CSE 512: Final Exam

Dec 9,10

Instructions:

- You are encouraged to use no more than 3 hours to complete the exam, but you may use as long as you wish within the 48 hour window.
- You are allowed 1 page (front and back) cheat sheet. The cheat sheet must be scanned (or photographed with high resolution) and submitted along with the exam solutions. The time begins when you flip this page.
- You are also allowed a simple calculator, or Wolfram Mathematica to simplify equations. You may also write simple scripts in python or MATLAB to aid with your calculation. However, you do not need to for solving the problem, and you do not need to submit any code.
- You may print the exam and write your solutions or use lyx/latex. If you need extra sheets of paper, please label them carefully as to which question they are answering. Make your final answer clear.
- If you choose to handwrite your solutions, you must make sure that the digital scan / photograph is of high enough quality that we can see everything clearly. Anything we can't read, we will not grade.
- You may not discuss any problem with any other student while the exam submission portal is still open. You may not look for answers on the internet or in any notes outside of your cheatsheet.

Name: ________________________

Student ID: ________________________

Scoring	
Q 1	__________ / 10
Q 2	__________ / 10
Q 3	__________ / 20
Q 4	__________ / 20
Q 5	__________ / 20
Q 6	__________ / 20
Total	__________ / 100

1. Classify the following as a supervised learning task (where training labels are needed) or unsupervised learning task (where training labels are not needed) **(2 point each)**

   (a) Principal component analysis

   Supervised **Unsupervised**

   (b) logistic regression

   Supervised **Unsupervised**

   (c) Gaussian mixture model

   Supervised **Unsupervised**

   (d) boosted decision trees

   Supervised **Unsupervised**

   (e) clustering

   Supervised **Unsupervised**

2. **True or False. (2 point each)**

   (a) A graphical model with on average 2 directed edges pointing to each node requires less training samples for inference, than a similar graphical model with on average 5 directed edges pointing to each node.

   True **False**

   (b) In a hidden Markov model, each hidden state is independent from all other hidden states

   True **False**

   (c) In multiclass logistic regression, the event that a sample x belongs to class i is independent of the event that x belongs to class j whenever $i \neq j$.

   True **False**

   (d) The difference between K-means and K-NN is that K-means is a supervised learning task and K-NN is an unsupervised learning task.

   True **False**

   (e) Ensembling different models presents more advantages if each model acts as independently as possible.

   True **False**

3. I am given a bunch of animals and asked to classify them. I have some training data, given below, and I will use it to construct a decision tree, which I will use to classify future animals. In this problem, round all values to the nearest 0.001, and use log base 2.

name	label	nose	ear shape	ear position
Arthur	aardvark	small	round	high
D.W.	aardvark	small	round	high
Buster	rabbit	small	long	high
Bud	rabbit	small	long	high
Bitzi	rabbit	small	long	high
Francine	monkey	big	round	high
Muffy	monkey	big	round	low
Neal	moose	big	long	low

(a) What is the entropy of the labels over the entire dataset? **(3pts)**

(b) For each feature, compute the information gain if that feature was to be used in the first split.

Nose (3pts)

Ear shape (3pts)

Ear position (3pts)

(c) Construct the decision tree, using node purity to pick which node to split next, and information gain to pick which feature/split to make. Draw the final decision tree here. **(6pts)**

(d) Using this tree, infer the labels of the animals in the following test set and report the test error. **(2pts)**

name	true label	inferred label (fill in)	nose	ear shape	ear location
Bambi	moose		small	round	high
Thumper	rabbit		big	long	low
George	monkey		big	round	low

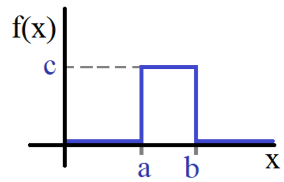
Test error =

4. **Gradient boosting using pulse functions** I am now faced with the task of estimating people's height based on their age. Again, I have some training data, which I will use to construct a boosted regression model. In this problem, round all values to the nearest 0.001.

training sample	name	age (x_i)	height in ft (y_i)
1	Alice	25	6
2	Brian	60	5.5
3	Carlos	15	5
4	Dianne	1	2
5	Esther	5	3
6	Freddy	35	6

