人工智能程序设计实验报告

白晋斌
171860607

目录

任务 1 回归（对应文件：data_akbilgic.csv ）..... 3

1. 使用可视化的方法观察数据之间的关联，推测该数据是否适合进行回归分析/线性回归分析。 3

2. 使用回归分析的方法（如线性回归）进行回归分析，并与你的推测结果进行对比和思考。（实验 过程中请注意评价指标、训练误差、泛化误差、测试数据划分等内容，并记录在实验报告中。） 6

3. （附加题）尝试使用降维前后的数据表示分别进行回归，并比较回归的结果，思考降维对该回归 任务的影响。 6

任务 2 二分类（对应文件：wdbc.csv） 9

1. 对检查数据进行处理并使用降维方法（如 PCA）进行降维（2 维或 3 维）。通过可视化观察降维结 果，并推测该数据是否适合进行分类学习。 10

2. 使用分类方法（如 logistic regression）对上述问题进行分类学习，并与你的推测结果进行对比和 思考。（实验过程中请注意评价指标、训练误差、泛化误差、测试数据划分等内容，并记录在实 验报告中。） 11

3. （附加题）尝试使用降维前后的数据表示分别进行分类，并比较分类的结果，思考降维对该分类 任务的影响。 11

任务 3 多分类 （对应文件：dataset.csv） 11

1. 用降维算法将数据降为 2 维或者 3 维，并以不同的颜色表示各类别进行可视化。 11

2. 尝试比较不同的降维算法（如自己实现的 PCA 算法、sklearn 中的 PCA方法，以及其他可能的降维方法等）的结果差异，如通过可视化结果进行比较等。 12

3. （附加题）尝试对此数据进行分类（分类方法、评估方法自行选择），并报告分类结果。 13

文件说明（readme） 14

1.py 文件分别对应第一题、第二题、第三题的全部代码 14

2.代码执行流程与修改过程详细记录于 jupyter 文件中，为方便读取，将该文件另存为 html 格式。 14

完成以下三个任务，提交源代码和实验报告。请在实验报告中记录你的实验过程、实验结果和思考。

任务 1 回归（对应文件：data_akbilgic.csv ）

任务描述：在给定的数据文件中，每一行代表一个开盘日中的股指交易涨跌值，第一列记录具体日 期，其后每一列代表一项股指数据，共九列，依次为：ISE(TL-based), ISE(usd), SP, DAX, FTSE, NIKKEI, BOVESPA, EU, EM。回归任务是通过后八项股指来对第一项股指（ISE(TL-based)）的数值 进行预测。请完成下列工作：

1. 使用可视化的方法观察数据之间的关联，推测该数据是否适合进行回归分析/线性回归分析。

此题目分析数据之间关联性，我们通过对数据的简单读取，处理为 Dataframe 格式，即可得到 pearson 相关系数、spearman 相关系数、kendall 相关系数以及协方差。

如下图。

pearson 相关系数：

	ISE(TL BASED)	ISE(USD BASED)	SP(imkb_x)	DAX	FTSE	\
ISE(TL BASED)	1.000000	0.942897	0.439489	0.602081	0.622948	
ISE(USD BASED)	0.942897	1.000000	0.449561	0.629218	0.648740	
SP(imkb_x)	0.439489	0.449561	1.000000	0.685843	0.657673	
DAX	0.602081	0.629218	0.685843	1.000000	0.867369	
FTSE	0.622948	0.648740	0.657673	0.867369	1.000000	
NIKKEI	0.260052	0.393225	0.131250	0.258538	0.255236	
BOVESPA	0.432898	0.446889	0.722069	0.585791	0.596287	
EU	0.655519	0.690761	0.687550	0.936393	0.948963	
EM	0.600295	0.701954	0.528243	0.665162	0.687543	

	NIKKEI	BOVESPA	EU	EM
ISE(TL BASED)	0.260052	0.432898	0.655519	0.600295
ISE(USD BASED)	0.393225	0.446889	0.690761	0.701954
SP(imkb_x)	0.131250	0.722069	0.687550	0.528243
DAX	0.258538	0.585791	0.936393	0.665162
FTSE	0.255236	0.596287	0.948963	0.687543
NIKKEI	1.000000	0.172752	0.283750	0.547288
BOVESPA	0.172752	1.000000	0.621704	0.688074
EU	0.283750	0.621704	1.000000	0.716502
EM	0.547288	0.688074	0.716502	1.000000

spearman 相关系数：

	ISE(TL BASED)	ISE(USD BASED)	SP(imkb_x)	DAX	FTSE	\
ISE(TL BASED)	1.000000	0.917323	0.381885	0.553468	0.597388	
ISE(USD BASED)	0.917323	1.000000	0.378619	0.581025	0.622367	
SP(imkb_x)	0.381885	0.378619	1.000000	0.593385	0.582349	
DAX	0.553468	0.581025	0.593385	1.000000	0.855156	
FTSE	0.597388	0.622367	0.582349	0.855156	1.000000	
NIKKEI	0.215731	0.352491	0.074585	0.212394	0.230699	
BOVESPA	0.369414	0.376916	0.649375	0.462921	0.467388	
EU	0.620426	0.656724	0.603499	0.931037	0.942116	
EM	0.572711	0.695162	0.447789	0.595062	0.636361	

