Statistical Machine Learning
Statistics GR 5241 — Spring 2022

Homework 3

Collaboration policy: Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 3 (a)"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only.
The following problems are due on Friday, March 25nd, 11:59pm.

1. **Hierarchical clustering.**

   (10 points) Perform single-linkage hierarchical clustering on the following 2-dimensional observations. Use the Manhattan distance between a pair of points: $d(A, B) = |X_{1A} - X_{1B}| + |X_{2A} - X_{2B}|$. Show your results after each step.

Obs.	X_1	X_2
A	1	4
B	1	3
C	3	4
D	5	2
E	3	2
F	3	0

2. **More about AdaBoost.**

Consider the binary-class classification task. Suppose we are given a dataset with m samples, $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$, where for each i, $y_i \in \{-1, +1\}$. The loss we are using here is the $0-1$ loss: $L = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(H(x_i) \neq y_i)$.

Recall that in class we have gone through the Boosting algorithm for this binary-class classification task, which follows the following procedure:

Algorithm 1: AdaBoost algorithms

while L *is minimized* **do**

initialize $D_1(i) = \frac{1}{m}$ for $i = 1, \ldots, m$;

for *t=1, ...,* T **do**

Using D_t, get weak learner h_t, where h_t can only take value -1 or $+1$;

Choose $\alpha_t \in \mathbb{R}$, update using the rule:

$$Z_t = \sum_{i=1}^{m} D_t(x_i) \exp\left(-\alpha_t y_i h_t(x_i)\right), \quad D_{t+1}(i) = \frac{D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t}$$

end

Final classifier

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x), \quad H(x) = \operatorname{sign}\left(f(x)\right),$$

and loss

$$L = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(H(x_i) \neq y_i).$$

end

In this question, we are going to figure out the value of α_t for $t = 1, \ldots, T$.

(a) (5 points) Find the minimizer $\alpha \in \mathbb{R}$ that minimizes $Z = (1-\epsilon)e^{-\alpha} + \epsilon e^{\alpha}$.

(b) (5 points) Show that for all values of x and either $y = -1$ or $y = +1$,

$$\mathbb{1}(H(x) \neq y) \leq \exp\left(-y f(x)\right) = \exp\left(-y\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

(c) (10 points) Show by induction that

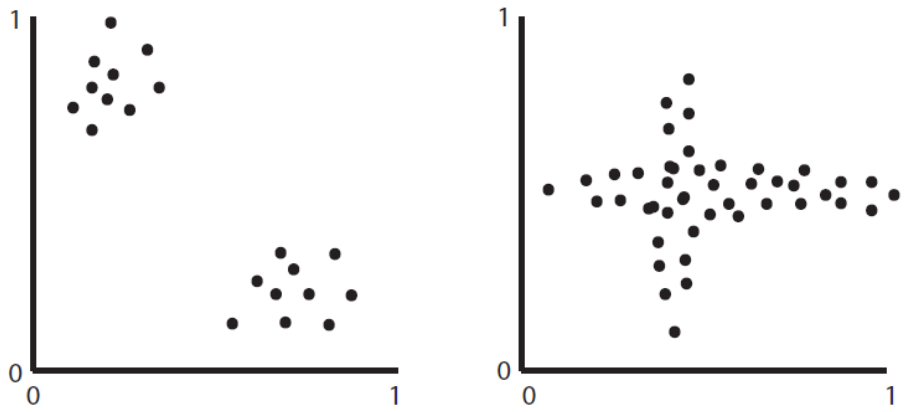
$$\frac{1}{m}\sum_{i=1}^{m} \exp\left(-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)\right) = \Pi_{t=1}^{T} Z_t.$$

Then together with part (a), conclude that choosing $\alpha_t = \frac{1}{2}\ln\frac{1-\epsilon_t}{\epsilon_t}$ can minimize the upper bound of loss L, where

$$\epsilon_t = \sum_{i=1}^{m} D_t(i)\mathbb{1}(h_t(x_i) \neq y_i).$$

3. **PCA : Principal Component Analysis**

(a) (5 points) For each of the two data sets shown below, draw vectors that correspond to the first and second principal components that would be produced by Principal Components Analysis (PCA). For each diagram, label which vector is the first component, and which is the second.