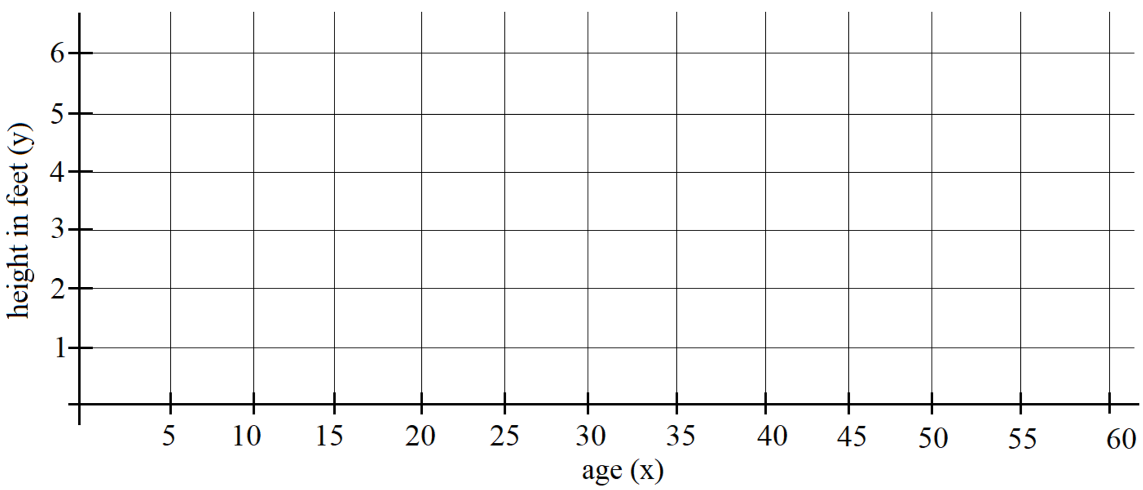
The goal is to return a function $f(\text{age}) \approx$ height. We will use gradient boosting, with the set of weak regressors as simple pulse functions, parametrized by scalars a, b, and c.

$$\mathcal{H} = \left\{ f : f(x) = \begin{cases} c & \text{if } a < x < b \\ 0 & \text{else.} \end{cases} \right\}$$



Formally, our goal is to minimize the squared error, e.g. solving

$$\begin{array}{ll} \underset{f_t \in \mathcal{H}}{\text{minimize}} & \displaystyle\sum_{i=1}^{m} (F(x_i) - y_i)^2 \\ \text{subject to} & F = f_1 + \cdots + f_T. \end{array}$$



(cont.)

(a) **(2pts)** For the training set, plot height vs age, using the graph on the previous page. Use $\times$ marks.

(b) **(1pts)** For the "zeroth" weak learner, we just compute the mean label

$$f^{(0)}(x_i) = y_0 = \frac{1}{m}\sum_{i=1}^{m} y_i.$$

Since this is the only weak learner thus far, it is also the strong learner; e.g. $F^{(0)}(x) = f^{(0)}(x)$. What is the mean squared error of picking the average height y_0 as the prediction for all the datapoints? (Round to nearest 0.01.)

(c) **Iteration 1** To find the first weak learner, we will solve the least squares problem, where $z_i^{(0)} = y_i - F^{(0)}(x_i)$ is the current residual.

$$\underset{f\in\mathcal{H}}{\text{minimize}} \quad \sum_{i=1}^{m}(f(x_i) - z_i^{(0)})^2. \tag{1}$$

i. **(2pts)** Frame problem (1) as an optimization problem over a, b, and c.

ii. **(2pts)** I have optimized this problem and found that the optimum values of a and b are $a = 0$ and $b = 5.5$. Find the best value of c, rounding to the nearest 0.1.

iii. **(2pts)** For the aggregate learner $F^{(1)}(x) = f^{(1)}(x) + f^{(0)}(x)$, draw $(x_i, F^{(1)}(x_i))$ on the plot in the previous page. Clearly label the curve drawn as $F^{(1)}$.

iv. **(2pts)** Compute the current loss and new residual vector $z^{(1)}$. Round to the nearest 0.01.

loss = , **residual** $z^{(1)} = ($ $)$

(cont.)

(d) **Iteration 2** To find the second weak learner, we will again solve (1), replacing $z^{(0)}$ with the newly computed residual vector $z^{(1)}$. Again, I have optimized this problem and found that the optimum values of a and c are $a = 1$ and $c = 0.4$.

i. **(2pts)** Find the best value of b, rounding to the nearest 0.1.

ii. **(3pts)** For the aggregate learner $F^{(2)}(x) = f^{(2)}(x) + f^{(1)}(x) + f^{(0)}(x)$, draw $(x_i, F^{(1)}(x_i))$ on the plot from before. Clearly label this new curve as $F^{(2)}$.

iii. **(2pts)** Compute the current loss and new residual vector $z^{(2)}$. Round to the nearest 0.01.

loss = , **residual** $z^{(2)} = ($ $)$

(e) **(2pts)** Using this final aggregated learner, predict and give the test set heights based on age, by filling in the table below. Round to the nearest 0.01. Report the test mean squared error.

person	age	true height	fitted height
Gary	12	5	
Harris	57	5	
Ivy	3	2	

Test mean squared error:

5. **Kmeans and Gaussian mixture models** I am on an alien planet, and there are three types of fruit. They are of an assortment of sizes, which I have measured and given below, in centimeters:

$$x_1 = 6, \quad x_2 = 1, \quad x_3 = 12, \quad x_4 = 14, \quad x_5 = 3, \quad x_6 = 5, \quad x_7 = 1, \quad x_8 = 1, \quad x_9 = 2$$

   (a) **(3pts)** Using the 3 initial cluster center values of $c_1 = 1$, $c_2 = 2$, $c_3 = 3$, list the points that are assigned to each cluster in this first iteration. (Write down x_i and the values of i.)

points in cluster 1	points in cluster 2	points in cluster 3

   (b) **(3pts)** Perform one iteration of Kmeans (using 2 norm distance) give the new cluster centers.

$$c_1 = \qquad\qquad , \quad c_2 = \qquad\qquad , \quad c_3 =$$

   (c) An alien approaches me and tells me that the fruits have names, called “Ajax”, ”Basic”, and ”C++”.

