HOMEWORK 5

Dario Placencio - 907 284 6018

Instructions: Use this latex file as a template to develop your homework. Submit your homework on time as a single pdf file. Please wrap your code and upload to a public GitHub repo, then attach the link below the instructions so that we can access it. Answers to the questions that are not within the pdf are not accepted. This includes external links or answers attached to the code implementation. Late submissions may not be accepted. You can choose any programming language (i.e. python, R, or MATLAB). Please check Piazza for updates about the homework. It is ok to share the experiments results and compare them with each other.

1 Clustering

1.1 K-means Clustering (14 points)

1. **(6 Points)** Given n observations $X_1^n = \{X_1, \ldots, X_n\}$, $X_i \in \mathcal{X}$, the K-means objective is to find k ($< n$) centres $\mu_1^k = \{\mu_1, \ldots, \mu_k\}$, and a rule $f : \mathcal{X} \to \{1, \ldots, K\}$ so as to minimize the objective

$$J(\mu_1^K, f; X_1^n) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}(f(X_i) = k)\|X_i - \mu_k\|^2 \tag{1}$$

Let $\mathcal{J}_K(X_1^n) = \min_{\mu_1^K, f} J(\mu_1^K, f; X_1^n)$. Prove that $\mathcal{J}_K(X_1^n)$ is a non-increasing function of K.

We want to show that for any given set of observations X_1^n, the K-means objective function:

$$\mathcal{J}_K(X_1^n) = \min_{\mu_1^K, f} \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbf{1}(f(X_i) = k)\|X_i - \mu_k\|^2$$

is non-increasing with respect to K. This means that:

$$\mathcal{J}_{K+1}(X_1^n) \leq \mathcal{J}_K(X_1^n)$$

Proof:

(a) Consider the optimal clustering given by μ_1^K and f for K clusters that achieves the minimum in $\mathcal{J}_K(X_1^n)$.

(b) When we move from K to $K+1$ clusters, we have two cases for the new set of centers μ_1^{K+1}:

i. The new center μ_{K+1} is equal to one of the existing centers, say μ_k. In this case, the assignment function f can remain the same, and the objective function does not change:

$$J(\mu_1^{K+1}, f; X_1^n) = J(\mu_1^K, f; X_1^n) = \mathcal{J}_K(X_1^n)$$

ii. The new center μ_{K+1} is different from all existing centers. In this case, at least one observation X_i that was assigned to some cluster k might now be closer to μ_{K+1}, and hence the assignment function f could change for some i, possibly reducing the overall objective function. Thus:

$$J(\mu_1^{K+1}, f; X_1^n) \leq J(\mu_1^K, f; X_1^n) = \mathcal{J}_K(X_1^n)$$

(c) Hence, in either case, we have:

$$\mathcal{J}_{K+1}(X_1^n) \leq \mathcal{J}_K(X_1^n)$$

Therefore, $\mathcal{J}_K(X_1^n)$ is a non-increasing function of K.

2. **(8 Points)** Consider the K-means (Lloyd's) clustering algorithm we studied in class. We terminate the algorithm when there are no changes to the objective. Show that the algorithm terminates in a finite number of steps.

   Let $J^{(t)}$ be the objective function value at iteration t.Then, $J^{(t+1)} \leq J^{(t)}$.

   At each iteration, two steps are performed:

   (1)Cluster Assignment Step: Assign each data point X_i to the nearest center μ_k.

   (2)Centroid Update Step: Update each μ_k to be the mean of all points assigned to cluster k.

   Since the mean minimizes the sum of squared distances:

$$J^{(t+1)} = \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbf{1}(f^{(t+1)}(X_i) = k)\|X_i - \mu_k^{(t+1)}\|^2$$
$$\leq \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbf{1}(f^{(t)}(X_i) = k)\|X_i - \mu_k^{(t)}\|^2 = J^{(t)}.$$

   Since there are a finite number of data points, there are a finite number of possible assignments of points to K clusters.

   Thus, the algorithm must terminate in a finite number of steps as the objective function J cannot decrease indefinitely.

