Machine Learning Reading Group **Spring 2020**

Chapter 1: Machine Learning Basis

Note: *LaTeX template courtesy of UC Berkeley EECS dept & CMU's convex optimization course taught by Ryan Tibshirani.*

Problem 1

Illustrate a classification problem that exists in the real world and answer the following questions.

1. What is the target variable of this problem? What are the potential features?
2. If we do not have these data now, where can we find them or how can we collectthem?
3. Does this problem solved comprehensively? How does people solve it usually?

Share your problem with your teammates.

Classify structured data with feature columns

使用Cleveland心脏病基金会提供的一个小型数据集。CSV文件每行描述一个患者， 每列描述一个属性。使用此信息来预测患者是否患有心脏病，在此数据集中这是一项二进制分类任务。

1. 此问题的target variable为判断患者是否患有心脏病。potential features有Age, Sex, Resting blood pressure, Serum cholestoral(胆固醇), fasting blood sugar(空腹血糖), ECG(心电图结果)等
2. Kaggle上提供有较多Dataset以供分析，或者使用API或爬虫等下载或收集
3. 没有。心脏疾病的确诊需要辅以超声心动，静脉造影，心脏导管等。体检数据仅能作为推测依据。当前分析一般基于医嘱或阈值(临界值)警示。

Problem 2

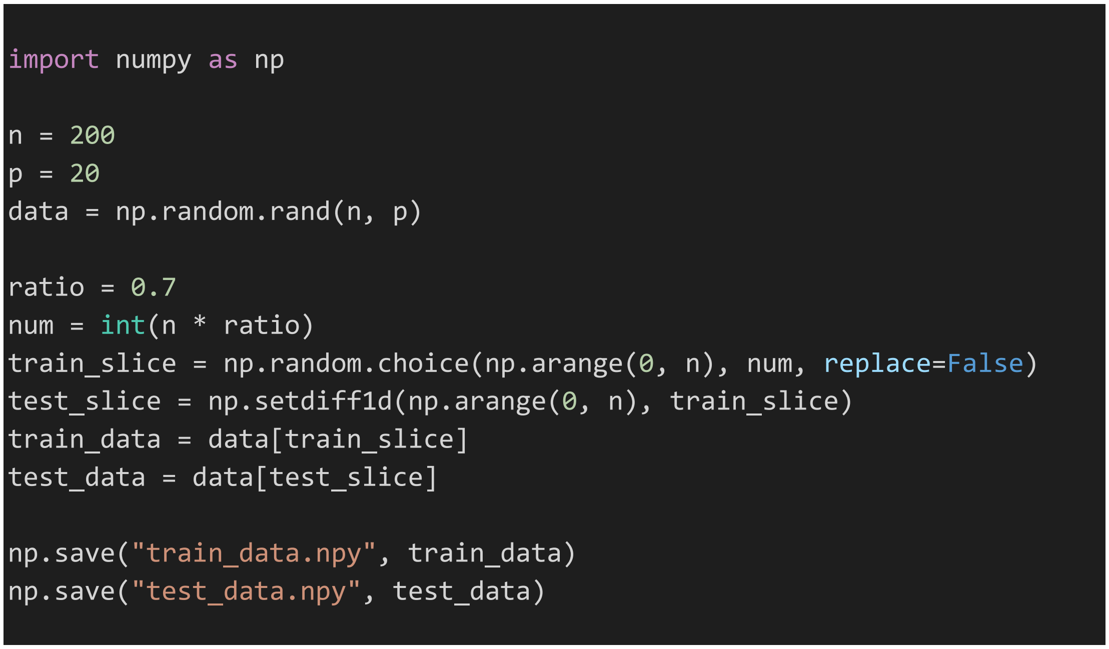
Use python to build a training and testing dataset following the steps below:

1. Initial the sample number $n = 200$ and the feature number $p = 20$.
2. Construct a $n * p$ random matrix and a length n vector as the raw feature matrix and the target variable vector respectively.

3. Randomly select $[n * 70\%]$ samples as the training dataset.

4. Use the remaining samples as the test dataset.

5. Write the code to save and load the training and test dataset.



```
import numpy as np

n = 200
p = 20
data = np.random.rand(n, p)

ratio = 0.7
num = int(n * ratio)
train_slice = np.random.choice(np.arange(0, n), num, replace=False)
test_slice = np.setdiff1d(np.arange(0, n), train_slice)
train_data = data[train_slice]
test_data = data[test_slice]

np.save("train_data.npy", train_data)
np.save("test_data.npy", test_data)
```