CS M148 Problem Set 4

Due Date: December 3, 2024 at 11:59 P.M.

For these questions, please show and explain your work and be able to do these calculations by hand. You can use calculators, Excel, or Python for intermediate calculations (just do not use sklearn on numpy to do all the calculations for you).

1. PCA [Adapted from Yu and Barter 2024, Ch. 6 Exercise 19]

   Suppose that after conducting principal component analysis, you obtained a singular value matrix, D, which is given by

   $$\begin{bmatrix} 30.6 & 0 & 0 & 0 \\ 0 & 16.2 & 0 & 0 \\ 0 & 0 & 11.2 & 0 \\ 0 & 0 & 0 & 6.08 \end{bmatrix}$$

   and a right-singular vector matrix, V, which is given by

   $$\begin{bmatrix} .45 & -.6 & -.64 & .15 \\ -.4 & -.8 & .42 & -.17 \\ .57 & -.01 & .23 & -.78 \\ .54 & -.08 & .59 & .58 \end{bmatrix}$$

   Based on these matrices, answer the following questions:

   (a) Is it possible to determine how many variables (columns) the original data contained? If so, how and how many variables are there?

   (b) Is it possible to determine how many observations (rows) the original data contained? If so, how and how many observations are there?

   (c) Is it possible to determine how many principal components (PC) were computed? If so, how and how many PC ?

   (d) Compute the proportion of variability explained for each principal component.

   (e) Write a formula (linear combination) for computing each of the first two principal components (using $x_1, x_2, x_3, \ldots$ etc... to represent the original variables).

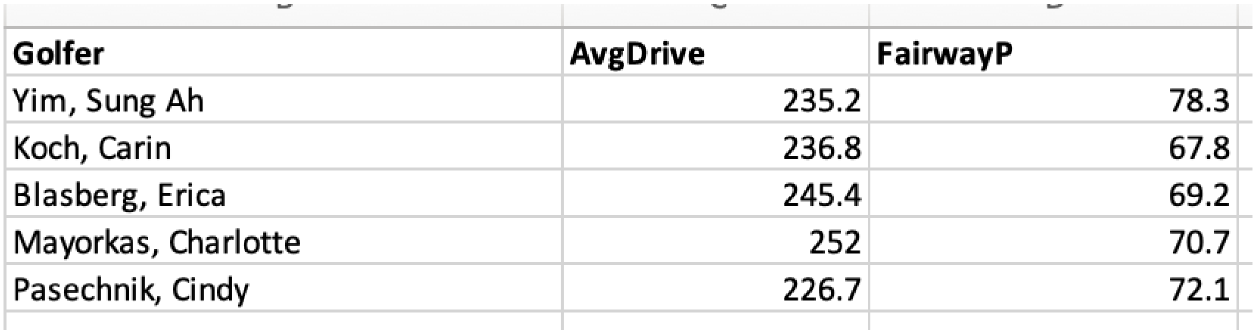
   (f) Compute the first two principal component values for an observation whose measurements are $(.65, -1.09, 1.21, .93)$

2. Clustering by hand
   In the this problem you will perform clustering by hand using the following subset of LPGA 2008 data from the Lecture 12 notebook below. You will create $k = 2$ clusters and use the Euclidean distance as a metric.

   (a) Perform k-means clustering using Cindy Pasechnik and Charlotte Mayorkas as the centroids. Assign the rest of the points to the clusters and re-calculate the centroids.

   (b) Perform hierarchical clustering on the data set to create two clusters of the players. Use complete linkage. Create the denodrogram for the clustering. Report the height at which the dendrogram is cut to create 2 clusters.

   (c) Calculate the Rand index for the k-means clustering and the hierarchical clustering after part (a) and (b).



Golfer	AvgDrive	FairwayP
Yim, Sung Ah	235.2	78.3
Koch, Carin	236.8	67.8
Blasberg, Erica	245.4	69.2
Mayorkas, Charlotte	252	70.7
Pasechnik, Cindy	226.7	72.1

Figure 1: LPGA 2008 Data

3. To Standardize or not standardize [Adapted from Yu and Barter 2024, Ch. 6 Exercise 16]
   For each of the following project goals, i) discuss whether it makes sense to mean-center and scale the data (using standard devation or range) if you were to apply a clustering method or PCA. Then ii) discuss how a clustering method and/or PCA would or would not help with achieving the goal.