	NIKKEI	BOVESPA	EU	EM
ISE(TL BASED)	0.215731	0.369414	0.620426	0.572711
ISE(USD BASED)	0.352491	0.376916	0.656724	0.695162
SP(imkb_x)	0.074585	0.649375	0.603499	0.447789
DAX	0.212394	0.462921	0.931037	0.595062
FTSE	0.230699	0.467388	0.942116	0.636361
NIKKEI	1.000000	0.107523	0.250710	0.503999
BOVESPA	0.107523	1.000000	0.483334	0.569039
EU	0.250710	0.483334	1.000000	0.661602
EM	0.503999	0.569039	0.661602	1.000000

kendall 相关系数:

	ISE(TL BASED)	ISE(USD BASED)	SP(imkb_x)	DAX	FTSE	\
ISE(TL BASED)	1.000000	0.767373	0.263867	0.391043	0.421748	
ISE(USD BASED)	0.767373	1.000000	0.263727	0.413254	0.443794	
SP(imkb_x)	0.263867	0.263727	1.000000	0.438132	0.428304	
DAX	0.391043	0.413254	0.438132	1.000000	0.681834	
FTSE	0.421748	0.443794	0.428304	0.681834	1.000000	
NIKKEI	0.145459	0.239743	0.049433	0.142665	0.156270	
BOVESPA	0.257360	0.262920	0.481145	0.329001	0.332743	
EU	0.443578	0.473276	0.446732	0.789306	0.805961	
EM	0.408341	0.511131	0.317292	0.432743	0.462533	

	NIKKEI	BOVESPA	EU	EM
ISE(TL BASED)	0.145459	0.257360	0.443578	0.408341
ISE(USD BASED)	0.239743	0.262920	0.473276	0.511131
SP(imkb_x)	0.049433	0.481145	0.446732	0.317292
DAX	0.142665	0.329001	0.789306	0.432743
FTSE	0.156270	0.332743	0.805961	0.462533
NIKKEI	1.000000	0.073067	0.168987	0.354544
BOVESPA	0.073067	1.000000	0.346206	0.412846
EU	0.168987	0.346206	1.000000	0.486124
EM	0.354544	0.412846	0.486124	1.000000

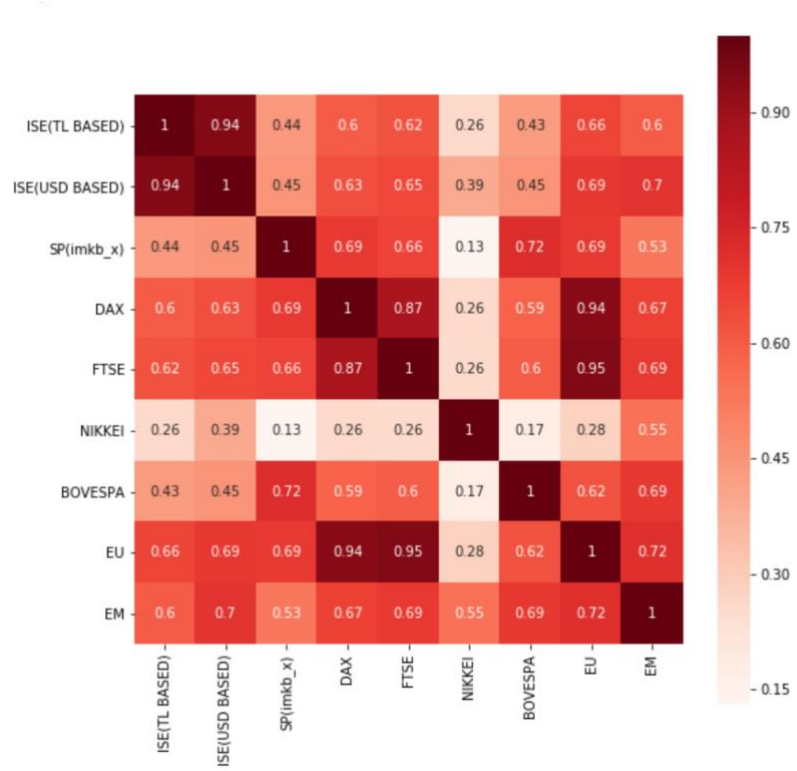
协方差:

