

# HOMEWORK 8

## LEARNING PARADIGMS

CMU 10-601: MACHINE LEARNING (SPRING 2021)

<https://mlcourse.org>

OUT: 04/30/2021

DUE: 5/07/2021, 11:59pm EDT

TAs: Amanda Coston, Young Kim, and Daniel Min

### START HERE: Instructions

Homework 8 covers topics on PAC-learning, Ensembles, RecSys and Graphical Models. The homework includes multiple choice, True/False, and short answer questions.

- **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 2.1”). 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. See the Academic Integrity Section on the course site for more information: <http://www.cs.cmu.edu/~mgormley/courses/10601/about.html>
- **Late Submission Policy:** See the late submission policy here: <http://www.cs.cmu.edu/~mgormley/courses/10601/about.html>
- **Submitting your work:**
  - **Gradescope:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, we will be using Gradescope (<https://gradescope.com/>). 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. Alternatively, submissions can be written in LaTeX. Regrade requests can be made, however this gives the TA the opportunity to regrade your entire paper, meaning if additional mistakes are found then points will be deducted. Each derivation/proof should be completed on a separate page. For short answer questions, you **should not** include your work in your solution. If you include your work in your solutions, your assignment may not be graded correctly by our AI assisted grader. In addition, please tag the problems to the corresponding pages when submitting your work.

For multiple choice or select all that apply questions, shade in the box or circle in the template document corresponding to the correct answer(s) for each of the questions. For L<sup>A</sup>T<sub>E</sub>X users, use  and  for shaded boxes and circles, and don't change anything else.

## Instructions for Specific Problem Types

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

**Select One:** Who taught this course?

- Matt Gormley
- 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?

- Matt Gormley
- 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-X601
---------

# 1 MLE/MAP [31 pts]

1. [3 pt] **True or False:** Consider a Boolean-valued random variable  $X \in \{0, 1\}$  where you wish to estimate  $\theta = P(X = 1)$ . You decide to use a MAP estimate based on a Beta prior plus your observed training data. Further suppose an adversary chooses “misleading”, but finite hyperparameters for your Beta prior in order to confuse your learning algorithm. As the number of training examples grows to infinity, the MAP estimate of  $\theta$  will still converge to the MLE estimate of  $\theta$ .

Select One:

- True
- False

2. [3 pt] In HW3, you have derived the closed form solution for linear regression. Now, we are coming back to linear regression, viewing it as a statistical model, and deriving the MLE and MAP estimate of the parameters in the following questions.

Assume we have data  $D = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$ , where  $\mathbf{x}^{(i)} = (x_1^{(i)}, \dots, x_M^{(i)})$ . So our data has  $N$  instances and each instance has  $M$  attributes/features. Each  $y^{(i)}$  is generated given  $\mathbf{x}^{(i)}$  with additive noise  $\epsilon^{(i)} \sim N(0, \sigma^2)$ , that is  $y^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)} + \epsilon^{(i)}$  where  $\mathbf{w}$  is the parameter vector of linear regression. Given this assumption, what is the distribution of  $y$ ?

Select one:

- $y^{(i)} \sim N(\mathbf{w}^T \mathbf{x}^{(i)}, \sigma^2)$
- $y^{(i)} \sim N(0, \sigma^2)$
- $y^{(i)} \sim Uniform(\mathbf{w}^T \mathbf{x}^{(i)} - \sigma, \mathbf{w}^T \mathbf{x}^{(i)} + \sigma)$
- None of the above

3. [4 pt] The next step is to learn the MLE of the parameters of the linear regression model. Which expression below is the correct conditional log likelihood  $\ell(\mathbf{w})$  with the given data?

Select one:

- $\sum_{i=1}^N [-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$
- $\sum_{i=1}^N [\log(\sqrt{2\pi\sigma^2}) + \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$
- $\sum_{i=1}^N [-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})]$
- $-\log(\sqrt{2\pi\sigma^2}) + \sum_{i=1}^N [-\frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$

4. [4 pt] Then, the MLE of the parameters is just  $\text{argmax}_{\mathbf{w}} \ell(\mathbf{w})$ . Among the following expressions, select ALL that can yield the correct MLE.

Select all that apply:

- $\text{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})]$
- $\text{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$
- $\text{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$

- $\operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\frac{1}{2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})]$
- $\operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\frac{1}{2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$

5. [3 pt] According to the above derivations, is the MLE for the conditional log likelihood equivalent to minimizing mean squared errors (MSE) for the linear regression model when making predictions? Why or why not?

