CIS5200: Machine Learning Spring 2023

Homework 2

Release Date: February 6, 2023 *Due Date: February 17, 2023*

- HW2 will count for 10% of the grade. This grade will be split between the written (40 points) and programming (40 points) parts.
- All written homework solutions are required to be formatted using LaTeX. Please use the template here. Do not modify the template. This is a good resource to get yourself more familiar with LaTeX, if you are still not comfortable.
- You will submit your solution for the written part of HW2 as a single PDF file via Gradescope. The deadline is **11:59 PM ET**. Contact TAs on Ed if you face any issues uploading your homeworks.
- Collaboration is permitted and encouraged for this homework, though each student must understand, write, and hand in their own submission. In particular, it is acceptable for students to discuss problems with each other; it is not acceptable for students to look at another student's written Solutions when writing their own. It is also not acceptable to publicly post your (partial) solution on Ed, but you are encouraged to ask public questions on Ed. If you choose to collaborate, you must indicate on each homework with whom you collaborated.
- **Bonus Questions**: We have added two bonus questions in this homework for extra credit. These are intended to be more challenging than the non-bonus homework questions.

Please refer to the notes and slides posted on the website if you need to recall the material discussed in the lectures.

Note: Corrections and clarifications appear in red.

Survey

The course feedback survey is worth up to 2 bonus points. Everyone will receive either 0, 1, or 2 points based on how many responses are received.

1 Written Questions (40 points + 8 bonus points)

Problem 1: k-means Clustering (15 points + 6 bonus points)

Recall the k-means clustering problem with data $x_1, \ldots, x_m$. Our goal was to find k clusters denoted by $C_1, \ldots, C_k \subseteq [m]$ such that $\cup_{i=1}^{k} C_i = [m]$ (cover all data points) and for all $i \neq j$, $C_i \cap C_j = \phi$ (and are disjoint).

1.1 (4 points) To measure the "goodness" of the cluster we defined

$$Z(C_1, \ldots, C_k) = \sum_{l=1}^{k} \frac{1}{2|C_l|} \sum_{i,j \in C_l} \|x_i - x_j\|_2^2.$$

Show that we can equivalently express Z as

$$Z(C_1, \ldots, C_k) = \sum_{l=1}^{k} \sum_{i \in C_l} \|x_i - \mu_l\|_2^2$$

where $\mu_l = \frac{1}{|C_l|} \sum_{i \in C_l} x_i$ is the centroid of cluster l.

Solution:

1.2 (3 points) In Lloyd's algorithm, for every cluster C, we choose the cluster center to be the centroid μ where $\mu = \frac{1}{|C|} \sum_{i \in C} x_i$. However, there may be other ways of choosing the cluster centers for k-means, for a cluster C. Let $Z(C, z) = \sum_{i \in C} \|x_i - z\|_2^2$ be the "goodness" of cluster C with center z. Prove that μ is the optimal center for the cluster C, that is,

$$Z(C, \mu) = \min_{z} Z(C, z).$$

Solution:

1.3 (3 points) Suppose we pick the center of the cluster z uniformly randomly from the points in cluster C, such that, $\Pr[z = x_i] = \frac{1}{|C|}$ for each $i \in C$. Let us denote this distribution over centers as ρ. Show that we do not lose too much by choosing a random cluster center according to ρ (similar to what k-means++ does) compared to choosing the centroid. In particular, prove that

$$\mathbb{E}_{z \sim \rho}[Z(C, z)] = 2Z(C, \mu).$$

Solution:

1.4 (5 points) Show that that the EM algorithm to solve Gaussian Mixture Models reduces to k-means if we set $\Sigma_l = \sigma^2 I$ for all $l \in [k]$ and take the limit $\sigma \to 0$.

Solution:

Bonus (6 points) The Lloyd's algorithm we discussed in class does not always find the optimal k-means solution, and it is hard in general to find the optimal solution efficiently. However, when the input is in 1-dimensions, the problem can be solved optimally and efficiently. Design an $O(km^2)$ dynamic programming algorithm for solving the k-means problem in single dimension.

Hint: If we sort the data points $x_1 \leq x_2 \leq \ldots \leq x_m$ in increasing order, then the optimal clusters correspond to intervals of the points, that is, cluster contains all points between some index i and j, $\{x_i, x_{i+1}, \ldots, x_j\}$.