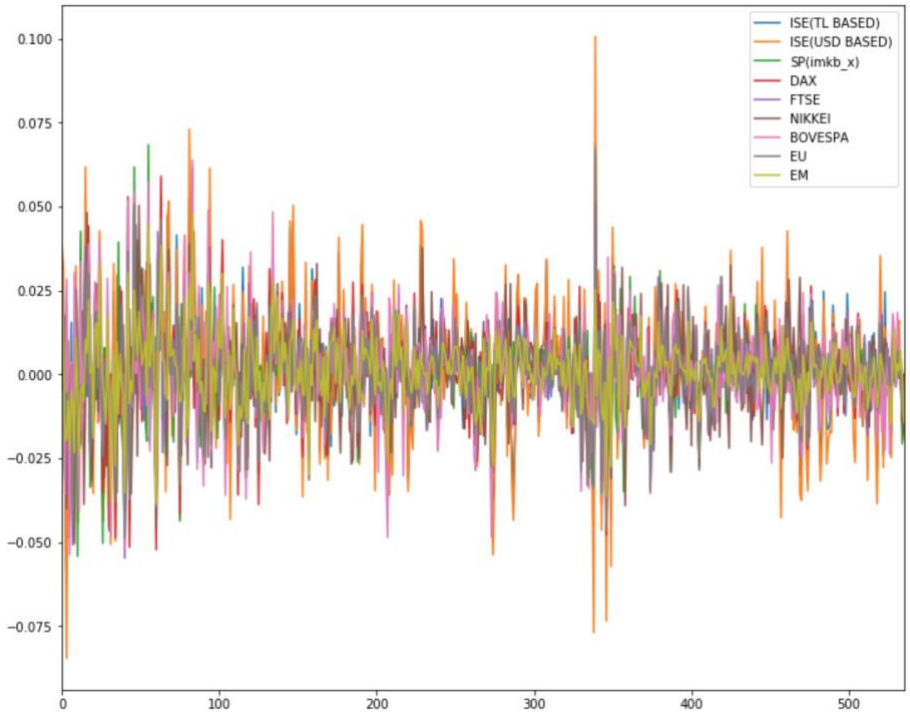
	ISE(TL BASED)	ISE(USD BASED)	SP(imkb_x)	DAX	FTSE	\
ISE(TL BASED)	0.000265	0.000324	0.000101	0.000143	0.000128	
ISE(USD BASED)	0.000324	0.000446	0.000134	0.000193	0.000173	
SP(imkb_x)	0.000101	0.000134	0.000199	0.000141	0.000117	
DAX	0.000143	0.000193	0.000141	0.000212	0.000160	
FTSE	0.000128	0.000173	0.000117	0.000160	0.000160	
NIKKEI	0.000063	0.000123	0.000027	0.000056	0.000048	
BOVESPA	0.000111	0.000149	0.000160	0.000134	0.000119	
EU	0.000138	0.000190	0.000126	0.000177	0.000156	
EM	0.000103	0.000156	0.000078	0.000102	0.000091	

	NIKKEI	BOVESPA	EU	EM
ISE(TL BASED)	0.000063	0.000111	0.000138	0.000103
ISE(USD BASED)	0.000123	0.000149	0.000190	0.000156
SP(imkb_x)	0.000027	0.000160	0.000126	0.000078
DAX	0.000056	0.000134	0.000177	0.000102
FTSE	0.000048	0.000119	0.000156	0.000091
NIKKEI	0.000221	0.000040	0.000055	0.000085
BOVESPA	0.000040	0.000248	0.000127	0.000114
EU	0.000055	0.000127	0.000169	0.000098
EM	0.000085	0.000114	0.000098	0.000110

考虑到数据不够直观，我们采取绘制相关性矩阵图与折现图的方式展现结果。

从下图可以看出，ISE(TL BASED)与某些值（比如说 ISE(USD BASED)、FTSE 等）相关性较高，与某些值（比如说 NIKKEI）相关性较低。因此适合做线性回归分析。



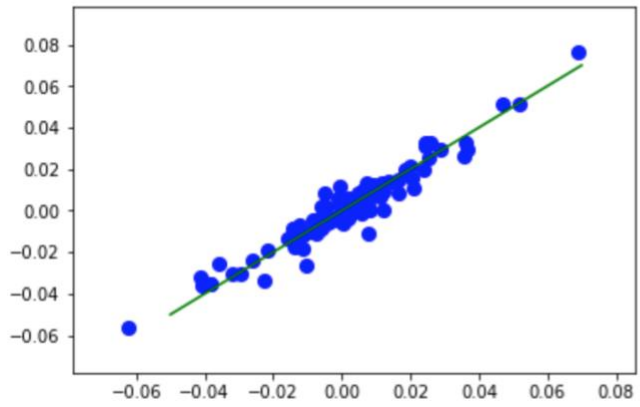


2. 使用回归分析的方法（如线性回归）进行回归分析，并与你的推测结果进行对比和思考。（实验 过程中请注意评价指标、训练误差、泛化误差、测试数据划分等内容，并记录在实验报告中。）

以下是采用线性回归模型，随机取 500 份数据进行训练，之后随机取 50 分数据进行测试。所得到的相关性系数，均方误差和方差分数。

并将预测结果与实际情况进行对比，我们发现较为接近 y=x，故预测结果良好。

```
Coefficients:
 [[ 0.79995676 -0.04388763  0.07256195  0.09232294 -0.1015813   0.09484352
  -0.10653872 -0.24966195]]
Mean squared error: 0.00
Variance score: 0.92
```

