# CM146, Winter 2023
# Problem Set 4: Boosting, Unsupervised learning
# Due March 19, 11:59pm (Math), March 24, 11:59pm (Coding)

## Submission instructions

- Submit your solutions electronically on the course Gradescope site as PDF files.
- If you plan to typeset your solutions, please use the LaTeX solution template. If you must submit scanned handwritten solutions, please use a black pen on blank white paper and a high-quality scanner app.
- First three math heavy questions are due Sunday March 19, 11:59pm. Suggested solutions will be released immediately after. Coding/Implementation questions will be due on Friday March 24, 11:59pm.

## 1 AdaBoost [5 pts]

In the lecture on ensemble methods, we said that in iteration $t$, AdaBoost is picking $(h_t, \beta_t)$ that minimizes the objective:

$$\begin{aligned}(h_t^*(\boldsymbol{x}), \beta_t^*) &= \arg\min_{(h_t(\boldsymbol{x}),\beta_t)} \sum_n w_t(n) e^{-y_n \beta_t h_t(\boldsymbol{x}_n)} \\ &= \arg\min_{(h_t(\boldsymbol{x}),\beta_t)} (e^{\beta_t} - e^{-\beta_t}) \sum_n w_t(n) \mathbb{I}[y_n \neq h_t(\boldsymbol{x}_n)] \\ &\qquad + e^{-\beta_t} \sum_n w_t(n)\end{aligned}$$

We define the weighted misclassification error at time t, $\epsilon_t$ to be $\epsilon_t = \sum_n w_t(n)\mathbb{I}[y_n \neq h_t(\boldsymbol{x}_n)]$. Also the weights are normalized so that $\sum_n w_t(n) = 1$.

(a) Take the derivative of the above objective function with respect to $\beta_t$ and set it to zero to solve for $\beta_t$ and obtain the update for $\beta_t$.

(b) Suppose the training set is linearly separable, and we use a hard-margin linear support vector machine (no slack) as a base classifier. In the first boosting iteration, what would the resulting $\beta_1$ be?

Parts of this assignment are adapted from course material by Jenna Wiens (UMich) and Tommi Jaakola (MIT).

## 2 K-means for single dimensional data [5 pts]

In this problem, we will work through K-means for a single dimensional data.

(a) Consider the case where $K = 3$ and we have 4 data points $x_1 = 1, x_2 = 2, x_3 = 5, x_4 = 7$. What is the optimal clustering for this data ? What is the corresponding value of the objective ?

(b) One might be tempted to think that Lloyd's algorithm is guaranteed to converge to the global minimum when $d = 1$. Show that there exists a suboptimal cluster assignment (*i.e.*, initialization) for the data in the above part that Lloyd's algorithm will not be able to improve (to get full credit, you need to show the assignment, show why it is suboptimal *and* explain why it will not be improved).

## 3 Gaussian Mixture Models [8 pts]

We would like to cluster data $\{x_1, \ldots, x_N\}$, $x_n \in \mathbb{R}^d$ using a Gaussian Mixture Model (GMM) with $K$ mixture components. To do this, we need to estimate the parameters $\boldsymbol{\theta}$ of the GMM, *i.e.*, we need to set the values $\boldsymbol{\theta} = \{\omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ where $\omega_k$ is the mixture weight associated with mixture component $k$, and $\boldsymbol{\mu}_k$ and $\Sigma_k$ denote the mean and the covariance matrix of the Gaussian distribution associated with mixture component $k$.

If we knew which cluster each sample $x_n$ belongs to (we had complete data), we showed in the lecture on Clustering that the log likelihood $l$ is what we have below and we can compute the maximum likelihood estimate (MLE) of all the parameters.