Select one:

- Yes, because the derivative of the negative conditional log-likelihood has the same form as the derivative of the MSE loss.
- Yes, because the parameters that maximize the conditional log-likelihood also minimize the MSE loss.
- No, because one is doing maximization and the other is doing minimization.
- No, because the MSE has an additional error term  $\epsilon^{(i)}$  in the expression whereas the quantity to be minimized in MLE does not.
- No, because the conditional log-likelihood has additional constant terms that do not appear in the MSE loss.

6. [3 pt] Now we are moving on to learn the MAP estimate of the parameters of the linear regression model. The MAP estimate is obtained through solving the following optimization problem (recall that  $D$  refers to the data, and  $w$  to the regression parameters (weights)).

$$w_{MAP} = \arg \max_w p(w|D) = \arg \max_w p(D, w)$$

Suppose are using a Gaussian prior distribution with mean 0 and variance  $\frac{1}{\lambda}$  for each element  $w_m$  of the parameter vector  $w$  ( $1 \leq m \leq M$ ), i.e.  $w_m \sim N(0, \frac{1}{\lambda})$ . Assume that  $w_1, \dots, w_M$  are mutually independent of each other. Which expression below is the correct log joint-probability of the data and parameters  $\log p(D, w))$ ?

(For simplicity, just use  $p(D|w)$  to denote the data likelihood.)

Select one:

- $\log p(D|w) - \sum_{m=1}^M \log(\sqrt{2\pi\lambda}) - \lambda(w_m)^2$
- $\log p(D|w) + \sum_{m=1}^M -\log(\sqrt{2\pi\lambda}) - \lambda(w_m)^2$
- $\log p(D|w) - \sum_{m=1}^M \log(\sqrt{\frac{2\pi}{\lambda}}) - \frac{\lambda}{2}(w_m)^2$
- $\log p(D|w) + \sum_{m=1}^M -\log(\sqrt{\frac{2\pi}{\lambda}}) - \frac{\lambda}{2}(w_m)^2$

7. [3 pt] A MAP estimator with a Gaussian prior  $\mathcal{N}(0, \sigma^2)$  you trained gives significantly higher test error than train error. What could be a possible approach to fixing this?

Select one:

- Increase variance  $\sigma^2$
- Decrease variance  $\sigma^2$
- Try MLE estimator instead
- None of the above

8. [4 pt] Maximizing the log posterior probability  $\ell_{MAP}(\mathbf{w})$  gives you the MAP estimate of the parameters. The MAP estimate with Gaussian prior is actually equivalent to a L2 regularization on the parameters of linear regression model in minimizing an objective function  $J(\mathbf{w})$  that consists of a term related to log conditional likelihood  $\ell(\mathbf{w})$  and a L2 regularization term. The following options specify the two terms in  $J(\mathbf{w})$  explicitly. Which one is correct based on your derived log posterior probability in the previous question?

Select one:

- $-\ell(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|_2$
- $-\ell(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$
- $-\ell(\mathbf{w}) + \lambda \|\mathbf{w}\|_2$
- $\ell(\mathbf{w}) - \frac{\lambda}{2} \|\mathbf{w}\|_2^2$

9. [4 pt] MAP estimation with what prior is equivalent to L1 regularization?

Note:

The pdf of a Uniform distribution over  $[a,b]$  is  $f(x) = \frac{1}{b-a}$  if  $x \in [a, b]$  and 0 otherwise.  
The pdf of an exponential distribution with rate parameter  $a$  is  $f(x) = a \exp(-ax)$  for  $x > 0$ .  
The pdf of a Laplace distribution with location parameter  $a$  and scale parameter  $b$  is  $f(x) = \frac{1}{2b} \exp\left(\frac{-|x-a|}{b}\right)$  for all  $x \in \mathbb{R}$ .

Select one:

- Uniform distribution over  $[-\mathbf{w}^T \mathbf{x}^{(i)}, \mathbf{w}^T \mathbf{x}^{(i)}]$
- Exponential distribution with rate parameter  $a = \frac{1}{2}$
- Exponential distribution with rate parameter  $a = \mathbf{w}^T \mathbf{x}^{(i)}$
- Laplace prior with location parameter  $a = 0$
- Laplace prior with location parameter  $a = \mathbf{w}^T \mathbf{x}^{(i)}$
- Uniform distribution over  $[-1, 1]$

## 2 Naive Bayes [27 pts]

1. [3 pt] I give you the following fact: for events A and B,  $P(A | B) = 2/3$  and  $P(A | \neg B) = 1/3$ , where  $\neg B$  denotes the complement of B. Do you have enough information to calculate  $P(B | A)$ ? If not, choose “not enough information”, if so, compute the value of  $P(B | A)$ .