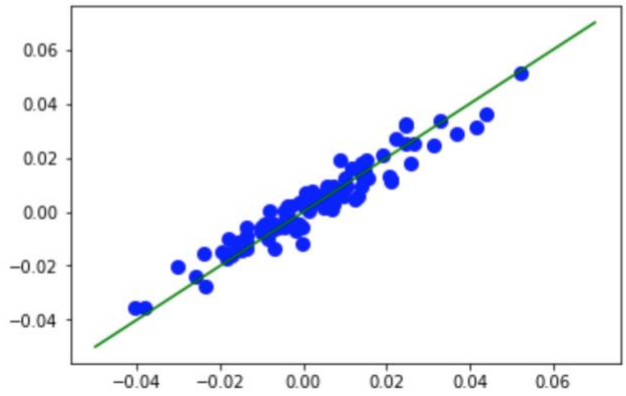


3. （附加题）尝试使用降维前后的数据表示分别进行回归，并比较回归的结果，思考降维对该回归 任务的影响。

我们分别将后八个维度的数据依次降到 7，6，5，4，3，2，1，结果如下图。可以得知，少量的降维并不会影响预测结果的精确性，但当维度降低至 3 维甚至更低，预测精度将大幅度下降。说明当面临高维数据时，特征降维对于机器学习任务非常必要，通过降维有效地消除无关和冗余特征，提高挖掘任务的效率，改善预测精确性等学习性能，增强学习结果的易理解性。但过度降维则会导致数据信息的缺失和精确性的下降。

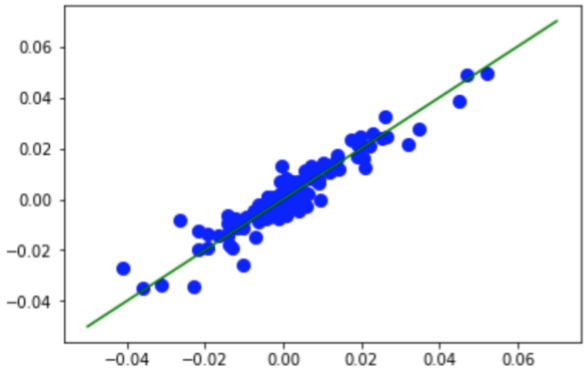
7 维：

```
Coefficients:
 [[ 0.38370899 -0.27821118 -0.50961705 -0.33678611 -0.10097013  0.04476996
  -0.30919855]]
Mean squared error: 0.00
Variance score: 0.92
```



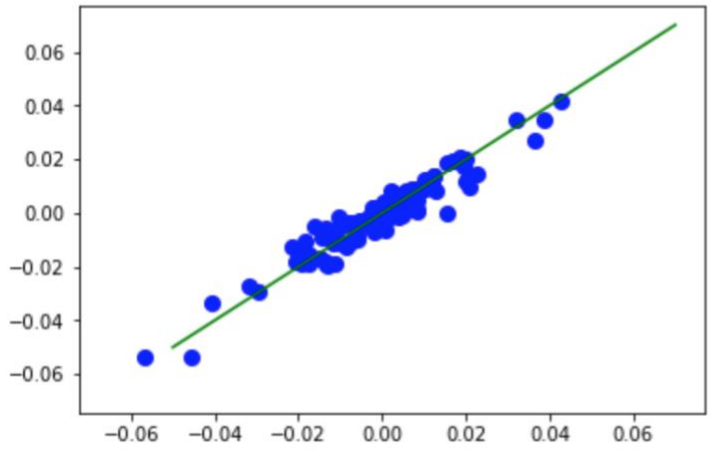
6 维:

```
Coefficients:
 [[ 0.3806577  -0.28159313 -0.50966251 -0.32530375 -0.08997049  0.02675432]]
Mean squared error: 0.00
Variance score: 0.89
```



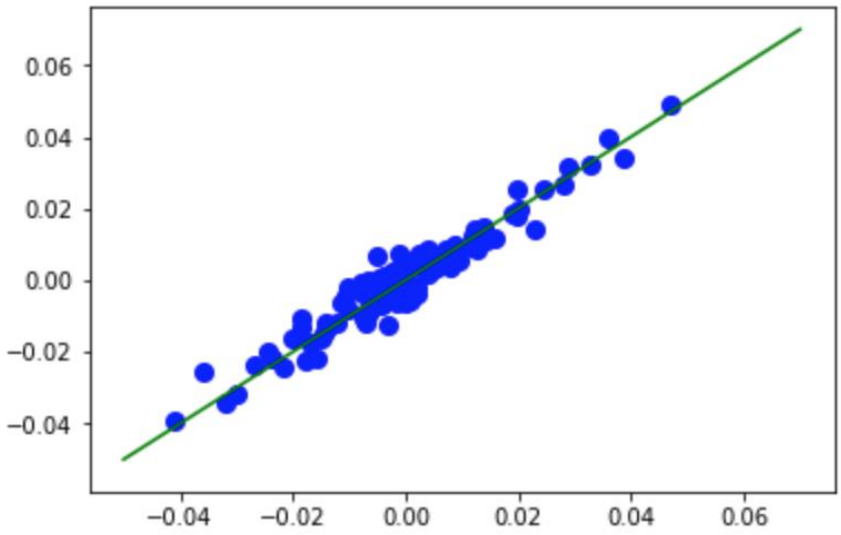
5 维:

```
Coefficients:
 [[ 0.38444185 -0.28375471 -0.49624242 -0.33538903 -0.09869503]]
Mean squared error: 0.00
Variance score: 0.91
```



