Matheus C. Fernandes
fernandes@g.harvard.edu
CS181-S17

Assignment #4
Due: 5:00pm March 31, 2017

Collaborators:
Bolei Deng
Callin Switzer
Joe Sedlak

Homework 4: Clustering and EM

This homework assignment focuses on different unsupervised learning methods from a theoretical and practical standpoint. In Problem 1, you will explore Hierarchical Clustering and experiment with how the choice of distance metrics can alter the behavior of the algorithm. In Problem 2, you will derive from scratch the full expectation-maximization algorithm for fitting a simple topic model. In Problem 3, you will implement K-Means clustering on a dataset of handwritten images and analyze the latent structure learned by this algorithm.

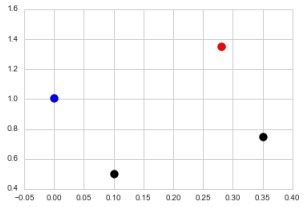
There is a mathematical component and a programming component to this homework. Please submit your PDF and Python files to Canvas, and push all of your work to your GitHub repository. If a question requires you to make any plots, please include those in the writeup.

Hierarchical Clustering [7 pts]

At each step of hierarchical clustering, the two most similar clusters are merged together. This step is repeated until there is one single group. We saw in class that hierarchical clustering will return a different result based on the pointwise-distance and cluster-distance that is is used. In this problem you will examine different choices of pointwise distance (specified through choice of norm) and cluster distance, and explore how these choices change how the HAC algorithm runs on a toy data set.

Problem 1

Consider the following four data points in $\mathbb{R}^2$, belonging to three clusters: the black cluster consisting of $\mathbf{x}_1 = (0.1, 0.5)$ and $\mathbf{x}_2 = (0.35, 0.75)$), the red cluster consisting of $\mathbf{x}_3 = (0.28, 1.35)$, and the blue cluster consisting of $\mathbf{x}_4 = (0, 1.01)$.



Different pointwise distances $d(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_p$ can be used. Recall the definition of the ℓ_1, ℓ_2, and ℓ_∞ norm:

$$\|\mathbf{x}\|_1 = \sum_{j=1}^{m} |x_i| \qquad \|\mathbf{x}\|_2 = \sqrt{\sum_{j=1}^{m} x_i^2} \qquad \|\mathbf{x}\|_\infty = \max_{j \in \{1,\ldots,m\}} |x_j|$$

Also recall the definition of min-distance, max-distance, centroid-distance, and average-distance between two clusters (where $\boldsymbol{\mu}_G$ is the center of a cluster G):

$$\begin{aligned} d_{\text{min}}(G, G') &= \min_{\mathbf{x} \in G, \mathbf{x}' \in G'} d(\mathbf{x}, \mathbf{x}') \\ d_{\text{max}}(G, G') &= \max_{\mathbf{x} \in G, \mathbf{x}' \in G'} d(\mathbf{x}, \mathbf{x}') \\ d_{\text{centroid}}(G, G') &= d(\boldsymbol{\mu}_G, \boldsymbol{\mu}_{G'}) \\ d_{\text{avg}}(G, G') &= \frac{1}{|G||G'|} \sum_{\mathbf{x} \in G} \sum_{\mathbf{x}' \in G'} d(\mathbf{x}, \mathbf{x}') \end{aligned}$$

1. Draw the 2D unit sphere for each norm, defined as $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| = 1\}$. Feel free to do it by hand, take a picture and include it in your pdf.
2. For each norm ($\ell_1, \ell_2, \ell_\infty$) and each clustering distance, specify which two clusters would be the first to merge.
3. Draw the complete dendrograms showing the order of agglomerations for the ℓ_2 norm and each of the clustering distances.