Select one:

- 1/2
- 2/3
- 1/3
- Not enough information

2. [3 pt] Instead if I give you for events A and B,  $P(A | B) = 2/3$ ,  $P(A | \neg B) = 1/3$  and  $P(B) = 1/3$  and  $P(A) = 4/9$ , where  $\neg B$  denotes the complement of B. Do you have information to calculate  $P(B | A)$ ? If not, choose “not enough information”, if so, compute the value of  $P(B | A)$ .

Select one:

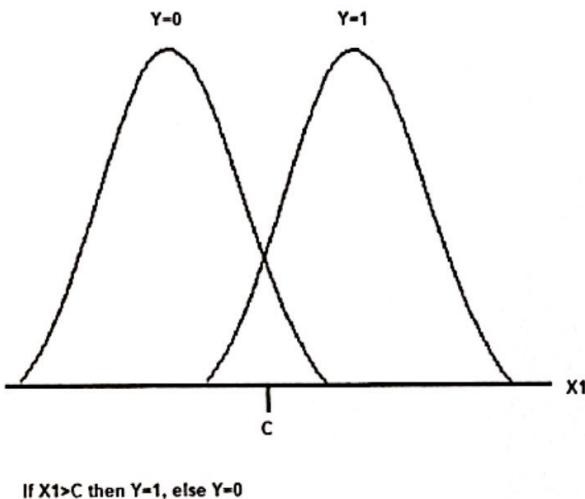
- 1/2
- 2/3
- 1/3
- Not enough information

3. [4 pt] Gaussian Naive Bayes in general can learn non-linear decision boundaries. Consider the simple case where we have just one real-valued feature  $X_1 \in \mathbb{R}$  from which we wish to infer the value of label  $Y \in \{0, 1\}$ . The corresponding generative story would be:

$$Y \sim \text{Bernoulli}(\phi)$$
$$X_1 \sim \text{Gaussian}(\mu_y, \sigma_y^2)$$

where the parameters are the Bernoulli parameter  $\phi = P(Y = 1)$  and the class-conditional Gaussian parameters  $\mu_0, \sigma_0^2$  and  $\mu_1, \sigma_1^2$  corresponding to  $Y = 0$  and  $Y = 1$ , respectively.

A linear decision boundary in one dimension, of course, can be described by a rule of the form “if  $X_1 > c$  then  $Y = 1$ , else  $Y = 0$ ”, where  $c$  is a real-valued threshold (see diagram provided). Is it possible in this simple one-dimensional case to construct a Gaussian Naive Bayes classifier with a decision boundary that cannot be expressed by a rule in the above form)?



If  $X_1 > C$  then  $Y=1$ , else  $Y=0$

Select one:

- Yes, this can occur if the Gaussians are of equal means and equal variances.
- Yes, this can occur if the Gaussians are of equal means and unequal variances.
- Yes, this can occur if the Gaussians are of unequal means and equal variances.
- No, this cannot occur regardless of the relationship of the means or variances.

4. [4 pt] Suppose that 0.3% people have cancer. Someone decided to take a medical test for cancer. The outcome of the test can either be positive (cancer) or negative (no cancer). The test is not perfect - among people who have cancer, the test comes back positive 97% of the time. Among people who don't have cancer, the test comes back positive 4% of the time. For this question, you should assume that the test results are independent of each other, given the true state (cancer or no cancer). What is the probability of a test subject having cancer, given that the subject's test result is positive?

If your answer is in decimals, answer with precision 4, e.g. (6.051, 0.1230, 1.234e+7)

Fill in the blank:

0.6480

5. [4 pt] In a Naive Bayes problem, suppose we are trying to compute  $P(Y | X_1, X_2, X_3, X_4)$ . Furthermore, suppose  $X_2$  and  $X_3$  are identical (i.e.,  $X_3$  is just a copy of  $X_2$ ). Which of the following are true in this case?

