Machine Learning - Assignment 1

Xie Jiaxin, 2020111142

Exercise1

Make a CV using R Markdown that involves your education, interests, future plan and what you expect to learn from this class. Feel free to add more sections and use more R markdown skills even they are not introduced. English and 中文 are both acceptable, but make sure to only use one of them :)

教育背景

上海财经大学，统计与管理学院，数据科学与大数据技术

- 专业课程：计算机编程，数据结构，数据库，金融建模，机器学习，数据分析与可视化
- 所获奖项：“正大杯”市场调查与分析大赛全国三等奖，上海财经大学数学建模竞赛二等奖

项目经历

基于SQL Server和Visual Basic的菜谱推荐系统

上海财经大学《数据库》课程结课项目，组长

- 通过SQL Server创建菜谱数据库，利用VB搭建前端界面，采取ADO.NET数据库访问技术连接两端，采用基于DataSet的数据查询与更新，实现用户注册与登录、菜谱的查阅、上传、评分、评论等功能
- 基于系统功能与数据需求抽象出用户、菜谱、论坛、访问和评分实体集与联系，绘制E-R图，并实现从实体集与联系到关系模式的转换，经过规范化处理后的关系模式满足BC范式
- 使用SQL语句实现建表、查询、创建视图，并通过创建触发器实现数据库的动态完整性控制

“傥屋未定，何以安家？”——上海市租赁社区发展现状及大众租房偏好研究

第十二届”正大杯”全国大学生市场调查与分析大赛，组长，全国三等奖

- 对上海市租赁社区发展现状进行调查并研究该模式下的大众租房偏好，建立模型定量分析租赁社区客户特征、挖掘潜在客户、研究大众租房偏好与人口学变量间的关系，最终与组员共同完成50,000字报告
- 通过相关研究文献确定模型并基于模型构建指标体系，进而设计问卷并发放；结合数据分析结果与租赁市场相关政策解读，采用SWOT模型对行业进行前瞻性预测和分析未来发展方向
- 利用Python构建二元Logistic模型探究租赁社区客户特征；采用SPSS分类数据主成分分析进行潜在客户挖掘；通过Python实现K-Means聚类将大众偏好分类，再利用随机森林给出人口学变量在影响偏好上的重要程度排序，以此探索人口学变量与所属偏好类型间的关系

技能与特长

- Microsoft Office、C++、SPSS、SQL、VB、Tableau、Python
- CET4

未来规划

短期：

- 一学期内熟练掌握R和Python，完成出色的期末报告
- 专心备考GRE和TOFEL，争取年底出分
- 找到满意的寒假实习

长期：

- 申到dream school
- 找到dream work
- 实现WLB

课程展望

- 提高代码能力，享受debug的过程
- 拓宽视野，关注前沿领域
- 找到兴趣所在，找准努力方向

Exercise2

- create the vector `1,1,1,1,1,2,2,2,2,2` with only `rep()` and name it `x1`.

```
x1<-rep(1:2,each=5)
x1
```

```
##  [1] 1 1 1 1 1 2 2 2 2 2
```

- create the vector `1,2,1,2,1,2,1,2,1,2` with only `rep()` and name it `x2`.

```
x2<-rep(1:2,times=5)
x2
```

```
##  [1] 1 2 1 2 1 2 1 2 1 2
```

- combine `x1` and `x2` into a matrix `x.col` by columns, i.e., `x1` and `x2` are the two columns of `x`. Hint: use `cbind()`.

```
x.col<-cbind(x1,x2)
x.col
```

```
##       x1 x2
##  [1,]  1  1
##  [2,]  1  2
##  [3,]  1  1
##  [4,]  1  2
##  [5,]  1  1
##  [6,]  2  2
##  [7,]  2  1
##  [8,]  2  2
##  [9,]  2  1
## [10,]  2  2
```

- combine `x1` and `x2` into a matrix `x.row` by rows, i.e., `x1` and `x2` are the two rows of `x`. Hint: use `rbind()`.

```
x.row<-rbind(x1,x2)
x.row
```

```
##    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## x1    1    1    1    1    1    2    2    2    2     2
## x2    1    2    1    2    1    2    1    2    1     2
```

- find two ways to calculate the sum of each column of `x.row`. Hint: use `apply()`.

```
sum1<-colSums(x.row)
sum1
```

```
##  [1] 2 3 2 3 2 4 3 4 3 4
```

```
sum2<-apply(x.row,2,sum)
sum2
```

```
##  [1] 2 3 2 3 2 4 3 4 3 4
```

Exercise3

This exercise involves the Boston housing data set. To begin, load in the Boston data set. The Boston data set is part of the `ISLR2` library. Of course, you should install the `ISLR2` package before using it. After `library(ISLR2)`, the data set is contained in the object `Boston`. Read about the data set with `?Boston`.

```
library(ISLR2)
boston_df<-Boston
```

