HOMEWORK 9: UNSUPERVISED LEARNING AND ENSEMBLE METHODS

10-301/10-601 Introduction to Machine Learning (Summer 2022)
https://www.cs.cmu.edu/~hchai2/courses/10601/
OUT: Wednesday, August 3
DUE: Tuesday, August 9 at 1:00 PM
TAs: Sana, Brendon, Neural, Ayush, Boyang (Jack), Chu

This is the final homework assignment. This assignment covers **Graphical Models**, **Reinforcement Learning**, **K-Means**, **PCA**, and **Ensemble Methods**.

START HERE: Instructions

- **Collaboration Policy**: Please read the collaboration policy here: https://www.cs.cmu.edu/~hchai2/courses/10601/#Syllabus
- **Late Submission Policy:** See the late submission policy here: https://www.cs.cmu.edu/~hchai2/courses/10601/#Syllabus
- **Submitting your work:** You will use Gradescope to submit answers to all questions and code. Please follow instructions at the end of this PDF to correctly submit all your code to Gradescope.
  - **Written:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, please use the provided template. Submissions can be handwritten onto the template, but should be labeled and clearly legible. If your writing is not legible, you will not be awarded marks. If your scanned submission misaligns the template, there will be a 5% penalty. Alternatively, submissions can be written in LaTeX. Each derivation/proof should be completed in the boxes provided. If you do not follow the template, your assignment may not be graded correctly by our AI-assisted grader.

Instructions for Specific Problem Types

For "Select One" questions, please fill in the appropriate bubble completely:

Select One: Who taught this course?

- ● Henry Chai
- ○ Marie Curie
- ○ Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

Select One: Who taught this course?

- ● Henry Chai
- ○ Marie Curie
- ⊗ Noam Chomsky

For "Select all that apply" questions, please fill in all appropriate squares completely:

Select all that apply: Which are scientists?

- ■ Stephen Hawking
- ■ Albert Einstein
- ■ Isaac Newton
- □ I don't know

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

Select all that apply: Which are scientists?

- ■ Stephen Hawking
- ■ Albert Einstein
- ■ Isaac Newton
- ⊠ I don't know

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

Fill in the blank: What is the course number?

10-601

10-6301

Written Questions (53 points)

1 LaTeX Bonus Point (1 points)

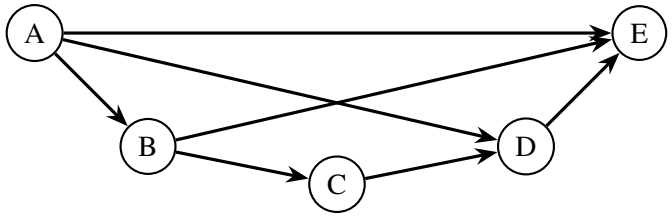
1. (1 point) **Select one:** Did you use LaTeX for the entire written portion of this homework?

   ○ Yes

   ● No

2 Graphical Models (4 points)

Consider the joint distribution over the binary random variables A, B, C, D, E represented by the Bayesian Network shown in the figure.



1. (1 point) Write the joint probability distribution for $P(A, B, C, D, E)$ factorized as much as possible according to the standard definition of a Bayesian Network using the conditional independence assumptions expressed by the above network.

   Your Answer

   $P(A)P(B|A)P(C|B)P(D|A,C)P(E|A,B,D)$

2. (1 point) Which nodes are in the Markov boundary of D? Note that the Markov boundary is the smallest possible Markov blanket.

   Your Answer

   A, B, C, E

3. (1 point) **True or False:** E is conditionally independent of C given $\{A, B, D\}$, i.e., $E \perp C \mid \{A, B, D\}$.