Solution

Problem 1

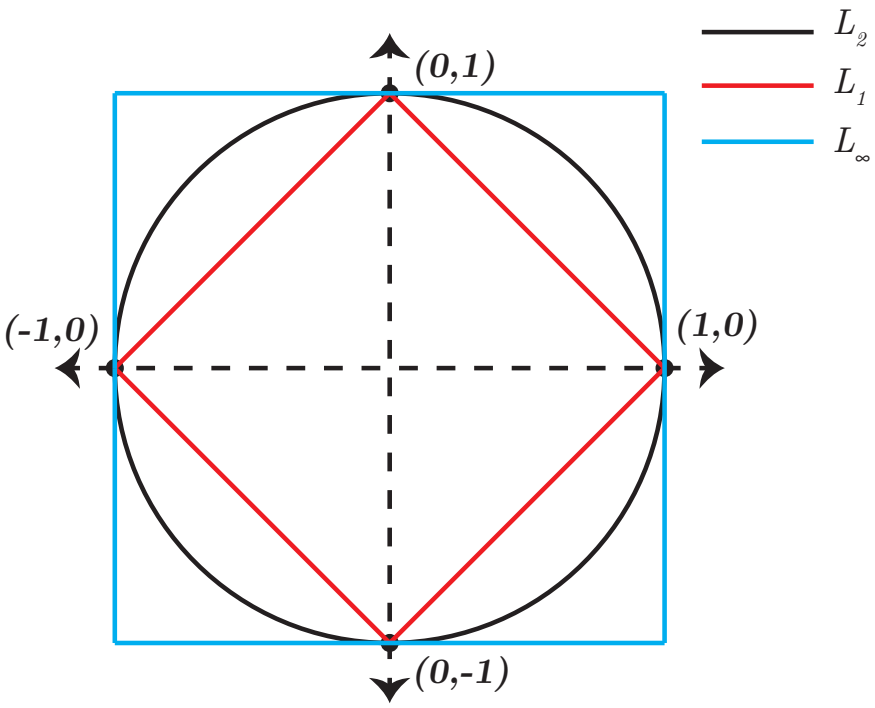


Figure 1

Problem 2

For this question we will subdivide it into the different distance categories, and will analyze the different norms for each of the distance definitions.

(1)

$$d_{\min}(G, G') = \min_{\mathbf{x} \in G, \mathbf{x}' \in G'} d(\mathbf{x}, \mathbf{x}')$$

Using the ℓ_1 norm $d(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_1$:
Picking the visually shortest points between the clusters we can boil down to options between black-red or black-blue, and particularly between points x2-x3 or x1-x4, respectively. To do this analysis we compared their norm distances to be:

X_a	X_b	$Dist$
1.	2.	0.5
1.	3.	1.03
1.	4.	0.61
2.	3.	0.67
2.	4.	0.61
3.	4.	0.62

$\boxed{Black - blue}$

Using the ℓ_2 norm $d(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_2$:

X_a	X_b	$Dist$
1.	2.	0.35355339
1.	3.	0.86884981
1.	4.	0.51971146
2.	3.	0.60406953
2.	4.	0.43600459
3.	4.	0.44045431

$\boxed{Black - blue}$

Using the Chebyshev norm (ℓ_∞) $d(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_\infty$:

X_a	X_b	$Dist$
1.	2.	0.25
1.	3.	0.85
1.	4.	0.51
2.	3.	0.6
2.	4.	0.35
3.	4.	0.34

$\boxed{Red - blue}$

(2)

$$d_{\max}(G, G') = \max_{\mathbf{x} \in G, \mathbf{x}' \in G'} d(\mathbf{x}, \mathbf{x}')$$

Using the ℓ_1 norm $d(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_1$:

X_a	X_b	$Dist$
1.	2.	0.5
1.	3.	1.03
1.	4.	0.61
2.	3.	0.67
2.	4.	0.61
3.	4.	0.62

$\boxed{Black - blue}$

Using the ℓ_2 norm $d(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_2$:

X_a	X_b	$Dist$
1.	2.	0.35355339
1.	3.	0.86884981
1.	4.	0.51971146
2.	3.	0.60406953
2.	4.	0.43600459
3.	4.	0.44045431

$\boxed{Red - blue}$

Using the Chebyshev norm (ℓ_∞) $d(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_\infty$:

X_a	X_b	$Dist$
1.	2.	0.25
1.	3.	0.85
1.	4.	0.51
2.	3.	0.6
2.	4.	0.35
3.	4.	0.34

$\boxed{Red - blue}$

(3)

$$d_{\text{centroid}}(G, G') = d(\boldsymbol{\mu}_G, \boldsymbol{\mu}_{G'})$$

Using the ℓ_1 norm $d(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_1$:

X_a	X_b	$Dist$
1.	2.	0.78
1.	3.	0.61
2.	3.	0.62

$\boxed{Black - blue}$

Using the ℓ_2 norm $d(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_2$:

$$\begin{bmatrix} X_a & X_b & Dist \\ 1. & 2. & 0.72708321 \\ 1. & 3. & 0.445926 \\ 2. & 3. & 0.44045431 \end{bmatrix}$$

$\boxed{Red - blue}$

Using the Chebyshev norm (ℓ_∞) $d(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_\infty$:

$$\begin{bmatrix} X_a & X_b & Dist \\ 1. & 2. & 0.725 \\ 1. & 3. & 0.385 \\ 2. & 3. & 0.34 \end{bmatrix}$$

$\boxed{Red - blue}$

(4)

$$d_{\text{avg}}(G, G') = \frac{1}{|G||G'|} \sum_{\mathbf{x} \in G} \sum_{\mathbf{x}' \in G'} d(\mathbf{x}, \mathbf{x}')$$

Using the ℓ_1 norm $d(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_1$:

$$\begin{bmatrix} X_a & X_b & Dist \\ 1. & 2. & 0.85 \\ 1. & 3. & 0.61 \\ 2. & 3. & 0.62 \end{bmatrix}$$

$\boxed{Black - blue}$

Using the ℓ_2 norm $d(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_2$:

$$\begin{bmatrix} X_a & X_b & Dist \\ 1. & 2. & 0.73645967 \\ 1. & 3. & 0.47785802 \\ 2. & 3. & 0.44045431 \end{bmatrix}$$

$\boxed{Red - blue}$

Using the Chebyshev norm (ℓ_∞) $d(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_\infty$:

$$\begin{bmatrix} X_a & X_b & Dist \\ 1. & 2. & 0.725 \\ 1. & 3. & 0.43 \\ 2. & 3. & 0.34 \end{bmatrix}$$

$\boxed{Red - blue}$

Problem 3

Topic Modeling [15 pts]

In this problem we will explore a simplified version of topic modeling in which each document has just a *single* topic. For this problem, we will assume there are c topics. Each topic $k \in \{1, \ldots, c\}$ will be associated with a vector $\boldsymbol{\beta}_k \in [0,1]^{|\mathcal{V}|}$ describing a distribution over the vocabulary $\mathcal{V}$ with $\sum_{j=1}^{|\mathcal{V}|} \beta_{k,j} = 1$.

Each document $\mathbf{x}_i$ will represented as a bag-of-words $\mathbf{x}_i \in (\mathbb{Z}^{\geq 0})^{|\mathcal{V}|}$, where $x_{i,j}$ is a non-negative integer representing the number of times word j appeared in document i. Document i has n_i word tokens in total. Finally let the (unknown) overall mixing proportion of topics be $\boldsymbol{\theta} \in [0,1]^c$, where $\sum_{k=1}^{c} \theta_k = 1$.

Our generative model is that each of the n documents has a single topic. We encode this topic as a one-hot vector $\mathbf{z}_i \in \{0,1\}^c$ over topics. This one-hot vector is drawn from $\boldsymbol{\theta}$; then, each of the words is drawn from $\boldsymbol{\beta}_{z_i}$. Formally documents are generated in two steps:

$$\begin{aligned} \mathbf{z}_i &\sim \text{Categorical}(\boldsymbol{\theta}) \\ \mathbf{x}_i &\sim \text{Multinomial}(\boldsymbol{\beta}_{\mathbf{z}_i}) \end{aligned}$$

Problem 2

1. Draw the graphical model representation of this problem. Be sure to use the plate notation to indicate repeated random variables and gray nodes to indicate observed variables.
2. **Complete-Data Log Likelihood** Define the complete data for this problem to be $D = \{(\mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^n$.
   - Write out the complete-data (negative) log likelihood.

     $$\mathcal{L}(\boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^c) = -\ln p(D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^c).$$