Select all that apply:

- Naive Bayes will learn identical parameter values for  $P(X_2|Y)$  and  $P(X_3|Y)$ .
- Naive Bayes will output probabilities  $P(Y|X_1, X_2, X_3, X_4)$  that are closer to 0 and 1 than they would be if we removed the feature corresponding to  $X_3$ .
- This will not raise a problem in the output  $P(Y|X_1, X_2, X_3, X_4)$  because the conditional independence assumption will correctly treat this situation.
- None of the above

6. [3 pt] Which of the following machine learning algorithms are probabilistic generative models?

Select one:

- Decision Tree
- K-nearest neighbors
- Perceptron
- Naive Bayes
- Logistic Regression
- Feed-forward neural network

7. [6 pt] Logistic Regression and Naive Bayes.

As discussed in the reading for class, if  $Y$  is Boolean and  $X = \langle X_1 \dots X_n \rangle$  is a vector of continuous variables, then the assumptions of the Gaussian Naive Bayes classifier imply that  $P(Y | X)$  is given by the logistic function with appropriate parameters  $w_i$  for all  $i$  and  $b$ . In particular:

$$P(Y = 1 | X) = \frac{1}{1 + \exp(b + \sum_{i=1}^n w_i X_i)}$$

and

$$P(Y = 0 | X) = \frac{\exp(b + \sum_{i=1}^n w_i X_i)}{1 + \exp(b + \sum_{i=1}^n w_i X_i)}$$

Consider instead the case where  $Y$  is Boolean and  $X = \langle X_1 \dots X_n \rangle$  is a vector of Boolean variables.

Since the  $X_i$  are Boolean variables, you need only one parameter to define  $P(X_i | Y = y_k)$ . Define  $\phi_{i1} \equiv P(X_i = 1 | Y = 1)$ , in which case  $P(X_i = 0 | Y = 1) = (1 - \phi_{i1})$ . Similarly, use  $\phi_{i0}$  to denote  $P(X_i = 1 | Y = 0)$ .

(a) Given that we can simplify

$$P(Y = 1 | X) = \frac{1}{1 + \exp(\ln(\frac{1-\pi}{\pi}) + \sum_i X_i \ln(\phi_{i0}) + \ln(1-\phi_{i0}) - X_i \ln(1-\phi_{i0}) - X_i \ln(\phi_{i1}) - \ln(1-\phi_{i1}) + X_i \ln(1-\phi_{i1}))}$$

Find expressions for  $P(Y = 1 | X)$  and  $P(Y = 0 | X)$  in terms of  $b$  and  $w_i$ . Explicitly define  $b$  and  $w_i$ .

$$b = \ln\left(\frac{1-\pi}{\pi}\right) + \sum_i \ln\left(\frac{1-\phi_{i0}}{1-\phi_{i1}}\right)$$

$$w_i = \ln\left(\frac{\phi_{i0}(1-\phi_{i1})}{\phi_{i1}(1-\phi_{i0})}\right)$$

$$P(Y=1|X) = \frac{1}{1 + \exp(b + \sum_{i=1}^n w_i X_i)}$$

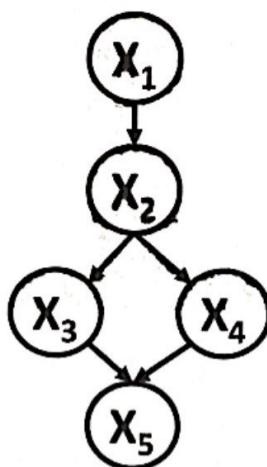
$$P(Y=0|X) = \frac{\exp(b + \sum_{i=1}^n w_i X_i)}{1 + \exp(b + \sum_{i=1}^n w_i X_i)}$$

### 3 Graphical Models [15pts]

In the Kingdom of Westeros, Summer has come. Jon Snow, the King in the North, has taken the responsibility to defeat the Giants and protect the realm.

If Jon Snow can get Queen Cersei and Daenerys Queen of the Dragons to help him Jon is likely to beat the giants. Cersei and Daenerys are powerful women who are skeptical of the existence of Giants and will most likely only consider joining Jon if they are shown evidence of an imminent Giant attack. They can only be shown of an attack if Jon captures a live Giant.

The Bayesian network that represents the relationship between the events described above is shown below. Use the following notation for your variables: Jon Snow captures a live Giant ( $X_1$ ), Jon shows Cersei and Daenerys a live Giant ( $X_2$ ), Cersei agrees to help ( $X_3$ ), Daenerys agrees to help ( $X_4$ ) and Giants defeated ( $X_5$ ).