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_n \log p(\boldsymbol{x}_n, z_n) \\ &= \sum_k \sum_n \gamma_{nk} \log \omega_k + \sum_k \left\{ \sum_n \gamma_{nk} \log N(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \end{aligned} \tag{1}$$

Since we do not have complete data, we use the EM algorithm. The EM algorithm works by iterating between setting each $\gamma_{nk}$ to the posterior probability $p(z_n = k|\boldsymbol{x}_n)$ (step 1 on slide 26 of the lecture on Clustering) and then using $\gamma_{nk}$ to find the value of $\boldsymbol{\theta}$ that maximizes $l$ (step 2 on slide 26). We will now derive updates for one of the parameters, *i.e.*, $\boldsymbol{\mu}_j$ (the mean parameter associated with mixture component $j$).

(a) To maximize $l$, compute $\nabla_{\boldsymbol{\mu}_j} l(\boldsymbol{\theta})$: the gradient of $l(\boldsymbol{\theta})$ with respect to $\boldsymbol{\mu}_j$.

(b) Set the gradient to zero and solve for $\boldsymbol{\mu}_j$ to show that $\boldsymbol{\mu}_j = \frac{1}{\sum_n \gamma_{nj}} \sum_n \gamma_{nj} \boldsymbol{x}_n$.

(c) Suppose that we are fitting a GMM to data using $K = 2$ components. We have $N = 5$ samples in our training data with $x_n, n \in \{1, \ldots, N\}$ equal to: $\{5, 15, 25, 30, 40\}$.

We use the EM algorithm to find the maximum likeihood estimates for the model parameters, which are the mixing weights for the two components, $\omega_1$ and $\omega_2$, and the means for the two components, $\mu_1$ and $\mu_2$. The standard deviations for the two components are fixed at 1.

| $\gamma_1$ | $\gamma_2$ |
|---|---|
| 0.2 | 0.8 |
| 0.2 | 0.8 |
| 0.8 | 0.2 |
| 0.9 | 0.1 |
| 0.9 | 0.1 |

Table 1: Entry in row $n$ and column $k$ of the table corresponds to $\gamma_{nk}$

Suppose that at the end of step 1 of iteration 5 in the EM algorithm, the soft assignment $\gamma_{nk}$ for the five data items are as shown in Table 1.

What are updated values for the parameters $\omega_1$, $\omega_2$, $\mu_1$, and $\mu_2$ at the end of step 2 of the EM algorithm?

# 4 Implementation: Clustering and PCA [32 pts]

Machine learning techniques have been applied to a variety of image interpretation problems. In this project, you will investigate facial recognition, which can be treated as a clustering problem ("separate these pictures of Joe and Mary").

For this project, we will use a small part of a huge database of faces of famous people (Labeled Faces in the Wild [LFW] people dataset[1]). The images have already been cropped out of the original image, and scaled and rotated so that the eyes and mouth are roughly in alignment; additionally, we will use a version that is scaled down to a manageable size of 50 by 37 pixels (for a total of 1850 "raw" features). Our dataset has a total of 1867 images of 19 different people. You will apply dimensionality reduction using principal component analysis (PCA) and explore clustering methods such as k-means and k-medoids to the problem of facial recognition on this dataset.

## Starter Files

---

code and data

- Code: CS146-Winter2023-PS4.ipynb – Code for the `Point`, `Cluster`, and `ClusterSet` classes, on which you will build the clustering algorithms and the main code for the project.
- Utility code: util.py – Utility methods for manipulating data, including through PCA.

Please use your *@g.ucla.edu* email id to access the code. Similar to previous problem sets, copy the colab notebook to your drive and make the changes. Mount the drive appropriately. To work on this HW: you need to download util.py from here. Then, copy/upload this file to your own Google drive.

The notebook has marked blocks where you need to code.

$\#\#\#$ ========= $TODO : START$ ========= $\#\#\#$

$\#\#\#$ ========= $TODO : END$ ========== $\#\#\#$

**Note: For the questions requiring you to complete a piece of code, you are expected to copy-paste your code as a part of the solution in the submission pdf. Tip: If you are using LaTeX, check out the Minted package (example) for code highlighting.**

Please note that you do not necessarily have to follow the skeleton code perfectly. We encourage you to include your own additional methods and functions. For this project, *you are not allowed to use any* `scikit-learn` *classes or functions other than those already imported in the skeleton code.*

### 4.1 PCA and Image Reconstruction [4 pts]

Before attempting automated facial recognition, you will investigate a general problem with images. That is, images are typically represented as thousands (in this project) to millions (more generally) of pixel values, and a high-dimensional vector of pixels must be reduced to a reasonably low-dimensional vector of features.