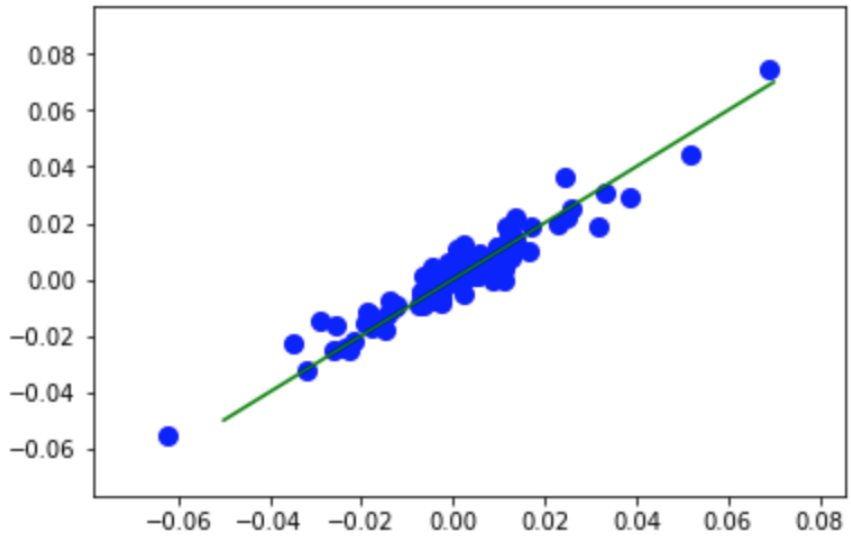
4 维:

```
Coefficients:
 [[ 0.38156    -0.27790259 -0.49513697 -0.32856101]]
Mean squared error: 0.00
Variance score: 0.93
```



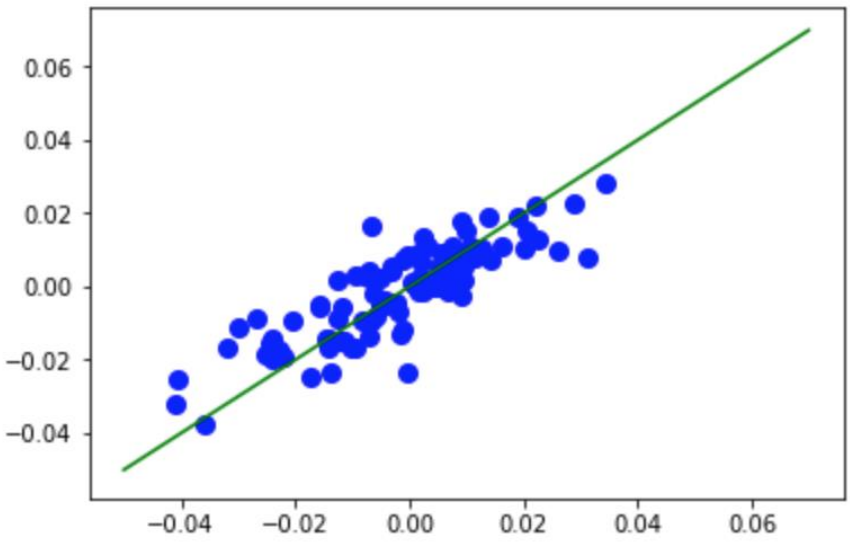
3 维:

```
Coefficients:
 [[ 0.3840152  -0.28669854 -0.49520195]]
Mean squared error: 0.00
Variance score: 0.91
```



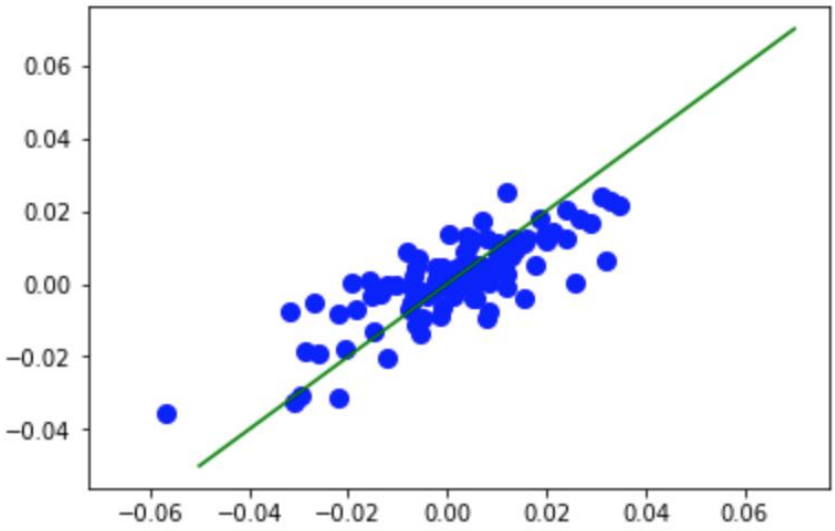
2 维:

```
Coefficients:
 [[ 0.3874866  -0.28653749]]
Mean squared error: 0.00
Variance score: 0.69
```



