CS 383 Assignment 1

Kevin Tayah (kst46@drexel.edu)

January 31, 2021

1. Theory Questions

a) Assuming the features are categorical, we will not have to take the standard deviation. Therefore we may continue with the calculations. Based on this, we can come up with the following set of subsets of our data based on whether feature 1 is either 0, 1, or 2 (the only options they can be).

$f_1 = 0$ we will have $\begin{bmatrix}0 & 1\\ 0 & 0\\ 0 & 0\end{bmatrix}$ $\begin{bmatrix}1\\ 1\\ 1\end{bmatrix}$ which has a probability, $P(y=0) = 0$

$f_1 = 1$ we will have $\begin{bmatrix}1 & 1\\ 1 & 1\\ 1 & 0\\ 1 & 0\\ 1 & 1\end{bmatrix}$ $\begin{bmatrix}1\\ 1\\ 0\\ 0\\ 0\end{bmatrix}$ which has a probability, $P(y=0) = \frac{3}{5}$

$f_1 = 2$ we will have $\begin{bmatrix}2 & 0\\ 2 & 1\end{bmatrix}$ $\begin{bmatrix}0\\ 0\end{bmatrix}$ which has a probability, $P(y=0) = 1$

We can then calculate the total entropy of feature 1 using the following formula:

$$H(P(v_1), ..., P(v_k)) = \sum_{i=1}^{K}(-P(v_i) log_K P(v_i))$$

$$H = \frac{5}{10} \cdot (\frac{-3}{5} log_2 \frac{3}{5} - \frac{2}{5} log_2 \frac{2}{5})$$

$$H = 0.4855$$

b) Based on the same logic as part a, we can split our data into subsets of our data based upon whether feature 1 is either 0, 1, or 2.

$f_2 = 0$ we will have $\begin{bmatrix}0 & 0\\ 0 & 0\\ 1 & 0\\ 1 & 0\\ 2 & 0\end{bmatrix}$ $\begin{bmatrix}1\\ 1\\ 0\\ 0\\ 0\end{bmatrix}$ which has a probability, $P(y=0) = \frac{3}{5}$

$f_2 = 1$ we will have $\begin{bmatrix}0 & 1\\ 1 & 1\\ 1 & 1\\ 1 & 1\\ 2 & 1\end{bmatrix}$ $\begin{bmatrix}1\\ 1\\ 1\\ 0\\ 0\end{bmatrix}$ which has a probability, $P(y=0) = \frac{3}{5}$

$f_2 = 2$ we do not have subsets with this criteria. Therefore $P(y = 0) = 0$ and $E = 0$

We can then calculate the total entropy of feature 2 using the same formula as before:

$$H(P(v_1), ..., P(v_k)) = \sum_{i=1}^{K}(-P(v_i)log_K P(v_i))$$
$$H = \frac{5}{10} \cdot (\frac{-3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5}) + \frac{5}{10} \cdot (\frac{-3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5})$$
$$H = 0.971$$

c) It seems that the most discriminating feature is feature 1 since it has a more deterministic behavior compared to feature 2 since it's entropy is closer to 0 while feature 2 is closer to 1.

d) Since we are now attempting to calculate the principle components, we will assume are features are continuous, therefore we must zscore our values. Using python functions, I was able to determine that for feature 1, $\mu = 0.9$ and $\sigma = 0.738$ and for feature 2, $\mu = 0.5$ and $\sigma = 0.527$. Therefore, if we are to subtract from all our features values by their respective mean and then divide by their respective standard deviation, we get the following standardized X matrix:

$$X = \begin{bmatrix} -1.21973567 & 0.0513167 \\ -1.21973567 & -0.9486833 \\ -0.21973567 & 0.0513167 \\ -1.21973567 & -0.9486833 \\ -0.21973567 & 0.0513167 \\ -0.21973567 & -0.9486833 \\ -0.21973567 & -0.9486833 \\ -0.21973567 & 0.0513167 \\ 0.78026433 & -0.9486833 \\ 0.78026433 & 0.0513167 \end{bmatrix}$$

Based upon the standardized X, I used python numpy's *cov* and *eig* tool to calculate the covariance matrix and eigenvalues and eigenvectors from the covariance matrix. The following values were are eigenvalues = $\begin{bmatrix} 0.5556 & 0.2667 \end{bmatrix}$ and our eigenvectors are $\begin{bmatrix} 0.9806 & -0.1961 \\ 0.1961 & 0.9806 \end{bmatrix}$. From these we can create our first, second, ..., kth principal component from performing the following formula: $Z = XW$ where W is our kth eigenvector ordered by it's corresponding eigenvalue. From this we get the following collection of pca's where the first column is the first projection and the second is the second projection.

$$\begin{pmatrix} -1.18598519 & 0.28953001 \\ -1.38210133 & -0.69105066 \\ -0.20540452 & 0.09341388 \\ -1.38210133 & -0.69105066 \\ -0.20540452 & 0.09341388 \\ -0.40152065 & -0.8871668 \\ -0.40152065 & -0.8871668 \\ -0.20540452 & 0.09341388 \\ 0.57906002 & -1.08328293 \\ 0.77517616 & -0.10270226 \end{pmatrix}$$

e) Each of these value pairs (feature 1 and feature 2) can be plotted on a 2D Cartesian coordinate system to represent a standardized version of of our original data. Since we calculated both pca values, we have reduced our dimensionality to 2, which is our original dimensionality. Our values that we got in part d are the same values (+/- rounding errors).

d) The first column of our calculated pca in part d is the 1-D reduction. So our new X would look like so:

$$\begin{pmatrix} -1.18598519 \\ -1.38210133 \\ -0.20540452 \\ -1.38210133 \\ -0.20540452 \\ -0.40152065 \\ -0.40152065 \\ -0.20540452 \\ 0.57906002 \\ 0.77517616 \end{pmatrix}$$