数据测试

来源：http://archive.ics.uci.edu/ml/datasets/Iris 【4】

Iris Data Set

Download: Data Folder, Data Set Description

Abstract: Famous database; from Fisher, 1936

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	736514

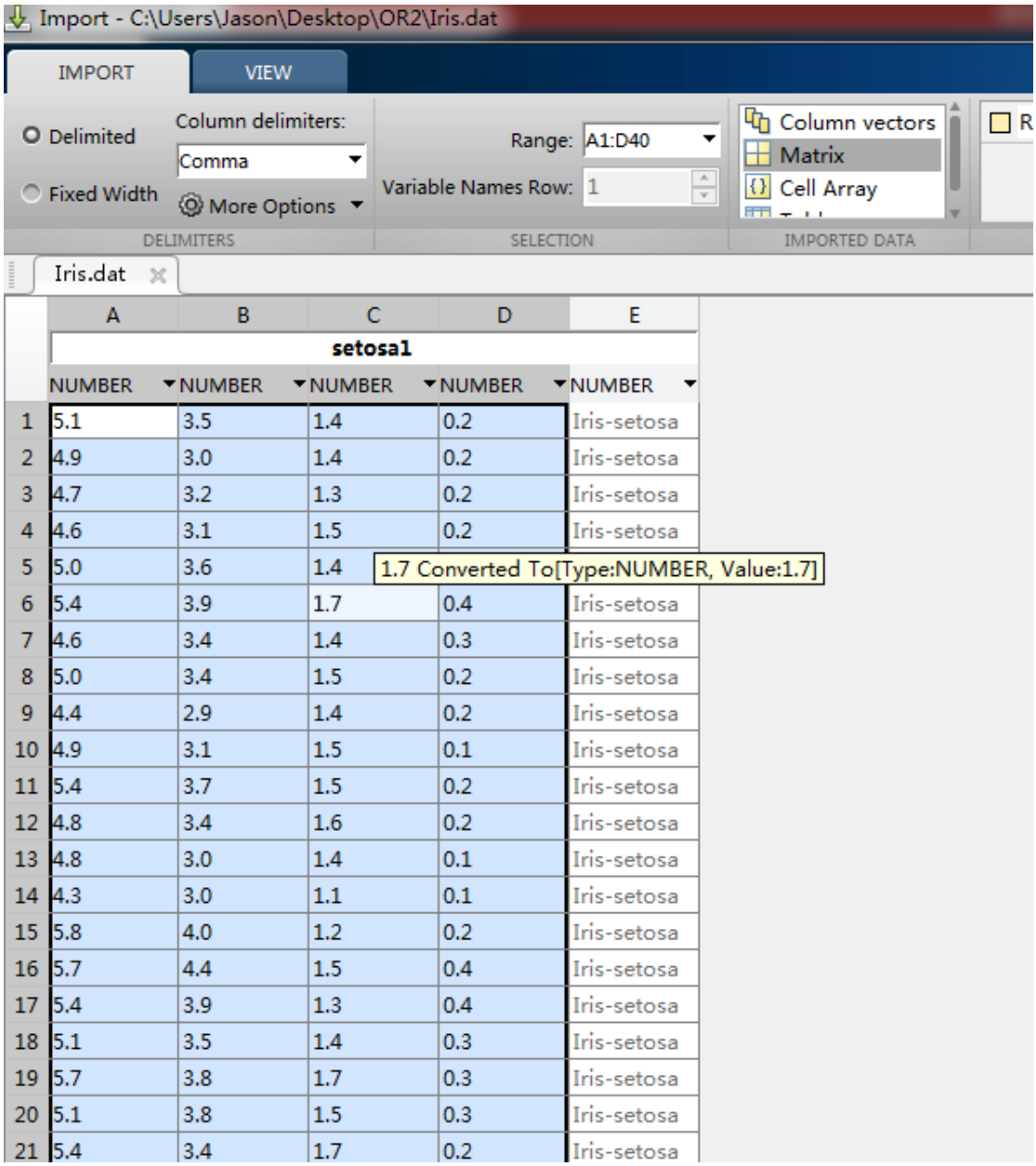
数据说明：我选择了 Iris 目录下 Iris.data 的数据。
Iris 是一种鸢尾花，其数据的属性是四维，分别表示花萼长度等花的特征属性。Iris.data 中共给出三种 Iris 花——setosa，vriginica 以及 versicolor，各 50 组数据，共 150 组数据。