1.2 Experiment (20 Points)

In this question, we will evaluate K-means clustering and GMM on a simple 2 dimensional problem. First, create a two-dimensional synthetic dataset of 300 points by sampling 100 points each from the three Gaussian distributions shown below:

$$P_a = \mathcal{N}\left(\begin{bmatrix}-1\\-1\end{bmatrix}, \sigma\begin{bmatrix}2, & 0.5\\0.5, & 1\end{bmatrix}\right), \quad P_b = \mathcal{N}\left(\begin{bmatrix}1\\-1\end{bmatrix}, \sigma\begin{bmatrix}1, & -0.5\\-0.5, & 2\end{bmatrix}\right), \quad P_c = \mathcal{N}\left(\begin{bmatrix}0\\1\end{bmatrix}, \sigma\begin{bmatrix}1 & 0\\0, & 2\end{bmatrix}\right)$$

Here, σ is a parameter we will change to produce different datasets.

First implement K-means clustering and the expectation maximization algorithm for GMMs. Execute both methods on five synthetic datasets, generated as shown above with $\sigma \in \{0.5, 1, 2, 4, 8\}$. Finally, evaluate both methods on *(i)* the clustering objective (**??**) and *(ii)* the clustering accuracy. For each of the two criteria, plot the value achieved by each method against σ.

Guidelines:

- Both algorithms are only guaranteed to find only a local optimum so we recommend trying multiple restarts and picking the one with the lowest objective value (This is (**??**) for K-means and the negative log likelihood for GMMs). You may also experiment with a smart initialization strategy (such as kmeans++).
- To plot the clustering accuracy, you may treat the 'label' of points generated from distribution P_u as u, where $u \in \{a, b, c\}$. Assume that the cluster id i returned by a method is $i \in \{1, 2, 3\}$. Since clustering is an unsupervised learning problem, you should obtain the best possible mapping from $\{1, 2, 3\}$ to $\{a, b, c\}$ to compute the clustering objective. One way to do this is to compare the clustering centers returned by the method (centroids for K-means, means for GMMs) and map them to the distribution with the closest mean.

Points break down: 7 points each for implementation of each method, 6 points for reporting of evaluation metrics.

2 Linear Dimensionality Reduction

2.1 Principal Components Analysis (10 points)

Principal Components Analysis (PCA) is a popular method for linear dimensionality reduction. PCA attempts to find a lower dimensional subspace such that when you project the data onto the subspace as much of the information is preserved. Say we have data $X = [x_1^\top; \ldots; x_n^\top] \in \mathbb{R}^{n\times D}$ where $x_i \in \mathbb{R}^D$. We wish to find a d ($< D$) dimensional subspace $A = [a_1, \ldots, a_d] \in \mathbb{R}^{D\times d}$, such that $a_i \in \mathbb{R}^D$ and $A^\top A = I_d$, so as to maximize $\frac{1}{n}\sum_{i=1}^{n} \|A^\top x_i\|^2$.

1. **(4 Points)** Suppose we wish to find the first direction a_1 (such that $a_1^\top a_1 = 1$) to maximize $\frac{1}{n}\sum_i (a_1^\top x_i)^2$. Show that a_1 is the first right singular vector of X.

   Given the data matrix $X \in \mathbb{R}^{n\times D}$, we want to find the first principal component $a_1 \in \mathbb{R}^D$ such that $a_1^\top a_1 = 1$ and it maximizes the quantity $\frac{1}{n}\sum_{i=1}^n (a_1^\top x_i)^2$.

   Using the singular value decomposition, we can write X as $X = U\Sigma V^\top$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{D\times D}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{n\times D}$ is a diagonal matrix with non-negative real numbers on the diagonal.

   The columns of V (the right singular vectors) are the eigenvectors of $X^\top X$, and the columns of U (the left singular vectors) are the eigenvectors of $XX^\top$.

   The optimization problem for the first principal component can be written as:

$$\max_{a_1 : a_1^\top a_1 = 1} \frac{1}{n}\sum_{i=1}^n (a_1^\top x_i)^2 = \max_{a_1 : a_1^\top a_1 = 1} a_1^\top \left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right) a_1. \tag{2}$$

   This is equivalent to the Rayleigh quotient, which is maximized when a_1 is the eigenvector corresponding to the largest eigenvalue of the covariance matrix $\frac{1}{n}X^\top X$.

   Since V contains the eigenvectors of $X^\top X$, the first column of V, which corresponds to the largest singular value (and thus the largest eigenvalue of $X^\top X$), is the solution to our optimization problem. Therefore, a_1 is the first right singular vector of X.

2. **(6 Points)** Given $a_1, \ldots, a_k$, let $A_k = [a_1, \ldots, a_k]$ and $\tilde{x}_i = x_i - A_k A_k^\top x_i$. We wish to find a_{k+1}, to maximize $\frac{1}{n}\sum_i (a_{k+1}^\top \tilde{x}_i)^2$. Show that a_{k+1} is the $(k+1)^{th}$ right singular vector of X.