   ● True

   ○ False

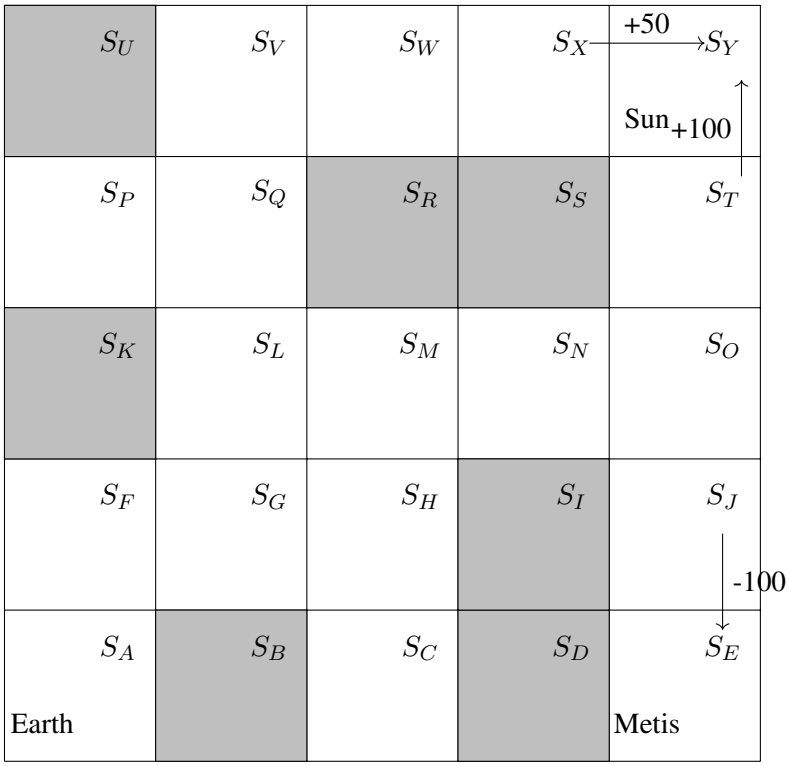
4. (1 point) How many parameters would we need to represent the joint distribution $P(A, B, C, D, E)$ **with** the conditional independence assumptions expressed by the Bayesian Network?

   Your Answer

   17

3 Reinforcement Learning, Revisited (9 points)

While attending the ML conference *The Fellowship of the Ring*, you meet Elon Musk, founder of SpaceX. He has a new idea for destroying the evil lord Sauron's precious ring: fly the ring directly into the Sun. Elon has asked you to develop a reinforcement learning agent capable of carrying out the space-flight from Earth to the Sun. You model this problem as a Markov decision process (MDP). The figure below depicts the state space.



Here are the details:

1. Each grid cell is a state $S_A, S_B, ..., S_Y$ corresponding to a position in the solar system.
2. The action space includes movement up/down/left/right. Transitions are deterministic. It is not possible to move into blocked states, which are shaded grey, since they contain other planets.
3. The start state is S_A (Earth). The terminal states include both the S_Y (Sun) and S_E (asteroid Metis, home to Sauron's cousin).
4. Non-zero rewards are depicted with arrows. Flying into the Sun from the left gives positive reward $R(S_X, \text{right}) = +50$. Flying into the Sun from below gives positive reward $R(S_T, \text{up}) = +100$. Flying to Metis is inadvisable and gives negative reward $R(S_J, \text{down}) = -100$. All other rewards are zero.
5. The discount factor is $\gamma = 0.5$.

Below, let $V^*(s)$ denote the value function for state s using the optimal policy $\pi^*(s)$. Let $Q^*(s, a)$ denote the Q function for π^*.

1. (1 point) What is the value $V^*(S_T)$?

Your Answer

100

2. (1 point) What is the value $V^*(S_O)$?

Your Answer

50

3. (1 point) What is the value $V^*(S_A)$?

Your Answer

0.78125

4. (1 point) What is the value $Q^*(S_T, \text{up})$?

Your Answer

100

5. (1 point) What is the value $Q^*(S_T, \text{down})$?

Your Answer

25

6. (1 point) What action does the optimal policy take from state S_Q (i.e. what is $\pi^*(S_Q)$)? (Note: If the optimal policy is not unique and there are multiple optimal actions, select them all.)

Select all that apply:

- [x] Up
- [ ] Down
- [ ] Left
- [ ] Right

Now suppose you employ Q-Learning to learn table values $Q(s, a)$ for each state s and action a. The table is initialized to all zeros. On the very first episode of training, you begin at state S_A (Earth), take eight steps, and arrive in state S_E (Metis). At each step, you perform a Q-Learning update of the appropriate entry in $Q(s, a)$. Assume a learning rate $\alpha = 1$.

7. (1 point) What is the new table value found in $Q(S_J, \text{down})$ after this episode?

Your Answer

100

8. (1 point) What is the new table value found in $Q(S_O, \text{down})$ after this episode?

Your Answer

0

9. (1 point) **True or False:** The Q-function is guaranteed to converge to the true Q-values in this environment given the specified initialization and assuming that the Q-Learning algorithm visits each state-action pair infinitely often.

