Group Exercise, Class 4: Preparing for Your Analysis

Part 1: Analytic Pipelines

In Assignment 4, you are asked to create, implement and compare two prediction pipelines using linear regression in R. Specifically, the instructions state:

- *Perform basic data cleaning. Note which features are continuous, which are categorical and ensure they are being stored that way in your R dataset (That is, if categorical variables have been read-in as continuous variables, convert them to factors)*
- *Partition data into training and testing (use a 70/30 split)*

1. *Fit two prediction models using different subsets of the features in the training data. Features can overlap in the two models, but the feature sets should not be exactly the same across models.*
2. *Apply both models within the test data and determine which model is the preferred prediction model using the appropriate evaluation metric(s).*

As a group, answer the following questions to prepare for your analysis:

1. List three things you may want to check as part of basic data cleaning.

2. When partitioning data into training and testing, name one thing you want to be sure is consistent across the two partitions to support a successful analysis.

3. What is an appropriate evaluation metric to use in this analysis?

Part 2: Dimension Reduction

Researchers are interested in understanding how information gained during routine clinical exams can reflect overall wellness in women. They compiled 9 measurements on 116 patients. All measurements are quantitative, continuous variables. They used PCA to describe the presence of underlying constructs that explain the shared variance of the clinical measurements. All features within their dataset are quantitative, continuous variables.

As a group, answer the following questions:

1. When preparing the data for analysis, what did the investigators likely do to the clinical measurements, after cleaning the data, to ensure appropriate results from their PCA analysis?

2. The following output was produced by the PCA analysis. Based on this output, how many components should the investigators retain in their analysis?

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     1.7489 1.2338 1.0805 1.0515 0.85002 0.81073 0.66449 0.54095 0.17894

Proportion of Variance 0.3398 0.1691 0.1297 0.1229 0.08028 0.07303 0.04906 0.03251 0.00356
Cumulative Proportion  0.3398 0.5090 0.6387 0.7615 0.84184 0.91487 0.96393 0.99644 1.00000
```

3. Using the following output, which component does glucose contribute to most? What about leptin?

Table: Factor Loadings on each principal component resulting from the PCA

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Age	0.1245739	0.06626166	-0.20670086	0.821387749	0.2529717	-0.29483811	0.30785182	-0.12712772	-0.028796676
BMI	0.2604297	0.49933879	0.42573480	-0.070921589	-0.2324569	-0.27652419	-0.05332906	-0.59945409	0.062178063
Glucose	0.4390227	-0.18594179	-0.13088534	0.125615405	0.1995197	-0.03011844	-0.80716167	-0.08449350	-0.201022847
Insulin	0.4439787	-0.38631334	0.09371342	-0.059771821	-0.2976113	0.12162117	0.39056498	-0.09436728	-0.613579380
HOMA	0.4928517	-0.37472788	-0.01219108	-0.005644021	-0.1391521	0.06763632	0.13142043	-0.01629459	0.758301332
Leptin	0.3314935	0.23364275	0.58320150	0.058353947	0.2878059	-0.02159401	0.05409473	0.63623826	-0.031145986
Adiponectin	-0.1726096	-0.48059810	0.28212343	-0.276866882	0.5292842	-0.48846279	0.10252955	-0.23175112	-0.004889575
Resistin	0.2817413	0.30361887	-0.28892346	-0.302703643	0.5975961	0.42081192	0.24436211	-0.24326258	-0.015424508
Triglyceride	0.2546307	0.21044940	-0.49675794	-0.359469872	-0.1188700	-0.63260295	0.08893578	0.30083068	-0.046497241

Part 3: Clustering

The built-in R dataset USArrests includes the crime statistics for each of the 50 US states in 1973. Incidence of arrest, per 100,000 residents for assault, murder and rape are included along with the proportion of the population that lives in urban communities. Using this dataset, investigators aimed to identify clusters of states based on their crime stats using k-means cluster analysis.

Based on subject-matter knowledge, the investigators hypothesized that there should be 5 clusters of states. They set k=5 and performed a k-means cluster analysis. The output describing the mean values of the input features by cluster are below.

Cluster	Murder	Assault	UrbanPop	Rape
1	-0.1642225	-0.3658283	-0.2822467	-0.11697538
2	0.7298036	1.1188219	0.7571799	1.32135653
3	1.5803956	0.9662584	-0.7775109	0.04844071
4	-1.1727674	-1.2078573	-1.0045069	-1.10202608
5	-0.6286291	-0.4086988	0.9506200	-0.38883734

Answer the following questions:

1. Describe cluster 2 and cluster 4 in words.

2. Describe a data-driven approach to determine the optimal number of k (i.e. number of clusters).