DataSci 347: Introduction to Machine Learning
Homework 2

Group Member 1
Group Member 2
Group Member 3

Due: 11:59 PM, Sunday, October 5th, 2024

1 Overview

Principal Component Analysis is widely used in data exploration, dimension reduction, and data visualization. The aim is to transform original data into uncorrelated linear combinations of the original data while keeping the information contained in the data. High dimensional data tends to show clusters in lower dimensional view.

Clustering Analysis is another form of EDA. Here we are hoping to group data points which are close to each other within the groups and far away between different groups. Clustering using PC's can be effective. Clustering analysis can be very subjective in the way we need to summarize the properties within each group.

Both PCA and Clustering Analysis are so called unsupervised learning. There is no response variables involved in the process.

For supervised learning, we try to find out how does a set of predictors relate to some response variable of the interest. Multiple regression is still by far, one of the most popular methods. We use linear models as a working model for its simplicity and interpretability. It is important that we use domain knowledge as much as we can to determine the form of the response as well as the function format of the factors.

1.1 Objectives

- PCA
- SVD
- Clustering Analysis
- Linear Regression

1.2 Review Materials

- Study Module 2: PCA
- Study Module 3: Clustering Analysis
- Study Lecture 4: Multiple regression

2 Case Study 1: Self-Esteem

Self-esteem generally describes a person's overall sense of self-worthiness and personal value. It can play significant role in one's motivation and success throughout the life. Factors that influence self-esteem can be inner thinking, health condition, age, life experiences etc. We will try to identify possible factors in our data that are related to the level of self-esteem.

In the well-cited National Longitudinal Study of Youth (NLSY79), it follows about 13,000 individuals and numerous individual-year information has been gathered through surveys. The survey data is open to public here. Among many variables we assembled a subset of variables including personal demographic variables in different years, household environment in 79, ASVAB test Scores in 81 and Self-Esteem scores in 81 and 87 respectively.

The data is stored in `NLSY79.csv`.

Here are the description of variables:

Personal Demographic Variables

- `Gender`: a factor with levels "female" and "male"
- `Education05`: years of education completed by 2005
- `HeightFeet05, HeightInch05`: height measurement. For example, a person of 5'10 will be recorded as HeightFeet05=5, HeightInch05=10.
- `Weight05`: weight in lbs.
- `Income87, Income05`: total annual income from wages and salary in 2005.
- `Job87, Job05`: job type in 1987 and 2005, including Protective Service Occupations, Food Preparation and Serving Related Occupations, Cleaning and Building Service Occupations, Entertainment Attendants and Related Workers, Funeral Related Occupations, Personal Care and Service Workers, Sales and Related Workers, Office and Administrative Support Workers, Farming, Fishing and Forestry Occupations, Construction Trade and Extraction Workers, Installation, Maintenance and Repairs Workers, Production and Operating Workers, Food Preparation Occupations, Setters, Operators and Tenders, Transportation and Material Moving Workers

Household Environment

- `Imagazine`: a variable taking on the value 1 if anyone in the respondent's household regularly read magazines in 1979, otherwise 0
- `Inewspaper`: a variable taking on the value 1 if anyone in the respondent's household regularly read newspapers in 1979, otherwise 0
- `Ilibrary`: a variable taking on the value 1 if anyone in the respondent's household had a library card in 1979, otherwise 0
- `MotherEd`: mother's years of education
- `FatherEd`: father's years of education
- `FamilyIncome78`

Variables Related to ASVAB test Scores in 1981

Test	Description
AFQT	percentile score on the AFQT intelligence test in 1981
Coding	score on the Coding Speed test in 1981
Auto	score on the Automotive and Shop test in 1981
Mechanic	score on the Mechanic test in 1981
Elec	score on the Electronics Information test in 1981
Science	score on the General Science test in 1981
Math	score on the Math test in 1981
Arith	score on the Arithmetic Reasoning test in 1981
Word	score on the Word Knowledge Test in 1981
Parag	score on the Paragraph Comprehension test in 1981
Numer	score on the Numerical Operations test in 1981

Self-Esteem test 81 and 87

We have two sets of self-esteem test, one in 1981 and the other in 1987. Each set has same 10 questions. They are labeled as Esteem81 and Esteem87 respectively followed by the question number. For example, Esteem81_1 is Esteem question 1 in 81.

The following 10 questions are answered as 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree

1. Esteem 1: ”I am a person of worth”
2. Esteem 2: ”I have a number of good qualities”
3. Esteem 3: ”I am inclined to feel like a failure”
4. Esteem 4: ”I do things as well as others”
5. Esteem 5: ”I do not have much to be proud of”
6. Esteem 6: ”I take a positive attitude towards myself and others”
7. Esteem 7: ”I am satisfied with myself”
8. Esteem 8: ”I wish I could have more respect for myself”
9. Esteem 9: ”I feel useless at times”
10. Esteem 10: ”I think I am no good at all”

2.1 Data Preparation

Load the data using `pandas.read_csv()`. Do a quick EDA to get familiar with the data set using `df.info()`, `df.describe()`, and `df.head()`. Pay attention to the unit of each variable. Are there any missing values? Use `df.isnull().sum()` to check.

