CIS 4190/5190: Applied Machine Learning — Fall 2023

Homework 4

Handed Out: October 11 — *Due: October 25, 8:00 p.m.*

- You are encouraged to format your solutions using LaTeX. You'll find some pointers to resources for learning LaTeX among the Canvas primers. Handwritten solutions are permitted, but remember that you bear the risk that we may not be able to read your work and grade it properly — do not count on providing post hoc explanations for illegible work. You will submit your solution manuscript for written HW3 as a single PDF file.
- The homework is **due at 8:00 PM** on the due date. We will be using Gradescope for collecting the homework assignments. Please submit your solution manuscript as a PDF file via Gradescope. Post on Ed Discussion and contact the TAs if you are having technical difficulties in submitting the assignment.

1 Multiple Choice & Written Questions

Note: You do not need to show work for multiple choice questions. If formatting your answer in LaTeX, use our LaTeX template `hw4_template.tex` (This is a read-only link. You'll need to make a copy before you can edit. Make sure you make only private copies.).

1. [Bias-Variance Tradeoff] (6 pts) State if the following prompts are true or false with reasoning.
   (a) Random forest classifiers are developed from multiple decision trees. Since deep decision trees can suffer from high variance, random forests similarly have high variance.
   (b) In Adaboost, training more "weak classifier" models results in a higher variance in the resulting model.
   (c) Boosting methods outperform their component classifiers by lowering bias in the predictions.

2. [Ensemble/Adagrad] (10 pts) In this problem, we will try to understand the intuition behind the AdaBoost algorithm by walking through the rationale behind the choice of β_t at each time step t. Recall that at each time step t, the AdaBoost algorithm receives a weak learner h_t, which is trained on the dataset (X, y) with instance weights w_t. The weighted training error of h_t is ϵ_t.

$$\epsilon_t = \sum_{i:y_i \neq h_t(x_i)} w_{t,i} \tag{1}$$

We set β_t by

$$\beta_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t}) \tag{2}$$

Then we update the instance weights w_{t+1} for the next time step by

$$w_{t+1,i} = w_{t,i} \cdot e^{-\beta_t y_i h_t(x_i)} / Z_t \tag{3}$$

where Z_t is a normalization factor, so that w_{t+1} sums up to one.

$$Z_t = \sum_{i=1}^{N} w_{t,i} \cdot e^{-\beta_t y_i h_t(x_i)} \tag{4}$$

(a) (3 pts) Show that Z_t is equivalent to the exponential loss of round t. In other words, show that

$$Z_t = e^{\beta_t} \epsilon_t + e^{-\beta_t}(1 - \epsilon_t) \tag{5}$$

(b) (5 pts) Next, we'll try to prove that our choice of β_t minimizes the value of Z_t. The way to prove this is to start from Equation. 5. Treat Z_t as a function of β_t, and take the derivative of Z_t with respect to β_t. Setting the derivative to zero, and solve for β_t. Verify that the β_t you solved is the same as Equation. 2 above.

(c) (2 pts) With the above two questions, we show that AdaBoost is choosing β_t to greedily minimizing the (weighted) average exponential loss on training examples at every iteration t. Briefly in a few sentences, describe how β_t is used to compose the decisions from the weak learners of each iteration h_t into a final hypothesis.

3. [PCA] (14 pts) *Note: You are expected to work out the entire question by hand, and not use any libraries/packages. Plots can be drawn manually/through drawing tools, but the labels and lines in the plots have to figured out manually.*
Bob wants to transmit the following set of four two-dimensional coordinates to his friend.

$$X = \begin{bmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{bmatrix} \tag{6}$$

However, due to transmission bandwidth, he is constrained to send only four one-dimensional coordinates. He learns about the PCA algorithm and wants to apply it.

(a) (6pts) Find the unit-vector principal components of X. Given that Bob is constrained to send four 1-D coordinates, which principal component would you suggest him to pick, and why? Show your work.

(b) (4pts) Bob plots the coordinates of X as in Figure 1. To obtain the 1-D transformation of his 2-D coordinates, he sketches the direction of the principal component and projects the four 2-D coordinates on this principle component. Show how this plot would look like. Label each of the projected points along with the value of the principal coordinate (note that these labels should be 1-D points).

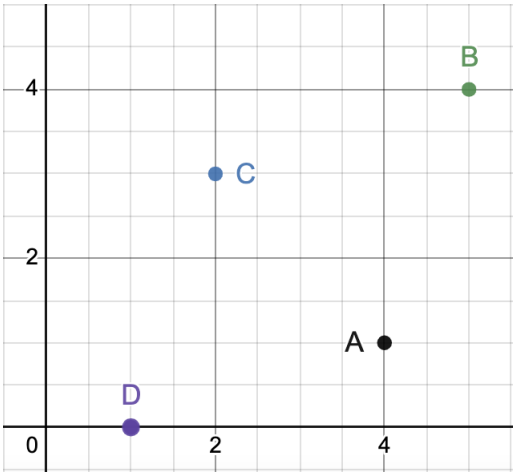

Figure 1: Bob's plot

(c) (4pts) Due to a mishap, all the coordinates in X get rotated by an angle of 30° anti-clockwise as in Figure 2. Bob applies PCA again on these points. Sketch the new direction of the principal component, project the four 2-D coordinates on this principal component and label the new principal coordinates. (note that these labels should be 1-D points)

Figure 2: Bob's plot after the mishap

Note: For (b) and (c), assume that origin in the original 2-D space i.e, $(0, 0)$ corresponds to zero in the new principal coordinate i.e, (0).

4. (8 pts) [k-Means] Work through the K-Means clustering algorithm for a dataset with 4 samples, with $K = 2$, and using the L_2 distance. The samples in the dataset are: $A = (2, 3), B = (4, 6), C = (5, 1)$, and $D = (10, 12)$. The initial centroids are chosen as: $(6, 9)$ for cluster 1 and $(8, 4)$ for cluster 2. Recall that in each iteration of K-Means, two

things happen: first, cluster assignments are updated, and second, cluster centroids are updated. Work through two such iterations. Report results for each iteration as:

- A: $d(A, 1)$, $d(A, 2)$
- B: $d(B, 1)$, $d(B, 2)$
- C: $d(C, 1)$, $d(C, 2)$
- D: $d(D, 1)$, $d(D, 2)$
- cluster 1 members: A, B, etc.
- cluster 1 updated centroid: (x, y)
- cluster 2 members: A, B, etc.
- cluster 2 updated centroid: (x, y)

where $d(S, c)$ is the L_2 distance from sample S to the cluster c centroid.