● True

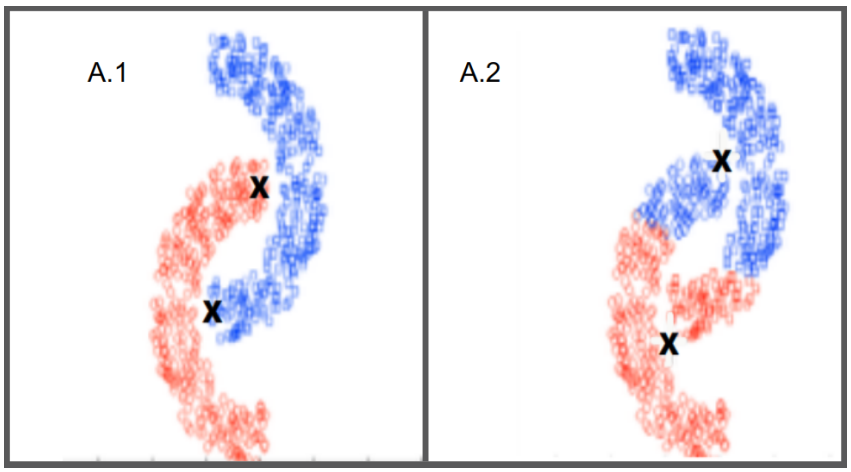
○ False

4 K-Means (11 points)

1. Consider the 3 datasets A, B and C. Each dataset is classified into k clusters, with centers marked X and cluster membership represented by different colors in the figure. For each dataset, exactly one clustering was generated by K-means with Euclidean distance. Select the image with clusters generated by K-means.

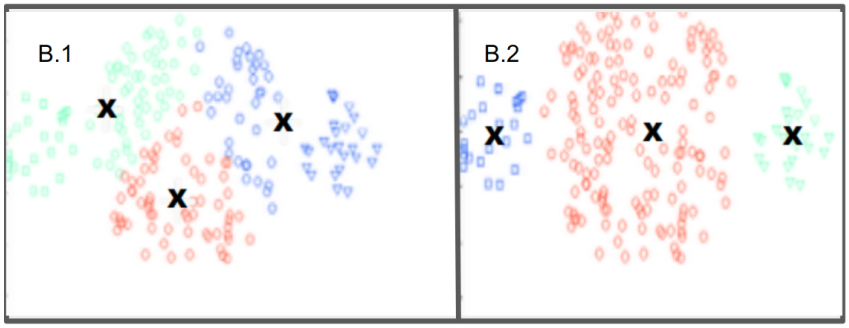
   (a) (1 point) Dataset A

   Select one:
   - ○ A.1
   - ● A.2



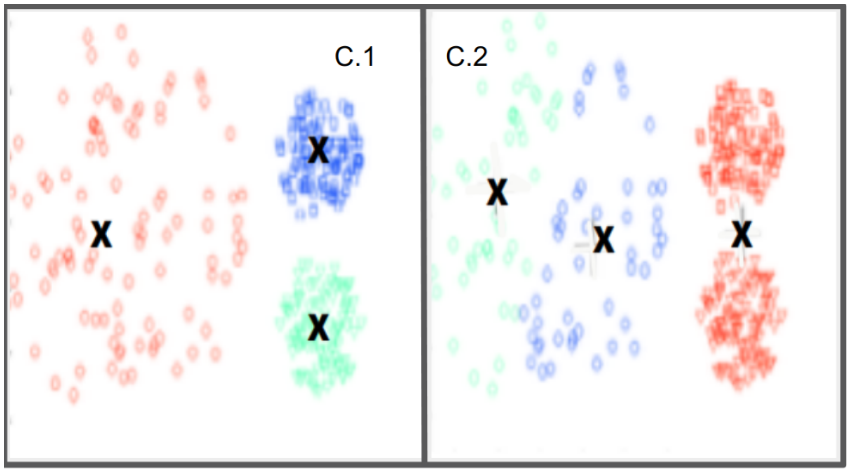
   (b) (1 point) Dataset B

   Select one:
   - ● B.1
   - ○ B.2



   (c) (1 point) Dataset C

   Select one:
   - ○ C.1
   - ● C.2



2. Consider a dataset $\mathcal{D}$ with 5 points as shown below. Perform K-means clustering on this dataset with $k = 2$ using Euclidean distance as the distance metric.