2.2 Self Esteem Evaluation

Let concentrate on Esteem scores evaluated in 87.

a) Reverse Esteem 1, 2, 4, 6, and 7 so that a higher score corresponds to higher self-esteem. (Hint: if we store the esteem data in `esteem_data`, then `esteem_data.iloc[:, [0, 1, 3, 5, 6]] = 5 - esteem_data.iloc[:, [0, 1, 3, 5, 6]]` to reverse the score.)

b) Write a brief summary with necessary plots about the 10 esteem measurements using `matplotlib` and `seaborn`.

c) Do esteem scores all positively correlated? Report the pairwise correlation table using `df.corr()` and write a brief summary.

d) PCA on 10 esteem measurements using `sklearn.decomposition.PCA` (centered but no scaling)

   i) Report the PC1 and PC2 loadings using `pca.components_`. Are they unit vectors? Are they uncorrelated? Use `np.linalg.norm()` to check unit vectors.

   ii) Are there good interpretations for PC1 and PC2? (If loadings are all negative, take the positive loadings for the ease of interpretation)

   iii) How is the PC1 score obtained for each subject? Write down the formula.

   iv) Are PC1 scores and PC2 scores in the data uncorrelated? Use `np.corrcoef()` to verify.

   v) Plot PVE (Proportion of Variance Explained) using `pca.explained_variance_ratio_` and summarize the plot.

   vi) Also plot CPVE (Cumulative Proportion of Variance Explained) using `np.cumsum()`. What proportion of the variance in the data is explained by the first two principal components?

   vii) PC's provide us with a low dimensional view of the self-esteem scores. Use a biplot with the first two PC's to display the data using `matplotlib`. Give an interpretation of PC1 and PC2 from the plot.

e) Apply k-means to cluster subjects on the original esteem scores using `sklearn.cluster.KMeans`

   i) Find a reasonable number of clusters using within sum of squared with elbow rules. Plot the elbow curve using `matplotlib`.

   ii) Can you summarize common features within each cluster using `df.groupby().mean()`?

   iii) Can you visualize the clusters with somewhat clear boundaries? You may try different pairs of variables and different PC pairs of the esteem scores using `plt.scatter()`.

f) We now try to find out what factors are related to self-esteem? PC1 of all the Esteem scores is a good variable to summarize one's esteem scores. We take PC1 as our response variable.

   i) Prepare possible factors:

      - Personal information: gender, education, log(income), job type, Body mass index as a measure of health (The BMI is defined as the body mass divided by the square of the body height, and is universally expressed in units of kg/m^2)
      - Household environment: Imagazine, Inewspaper, Ilibrary, MotherEd, FatherEd, FamilyIncome78. Do set indicators Imagazine and Ilibrary as factors using `pd.Categorical()`
      - Use PC1 of ASVAB as level of intelligence

   ii) Run a few regression models between PC1 of all the esteem scores and factors listed in a) using `sklearn.linear_model.LinearRegression` or `statsmodels.api`. Find a final best model with your own criterion.

   iii) How did you land this model? Run a model diagnosis to see if the linear model assumptions are reasonably met using residual plots.

   iv) Write a summary of your findings. In particular, explain what and how the variables in the model affect one's self-esteem.

3 Case Study 2: Breast Cancer Sub-type

The Cancer Genome Atlas (TCGA), a landmark cancer genomics program by National Cancer Institute (NCI), molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. The genome data is open to public from the Genomic Data Commons Data Portal (GDC).

In this study, we focus on 4 sub-types of breast cancer (BRCA): basal-like (basal), Luminal A-like (lumA), Luminal B-like (lumB), HER2-enriched. The sub-type is based on PAM50, a clinical-grade luminal-basal classifier.

- Luminal A cancers are low-grade, tend to grow slowly and have the best prognosis.
- Luminal B cancers generally grow slightly faster than luminal A cancers and their prognosis is slightly worse.
- HER2-enriched cancers tend to grow faster than luminal cancers and can have a worse prognosis, but they are often successfully treated with targeted therapies aimed at the HER2 protein.
- Basal-like breast cancers or triple negative breast cancers do not have the three receptors that the other sub-types have so have fewer treatment options.

We will try to use mRNA expression data alone without the labels to classify 4 sub-types. Classification without labels or prediction without outcomes is called unsupervised learning. We will use K-means and spectral clustering to cluster the mRNA data and see whether the sub-type can be separated through mRNA data.

We first read the data using `pd.read_csv()` which is efficient for reading large datasets.