- How many rows are in this data set? How many columns? What do the rows and columns represent?

```
nrow(boston_df)
```

```
## [1] 506
```

```
ncol(boston_df)
```

```
## [1] 13
```

```
names(boston_df)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "lstat"   "medv"
```

- CRIM–城镇人均犯罪率——【城镇人均犯罪率】
- ZN - 占地面积超过25,000平方英尺的住宅用地比例——【住宅用地所占比例】
- INDUS - 每个城镇非零售业务的比例——【城镇中非商业用地占比例】
- CHAS - Charles River虚拟变量（如果是河道，则为1;否则为0——【查尔斯河虚拟变量，用于回归分析】
- NOX - 一氧化氮浓度（每千万份）——【环保指标】
- RM - 每间住宅的平均房间数——【每栋住宅房间数】
- AGE - 1940年以前建造的自住单位比例——【1940年以前建造的自住单位比例 】
- DIS -波士顿的五个就业中心加权距离——【与波士顿的五个就业中心加权距离】
- RAD - 径向高速公路的可达性指数——【距离高速公路的便利指数】
- TAX - 每10,000美元的全额物业税率——【每一万美元的不动产税率】
- PTRATIO - 城镇的学生与教师比例——【城镇中教师学生比例】
- LSTAT - 人口状况下降%——【房东属于低等收入阶层比例】
- MEDV - 自有住房的中位数报价, 单位1000美元——【自住房屋房价中位数】

- Which of the predictors are quantitative, and which are qualitative?

- qualitative : chas
- quantitative : others

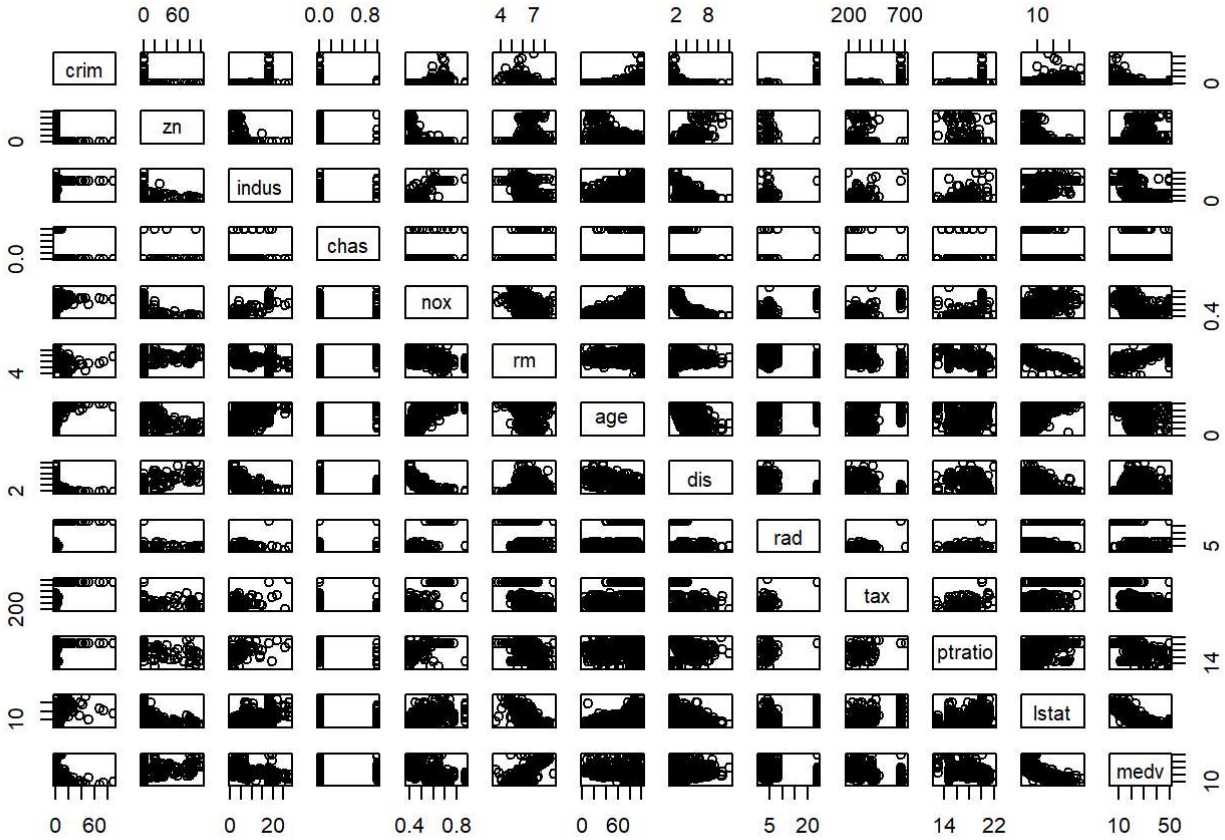
- What is the range of each quantitative predictor? You can answer this using the `range()` function.

```
range_mat<-matrix(0,13,2)

for (i in 1:13) {
  range_mat[i,]<-range(boston_df[i])
}

rownames(range_mat)<-names(boston_df)
colnames(range_mat)<-c('min','max')
range_mat
```

```
##                min       max
## crim       0.00632   88.9762
## zn         0.00000  100.0000
## indus      0.46000   27.7400
## chas       0.00000    1.0000
## nox        0.38500    0.8710
## rm         3.56100    8.7800
## age        2.90000  100.0000
## dis        1.12960   12.1265
## rad        1.00000   24.0000
## tax      187.00000  711.0000
## ptratio   12.60000   22.0000
## lstat      1.73000   37.9700
## medv       5.00000   50.0000
```

- Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

```
pairs(boston_df)
```



房价与低等收入阶层比例、房间数、环保指标、非商业用地比例、住宅用地所占比例等具有一定的线性关系；环保指标与五个就业中心加权距离、房屋年代有一定相关性。

- Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

```
cor(boston_df)
```

```
##                crim          zn       indus         chas         nox
## crim     1.00000000 -0.20046922  0.40658341 -0.055891582  0.42097171
## zn      -0.20046922  1.00000000 -0.53382819 -0.042696719 -0.51660371
## indus    0.40658341 -0.53382819  1.00000000  0.062938027  0.76365145
## chas    -0.05589158 -0.04269672  0.06293803  1.000000000  0.09120281
## nox      0.42097171 -0.51660371  0.76365145  0.091202807  1.00000000
## rm      -0.21924670  0.31199059 -0.39167585  0.091251225 -0.30218819
## age      0.35273425 -0.56953734  0.64477851  0.086517774  0.73147010
## dis     -0.37967009  0.66440822 -0.70802699 -0.099175780 -0.76923011
## rad      0.62550515 -0.31194783  0.59512927 -0.007368241  0.61144056
## tax      0.58276431 -0.31456332  0.72076018 -0.035586518  0.66802320
## ptratio  0.28994558 -0.39167855  0.38324756 -0.121515174  0.18893268
## lstat    0.45562148 -0.41299457  0.60379972 -0.053929298  0.59087892
## medv    -0.38830461  0.36044534 -0.48372516  0.175260177 -0.42732077
##                  rm         age         dis          rad         tax    ptratio
## crim    -0.21924670  0.35273425 -0.37967009  0.625505145  0.58276431  0.2899456
## zn       0.31199059 -0.56953734  0.66440822 -0.311947826 -0.31456332 -0.3916785
## indus   -0.39167585  0.64477851 -0.70802699  0.595129275  0.72076018  0.3832476
## chas     0.09125123  0.08651777 -0.09917578 -0.007368241 -0.03558652 -0.1215152
## nox     -0.30218819  0.73147010 -0.76923011  0.611440563  0.66802320  0.1889327
## rm       1.00000000 -0.24026493  0.20524621 -0.209846668 -0.29204783 -0.3555015
## age     -0.24026493  1.00000000 -0.74788054  0.456022452  0.50645559  0.2615150
## dis      0.20524621 -0.74788054  1.00000000 -0.494587930 -0.53443158 -0.2324705
## rad     -0.20984667  0.45602245 -0.49458793  1.000000000  0.91022819  0.4647412
## tax     -0.29204783  0.50645559 -0.53443158  0.910228189  1.00000000  0.4608530
## ptratio -0.35550149  0.26151501 -0.23247054  0.464741179  0.46085304  1.0000000
## lstat   -0.61380827  0.60233853 -0.49699583  0.488676335  0.54399341  0.3740443
## medv     0.69535995 -0.37695457  0.24992873 -0.381626231 -0.46853593 -0.5077867
##              lstat       medv
## crim     0.4556215 -0.3883046
## zn      -0.4129946  0.3604453
## indus    0.6037997 -0.4837252
## chas    -0.0539293  0.1752602
## nox      0.5908789 -0.4273208
## rm      -0.6138083  0.6953599
## age      0.6023385 -0.3769546
## dis     -0.4969958  0.2499287
## rad      0.4886763 -0.3816262
## tax      0.5439934 -0.4685359
## ptratio  0.3740443 -0.5077867
## lstat    1.0000000 -0.7376627
## medv    -0.7376627  1.0000000
```

犯罪率与径向高速公路的可达性、物业税率有一定正相关关系，高速公路可达性强意味着交通便利，导致犯罪分子便于犯罪和逃离；物业税率可以反映该地的财务水平，税率高意味着以高档住宅为主，因此容易被犯罪分子觊觎。