1 维:

```
Coefficients:
 [[0.39058626]]
Mean squared error: 0.00
Variance score: 0.62
```



任务 2 二分类（对应文件：wdbc.csv）

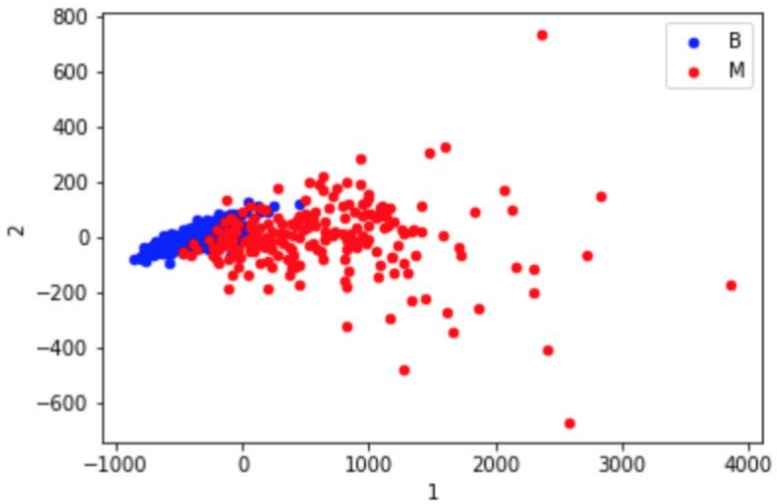
任务描述: 该数据为乳腺癌检查的医疗检测数据，每行对应一个案例(以逗号分隔每个数据)，其中第 一列为案例编号，第二列为诊断结果(M，malignant; B，benign)，其后十列为各项检查数据。分类 任务是通过后十项的检查数据来预测诊断结果。请完成下列工作:

1. 对检查数据进行处理并使用降维方法（如 PCA）进行降维（2 维或 3 维）。通过可视化观察降维结 果，并推测该数据是否适合进行分类学习。

将原数据降到二维后，发现贡献率主要由第一维提供，故后几个维度对数据影响较小，适合分类学习。

此外，观察图片发现，B 较为集中，M 叫维分散，但二者依旧有比较鲜明的分界线，故该数据适合进行分类学习。

```
[0.98204467 0.01617649]
[[1160.1425737  -293.91754364]
 [1269.12244319   15.63018184]
 [ 995.79388896   39.15674324]
 ...
 [ 314.50175618   47.55352518]
 [1124.85811531   34.12922497]
 [-771.52762188  -88.64310636]]
```



2. 使用分类方法（如 logistic regression）对上述问题进行分类学习，并与你的推测结果进行对比和 思考。（实验过程中请注意评价指标、训练误差、泛化误差、测试数据划分等内容，并记录在实 验报告中。）

使用 logistic 模型对数据进行训练，预测。因为数据无时间关系，故我们采取后 50 条数据为预测使用，前面的数据做训练使用。得到的相关参数如下图所示。对我们的模型做了一个简单的评估，均方误差 2%，方差分数 88%。还可以。

```
降维后：
Coefficients:
 [[-0.01186346  0.0309355 ]]
Mean squared error: 0.02
Variance score: 0.88
```

3. （附加题）尝试使用降维前后的数据表示分别进行分类，并比较分类的结果，思考降维对该分类 任务的影响。

这里我们再用降维前的数据进行训练预测，得到的系数与预测结果如下图所示。对我们的模型做了一个简单的评估，均方误差 4%，方差分数 75%，低于降维后数据，说明在本次实验里降维提高了模型预测的精确性。

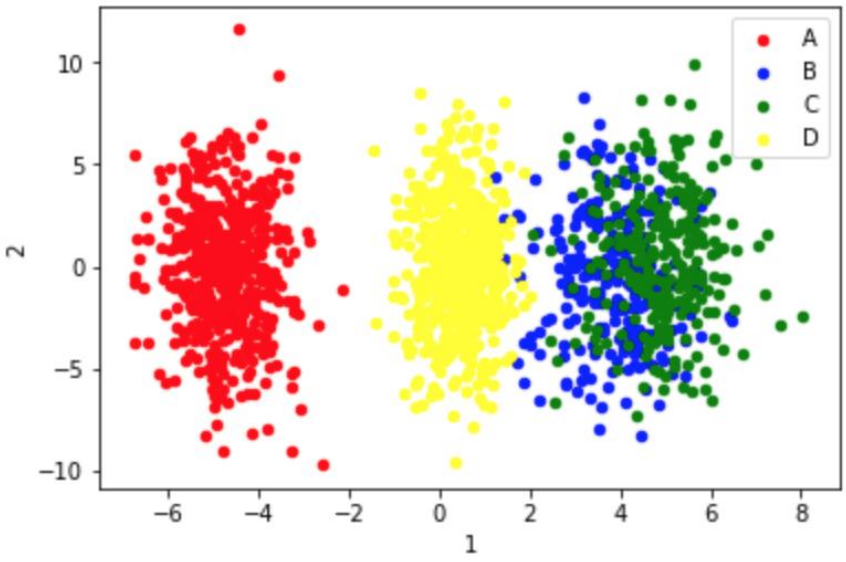
```
Coefficients:
 [[ 2.13934056  0.02187034 -0.0451831  -0.00290226 -0.15032753 -0.40682136
  -0.60928074 -0.32921762 -0.23015434 -0.03013244 -0.02521619  1.26683794
  -0.01595829 -0.09034333 -0.01844614 -0.00252074 -0.03933544 -0.04051865
  -0.04618911  0.00477251  1.24695724 -0.29316114 -0.13413793 -0.02504585
  -0.29164725 -1.1488553  -1.50043226 -0.64450727 -0.70108917 -0.11852539]]
Mean squared error: 0.04
Variance score: 0.75
```