   - Explain in one sentence why we cannot directly optimize this loss function.
3. **Expectation Step** Our next step is to introduce a mathematical expression for $\mathbf{q}_i$, the posterior over the hidden topic variables $\mathbf{z}_i$ conditioned on the observed data $\mathbf{x}_i$ with fixed parameters, i.e $p(\mathbf{z}_i | \mathbf{x}_i; \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^c)$.
   - Write down and simplify the expression for $\mathbf{q}_i$.
   - Give an algorithm for calculating $\mathbf{q}_i$ for all i, given the observed data $\{\mathbf{x}_i\}_{i=1}^n$ and settings of the parameters $\boldsymbol{\theta}$ and $\{\boldsymbol{\beta}_k\}_{k=1}^c$.
4. **Maximization Step** Using the $\mathbf{q}_i$ estimates from the Expectation Step, derive an update for maximizing the expected complete data log likelihood in terms of $\boldsymbol{\theta}$ and $\{\boldsymbol{\beta}_k\}_{k=1}^c$.
   - Derive an expression for the expected complete-data log likelihood in terms of $\mathbf{q}_i$.
   - Find an expression for $\boldsymbol{\theta}$ that maximizes this expected complete-data log likelihood. You may find it helpful to use Lagrange multipliers in order to force the constraint $\sum \theta_k = 1$. Why does this optimized $\boldsymbol{\theta}$ make intuitive sense?
   - Apply a similar argument to find the value of the $\boldsymbol{\beta}_k$'s that maximizes the expected complete-data log likelihood.

Solution

Problem 1

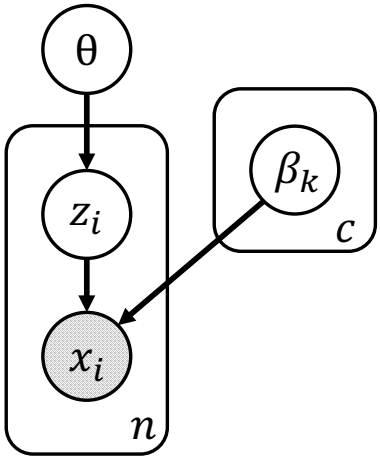


Figure 2: Graphical Representation of Problem 2

Problem 2

The complete-data (negative) log likelihood is given as:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^c) &= -\ln p(D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^c) \\
&= -\ln \prod_{i=1}^{n} p(\{\mathbf{x}_i, \mathbf{z}_i\} | \theta, \{\boldsymbol{\beta}_k\}_{k=1}^c) \\
&= -\ln \prod_{i=1}^{n} p(\mathbf{x}_i | \mathbf{z}_i; \theta, \{\boldsymbol{\beta}_k\}_{k=1}^c)) p(\mathbf{z}_i; \theta, \{\boldsymbol{\beta}_k\}_{k=1}^c)) \\
&= -\ln \prod_{i=1}^{n} \left(\prod_{j=1}^{|\mathcal{V}|} \prod_{k=1}^{c} \beta_{k,j}^{x_{ij} z_{ik}} + \prod_{k=1}^{c} \theta_k^{z_{ik}} \right) \\
&= -\sum_{i=1}^{n} \left(\sum_{j=1}^{|\mathcal{V}|} \sum_{k=1}^{c} \ln \beta_{k,j}^{x_{ij} z_{ik}} + \sum_{k=1}^{c} \ln \theta_k^{z_{ik}} \right) \\
&= -\sum_{i=1}^{n} \left(\sum_{j=1}^{|\mathcal{V}|} \sum_{k=1}^{c} x_{ij} z_{ik} \ln \beta_{k,j} + \sum_{k=1}^{c} z_{ik} \ln \theta_k \right) \\
&\boxed{= -\sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \left(\sum_{j=1}^{|\mathcal{V}|} x_{ij} \ln \beta_{k,j} + \ln \theta_k \right)}
\end{aligned}
$$

The reason why we cannot directly optimize this loss function is we don't know the latent variable z_i.

Problem 3

For this problem we want to know the posterior $\mathbf{q}_i$ which is given as

$$\mathbf{q}_i = p(\mathbf{z}_i|\mathbf{x}_i; \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^c).$$

However, it may be more useful to work with each individual element of $\mathbf{q}_i$ as q_{ik}. To do this, we can take the solution from above and normalize it by the sum of all of the values such that:

$$q_{ik} = \frac{\theta_k \prod_{j=1}^{|\mathcal{V}|} \beta_{k,j}^{x_{ij}}}{\sum_{k=1}^c \theta_k \prod_{j=1}^{|\mathcal{V}|} \beta_{k,j}^{x_{ij}}} \quad \text{where } \mathbf{q}_i = q_{ik} \forall k \in [1, c]$$

Because each value q_{ik} depends on every other value of q, that means we cannot compute this naively as is. To avoid this issue, I would recommend an algorithm that solves for all values of q_{ik} without the denominator, then computes the actual sum and divides all of the values by that sum. In other words:

1. Compute the numerator for all q_{ik}
2. Compute the sum of the denominator with the values from the previous step
3. Divide the numerator for all q_{ik} by the denominator sum computed in the previous step.