   Remember that in the K-means algorithm, one iteration consists of following two steps: first, we assign each data point to its nearest cluster center; second, we recompute each center as the average of the data points assigned to it. Initially, the 2 cluster centers are chosen randomly as μ_0 = (5.3, 3.5), μ_1 = (5.1, 4.2). Parts (a) through (d) refer only to the first iteration of K-means clustering performed on $\mathcal{D}$.

$$\mathcal{D} = \begin{bmatrix} 5.5 & 3.1 \\ 5.1 & 4.8 \\ 6.6 & 3.0 \\ 5.5 & 4.6 \\ 6.8 & 3.8 \end{bmatrix}$$

   (a) (1 point) **Select one:** Which of the following points will be the new center for cluster 0?

   ◯ (5.7 , 4.1)

   ◯ (5.6 , 4.8)

   ● (6.3 , 3.3)

   ◯ (6.7 , 3.4)

   (b) (1 point) **Select one:** Which of the following points will be the new center for cluster 1?

   ◯ (6.1 , 3.8)

   ◯ (5.5 , 4.6)

   ◯ (5.4 , 4.7)

   ● (5.3 , 4.7)

   (c) (1 point) How many points will belong to cluster 0, using the new centers?

   Answer

   3

   (d) (1 point) How many points will belong to cluster 1, using the new centers?

   Answer

   2

3. Consider the following brute-force algorithm for minimizing the K-means objective: Iterate through each possible assignment of the points to k clusters, $\mathbf{z} = [z^{(1)}, \ldots, z^{(N)}]$ and for each assignment $\mathbf{z} \in \{1, \ldots, k\}^N$ evaluate the following objective function:

$$J(\mathbf{z}) = \underset{\mathbf{c}_1, \ldots, \mathbf{c}_k}{\operatorname{argmin}} \sum_{n=1}^{N} ||\mathbf{x}^{(n)} - \mathbf{c}_{z^{(n)}}||_2^2$$

At the end, pick the assignment $\mathbf{z}$ that minimizes $J(\mathbf{z})$.

(a) (1 point) Suppose we have N points and k clusters. How many possible assignments $\mathbf{z}$ does the brute force algorithm have to evaluate $J(\mathbf{z})$ for? Express your answer in terms of k and N only.

Answer

k^N

(b) (1 point) Suppose $N = 1000$, $k = 10$, and it takes us 0.01 seconds to evaluate $J(\mathbf{z})$ for a single assignment $\mathbf{z}$. How many seconds will the brute force algorithm take to check all assignments?

Answer

10^{998}

4. (2 points) In 1-2 concise sentences, explain why using k-means++ initialization is more likely to choose one sample from each cluster than random initialization when given the dataset in Figure 1 below.

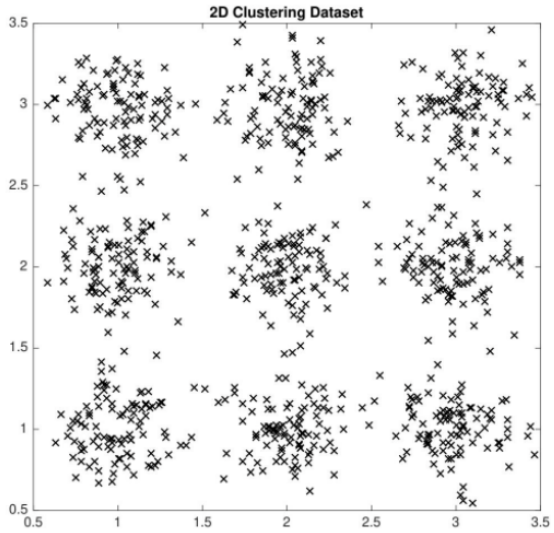


Figure 1: 2D Dataset

Your Answer

k-means++ chooses samples from a distance based on previously chosen samples. Thus, subsequent samples chosen are likely to be spread out and represent each cluster more so than random initializa