   Given the data matrix $X \in \mathbb{R}^{n\times D}$, suppose we have already found the first k principal components $a_1, \ldots, a_k$ and constructed the matrix $A_k = [a_1, \ldots, a_k]$. For each data point x_i, we define the residual after projecting onto the subspace spanned by A_k as $\tilde{x}_i = x_i - A_k A_k^\top x_i$. We wish to find the next principal component a_{k+1} such that it maximizes the variance of the residuals, i.e.,

$$\max_{a_{k+1} : a_{k+1}^\top a_{k+1} = 1} \frac{1}{n}\sum_{i=1}^n (a_{k+1}^\top \tilde{x}_i)^2. \tag{3}$$

   This is equivalent to finding the eigenvector associated with the largest eigenvalue of the covariance matrix of the residuals, $\frac{1}{n}\tilde{X}^\top \tilde{X}$, where $\tilde{X}$ is the matrix with rows $\tilde{x}_i^\top$.

   Note that $\tilde{X}$ can be written as $\tilde{X} = X - A_k A_k^\top X$. Using the fact that A_k is orthogonal to the residual space and $A_k^\top A_k = I_k$, we can perform an SVD on $\tilde{X}$ to find its right singular vectors.

   The $(k+1)^{th}$ principal component a_{k+1} will then be the right singular vector of $\tilde{X}$ associated with its largest singular value, which is not in the span of A_k. Since the singular vectors of $\tilde{X}$ and X are the same beyond the first k vectors, a_{k+1} is also the $(k+1)^{th}$ right singular vector of X.

   Thus, we conclude that a_{k+1} is the $(k+1)^{th}$ right singular vector of X.

2.2 Dimensionality reduction via optimization (22 points)

We will now motivate the dimensionality reduction problem from a slightly different perspective. The resulting algorithm has many similarities to PCA. We will refer to method as DRO.
As before, you are given data $\{x_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^D$. Let $X = [x_1^\top; \ldots x_n^\top] \in \mathbb{R}^{n\times D}$. We suspect that the data actually lies approximately in a d dimensional affine subspace. Here $d < D$ and $d < n$. Our goal, as in PCA, is to use this dataset to find a d dimensional representation z for each $x \in \mathbb{R}^D$. (We will assume that the span of the data has dimension larger than d, but our method should work whether $n > D$ or $n < D$.)
Let $z_i \in \mathbb{R}^d$ be the lower dimensional representation for x_i and let $Z = [z_1^\top; \ldots; z_n^\top] \in \mathbb{R}^{n\times d}$. We wish to find parameters $A \in \mathbb{R}^{D\times d}$, $b \in \mathbb{R}^D$ and the lower dimensional representation $Z \in \mathbb{R}^{n\times d}$ so as to minimize

$$J(A, b, Z) = \frac{1}{n}\sum_{i=1}^n \|x_i - Az_i - b\|^2 = \|X - ZA^\top - \mathbf{1}b^\top\|_F^2. \tag{4}$$

Here, $\|A\|_F^2 = \sum_{i,j} A_{ij}^2$ is the Frobenius norm of a matrix.

1. **(3 Points)** Let $M \in \mathbb{R}^{d\times d}$ be an arbitrary invertible matrix and $p \in \mathbb{R}^d$ be an arbitrary vector. Denote, $A_2 = A_1M^{-1}$, $b_2 = b_1 - A_1M^{-1}p$ and $Z_2 = Z_1M^\top + \mathbf{1}p^\top$. Show that both (A_1, b_1, Z_1) and (A_2, b_2, Z_2) achieve the same objective value J (**??**).

Therefore, in order to make the problem determined, we need to impose some constraint on Z. We will assume that the z_i's have zero mean and identity covariance. That is,

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}Z^\top \mathbf{1}_n = 0, \qquad S = \frac{1}{n}\sum_{i=1}^{n} z_i z_i^\top = \frac{1}{n}Z^\top Z = I_d$$

Here, $\mathbf{1}_d = [1, 1 \ldots, 1]^\top \in \mathbb{R}^d$ and I_d is the $d \times d$ identity matrix.