- What is the mean and standard deviation of each quantitative predictor?

```
mean_sd_mat<-matrix(0,13,2)

for (i in 1:13) {
  mean_sd_mat[i,1]<-mean(boston_df[,i])
  mean_sd_mat[i,2]<-sd(boston_df[,i])
}

rownames(mean_sd_mat)<-names(boston_df)
colnames(mean_sd_mat)<-c('mean','sd')
mean_sd_mat
```

```
##                mean          sd
## crim       3.61352356   8.6015451
## zn        11.36363636  23.3224530
## indus     11.13677866   6.8603529
## chas       0.06916996   0.2539940
## nox        0.55469506   0.1158777
## rm         6.28463439   0.7026171
## age       68.57490119  28.1488614
## dis        3.79504269   2.1057101
## rad        9.54940711   8.7072594
## tax      408.23715415 168.5371161
## ptratio   18.45553360   2.1649455
## lstat     12.65306324   7.1410615
## medv      22.53280632   9.1971041
```

- How many of the census tracts in this data set bound the Charles river?

```
nrow(boston_df[boston_df$chas==1,])
```

```
## [1] 35
```

- What is the median pupil-teacher ratio among the towns in this data set?

```
median(boston_df$ptratio)
```

```
## [1] 19.05
```

- Which census tract of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

```
boston_df[boston_df$medv==min(boston_df$medv),]
```

	crim <dbl>	zn <dbl>	indus <dbl>	chas <int>	nox <dbl>	rm <dbl>	age <dbl>	dis <dbl>	rad <int>
399	38.3518	0	18.1	0	0.693	5.453	100	1.4896	24
406	67.9208	0	18.1	0	0.693	5.683	100	1.4254	24

2 rows | 1-10 of 14 columns

房价最低的地方有几个共同点:

- 不是大型住宅区，'zn'为最小值
- 不靠近河流，一氧化碳浓度较高，环境一般
- 都是老住宅，'age'为最大值
- 高速公路可达性指数高，'rad'位最大值
- 物业税率较高，'tax'高于上四分位
- 学术与教师比例高，'ptratio'高于上四分位
- 房东属于低等收入阶层比例较高

- In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.

```
boston_df[boston_df$rm>=7,];
```

	crim <dbl>	zn <dbl>	indus <dbl>	chas <int>	nox <dbl>	rm <dbl>	age <dbl>	dis <dbl>	rad <int>
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3
41	0.03359	75.0	2.95	0	0.4280	7.024	15.8	5.4011	3
56	0.01311	90.0	1.22	0	0.4030	7.249	21.9	8.6966	5
65	0.01951	17.5	1.38	0	0.4161	7.104	59.5	9.2229	3
89	0.05660	0.0	3.41	0	0.4890	7.007	86.3	3.4217	2
90	0.05302	0.0	3.41	0	0.4890	7.079	63.1	3.4145	2
98	0.12083	0.0	2.89	0	0.4450	8.069	76.0	3.4952	2
99	0.08187	0.0	2.89	0	0.4450	7.820	36.9	3.4952	2
100	0.06860	0.0	2.89	0	0.4450	7.416	62.5	3.4952	2

1-10 of 64 rows | 1-10 of 14 columns Previous **1** 2 3 4 5 6 7 Next

```
boston_df[boston_df$rm>=8,]
```

	crim <dbl>	zn <dbl>	indus <dbl>	chas <int>	nox <dbl>	rm <dbl>	age <dbl>	dis <dbl>	rad <int>
98	0.12083	0	2.89	0	0.4450	8.069	76.0	3.4952	2
164	1.51902	0	19.58	1	0.6050	8.375	93.9	2.1620	5
205	0.02009	95	2.68	0	0.4161	8.034	31.9	5.1180	4
225	0.31533	0	6.20	0	0.5040	8.266	78.3	2.8944	8
226	0.52693	0	6.20	0	0.5040	8.725	83.0	2.8944	8

	crim <dbl>	zn <dbl>	indus <dbl>	chas <int>	nox <dbl>	rm <dbl>	age <dbl>	dis <dbl>	rad <int>
227	0.38214	0	6.20	0	0.5040	8.040	86.5	3.2157	8
233	0.57529	0	6.20	0	0.5070	8.337	73.3	3.8384	8
234	0.33147	0	6.20	0	0.5070	8.247	70.4	3.6519	8
254	0.36894	22	5.86	0	0.4310	8.259	8.4	8.9067	7
258	0.61154	20	3.97	0	0.6470	8.704	86.9	1.8010	5

1-10 of 13 rows | 1-10 of 14 columns Previous **1** 2 Next

多数住宅的房间数处于7-8之间。 平均房间数超过8的住宅有以下特点：

- 犯罪率低
- 环境质量好
- 房龄多数较大
- 房东多不属于低收入阶层
- 房价较高