(b) Let $3 \times N$ matrix X be a dataset with 3 variables and N samples.
Now consider drawing the columns of $X = [X_1, \ldots, X_N]$ from the given multivariate Gaussian distribution:

$$X_n \sim \mathcal{N}\left(\mu = \begin{bmatrix} 5 \\ 5 \\ 10 \end{bmatrix}, \Sigma = \begin{bmatrix} 6 & -3 & 0 \\ -3 & 6 & 0 \\ 0 & 0 & 10 \end{bmatrix}\right)$$

- (5 points) What are the ordered variances $\sigma_1, \sigma_2, \sigma_3$ (high to low) for the ordered principal components?
- (5 points) What are the ordered basis vectors $u_1, u_2, u_3 \in R^3$ for the ordered principal components? (You can use any software for calculating this).

4. **Mixture models.**

We have introduced the EM-algorithm for mixture models in class. In this problem, we will apply the Gibbs sampler to the same problem. Our goal is to find the clusters in our data $\mathcal{X} = \{x_1, \cdots, x_n\}$ with $x_i \in \mathbb{R}$. We use the mixture of three Gaussians to model the population distribution:

$$p(x) = \sum_{k=1}^{3} c_k \phi(x; \mu_k, \sigma^2), \tag{1}$$

where $\phi(\cdot; \mu, \sigma^2)$ is the one-dimensional Gaussian density function with KNOWN variance σ^2 and takes the form

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right),$$

For each data point x_i, we introduce a latent variable Z_i which indicates the component x_i belongs to. This corresponds to the hard assignment in the EM-algorithm. Z_i is a discrete random variable, and can take values from the set $\{1, 2, 3\}$. Now, we consider an approximation to model (1):

$$p(x_i) = \sum_{k=1}^{3} \mathbb{1}_{\{Z_i=k\}} \phi(x_i; \mu_k, \sigma^2), \tag{2}$$

where $\mathbb{1}_{\{Z=k\}}$ is the indicator function taking value 1 when $Z = k$ and being 0 otherwise.

Under the Bayesian framework, the parameters and latent variables of model (2) are μ_k for $k = 1, 2, 3$, and Z_i for $i = 1, \cdots, n$. We impose the following prior on the parameters:

$$\mu_k \sim \mathcal{N}(0, \tau^2) \quad \text{for } k = 1, 2, 3,$$

and the prior on Z_i are independent and uniform across the three components, namely $\pi(Z_i = k) = 1/3$ for $k = 1, 2, 3$.

Now we use the Gibbs sampler to sample from the posterior $p(\mu_1, \mu_2, \mu_3, Z_1, \cdots, Z_n | \mathcal{X})$.

(a) (5 points) Given all the parameters and latent variables, what is the likelihood of our data $\mathcal{X}$.

(b) (10 points) Now we consider the posterior distribution. What is the conditional distribution of each μ_k given the other parameters and latent variables? Does it have a closed-form expression? Are the conditional distributions of μ_1, μ_2, and μ_3 independent given Z_i's?

(c) (10 points) What is the conditional distribution of each Z_i given μ_1, μ_2, μ_3 and the other latent variables? Are they independent given the means (μ_1, μ_2, μ_3) of the three components?

5. **Clustering.**

In this problem, there are K components $\{C_1, C_2, .., C_K\}$ and each component is generated from a normal distribution $\sim \mathcal{N}(\mu_i, \Sigma_i)$. Thus each data point is generated as follows :

- Choose component i with a probability $\pi_i = P(y = i)$
- Further each datapoint can be generated as $x \sim \mathcal{N}(\mu_i, \Sigma_i)$

Now we can write the following,

$$p(x|y = i) = \mathcal{N}(\mu_i, \Sigma_i), \quad p(x) = \sum_{i=1}^{K} (p(x|y = i)P(y = i)$$

At test time, we wish to assign a cluster to each x, we can do so by finding the component that yields the maximum probability for the test point x

$$\text{argmax}_k P(y = k|x)$$

We are given N examples, each example consists of a d dimensional vector, more formally data $D = \{x_1, x_2, x_3...x_N\}$ and each $x_i \in \mathbb{R}^d$. The task is to assign each example into one of the K clusters using the above formulation.
The parameters for the above formulation are $\theta = (\pi_1, \pi_2....\pi_K, \mu_1, \mu_2....\mu_K, \Sigma_1, \Sigma_2....\Sigma_K)$, hence the problem reduces to estimating these parameters.

