Homework 3

Due Date: 2025-10-29 at 3 PM ET

Theoretical Component

Complete the following problems and email your solutions to Carter Price (price@rand.org) and Gabriel Hassler (ghassler@rand.org) by the due date. The subject line should read: "Intro to ML HW 03". Complete the problems manually (i.e., do not use Python code).

1. Do the first three iterations of gradient descent $x_{k+1} = x_k - hf^1\left(x_k\right)$, where $f^1\left(x_k\right)$ is the first derivative of $f\left(x\right)$ for the function $f(x) = \ x^2$ with step size h=0.5 and starting point $x_0 = 0.5$.

2. The first two eigenvectors from a Principal Component Analysis are $x_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ and $x_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$.

    a. Show that x_1 and x_2 are orthogonal (that is to say, show that $x_1^T x_2 = 0$).
    b. If the variables for this are blood pressure, body mass index, and cholesterol level, respectively, describe in words how to practically interpret x_1 and x_2. In words, how would you describe the meaning of those eigenvectors, so "a big value for the first principal component means that…".

3. Why might dimension reduction be useful for visualization? What is a concern about dimension reduction for visualization?

Programming Component

Submit this component of the homework via GitHub by the due date.

Preparation

1. Add a 'processed' directory to the 'data' folder in your repository.
2. Add a line at the end of the HW_02.ipynb file that saves the merged and cleaned dataset to 'data/processed' directory.
3. Run the HW_02.ipynb file to create the processed data file.
4. Commit and push your changes to GitHub.

NOTE: Do not include the data files in your GitHub repository. Make sure your .gitignore file is set up to ignore data files.

Homework - Principal Component Analysis

The CDC Social Vulnerability Index (SVI) takes multiple differen population-level inputs (e.g., % of the population living in poverty, % of the population without health insurance) to identify particularly vulnerable counties. While the CDC SVI scores rely on adding up the percentiles of various characteristics, there are alternative indexes (e.g., University of South Carolina SoVI index) that use methods like PCA. Here, we are going to use the CDC SVI data to create an alternative index based on PCA.

1. The following variables are used in the SVI: `EP_POV150, EP_UNEMP, EP_HBURD, EP_NOHSDP, EP_UNINSUR, EP_AGE65, EP_AGE17, EP_DISABL, EP_SNGPNT, EP_LIMENG, EP_MINRTY, EP_MUNIT, EP_MOBILE, EP_CROWD, EP_NOVEH, EP_GROUPQ, EP_NOINT`
    a. Subset the merged dataset to only include the variables above and look at the pattern of missing data. Are missing observations scattered throughout the data or are entire rows or columns missing?
    b. PCA cannot handle missing values by default. There are several options for handling missing data generally, including imputation, removing rows with missing data, or removing columns with missing data. Deal with the missing data in a way that makes sense for the pattern of missing data and the goals of the analysis. Explain why you made this decision. *Note: How you handle this is specific to the missing data pattern and the goals of the analysis. For example, when entire rows or columns are missing, imputation may not be appropriate and dropping those rows or columns is usually the best option. Conversely, if you have a general missingness pattern where missing observations are scattered throughout the data, imputation is likely the best option.*
    c. After dealing with the missing data, perform PCA on the SVI variables.
2. Plot the eigenvectors or loadings associated of the first three principal components. Make sure that the axis labels correspond to the variable names and not the indices of the variables. How would you interpret the first three prinicpal components? *Note: you can find the documentation for the SVI variables here.*
3. People often use PCA in downstream analyses (e.g., regression). When doing this, they need to choose the number of principal components to use in the analysis. There are several different ways to determine the number of principal components to retain. One common method is to retain principal components that explain a certain percentage of the variance in the data.
    a. How many principal components are needed to explain 80% of the variance in the data?
    b. How many principal components are needed to explain 90% of the variance in the data?
4. An alternative approach is to plot the eigenvalues of the principal components and retain the components that are above the “elbow” in the plot. In other words the eigenvalues that are substantially larger than the rest.
    a. Create a scree plot of the eigenvalues of the principal components.
    b. How many principal components should be retained based on the scree plot? This video may help: PCA Scree Plot
5. Plot the first principal component score on a map of the US counties. Briefly describe any spatial patterns you see.