- [1pt] Write down the factorization of the above directed graphical model that represents the joint probability distribution  $P(X_1, X_2, X_3, X_4, X_5)$ .

$$P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_2)P(X_5|X_3, X_4)$$

- [5pts] For the following questions fill in the blank with the smallest set  $S$  of random variables needed to be conditioned on in order for the independence assumption to hold. For example  $X_i \perp X_j | S$ . What is the smallest set  $S$  that makes this statement true? The empty set  $\emptyset$  is a valid answer, additionally if the independence assumption cannot be satisfied no matter what we condition on then your answer should be 'Not possible'.

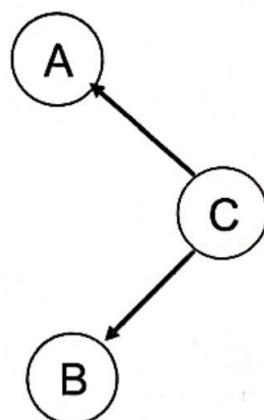
(a) [1pt]  $X_1 \perp X_3 |$

(b) [1pt]  $X_1 \perp X_5 |$

(c) [1pt]  $X_2 \perp X_4 |$  Not Possible

(d) [1pt]  $X_3 \perp X_4 |$   $X_2$

(e) [1pt]  $X_2 \perp X_5 |$   $X_3, X_4$



3. [2pt] Prove that A is conditionally independent from B, given C, in the tail to tail subgraph shown above. (i.e. prove  $p(a, b | c) = p(a | c)p(b | c)$ )

$$p(a, b | c) = \frac{p(a, b, c)}{p(c)} = \frac{p(c)p(a|c)p(b|c)}{p(c)} = p(a|c)p(b|c) \quad \text{Q.e.d.}$$

## 4 Learning Theory [18 pts]

1. [3 pt] Let  $\delta = |H|e^{-\epsilon m}$ . According to the PAC theorems discussed in class, which of the following is correct? Select one.

Select one:

- With probability at least  $1 - \delta$ , every hypothesis with training error at most  $\epsilon$  has true error 0.
  - With probability at least  $1 - \epsilon$ , a random hypothesis with training error 0 has true error at most  $\delta$ .
  - With probability at least  $1 - \delta$ , every hypothesis with training error 0 has true error at most  $\epsilon$ .
  - With probability at least  $1 - \epsilon$ , a random hypothesis with true error 0 has training error at most  $\delta$ .
2. [3 pt] Consider a decision tree learner applied to data where each example is described by 10 boolean variables  $X_1, X_2, \dots, X_{10}$ . What is the VC dimension of the hypothesis space used by this decision tree learner?

Fill in the blank:

1024

3. [4 pt] Consider instance space  $X$  which is the set of real numbers. What is the VC dimension of hypothesis class  $H$ , where each hypothesis  $h$  in  $H$  is of the form "if  $a < x < b$  or  $c < x < d$  then  $y = 1$ ; otherwise  $y = 0$ "? (i.e.,  $H$  is an infinite hypothesis class where  $a$ ,  $b$ ,  $c$ , and  $d$  are arbitrary real numbers.

Select one:

- 2
- 3
- 4
- 5
- 6

4. [3 pt] Alex is given a classification task to solve. He has no idea where to start, so he decided to try out a decision tree learner with 2 binary features  $X_1$  and  $X_2$ . He recently learned about PAC learning, and would like to know what is the minimum number ( $N$ ) of data points that would suffice for the PAC criterion with  $\epsilon = 0.1$  and  $\delta = 0.01$ .

Notice that a valid decision tree may or may not be full, meaning it doesn't have to split on all features.

Fill in the blank:

404

$$\begin{aligned} & \frac{1}{2\epsilon^2} [\log(1/\epsilon) + \log(\frac{2}{\delta})] \\ &= \frac{1}{2 \cdot 0.01} [\log 16 + \log 200] \end{aligned}$$

5. [2 pt] Sally thinks Alex shouldn't have used a decision tree with 2 binary features. Instead, she thinks it would be best to use logistic regression with 16 real-valued features in addition to a bias term. Sally overheard Alex talking about this cool concept called PAC learning and she too would like to use it to analyze her method. She first trains her logistic regression model on  $N$  examples to obtain a training error  $\hat{R}$ . What is the upper bound on the true error  $R$  in terms of  $\hat{R}$ ,  $\delta$ , and  $N$ . You may use big- $\mathcal{O}$  notation.