[1] http://vis-www.cs.umass.edu/lfw/

(a) As always, the first thing to do with any new dataset is to look at it. Use `get_lfw_data(...)` to get the LFW dataset with labels, and plot a couple of the input images using `show_image(...)`. Then compute the mean of all the images, and plot it. (Remember to include all requested images in your writeup.) Comment briefly on this "average" face.

(b) Perform PCA on the data using `util.PCA(...)`. This function returns a matrix `U` whose columns are the principal components, and a vector `mu` which is the mean of the data. If you want to look at a principal component (referred to in this setting as an eigenface), run `show_image(vec_to_image(v))`, where `v` is a column of the principal component matrix. (This function will scale vector `v` appropriately for image display.) Show the top twelve eigenfaces:

```
plot_gallery([vec_to_image(U[:,i]) for i in range(12)])
```

Comment briefly on your observations. Why do you think these are selected as the top eigenfaces?

(c) Explore the effect of using more or fewer dimensions to represent images. Do this by:

- Finding the principal components of the data
- Selecting a number $l$ of components to use
- Reconstructing the images using only the first $l$ principal components
- Visually comparing the images to the originals

To perform PCA, use `apply_PCA_from_Eig(...)` to project the original data into the lower-dimensional space, and then use `reconstruct_from_PCA(...)` to reconstruct high-dimensional images out of lower dimensional ones. Then, using `plotGallery(...)`, submit a gallery of the first 12 images in the dataset, reconstructed with $l$ components, for $l = 1, 10, 50, 100, 500, 1288$. Comment briefly on the effectiveness of differing values of $l$ with respect to facial recognition.

We will revisit PCA in the last section of this project.

### 4.2 $K$-Means and $K$-Medoids [16 pts]

Next, we will explore clustering algorithms in detail by applying them to a toy dataset. In particular, we will investigate $k$-means and $k$-medoids (a slight variation on $k$-means).

(a) In $k$-means, we attempt to find $k$ cluster centers $\boldsymbol{\mu}_j \in \mathbb{R}^d$, $j \in \{1, \ldots, k\}$ and $n$ cluster assignments $c^{(i)} \in \{1, \ldots, k\}$, $i \in \{1, \ldots, n\}$, such that the total distance between each data point and the nearest cluster center is minimized. In other words, we attempt to find $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ and $c^{(1)}, \ldots, c^{(n)}$ that minimizes

$$J(\boldsymbol{c}, \boldsymbol{\mu}) = \sum_{i=1}^{n} ||\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{c^{(i)}}||^2.$$

To do so, we iterate between assigning $\boldsymbol{x}^{(i)}$ to the nearest cluster center $c^{(i)}$ and updating each cluster center $\boldsymbol{\mu}_j$ to the average of all points assigned to the $j^{\text{th}}$ cluster.

Instead of holding the number of clusters $k$ fixed, one can think of minimizing the objective function over $\boldsymbol{\mu}$, $\boldsymbol{c}$, *and* $k$. Show that this is a bad idea. Specifically, what is the minimum possible value of $J(\boldsymbol{c}, \boldsymbol{\mu}, k)$? What values of $\boldsymbol{c}$, $\boldsymbol{\mu}$, and $k$ result in this value?

(b) To implement our clustering algorithms, we will use Python classes to help us define three abstract data types: `Point`, `Cluster`, and `ClusterSet`. Read through the documentation for these classes. (You will be using these classes later, so make sure you know what functionality each class provides!) Some of the class methods are already implemented, and other methods are described in comments. Implement all of the methods marked `TODO` in the `Cluster` and `ClusterSet` classes.