本实验用 setosa 的前 40 组点作为样本训练，后 10 组点作为同分类检验样本，检查判断正确的百分比；再从 virginica 和 versicolor 中选择非同分类，检查判断非此分类的百分比。

首先，将 MATLAB 切换至文件路径下，

```
>> cd c:\Users\Jason\desktop\OR2
```

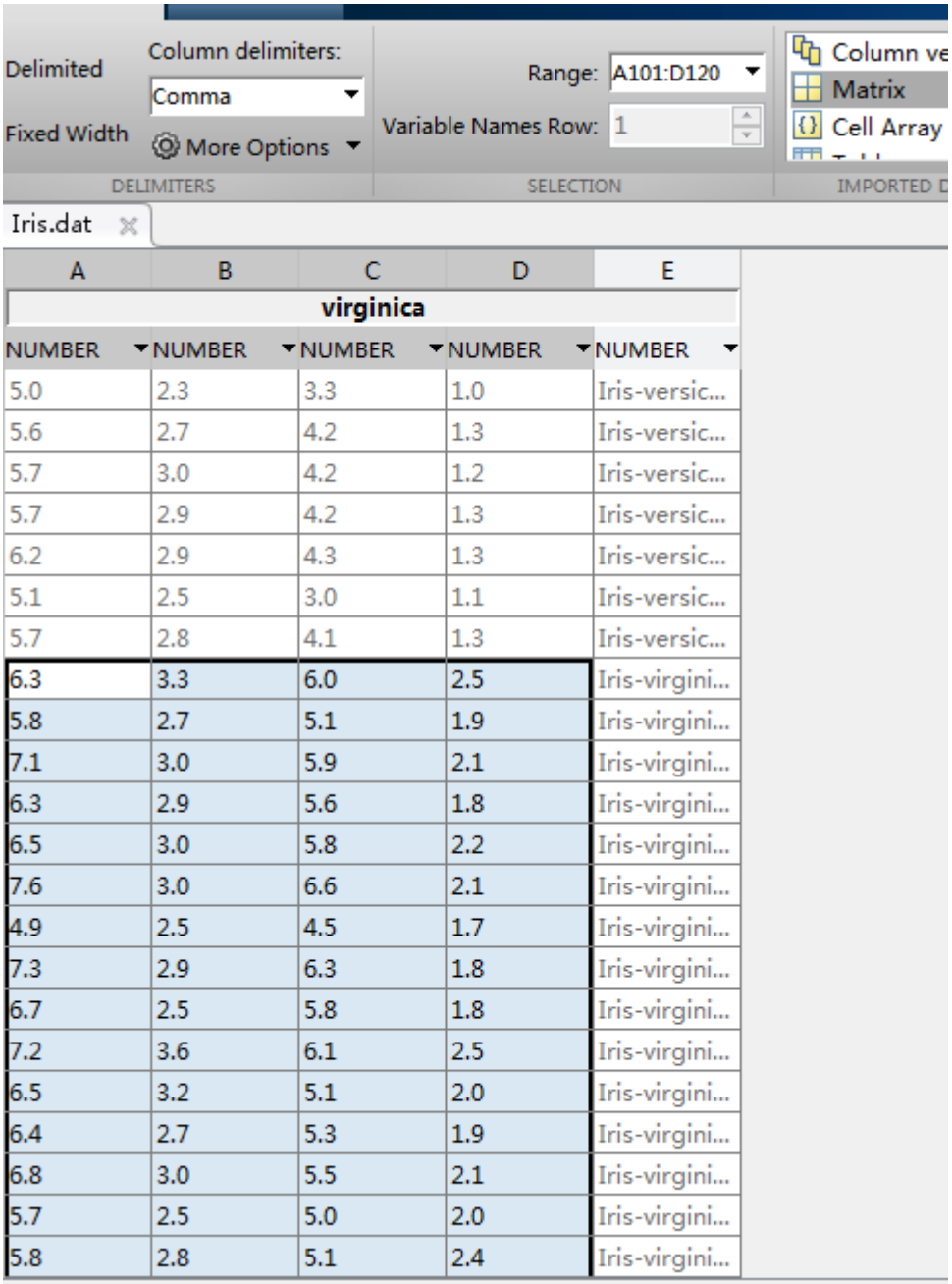
导入训练数据：



将 setosa 类别下的前 40 个点导入，作为训练点集，每个点有 4 个属性。将此 40*4 的矩阵命名为 setosa1；
同理，将 setosa 剩下的 10 个点集作为测试点，导入 10*4 的矩阵，命名为 setosa2.
键入代码，将 setosa1 命名为 data1，setosa2 命名为 data2.
定义想要训练的维度为 2，调用 svdd 模块进行训练，调用 test 模块进行检验，发现同分类下的准确度可以达到 90%。

```
>> cd c:\Users\Jason\desktop\OR2
>> data1=setosa1;
>> data2=setosa2;
>> dimension=2;
>> svdd;
>> test;
9 /10 samples is included in the classification.
1 /10 samples is excluded in the classification.
```

随后，导入 virginica 类中的 20 个点进行排他性测试。



A	B	C	D	E
virginica				
NUMBER	NUMBER	NUMBER	NUMBER	NUMBER
5.0	2.3	3.3	1.0	Iris-versic...
5.6	2.7	4.2	1.3	Iris-versic...
5.7	3.0	4.2	1.2	Iris-versic...
5.7	2.9	4.2	1.3	Iris-versic...
6.2	2.9	4.3	1.3	Iris-versic...
5.1	2.5	3.0	1.1	Iris-versic...
5.7	2.8	4.1	1.3	Iris-versic...
6.3	3.3	6.0	2.5	Iris-virgini...
5.8	2.7	5.1	1.9	Iris-virgini...
7.1	3.0	5.9	2.1	Iris-virgini...
6.3	2.9	5.6	1.8	Iris-virgini...
6.5	3.0	5.8	2.2	Iris-virgini...
7.6	3.0	6.6	2.1	Iris-virgini...
4.9	2.5	4.5	1.7	Iris-virgini...
7.3	2.9	6.3	1.8	Iris-virgini...
6.7	2.5	5.8	1.8	Iris-virgini...
7.2	3.6	6.1	2.5	Iris-virgini...
6.5	3.2	5.1	2.0	Iris-virgini...
6.4	2.7	5.3	1.9	Iris-virgini...
6.8	3.0	5.5	2.1	Iris-virgini...
5.7	2.5	5.0	2.0	Iris-virgini...
5.8	2.8	5.1	2.4	Iris-virgini...

重新定义 data2，并调用 test 模块。