Fill in the blank:

$\hat{R} + O(\sqrt{\frac{1}{N}[\ln(1/\delta)]})$

6. [3 pt] Sally wants to argue her method has lower bound on the true error. Assuming Sally has obtained enough data points to satisfy PAC criterion with  $\epsilon = 0.1$  and  $\delta = 0.01$ . Which of the following is true?

Select one:

- Sally is wrong. Alex's method will always classify unseen data more accurately since it is simpler as it only needs 2 binary features.
- She must first regularize her model by removing 14 features to make any comparison at all.
- It is sufficient to show that the VC Dimension of her classifier is higher than Alex's, therefore having lower bound for the true error.
- It is necessary to show that the training error she achieves is lower than the training error Alex achieves.

## 5 Ensemble Methods [22pt]

1. [3pts] In the AdaBoost algorithm, if the final hypothesis makes no mistakes on the training data, which of the following is correct?

Select all that apply:

- Additional rounds of training can help reduce the errors made on unseen data.
- Additional rounds of training have no impact on unseen data.
- The individual weak learners also make zero error on the training data.
- Additional rounds of training always leads to worse performance on unseen data.

2. [3pts] Which of the following is true about ensemble methods?

Select all that apply:

- Ensemble methods combine together many simple, poorly performing classifiers in order to produce a single, high quality classifier.
- Neural networks can be used in the ensemble methods.
- For the weighted majority algorithm, the weak classifiers are learned along the way.
- For the weighted majority algorithm, we want to give higher weights to better performing models.

3. [2pt] True or False: In AdaBoost weights of the misclassified examples go up by the same multiplicative factor.

- True
- False

4. [2pt] True or False: AdaBoost will eventually give zero training error regardless of the type of weak classifier it uses, provided enough iterations are performed.

- True
- False

5. [12pts] In the last semester, someone used AdaBoost to train some data and recorded all the weights throughout iterations but some entries in the table are not recognizable. Clever as you are, you decide to employ your knowledge of Adaboost to determine some of the missing information.

Above, you can see part of table that was used in the problem set. There are columns for the Round # and for the weights of the six training points (A, B, C, D, E, and F) at the start of each round. Some of the entries, marked with "?", are impossible for you to read.

In the following problems, you may assume that non-consecutive rows are independent of each other, and that a classifier with error less than  $\frac{1}{2}$  was chosen at each step.

Round	$D_t(A)$	$D_t(B)$	$D_t(C)$	$D_t(D)$	$D_t(E)$	$D_t(F)$
1	? $\frac{1}{6}$	? $\frac{1}{6}$	$\frac{1}{6}$	? $\frac{1}{6}$	? $\frac{1}{6}$	? $\frac{1}{6}$
2	?	?	?	?	?	?
...						
219	?	?	?	?	?	?
220	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{7}{14}$	$\frac{1}{14}$	$\frac{2}{14}$	$\frac{2}{14}$
221	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{7}{20}$	$\frac{1}{20}$	$\frac{1}{4}$	$\frac{1}{10}$
...						
3017	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	0
...						
8888	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$$\begin{aligned}\varepsilon_1 &= \frac{1}{3} \\ d_1 &= \frac{1}{2} \ln\left(\frac{1-\frac{1}{3}}{\frac{1}{3}}\right) = \frac{1}{2} \ln 2 \\ e^{d_1} &= \sqrt{2}, e^{-d_1} = \frac{\sqrt{2}}{2} \\ x+x+x+X+2x+2X &= 1 \\ 117122 \\ 55142104\end{aligned}$$

- (a) [3pts] The weak classifier chosen in Round 1 correctly classified training points A, B, C, and E but misclassified training points D and F. What should the updated weights have been in the following round, Round 2? Please complete the form below.

Round	$D_2(A)$	$D_2(B)$	$D_2(C)$	$D_2(D)$	$D_2(E)$	$D_2(F)$
2	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$

- (b) [3pts] During Round 219, which of the training points (A, B, C, D, E, F) must have been misclassified, in order to produce the updated weights shown at the start of Round 220? List all the points that were misclassified. If none were misclassified, write 'None'. If it can't be decided, write 'Not Sure' instead.