   (a) Your goal is to develop an algorithm that will predict how much houses will sell for in your city. Your data consists of a large number of numeric features (e.g., the area, quality, and number of bedrooms) for several thousand houses that have recently been sold, and you aim to create a simpler lower-dimensional dataset that you can use as the input for your predictive algorithm.

   (b) You are hoping to learn about public opinion on autonomous vehicles. Your goal is to create a simple visualization, such as a scatterplot or a histogram, that can be used to visualize the different categories of people based on their answers to a survey. The survey is conducted on random members of the public and asks dozens of questions whose answers are each on a scale of 0 to 5, such as “How safe do you think autonomous vehicles are?” and “How much do you know about autonomous vehicle technology?”

4. Backpropagation

   Consider a neural network with the following architecture:

   - Input $x \in \mathbb{R}^2$
   - Hidden layer: linear layer with sigmoid activation with 3 neurons
   - Output layer: linear layer with a single neuron with sigmoid activation

   The network has the following parameters:

$$\begin{aligned} W_1 &= \text{weights from input to hidden neurons} \\ b_1 &= \text{bias of hidden neurons} \\ W_2 &= \text{weights from hidden neurons to output} \\ b_2 &= \text{bias of output} \end{aligned}$$

   A linear layer with sigmoid activation means that the output of the linear transformation of the input is inputted to the sigmoid function. For example, for the hidden neuron h in the above two-layer neural network is calculated using: $h = \sigma(z_1)$ where $z_1 = W_1 x + b_1$.

   (Hints: (1) The derivative of the sigmoid: $\frac{d\sigma(z)}{dz} = \sigma(z)(1 - \sigma(z))$ (2) Remember that $\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial W_2}$

   Problem:

(a) Draw this neural network, labeling the weights and bias. What are the dimensions of each of the parameters $W_1, b_1, W_2,$ and b_2 ?

(b) Expand out the formula for $\hat{y}$ so that it's a closed-form expression using only variables $W_1, b_1, W_2, b_2, \sigma$, and the input x.

(c) Using the squared loss $L = (y - \hat{y})^2$, where y is the true value we are trying to predict, we will complete one update of the weights of the neural network using one sample to trace through backpropagation and stochastic gradient descent. Here are the initial parameters:

$$W_1 = \begin{bmatrix} 0.13315865 & 0.0715279 \\ -0.15454003 & -0.00083838 \\ 0.0621336 & -0.07200856 \end{bmatrix}$$

$$b_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$W_2 = \begin{bmatrix} 0.02655116 & 0.01085485 & 0.00042914 \end{bmatrix}$$

$$b_2 = \begin{bmatrix} 0 \end{bmatrix}$$

This is our sample:

$$x = \begin{bmatrix} -0.01746002 & 0.04330262 \end{bmatrix}$$

$$y = \begin{bmatrix} 1 \end{bmatrix}$$

i. Using the forward algorithm (or the closed-from above), calculate $\hat{y}$ for the sample. (Remember that you may want to reuse some of these calculations for backprop.)

ii. Using backpropagation, derive the gradients of the loss with respect to each parameter. (Remember to use the chain rule and what you can save and reuse from the calculations.) Write down each partial derivative formula in terms of y, $\hat{y}$, activations, net inputs, and x. (Your final answer should have no partial derivatives left that are not simplified.)

iii. Calculate the update for the weight, $w_{1,1}$ in the matrix W_1. To do so, first calculate the value of the partial derivative using the sample and initial parameters. Update the weight using the gradient and the learning rate, $\eta = .1$ Please show and explain your work. (You may use Python or other coding to do these calculations for you. The MNIST NN code from lecture can help with this. To update all the weights, you would calculate the value of each partial derivative using the sample and initial parameters and update all the weights for the model using the gradients and learning rate. However, it is not necessary to show how to do all of them for this problem.)

Citations:
Yu, B., & Barter, R. L. (2024). Veridical data science: The practice of responsible data analysis and decision making. The MIT Press.