Solution:

Problem 2: PCA (15 points)

Consider data points $x_1, \ldots, x_m$ such that each feature has been normalized to have mean 0 and variance 1 (as discussed in class). One way to look at PCA is to take the maximizing variance point of view, as we did in class. Here we showed that to find the best direction to project the data into is given by finding u such that it

$$\max_{\|u\|_2=1} \frac{1}{m} \sum_{i=1}^{m} (x_i^\top u)^2 = u^\top \hat{\Sigma} u$$

where $\hat{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} x_i x_i^\top$ is the empirical covariance matrix.

Hint: Use the eigenvalue decomposition for $\hat{\Sigma}$.

2.1 (3 points) Show that

$$\max_{u:\|u\|_2=1} \frac{1}{m} u^\top \hat{\Sigma} u = \lambda_1$$

where λ_1 is the principle eigenvalue (largest eigenvalue) of $\hat{\Sigma}$.

Solution:

2.2 (5 points) Consider another alternate view of finding a useful direction u by instead minimizing the following reconstruction error

$$\min_{u:\|u\|_2=1} \frac{1}{m} \sum_{i=1}^{m} \|x_i - (u^\top x_i) u\|_2^2.$$

Show that this is equivalent to maximizing variance, that is

$$\arg\max_{u:\|u\|_2=1} \frac{1}{m} \sum_{i=1}^{m} (x_i^\top u)^2 = \arg\min_{u:\|u\|_2=1} \frac{1}{m} \sum_{i=1}^{m} \|x_i - (u^\top x_i) u\|_2^2.$$

Solution:

2.3 (7 points) Let us generalize this to k useful directions. In particular, show that the following two optimizations are equivalent (they have the same optimal solution $U^\star$). Also find the optimal solution $U^\star$.

$$\max_{U \in \mathbb{R}^{k \times d}: UU^\top = I} \frac{1}{m} \sum_{i=1}^{m} \|Ux_i\|_2^2 \qquad (\textit{maximize variance})$$

$$\min_{U \in \mathbb{R}^{k \times d}: UU^\top = I} \frac{1}{m} \sum_{i=1}^{m} \|x_i - U^\top U x_i\|_2^2 \qquad (\textit{minimize reconstruction error})$$

Solution:

Problem 3: Duality (10 points + 2 bonus points)

3.1 (4 points) Consider the following optimization problem:

$$\begin{aligned} \text{minimize over } w \quad & J(w) \\ \text{such that} \quad & c_i(w) \leq 0, i \in [k] \\ & e_j(w) = 0, j \in [l] \end{aligned}$$

Consider the dual problem

$$D(\alpha, \beta) = \min_w \mathcal{L}(w, \alpha, \beta) := \min_w \left(J(w) + \sum_{i=1}^{k} \alpha_i c_i(w) + \sum_{j=1}^{l} \beta_j e_j(w) \right)$$

for $\alpha_i \geq 0$ and $\beta_j \in \mathbb{R}$.

Show that for all feasible α, β, w,

$$D(\alpha, \beta) \leq J(w).$$

This property is known as *weak duality*.

Solution:

3.2 (6 points) Consider the following optimization problem:

$$\begin{aligned} \text{minimize over } w \quad & c^\top w \\ \text{such that} \quad & Aw = b \\ & w \geq 0 \end{aligned}$$

Show that the dual of this problem is

$$\begin{aligned} \text{maximize over } \beta \quad & \beta^\top b \\ \text{such that} \quad & A^\top \beta \leq c \end{aligned}$$

Recall that the dual is $\max_{\alpha \geq 0, \beta} D(\alpha, \beta)$ where $D(\alpha, \beta) = \min_w \mathcal{L}(w, \alpha, \beta)$.

Hint: For Linear optimization problems the optimal α^*, β^*, w^* satisfy all the KKT conditions. Use these to derive the dual form.

Solution:

Bonus (2 points) Express the following problem as a linear program:

$$\begin{aligned} \text{minimize over } w \quad & \|w\|_1 \\ \text{such that} \quad & Aw = b \end{aligned}$$

Solution:

2 Programming Questions (40 points)

Use the link here to access the Google Colaboratory (Colab) file for this homework. Be sure to make a copy by going to "File", and "Save a copy in Drive". As with the previous homeworks, this assignment uses the PennGrader system for students to receive immediate feedback. As noted on the notebook, please be sure to change the student ID from the default '99999999' to your 8-digit PennID.

Instructions for how to submit the programming component of HW 2 to Gradescope are included in the Colab notebook. You may find this PyTorch linear algebra reference and this general PyTorch reference to be helpful in perusing the documentation and finding useful functions for your implementation.