```
>> data2=virginica;
>> test;
0 /20 samples is included in the classification.
20 /20 samples is excluded in the classification.
```

发现，在做异分类排他性问题时，准确率可以达到 100%。

下面进行维度的升高。

重新定义 dimension，使其为 3 为以及 4 维。然后分别做同分类与异分类的检测。
同分类：

```
>> data2=setosa2;
>> dimension=3;
>> test;
6 /10 samples is included in the classification.
4 /10 samples is excluded in the classification.
>> dimension=4;
>> test;
6 /10 samples is included in the classification.
4 /10 samples is excluded in the classification.
```

异分类：

```
>> data2=virginica;
>> dimension=3;
>> test;
0 /20 samples is included in the classification.
20 /20 samples is excluded in the classification.
>> dimension=4;
>> test;
0 /20 samples is included in the classification.
20 /20 samples is excluded in the classification.
```

可以发现，进行维度升高后，同分类的准确度有所降低，而异分类的准确度依旧很高。

交叉检验;
在大作业完成过程中，我询问助教哥哥有关检验方法的问题。比如同一类别的 50 个点如何选择被训练点与测试点。助教给出了交叉检验的思路。
下面，我就进行交叉检验：
重新选择检验样本，选取 setosa 的后 40 个点，然后选取前 10 个点用作检验：
重新导入 setosa1 和 setosa2,

```
>> dimension=2;
>> data1=setosa1;
data2=setosa2;
>> svdd;
>> test;
10 /10 samples is included in the classification.
0 /10 samples is excluded in the classification.
```

升高维度：

```
>> dimension=3;
>> svdd;
>> test;
10 /10 samples is included in the classification.
0 /10 samples is excluded in the classification.
>> dimension=4;
>> svdd;
>> test;
10 /10 samples is included in the classification.
0 /10 samples is excluded in the classification.
```

可见，这一组的测试点特性比较好，升维后准确度依旧可以达到 100%