To show that both (A_1, b_1, Z_1) and (A_2, b_2, Z_2) achieve the same objective value J, we must demonstrate that the transformation involving M and p does not change the Frobenius norm of the error matrix.
Let's consider the transformed variables:

$$\begin{aligned} A_2 &= A_1M^{-1}, \\ b_2 &= b_1 - A_1M^{-1}p, \\ Z_2 &= Z_1M^\top + \mathbf{1}p^\top. \end{aligned}$$

The objective function $J(A, b, Z)$ is defined as:

$$J(A, b, Z) = \frac{1}{n}\sum_{i=1}^{n} \|x_i - Az_i - b\|^2 = \|X - ZA^\top - \mathbf{1}b^\top\|_F^2.$$

Now, we substitute A_2, b_2, and Z_2 into the objective function and simplify:

$$\begin{aligned} J(A_2, b_2, Z_2) &= \|X - Z_2A_2^\top - \mathbf{1}b_2^\top\|_F^2 \\ &= \|X - (Z_1M^\top + \mathbf{1}p^\top)(M^{-1})^\top A_1^\top - \mathbf{1}(b_1 - A_1M^{-1}p)^\top\|_F^2 \\ &= \|X - Z_1A_1^\top - \mathbf{1}p^\top A_1^\top - \mathbf{1}b_1^\top + \mathbf{1}p^\top A_1^\top\|_F^2 \\ &= \|(X - Z_1A_1^\top - \mathbf{1}b_1^\top) + \mathbf{1}p^\top A_1^\top - \mathbf{1}p^\top A_1^\top\|_F^2 \\ &= \|X - Z_1A_1^\top - \mathbf{1}b_1^\top\|_F^2 \\ &= J(A_1, b_1, Z_1). \end{aligned}$$

Thus, the objective value remains unchanged under the transformation using M and p, showing that $J(A_1, b_1, Z_1) = J(A_2, b_2, Z_2)$. This means that the solution is not unique and that any set of parameters related by an invertible linear transformation will yield the same objective value. Therefore, constraints must be imposed on Z to make the problem well-defined, such as requiring that Z has zero mean and identity covariance.

2. **(16 Points)** Outline a procedure to solve the above problem. Specify how you would obtain A, Z, b which minimize the objective and satisfy the constraints.

   Hint: The rank k approximation of a matrix in Frobenius norm is obtained by taking its SVD and then zeroing out all but the first k singular values.

   To solve the given problem, we can leverage the fact that the best rank k approximation in the Frobenius norm of a matrix is given by its Singular Value Decomposition (SVD) truncated to the first k singular values. The optimization process can be outlined as follows:

   1. Center the Data: Subtract the mean of the data points from each point to ensure that the data is centered around the origin. Let $\bar{x}$ be the mean of $\{x_i\}_{i=1}^n$, then we update x_i to $x_i - \bar{x}$ for each i. This centers X and allows us to ignore b in the initial steps since it's now zero.

   2. SVD of the Centered Data: Compute the SVD of the centered data matrix X. The SVD is given by $X = U\Sigma V^\top$, where U and V are orthogonal matrices, and Σ is a diagonal matrix with non-negative real numbers on the diagonal (singular values).

   3. Truncate the SVD: To get a rank d approximation of X, retain only the first d columns of U and V, and the first d singular values in Σ. This gives the matrices U_d, Σ_d, and V_d.

4. Compute the Lower Dimensional Representation Z: The matrix Z can be computed as $Z = U_d\Sigma_d$. This is the projection of the data onto the d-dimensional subspace that captures the most variance.

5. Reconstruct A: The matrix A can be computed as the first d columns of V (i.e., V_d).

6. Recover b: Now that we have Z and A, we can find b by solving the equation $\bar{x} = A\bar{z} + b$, where $\bar{z}$ is the mean of the projections Z. Since Z has zero mean by construction, b is simply the original mean $\bar{x}$ of the data.

7. Enforce the Constraints: To ensure that Z has zero mean and identity covariance, we can further orthogonalize Z using QR decomposition if necessary. However, if the SVD is computed correctly, Z should already satisfy these constraints.

The procedure outlined above provides a solution to the dimensionality reduction problem as specified, yielding the parameters A, Z, and b that minimize the objective while satisfying the constraints of zero mean and identity covariance for Z.

3. **(3 Points)** You are given a point x_* in the original D dimensional space. State the rule to obtain the d dimensional representation z_* for this new point. (If x_* is some original point x_i from the D–dimensional space, it shoud be the d–dimensional representation z_i.)