(c) Next, implement `random_init(...)` and `kMeans(...)` based on the specifications provided in the code.

(d) Now test the performance of $k$-means on a toy dataset.

Use `generate_points_2d(...)` to generate three clusters each containing 20 points. (You can modify `generate_points_2d(...)` to test different inputs while debugging your code, but be sure to return to the initial implementation before creating any plots for submission.) You can plot the clusters for each iteration using the `plot_clusters(...)` function.

In your writeup, include plots for the $k$-means cluster assignments and corresponding cluster "centers" *for each iteration* when using random initialization and "k = 3".

(e) Implement `kMedoids(...)` based on the provided specification.

*Hint*: Since $k$-means and $k$-medoids are so similar, you may find it useful to refactor your code to use a helper function `kAverages(points, k, average, init='random', plot=False)`, where `average` is a method that determines how to calculate the average of points in a cluster (so it can take on values `ClusterSet.centroids` or `ClusterSet.medoids`).[2]

As before, include plots for $k$-medoids clustering *for each iteration* when using random initialization and "k = 3".

(f) Finally, we will explore the effect of initialization. Implement `cheat_init(...)`.

Now compare clustering by initializing using `cheat_init(...)`. Include plots for $k$-means and $k$-medoids using "k = 3" *for each iteration.*

### 4.3 Clustering Faces [12 pts]

Finally (!), we will apply clustering algorithms to the image data. To keep things simple, we will only consider data from four individuals. Make a new image dataset by selecting 40 images each from classes 4, 6, 13, and 16, then translate these images to (labeled) points: [3]

```
X1, y1 = util.limit_pics(X, y, [4, 6, 13, 16], 40)
points = build_face_image_points(X1, y1)
```

(a) Apply $k$-means and $k$-medoids to this new dataset with $k = 4$ and initializing the centroids randomly. Evaluate the performance of each clustering algorithm by computing the average

[2] In Python, if you have a function stored to the variable `func`, you can apply it to parameters `arg` by callling `func(arg)`. This works even if `func` is a class method and `arg` is an object that is an instance of the class.

[3] There is a bug in fetch_lfw version 0.18.1 where the results of the loaded images are not always in the same order. This is not a problem for the previous parts but can affect the subset selected in this part. Thus, you may see varying results. Results that show the correct qualitative behavior will get full credit.

cluster purity with `ClusterSet.score(...)`. As the performance of the algorithms can vary widely depending upon the initialization, run both clustering methods 10 times and report the average, minimum, and maximum performance along with runtime. How do the clustering methods compare in terms of clustering performance and runtime?

| | average purity | min purity | max purity | average time | min time | max time |
|---|---|---|---|---|---|---|
| $k$-means | | | | | | |
| $k$-medoids | | | | | | |

Now construct another dataset by selecting 40 images each from two individuals 4 and 13.

(b) Explore the effect of lower-dimensional representations on clustering performance. To do this, compute the principal components for the entire image dataset, then project the newly generated dataset into a lower dimension (varying the number of principal components), and compute the scores of each clustering algorithm.

So that we are only changing one thing at a time, use `init='cheat'` to generate the same initial set of clusters for $k$-means and $k$-medoids. For each value of $l$, the number of principal components, you will have to generate a new list of points using `build_face_image_points(...)`.)

Let $l = 1, 3, 5, \ldots, 49$. The number of clusters $K = 2$. Then, on a single plot, plot the clustering score versus the number of components for each clustering algorithm (be sure to label the algorithms). Discuss the results in a few sentences.

Some pairs of people are more similar to one another and some more different.

(c) Experiment with the data to find a pair of individuals that clustering can discriminate very well and another pair that it finds very difficult (assume you have 40 images for each individual, projected to the top 50 principal components space). Describe your methodology (you may choose any of the clustering algorithms you implemented). Report these twp pairs of individuals (most similar pair and most discriminative pair) in your writeup (display each pair of images using `plot_representative_images`), and comment briefly on the results.