Not Sure

- (c) [3pts] During Round 220, which of the training points (A, B, C, D, E, F) must have been misclassified in order to produce the updated weights shown at the start of Round 221? List all the points that were misclassified. If none were misclassified, write 'None'. If it can't be decided, write 'Not Sure' instead.

A, B, E

- (d) [3pts] You observe that the weights in round 3017 or 8888 (or both) cannot possibly be right. Which one is incorrect? Why? Please explain in one or two short sentences.

- Round 3017 is incorrect.
- Round 8888 is incorrect.
- Both rounds 3017 and 8888 are incorrect.

NOTE: Please do not change the size of the following text box, and keep your answer in it. Thank you!

Your answer.

3017: Since  $0 \leq \varepsilon_i \leq 1$ , then  $\alpha_i^t$  ranges from all real values, thus  $e^{\alpha_i^t}$  and  $e^{-\alpha_i^t}$  are always positive. It's impossible for  $D_t(F)$  to be reduced to 0.

8888:  $\frac{1}{8} + \frac{3}{8} + \frac{1}{8} + \frac{2}{8} + \frac{3}{8} + \frac{1}{8} = \frac{11}{8} > 1$ , which makes no sense all the weights must sum to 1.  $\sum_i D_t(i) = 1$ .

## 6 Recommender Systems [10pt]

1. [4pts] In which of the following situations will a collaborative filtering system be the most appropriate learning algorithm compared to linear or logistic regression?

Select all that apply:

- You manage an online bookstore and you have the book ratings from many users. For each user, you want to recommend other books she will enjoy, based on her own ratings and the ratings of other users.
- You run an online news aggregator, and for every user, you know some subset of articles that the user likes and some different subset that the user dislikes. You'd want to use this to find other articles that the user likes.
- You've written a piece of software that has downloaded news articles from many news websites. In your system, you also keep track of which articles you personally like vs. dislike, and the system also stores away features of these articles (e.g., word counts, name of author). Using this information, you want to build a system to try to find additional new articles that you personally will like.
- You manage an online bookstore and you have the book ratings from many users. You want to learn to predict the expected sales volume (number of books sold) as a function of the average rating of a book.

2. [3pts] What is the basic intuition behind matrix factorization?

Select all that apply:

- That content filtering and collaborative filtering are just two different factorizations of the same rating matrix.
- That factoring user and item matrices can partition the users and items into clusters that can be treated identically, reducing the complexity of making recommendations.
- The user-user and item-item correlations are more efficiently computed by factoring matrices.
- That user-item relations can be well described in a low dimensional space that can be computed from the rating matrices.

3. [3pts] When building a recommender system using matrix factorization, the regularized objective function we wish to minimize is:

$$J(, ) = \sum_{u,i \in S} (v_{ui} - u^T i)^2 + \lambda (\sum_u \|u\|^2 + \sum_i \|i\|^2)$$

where  $u$  is the  $u$ th row of  $H$  and the vector representing user  $u$ ;  $i$  is the  $i$ th row of  $H$  and the vector representing item  $i$ ;  $S$  is the index set of observed user/item ratings in the training set; and  $\lambda$  is the weight of the L2 regularizer. One method of solving this optimization problem is to apply Block Coordinate Descent. The algorithm proceeds as shown below:

- while not converged:
  - for  $u$  in  $\{1, \dots, N_u\}$ :
    - \*  $u' \leftarrow \operatorname{argmin}_{u'} J(, )$
  - for  $i$  in  $\{1, \dots, N_i\}$ 
    - \*  $i' \leftarrow \operatorname{argmin}_{i'} J(, )$

Doing so yields an algorithm called Alternating Least Squares (ALS) for matrix factorization. Which of the following is equal to the transpose of  $\operatorname{argmin}_{u'} J(, )$ ?

Select one:

- $v_u H(H^T H + \lambda I)^{-1}$
- $(H^T H + \lambda I)^{-T} v_u H$
- $v_u H(H^T H + \lambda I)^{-T}$
- $v_u H(H^T H)^{-1}$

**Collaboration Questions** Please answer the following:

After you have completed all other components of this assignment, report your answers to the collaboration policy questions detailed in the Academic Integrity Policies 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.

**Solution:**

I checked Piazza and used this page for calculating the cardinality of hypothesis space of decision trees.

<https://piazza.com/class/kjvu0xh54r72d1?cid=1162>

I also used the following slides also from CMU to review the algorithm of Adaboost

<https://www.cs.cmu.edu/~aarti/Class/10701/slides/Lecture10.pdf>