。

Fin\_Factor\_Analyze.py  进行因子分析，判断是否可以重新调整变量

Fin\_Classify.py     利用SVM进行多分类。

**四、备注：**

部分处理是在Excel中进行的，在word中并没有列出。