   - Ajax is the smallest type of fruit, with a mean size of 30cm and a variance of 25cm.
   - Basic is a medium sized fruit, with mean size 50cm, and a variance of 5cm.
   - C++ is a large sized fruit, with mean size 60cm, and a variance of 100cm.

   Additionally, Ajax is a super common fruit, occuring 3x as often as the other two, combined. Basic occurs 4x as often as C++.

   (4pts) Fit a Gaussian mixture model to this story. That is, write out the PDF to model the size of a berry picked at random from this planet.

$$f_{\text{fruit size}}(X) =$$

(d) **(5pts)** Using the above model, return the probability that a berry of size 40cm is Ajax. Hint: Remember Bayes' rule. Round to the nearest 0.01.

(e) **(5pts)** Ajax and C++ are poisonous, and Basic is not. I find a fruit of size 50cm and eat it. What is the probability that I have been poisoned? Round to the nearest 0.01.

6. **Teamwork.** Alice, Bob, and Carlos are all working on their problem set together. Each question is a True/False question. The group works by each guessing the answer, and writing down the majority vote answer.

   (a) **(3pts)** Given that each friend is independently 80% sure that the answer is False, what is the probability that they will write down the answer True? Round to the nearest 0.001.

   (b) **(3pts)** Suppose that each friend independently has p chance of being correct, where p is a probability value between 0 and 1. Given the guesses of each three friends, we define the maximum likelihood guess to be

   $$\hat{s}_{MLE} = \underset{\hat{s}\in\{T,F\}}{\operatorname{argmax}} \mathbf{Pr}(\text{ answer is } \hat{s}|A, B, C).$$

   Show that as long as $p > 0.5$, then $\hat{s}_{MLE}$ is in fact the guess given by majority vote.

   Hint 1: show that the probability that the majority vote is correct is also > 0.5.

   Hint 2: In general, $f(x) > c$ for all $x > b$ if f is strictly monotonically increasing and $f(b) \geq c$.

   (c) **(3pts)** Give the mean and variance of $\hat{s}_{MLE}$, given that Alice, Bob, and Carlos are all simultaneously and independently p percent sure the answer is True. Use the numerical encoding $\hat{s} = 1$ if the answer is True and $\hat{s} = 0$ if the answer is False. (No need to simplify the equations.)

   Hint: recall the mathematical definition of expectations.

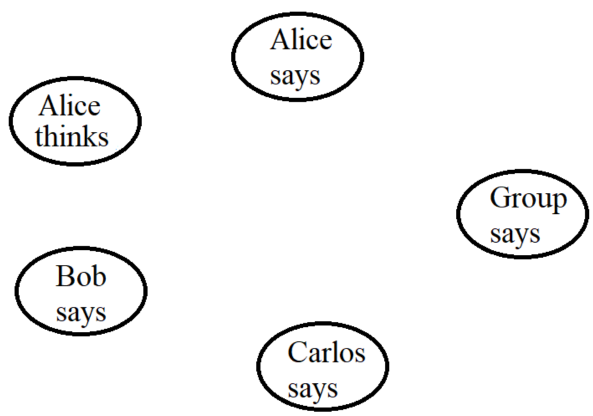
(cont.)

(d) **(4pts)** Again, Alice, Bob, and Carlos all simultaneously and independently believe the answer is True with probability p. The professor walks in. The three ask her for the answer. She mumbles "I think it's True". The professor really isn't that bright, though, so rather than trusting her completely, the students decide on a modified majority vote scheme: if all of the students believe the answer is False, then they write False. But, otherwise, the students will write down True, following the advice of the professor. We call what the students write down $\hat{s}_{\text{MAP}}$, where $\hat{s}_{\text{MAP}} = 1$ if they write down True, and $\hat{s}_{\text{MAP}} = 0$ if they write down False. What is the mean and variance of $\hat{s}_{\text{MAP}}$?

(e) Now it is the next day, and the professor has gone on vacation; we must rely only on the guesses of the three students. While they are all good friends, the score overall is curved, so there is incentive for friends to sabotage each other, in order to get better scores. After doing some problems, Bob and Carlos start to suspect Alice of sabotaging.

If Alice is sabotaging, then her strategy is simply to flip her answers with probability q, where $0 \leq q \leq 1$.

i. **(3pts)** Draw directed edges on the graphical model corresponding to this scenario, and write the conditional probability on each edge.



ii. **(4pts)** Given that Alice, Bob, and Carlos are independently 70% correct on any given problem, and their majority voting scheme is also 70% correct, find q.