5 PCA (9 points)

PCA Theory

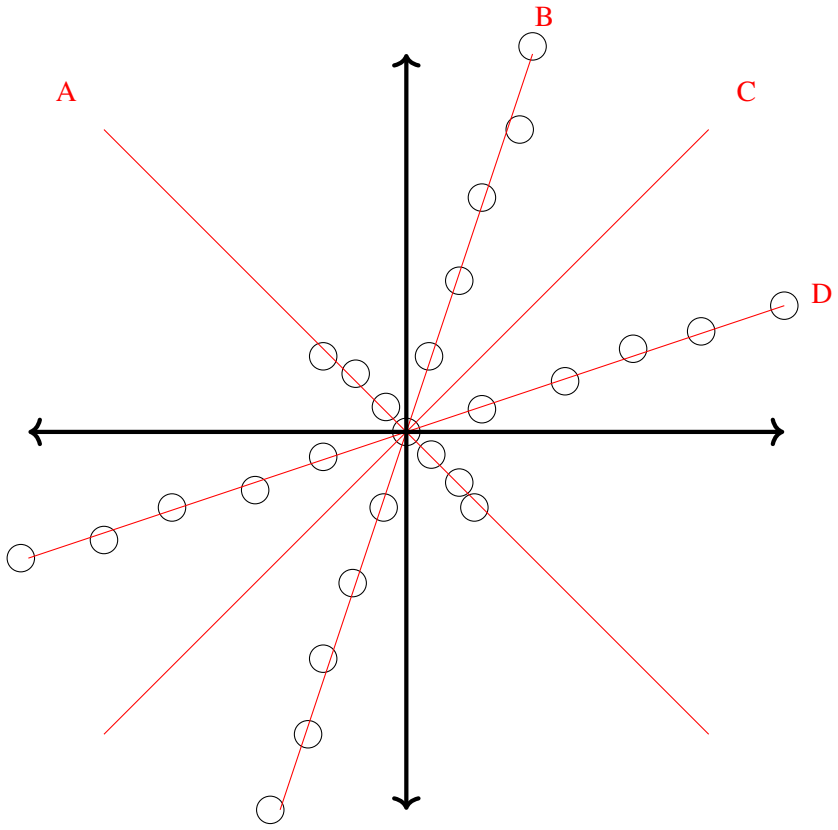
1. (1 point) **Select one:** Assume we apply PCA to a matrix $X \in \mathbb{R}^{n \times 2}$ and obtain two sets of PCA feature scores, $Z_1, Z_2 \in \mathbb{R}^n$, where Z_1 corresponds to the first principal component and Z_2 corresponds to the second principal component. Comparing the values of features in Z_1 and Z_2, which is most common in the training data:

   - ● a point with small feature values in Z_2 and large feature values in Z_1
   - ○ a point with large feature values in Z_2 and small feature values in Z_1
   - ○ a point with large feature values in Z_2 and large feature values in Z_1
   - ○ a point with small feature values in Z_2 and small feature values in Z_1

2. (2 points) For the dataset shown below, list the principal components from first to last.

Your Answer

C, A



PCA in Practice

For this section, refer to the PCA demo linked here. In this demonstration, we have performed PCA for you on a simple four-feature dataset. The questions below have also been added to the Colab notebook linked for ease of access. Run the code in the notebook, then answer the questions based on the results.

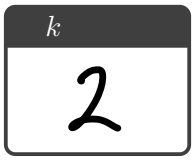
3. (1 point) **Select one:** Do you see any special relationships between any of the features? In particular, take a look at the `petal_length` feature. How would you describe its association with each of the **other features**?

   - ● The features are highly correlated: we observe linearly proportional relationships where increases in `petal_length` often correspond to increases in another feature.
   - ○ The features are highly correlated: we observe that the color classes can be separated with decision boundaries along the `petal_length` axis.
   - ○ The features are uncorrelated: we observe random noise as if the features were generated from independent distributions.
   - ○ The features are uncorrelated: we observe the "default $y = x$" relationship between features.

4. (2 points) **Select all that apply:** To get the principal components of the features, we calculate the eigenvectors of the covariance matrix, which are orthogonal, along with their corresponding eigenvalues. Which of the following are consequences of the principal components being orthogonal to each other?

   - ■ The variance of the data is maximized.
   - ■ The reconstruction error is minimized.
   - □ The dot product of any two principal components will be 1.0.
   - □ We can attribute certain variations in the data to unique principal components.
   - □ In the dimensionality-reduced space, the covariance of the first and second dimensions will always be zero.
   - □ It ensures that our lower-dimensional data will be linearly separable.
   - □ None of the above.

5. (1 point) If we wanted to find k principal components such that we preserve **at least** 95% of the variance in the data, what would be the value of k? *Hint:* it is helpful here to look at the cumulative variance in the first k components, which we have calculated for you.