We will use a synthetic dataset , where each row is a unique vector of 5 dimensions. There are 3000 unique examples. Hence $N = 3000$, $d = 5$, and we want to group these points into three clusters ($K = 3$). They are generated as per the generation process outlined above. To visualize the dataset, one can plot the first 2 dimensions (i.e x_{i1}, x_{i2}) of each i from 1 to 3000 (see Figure 2).

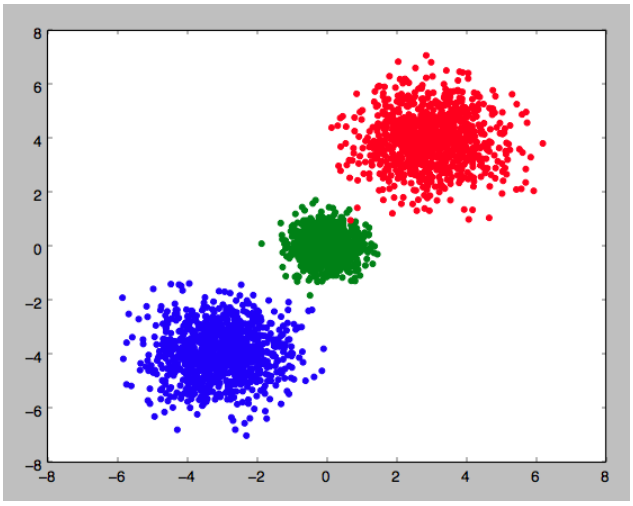


Figure 2: The first two dimensions of the dataset, different colors denote different clusters

For consistency, we would denote the leftmost cluster (i.e one with the least μ_{i1}) as Blue cluster (denoted by blue color), the middle cluster as Green cluster (denoted by green color), and the rightmost (the one with the largest μ_{i1}) cluster is Red cluster (denoted by red color). Figure 2 also adheres to this convention. Now, please answer the following questions :

- (4 points) Assuming that $\Sigma_i = \sigma_i^2 I$, compute and write below the means μ_1, μ_2, μ_3 and the standard deviations $\sigma_1, \sigma_2, \sigma_3$. We highly recommend that you use the GaussianMixture model from sklearn. (Please use random initialization for this part, and a default convergence threshold/tolerance of 0.001)
- (4 points) How many iterations did it take to converge with a random initialization versus pre initialization from K-Means?
- (6 points) From these computed means and std. deviations, cluster all the 3000 points, and plot three figures
  - Plot the first two dimensions (x_{i1}, x_{i2}) with their cluster assignments. (Similar to figure 2)
  - Plot the third and fourth dimensions (x_{i3}, x_{i4}) with their cluster assignments.
  - Plot the fourth and fifth dimensions (x_{i4}, x_{i5}) with their cluster assignments.

  Are the computed means consistent with the graphs?
- (10 points) For this subquestion, you are not permitted to use any library function that performs the EM algorithm, and you are instead supposed to write your own EM algorithm to estimate the means μ_1, μ_2, μ_3, however, to simplify things you can assume that values of π_1, π_2, π_3 and $\sigma_1, \sigma_2, \sigma_3$ are known and reuse the values computed from the first part. (Please use convergence threshold/tolerance of 0.001)
- (6 points) We relax the assumption that π_1, π_2, π_3 are known, now your EM algorithm is supposed to simultaneously estimate means μ_1, μ_2, μ_3 and the component probabilities π_1, π_2, π_3. You can still assume that σ_1, σ_2 and σ_2 are known and use their values from the solution above. (Please use convergence threshold/tolerance of 0.001) [For this part too, you are not permitted to use library functions that perform EM algorithm]