Final project:

第一题

数据介绍

这个数据是关于一个葡萄牙银行向客户进行电话营销的记录。每一行都是关于一个具体客户的详细信息，包括年龄、婚姻状况等等，同时还包括与营销相关的一些数据，如持续时间，客户被营销次数，最后还有一些该营销进行时，葡萄牙整体的宏观经济数据。最后一列是表示的是该客户是否会在该银行注册一份定期存款，用二元变量y表示，1代表会0则代表不会。

1、 年龄

2、 职业

3、 婚姻状况

4、 教育程度

5、 是否有违约记录

6、 是有用房贷记录

7、 是否有个人贷款

8、 联络方式

9、 最近一次联系的月份

10、 最近一次练习的日期

11、 通话持续时间

12、 联系的次数

13、 离上一次电话营销的天数

14、 在此次电话营销前与该客户的联系次数

15、 上一次对该客户的营销结果

16、 就业情况（宏观指标）

17、 消费者价格指数（宏观指标）

18、 消费者信心指数（宏观指标）

19、 欧洲银行间欧元同业拆借利率（宏观指标）

20、 就业人数（宏观指标）

二、 任务介绍

搭建一个机器学习的预测模型，可以通过输入客户的各项信息，得到客户是否会注册定期存款的预测。

三、 任务提示：

（1）本次考核最低要求为学员能够搭建简单的logistic regression模型，当然也可以应用更复杂、预测效果更好的如神经网络等模型。

（2）可能会用到的python 包：pandas、numpy、sklearn、matplotlib、seaborn等。

（3）并不是所有自变量都与预测变量相关，应当根据商业和日常经验找到相关的自变量，舍弃不相关的自变量。

（4）对数据的预处理是必要的，比如对缺失值的处理等等。

第二题

In this problem we will apply discriminant analysis to recognize the digits in the MNIST data set (http://yann.lecun.com/exdb/mnist/). We will train our model using the training data sets ("train-images-idx3-ubyte.gz" and "train-labels-idx1-ubyte.gz") and test the performance using the test data set ("t10k-images-idx3-ubyte.gz" and "t10k-labels-idx1-ubyte.gz").

1. The images are 28 x 28 pixels in gray-scale. The categories are 0, 1, ... 9. We concatenate the image rows into a 28 x 28 vector and treat this as our feature, and assume the feature vectors in each category in the training data "train-images-idx3-ubyte.gz") have Gaussian distribution. Draw the mean and standard deivation of those features for the 10 categories as 28 x 28 images using the training images ("train-images-idx3-ubyte.gz"). There should be 2 images for each of the 10 digits, one for mean and one for standard deviation.

2. Classify the images in the testing data set ("t10k-images-idx3-ubyte.gz") using 0-1 loss function and Bayesian decision rule and report the performance. Why it doesn't perform as good as many other methods on LeCuns web page?

3. Write code to train a multi-class support vector classifier with dot-product kernel and 1-norm soft margin using the MNIST training data

set. Then reporting the performance using MNIST test data set. There is a hyper-parameter that sets the trade-off between the margin and the training error --- tune this hyper-parameter through cross-validation.

第三题

从数据收集、数据挖掘，乃至商业解释都需要发挥各位各自的长处。学习的目的是为了应用，现实往往更复杂，希望大家调动一切脑筋，解决他。
问题的简单描述：
中国社会的信任水平在不断的降低，根据社科院一份今年报告，中国社会人与人之间的信任度已经到了危机边缘。许多学者认为中国进入了幸运危机。是什么造成中国社会如此之低的信任度？这次的任务就要给出一个尽可能合理的解释。
提示：解释变量是不断降低的信任度，如何量化呢？第一种你可以去获得大量的调查问卷；第二种，通过大数据文本分析，比如微博平台代表的各类SN平台，对信任做与语言挖掘，信任一次是否越来越的多的与更多的负面词相练习，具体这个模型你们要自己去找或建立，那么建立好之后，我么可以得到一段时间内的信任的指数，这就是我们的应变量。
至于自变量，我可以提供给你们一些经典的假设，比如说社会贫富差距、比如说城市化进程、比如说交通便捷性等等，这些数据都比较容易获得，你们也可以根据自身的才智，去识别也有可能对信任产生其他影响的变量，都放到数据中来。
有了这些数据，你们可以调用各种算法，去实现对信任的解析。
这套题考察了至少四种能力：数据搜集、文本分析、商业逻辑、算法模型应用等。
希望大家勇敢尝试！！！