To obtain the d-dimensional representation z_* for a new point x_* in the original D-dimensional space, we could follow these steps:

1. Center the New Point: Subtract the mean $\bar{x}$ of the original data points from x_* to center it in the same way as the original data. This gives the centered point $x_{\text{centered}} = x_* - \bar{x}$.

2. Project onto the New Subspace: Use the matrix A obtained from the dimensionality reduction process to project the centered point onto the new d-dimensional subspace. This is done by calculating $z_* = A^\top x_{\text{centered}}$.

The resulting vector z_* is the d-dimensional representation of x_*. This representation will be consistent with the representations obtained for the original data points, meaning that if x_* was an original data point x_i, then z_* should correspond to the d-dimensional representation z_i obtained during the dimensionality reduction process.

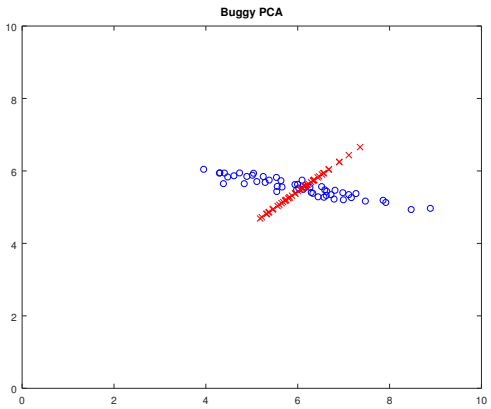
2.3 Experiment (34 points)

Here we will compare the above three methods on two data sets.

- We will implement three variants of PCA:
  1. "buggy PCA": PCA applied directly on the matrix X.
  2. "demeaned PCA": We subtract the mean along each dimension before applying PCA.
  3. "normalized PCA": Before applying PCA, we subtract the mean and scale each dimension so that the sample mean and standard deviation along each dimension is 0 and 1 respectively.
- One way to study how well the low dimensional representation Z captures the linear structure in our data is to project Z back to D dimensions and look at the reconstruction error. For PCA, if we mapped it to d dimensions via $z = Vx$ then the reconstruction is $V^\top z$. For the preprocessed versions, we first do this and then reverse the preprocessing steps as well. For DRO we just compute $Az + b$. We will compare all methods by the reconstruction error on the datasets.
- Please implement code for the methods: Buggy PCA (just take the SVD of X) , Demeaned PCA, Normalized PCA, DRO. In all cases your function should take in an $n \times d$ data matrix and d as an argument. It should return the the d dimensional representations, the estimated parameters, and the reconstructions of these representations in D dimensions.
- You are given two datasets: A two Dimensional dataset with 50 points `data2D.csv` and a thousand dimensional dataset with 500 points `data1000D.csv`.
- For the $2D$ dataset use $d = 1$. For the $1000D$ dataset, you need to choose d. For this, observe the singular values in DRO and see if there is a clear "knee point" in the spectrum. Attach any figures/ Statistics you computed to justify your choice.

- For the $2D$ dataset you need to attach the a plot comparing the orignal points with the reconstructed points for all 4 methods. For both datasets you should also report the reconstruction errors, that is the squared sum of differences $\sum_{i=1}^{n} \|x_i - r(z_i)\|^2$, where x_i's are the original points and $r(z_i)$ are the D dimensional points reconstructed from the d dimensional representation z_i.
- **Questions:** After you have completed the experiments, please answer the following questions.
  1. Look at the results for Buggy PCA. The reconstruction error is bad and the reconstructed points don't seem to well represent the original points. Why is this?
     Hint: Which subspace is Buggy PCA trying to project the points onto?
  2. The error criterion we are using is the average squared error between the original points and the reconstructed points. In both examples DRO and demeaned PCA achieves the lowest error among all methods. Is this surprising? Why?
- Point allocation:
  - Implementation of the three PCA methods: **(6 Points)**
  - Implementation of DRO: **(6 points)**
  - Plots showing original points and reconstructed points for 2D dataset for each one of the 4 methods: **(10 points)**
  - Implementing reconstructions and reporting results for each one of the 4 methods for the 2 datasets: **(5 points)**
  - Choice of d for $1000D$ dataset and appropriate justification: **(3 Points)**
  - Questions **(4 Points)**

Answer format:
The graph bellow is in example of how a plot of one of the algorithms for the 2D dataset may look like:



The blue circles are from the original dataset and the red crosses are the reconstructed points.

And this is how the reconstruction error may look like for Buggy PCA for the 2D dataset: 0.886903