3.1 Summary and Transformation

a) How many patients we have in each sub-type? Use `df['brca_subtype'].value_counts()`.

b) Randomly pick 5 genes and plot the histogram by each sub-type using `seaborn.histplot()`.

c) Remove genes with zero count and no variability using boolean indexing. Then apply logarithmic transform using `np.log1p()`.

d) Apply kmeans on the transformed dataset with 4 centers using `sklearn.cluster.KMeans` and output the discrepancy table between the real sub-type `brca_subtype` and the cluster labels using `pd.crosstab()`.

3.2 Spectral Clustering: To Scale or Not to Scale?

a) Apply PCA on the centered and scaled dataset using `sklearn.decomposition.PCA` and `sklearn.preprocessing.StandardScaler`. How many PCs should we use and why? You are encouraged to use `sklearn.decomposition.TruncatedSVD` for large datasets.

b) Plot PC1 vs PC2 of the centered and scaled data and PC1 vs PC2 of the centered but unscaled data side by side using `matplotlib.pyplot.subplots()`. Should we scale or not scale for clustering purposes? Why?

3.3 Spectral Clustering: Center but Do Not Scale the Data

a) Use the first 4 PCs of the centered and unscaled data and apply kmeans. Find a reasonable number of clusters using within sum of squared with the elbow rule.

b) Choose an optimal cluster number and apply kmeans. Compare the real sub-type and the clustering label as follows: Plot scatter plot of PC1 vs PC2 using `plt.scatter()`. Use point color to indicate the true cancer type and point shape to indicate the clustering label. Plot the kmeans centroids with black dots. Summarize how good is clustering results compared to the real sub-type.

c) Compare the clustering result from applying kmeans to the original data and the clustering result from applying kmeans to 4 PCs. Does PCA help in kmeans clustering? What might be the reasons if PCA helps?

d) Now we have a patient with breast cancer but with unknown sub-type. We have this patient's mRNA sequencing data. Project this patient to the space of PC1 and PC2. (Hint: remember we remove some genes with no counts or no variability, take log and center) Plot this patient in the plot in iv) with a black dot. Calculate the Euclidean distance between this patient and each centroid of the cluster using `scipy.spatial.distance.euclidean`. Can you tell which sub-type this patient might have?

4 Case Study 3: Auto Dataset

This question utilizes the Auto dataset. The original dataset contains 408 observations about cars. We'll use this dataset to practice the methods learned so far. You can download the dataset from: `https://archive.ics.uci.edu/ml/datasets/auto+mpg`

Get familiar with this dataset first using `df.info()` and `df.describe()`.

4.1 EDA

Explore the data, with particular focus on pairwise plots using `seaborn.pairplot()` and summary statistics. Briefly summarize your findings and any peculiarities in the data.

4.2 What Effect Does Time Have on MPG?

a) Start with a simple regression of mpg vs. year using `sklearn.linear_model.LinearRegression` or `statsmodels.formula.api.ols` and report the summary output. Is year a significant variable at the .05 level? State what effect year has on mpg, if any, according to this model.

b) Add horsepower on top of the variable year to your linear model. Is year still a significant variable at the .05 level? Give a precise interpretation of the year's effect found here.

c) The two 95% CI's for the coefficient of year differ among (i) and (ii). How would you explain the difference to a non-statistician?

d) Create a model with interaction by fitting `mpg ~ year * horsepower`. Is the interaction effect significant at .05 level? Explain the year effect (if any).

4.3 Categorical Predictors

Remember that the same variable can play different roles! Take a quick look at the variable cylinders, and try to use this variable in the following analyses wisely. We all agree that a larger number of cylinders will lower mpg. However, we can interpret cylinders as either a continuous (numeric) variable or a categorical variable.

a) Fit a model that treats cylinders as a continuous/numeric variable. Is cylinders significant at the 0.01 level? What effect does cylinders play in this model?

b) Fit a model that treats cylinders as a categorical/factor using `pd.get_dummies()`. Is cylinders significant at the .01 level? What is the effect of cylinders in this model? Describe the cylinders effect over mpg.

c) What are the fundamental differences between treating cylinders as a continuous and categorical variable in your models?

d) Can you test the null hypothesis: fit0: mpg is linear in cylinders vs. fit1: mpg relates to cylinders as a categorical variable at .01 level? Use F-test or likelihood ratio test.

4.4 Results

a) Final modeling question: we want to explore the effects of each feature as best as possible. You may explore interactions, feature transformations, higher order terms, or other strategies within reason. The model(s) should be as parsimonious (simple) as possible unless the gain in accuracy is significant from your point of view.

b) Describe the final model. Include diagnostic plots with particular focus on the model residuals and diagnoses using `matplotlib` and `scipy.stats`.

c) Summarize the effects found.

d) Predict the mpg of the following car: A red car built in the US in 1983 that is 180 inches long, has eight cylinders, displaces 350 cu. inches, weighs 4000 pounds, and has a horsepower of 260. Also give a 95% CI for your prediction using bootstrap or analytical methods.