ECON 425 HW9

March 5, 2024

Due Thu, March 14, 6pm in Bruinlearn

Problem 1 (PC1 is the direction of maximal variation) Prove that the first principal component w_1 of the centered data $X \in \mathbb{R}^{n \times k}$ is the direction of maximum variance in the sense that

$$w_1 = \arg\max_{w \in \mathbb{R}^k} \widehat{\text{Var}}\left[w'x\right] \text{ s.t. } \|w\| = 1,$$

where $\|\cdot\|$ is the Euclidean norm and

$$\widehat{\text{Var}}\left[w'x\right] = \frac{1}{n}\sum_{i=1}^{n}(w'x_i)^2.$$

Hint. Set the derivative of the Lagrangian to zero and use this equality to prove that the Lagrange multiplier is the largest eigenvalue of the sample covariance matrix $\hat{\Sigma} = \frac{1}{n}X'X$.

Problem 2 (K-means clustering)

In this problem, you will perform K-means clustering manually, with $K = 2$, on a small example with $n = 6$ observations and $p = 2$ features. The data matrix is

$$X = \begin{pmatrix} 1 & 4 \\ 1 & 3 \\ 0 & 4 \\ 5 & 1 \\ 6 & 2 \\ 4 & 0 \end{pmatrix}$$

(a) Plot the observations.

(b) Randomly assign a cluster label to each observation. You can use the np.random.choice() function to do this. Report the cluster labels for each observation.

(c) Compute the centroid for each cluster.

(d) Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.

(e) Repeat (c) and (d) until the answers obtained stop changing.

(f) In your plot from (a), color the observations according to the cluster labels obtained.

Problem 3 (PCA on Olivetti faces)

Use the Olivetti faces dataset available through sklearn to do the following.

(a) Fetch and load the data with the fetch_olivetti_faces method from sklearn.datasets.

(b) Demean each face in the data set (no need to divide by standard deviation as every dimension is a number between a fixed range representing a pixel).

(c) Compute and display the first 9 *eigenfaces*. The k-th eigenface of a given face is an image based on the first k principal components only.

(d) Any given face in the data set can be represented as a linear combination of the eigenfaces. For any face in the data set, show how it progresses as we combine $1, 51, 101, ...$ eigenfaces, until the full image is recovered.