任务 3 多分类 （对应文件：dataset.csv）

任务描述：有 1500 个样例数据（每一行为一个样例，每个样例中有十一列数据，其中第一列为样例 的类别，共四种 A、B、C、D，其后十列为样例的输入特征）。分类任务是通过输入特征来预测样例 的类别。请完成下列工作:

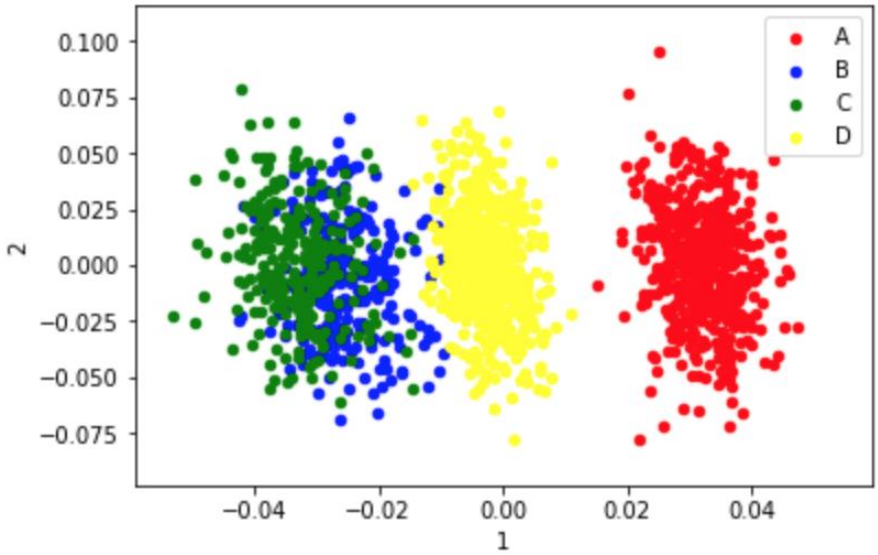
1. 用降维算法将数据降为 2 维或者 3 维，并以不同的颜色表示各类别进行可视化。

这里采用 PCA 法将数据降低到 2 维，绘图如下。



2. 尝试比较不同的降维算法（如自己实现的 PCA 算法、sklearn 中的 PCA方法，以及其他可能的降维方法等）的结果差异，如通过可视化结果进行比较等。

这里我们再次用 ICA 法将数据降到 2 维，结果如下图。



与第一问产生差别的原因是：独立成分分析法（ICA）是基于信息理论的一种常用的降维方法，其与 PCA 主要的不同是 PCA 是寻找不相关的变量，而 ICA 是挑选独立变量。同时 PCA 主要对于高斯分布数据比较有效，而 ICA 适用于其他分布。