6. (2 points) If we wanted to perform dimensionality reduction to have just two features, we could pick any two features from the dataset and train a classifier on just those. What is one reason we could prefer the PCA features to just choosing two of the original features to represent our data?

Your Answer

Choosing two features at random to train the data may not yield significant results whereas using PCA features focuses on the strongest correlations in the data before dimensionality reduction.

6 Random Forests (10 points)

1. (1 point) **True or False:** In a random forest, it is generally better for the trees to be highly correlated, as this reduces variability.

   ◯ True

   ● False

2. Recall the following dataset $\mathcal{D}$ from Homework 2; $\mathcal{D}$ consists of 8 examples, each with 3 attributes, (A, B, C), and a label, Y. However, instead of training a decision tree to classify the data, you will train a random forest instead.

A	B	C	Y
1	2	0	1
0	1	0	0
0	0	1	0
0	2	0	1
1	1	0	1
1	0	1	0
1	2	1	0
1	1	0	1

   Suppose you set $B = 3, \rho = 2$, and create the datasets $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$. Instead of full decision trees, you decide to train decision stumps, i.e., each tree will have depth 1. For each bootstrapped dataset, you sample the two features S_1, S_2, S_3 as shown below.

A	B	C	Y
0	1	0	0
0	1	0	0
0	0	1	0
0	2	0	1
1	1	0	1
1	0	1	0
1	1	0	1
1	1	0	1

$\mathcal{D}_1$
$S_1 = \{A, B\}$

A	B	C	Y
0	1	0	0
0	1	0	0
0	0	1	0
0	2	0	1
0	2	0	1
1	0	1	0
1	0	1	0
1	1	0	1

$\mathcal{D}_2$
$S_2 = \{B, C\}$

A	B	C	Y
1	2	0	1
1	2	0	1
1	2	0	1
0	1	0	0
0	2	0	1
1	1	0	1
1	1	0	1
1	2	1	0

$\mathcal{D}_3$
$S_3 = \{A, C\}$

   (a) (3 points) Calculate the in-sample training error for the decision stumps trained on $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$.

Error on $\mathcal{D}_1$	Error on $\mathcal{D}_2$	Error on $\mathcal{D}_3$
0.25	0.125	0.125

(b) (2 points) Calculate the out-of-bag (OOB) error for the random forest. Aggregated predictions for some rows in $\mathcal{D}$ have been provided.

A	B	C	Y	$\hat{Y}$
0	1	0	0	1
0	2	0	1	1
1	1	0	1	0
1	1	0	1	1

OOB Error

0.5

Work

3. (2 points) Briefly explain the difference between the calculation of OOB error and cross-validation error.

Explanation

OOB error does not use a validation set or held out data to influence the model/trees at any point of the training.

4. (2 points) **Select all that apply:** Which of the following are hyperparameters that can be tuned in a random forest?

- ■ Number of trees trained
- ■ Number of points used to train each decision tree
- ■ Size of feature subsets
- □ Features used for splits in each decision tree
- □ None of the above

7 Ensemble Methods (9 points)

1. In the following question, we will explore an ensemble method known as the **halving algorithm**:

   We first maintain a list of n weak classifiers $h_1, ..., h_n$ which have not yet made mistakes. The training labels and classifier predictions are in $\{-1, 1\}$. For each training sample (x, y), we make prediction based on the majority vote of weak classifiers $\hat{y} = \text{sign}(\sum_{i=1}^{n} h_i(x))$. If the majority vote prediction does not equal to the label, we will eliminate all the h_i such that $h_i(x) \neq y$. The final aggregated classifier will be the ensemble of all the remaining classifiers.

   (a) (1 point) Assume we start with a total of n classifiers, and in the end, we are left with at least one classifier. What would be the big-O bound of the total number of mistakes made by the aggregated classifier in terms of n?

   Your Answer

   $O(\log_2 |n|)$

   (b) (1 point) In one sentence, which weak classifiers are guaranteed to be kept by the halving algorithm?

   Your Answer

   The ones that have zero training error

2. (1 point) **True or False:** Consider some training data point $(x^{(n)}, y^{(n)})$ used during a run of the AdaBoost algorithm. If for all t, the weak learner h_t learned during training in iteration t correctly classifies $h_t(x^{(n)}) = y^{(n)}$, the weight assigned to $x^{(n)}$ in the training distribution $\mathcal{D}_t$ will reach exactly 0 in a finite number of iterations of AdaBoost.