Problem 4

We know that for this problem we cannot solve directly for $\mathbf{z}_i$ in the complete-data log likelihood, thus we must use our expectation step to infer the $\mathbf{z}_i$ as $\mathbf{q}_i$.

$$\mathcal{L}(\boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^c) = -\sum_{i=1}^n \sum_{k=1}^c z_{ik} \left(\sum_{j=1}^{|\mathcal{V}|} x_{ij} \ln \beta_{k,j} + \ln \theta_k \right)$$

$$\sum_{i=1}^n \sum_{k=1}^c q_{ik} \left(\sum_{j=1}^{|\mathcal{V}|} x_{ij} \ln \beta_{k,j} + \ln \theta_k \right)$$

K-Means [15 pts]

For this problem you will implement K-Means clustering from scratch. Using `numpy` is fine, but don't use a third-party machine learning implementation like `scikit-learn`. You will then apply this approach to clustering of image data.

We have provided you with the MNIST dataset, a collection of handwritten digits used as a benchmark of image recogntion (you can learn more about the data set at http://yann.lecun.com/exdb/mnist/). The MNIST task is widely used in supervised learning, and modern algorithms with neural networks do very well on this task.

Here we will use MNIST unsupervised learning. You have been given representations of 6000 MNIST images, each of which are 28×28 greyscale handwritten digits. Your job is to implement K-means clustering on MNIST, and to test whether this relatively simple algorithm can cluster similar-looking images together.

Problem 3

The given code loads the images into your environment as a 6000x28x28 array.

- Implement K-means clustering from different random initializations and for several values of K using the ℓ_2 norm as your distance metric. (You should feel free to explore other metrics than the ℓ_2 norm, but this is strictly optional.) Compare the K-means objective for different values of K and across random initializations.
- For three different values of K, and a couple of random restarts for each, show the mean images for each cluster (i.e., for the cluster prototypes), as well as the images for a few representative images for each cluster. You should explain how you selected these representative images. To render an image, use the pyplot `imshow` function.
- Are the results wildly different for different restarts and/or different values of K? For one of your runs, plot the K-means objective function as a function of iteration and verify that it never increases.

As in past problem sets, please include your plots in this document. (There may be several plots for this problem, so feel free to take up multiple pages.)

Solution

Problem 4 (Calibration, 1pt)
Approximately how long did this homework take you to complete?

~ 10 hours

Problem 3 Code – K-Means Clustering

```
# CS 181, Spring 2017
# Homework 4: Clustering
# Name:
# Email:

import numpy as np
import matplotlib.pyplot as plt
import matplotlib.image as mpimg

class KMeans(object):
  # K is the K in KMeans
        def __init__(self, K):
    self.K = K

  # X is a (N x 28 x 28) array where 28x28 is the dimensions of each of the N images.
  def fit(self, X):
    pass

  # This should return the arrays for K images. Each image should represent the mean of each of the fitted clusters.
  def get_mean_images(self):
    pass

  # This should return the arrays for D images from each cluster that are representative of the clusters.
  def get_representative_images(self, D):
    pass

  # img_array should be a 2D (square) numpy array.
  # Note, you are welcome to change this function (including its arguments and return values) to suit your needs.
  # However, we do ask that any images in your writeup be grayscale images, just as in this example.
  def create_image_from_array(self, img_array):
    plt.figure()
    plt.imshow(img_array, cmap='Greys_r')
    plt.show()
    return

# This line loads the images for you. Don't change it!
pics = np.load("images.npy", allow_pickle=False)

# You are welcome to change anything below this line. This is just an example of how your code may look.
# That being said, keep in mind that you should not change the constructor for the KMeans class,
# though you may add more public methods for things like the visualization if you want.
# Also, you must cluster all of the images in the provided dataset, so your code should be fast enough to do that.
K = 10
KMeansClassifier = KMeans(K=10, useKMeansPP=False)
KMeansClassifier.fit(pics)
KMeansClassifier.create_image_from_array(pics[1])
```

Listing 1: Code for problem 3