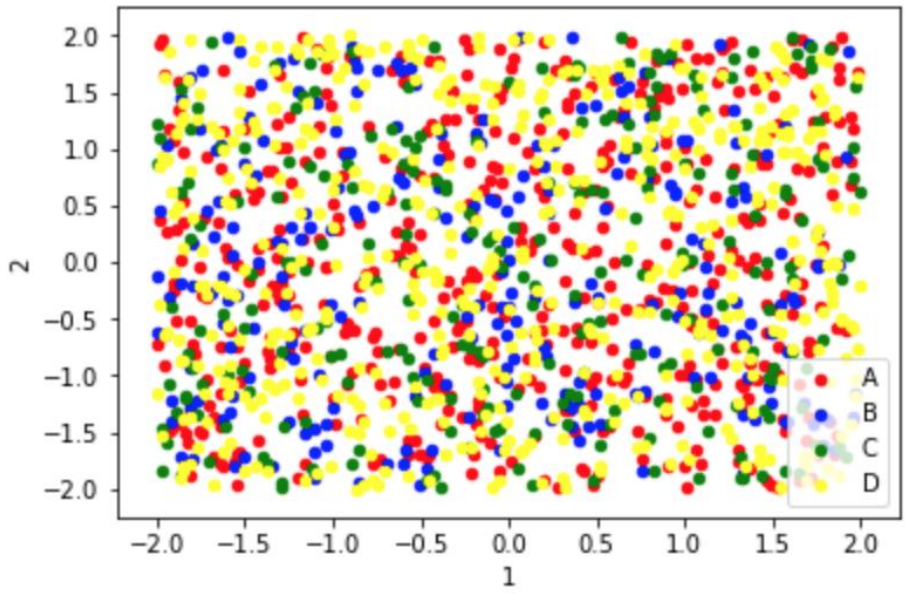
此外，我还尝试了多种基础降维方式，如：

低方差滤波（Low Variance Filter）

如果我们有一个数据集，其中某列的数值基本一致，也就是它的方差非常低，那么这个变量还有价值吗？和上一种方法的思路一致，我们通常认为低方差变量携带的信息量也很少，所以可以把它直接删除。

放到实践中，就是先计算所有变量的方差大小，然后删去其中最小的几个。需要注意的一点是：方差与数据范围相关的，因此在采用该方法前需要对数据做归一化处理。

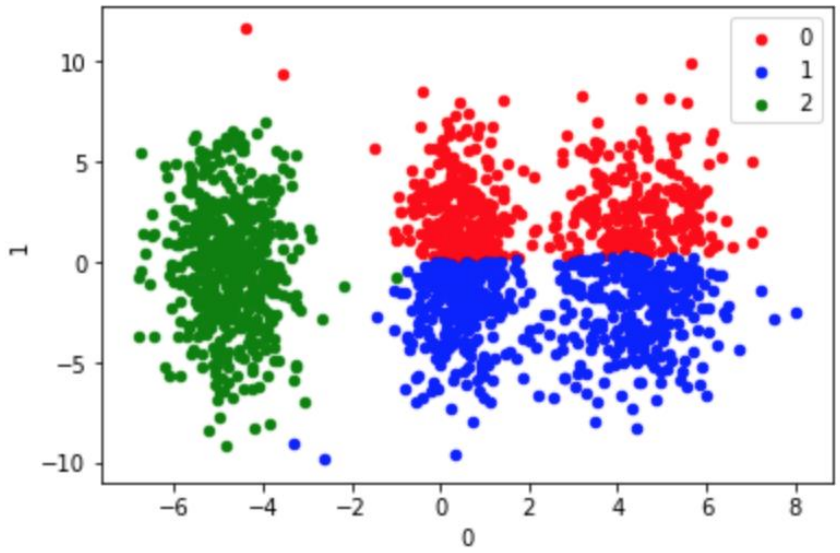
但是实践证明，低方差滤波对本数据集的分类效果并不好。



3. （附加题）尝试对此数据进行分类（分类方法、评估方法自行选择），并报告分类结果。

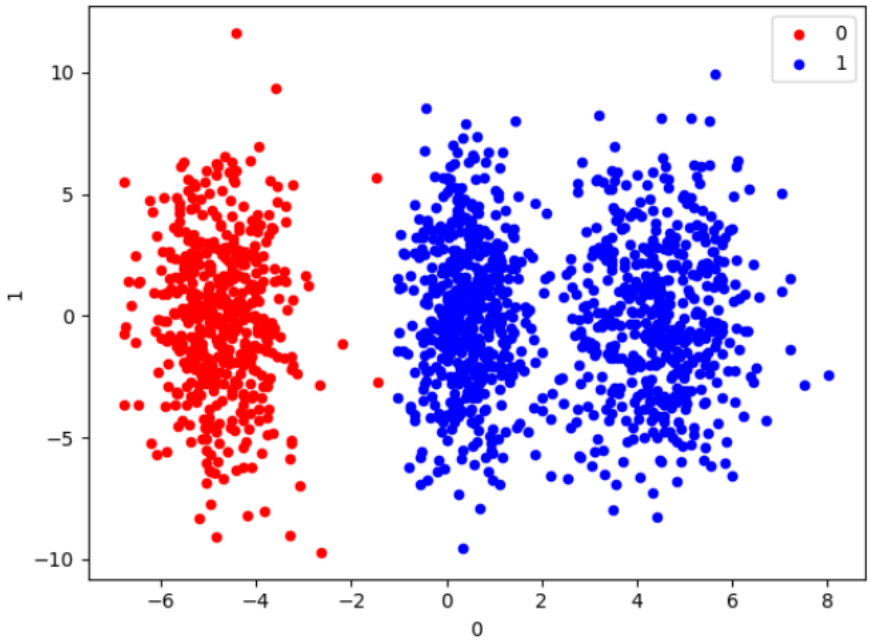
采用 kmeans 法进行分类。

分三类结果如下：



可以看出，第二部分分类结果较好，第 0，1 部分分类结果较迷。原因可能是原始标签中的 BCD 混在一块难以分清。

分成两类则效果较好。



因此，就个人思考，要想像原始数据一样分得开四类，则降维后的数据至少要有三维或更多。

文件说明（readme）

1.py 文件分别对应第一题、第二题、第三题的全部代码

2.代码执行流程与修改过程详细记录于 jupyter 文件中，为方便读取，将该文件另存为 html 格式。