   ○ True

   ● False

3. (1 point) **True or False:** If the ensemble classifier learned by AdaBoost reaches 0 training error, all weak learners created in subsequent iterations will be identical (i.e., they will produce the same output on any input). Assume we are using deterministically selected weak learners.

   ○ True

   ● False

4. In the following question, we will examine the generalization error of AdaBoost using a concept known as the *classification margin*.

   Throughout the question, use the following definitions:

   - N: The number of training samples.
   - $S = \{(x^{(1)}, y^{(1)}), \cdots, (x^{(N)}, y^{(N)})\}$: The training samples with binary labels ($y^{(i)} \in \{-1, +1\}$).
   - $\mathcal{D}_t(i)$: The weight assigned to training example i at time t. Note that $\sum_i \mathcal{D}_t(i) = 1$.
   - h_t: The weak learner constructed at time t.
   - ϵ_t: The error of h_t on $\mathcal{D}_t$.
   - Z_t: The normalization factor for the distribution update at time t.
   - α_t: The weight assigned to the learner h_t in the composite hypothesis.
   - $f_t(x) = \left(\sum_{t'=1}^{t} \alpha_{t'} h_{t'}(x)\right) / \left(\sum_{t'=1}^{t} \alpha_{t'}\right)$: The aggregated vote of the weak learners, rescaled based on the total weight.

   For a binary classification task, assume that we use a probabilistic classifier that provides a probability distribution over the possible labels (i.e. $p(y|x)$ for $y \in \{+1, -1\}$). The classifier output is the label with highest probability. We define the *classification margin* for an input as the signed difference between the probability assigned to the correct label and the incorrect label $p_{correct} - p_{incorrect}$, which takes on values in the range $[-1, 1]$. Recall from recitation that $\text{margin}_t(x^{(i)}, y^{(i)}) = y^{(i)} f_t(x^{(i)})$.

   (a) (2 points) Recall the update AdaBoost performs on the distribution of weights:

   - $\mathcal{D}_1(i) = 1/N$
   - $\mathcal{D}_{t+1}(i) = \mathcal{D}_t(i) \frac{\exp(-y^{(i)} \alpha_t h_t(x^{(i)}))}{Z_t} = \frac{1}{N} \left(\prod_{t'=1}^{t} \frac{1}{Z_{t'}}\right) \exp(-\sum_{t'=1}^{t} y^{(i)} \alpha_{t'} h_{t'}(x^{(i)}))$

   We define $C_{t+1} = \frac{1}{N} \left(\prod_{t'=1}^{t} \frac{1}{Z_{t'}}\right)$ and $M_{t+1}(i) = -\sum_{t'=1}^{t} y^{(i)} \alpha_{t'} h_{t'}(x^{(i)})$. We then have

   $$\mathcal{D}_{t+1}(i) = C_{t+1} \exp(M_{t+1}(i))$$

   Let $\alpha = \sum_{t'=1}^{t} \alpha_{t'}$. Rewrite $M_{t+1}(i)$ in terms of $\text{margin}_t(x^{(i)}, y^{(i)})$ and α. (Hint: first rewrite $M_{t+1}(i)$ in terms of $y^{(i)}, \alpha, f_t, x^{(i)}$, then apply our given formula for the margin).

   Your Answer

   $-\text{margin}_t(x^{(i)}, y^{(i)})\alpha$

(b) (1 point) Note that C_{t+1}, α are treated as positive constants with respect to the input points. Using the classification margin and the above formulation of the weights assigned by AdaBoost, fill in the blanks to describe which points AdaBoost assigns high weight to at time t.

At time t, AdaBoost assigns higher weight to points $x^{(i)}$ with _____ _____ value of margin on the current ensemble classifier (i.e., $\text{margin}_t(x^{(i)}, y^{(i)})$).

Select one:

- ◯ higher absolute
- ◯ higher signed
- ◯ lower absolute
- ● lower signed

(c) (2 points) How does this weighting behavior explain the empirical result of test error continuing to decrease after training error has converged?

Your Answer

weights will continue to update even though classifier correctly classifies points. Thus, the margin will continue to increase and make more confident predictions.

8 Collaboration Questions

After you have completed all other components of this assignment, report your answers to these questions regarding the collaboration policy. Details of the policy can be found here.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details.
2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details.
3. Did you find or come across code that implements any part of this assignment? If so, include full details.

Your Answer

1) no

2) no

3) no