Sample Final Exam (with Solutions)

CSCI 1420 - Spring 2021

The instructions below are just to illustrate the format of the true final. So you can post on Piazza, etc. about the sample final.

Instructions

Timeline: The final will be posted on the course homepage no later than noon Eastern U.S. time on Monday, April 19. It must be submitted on Gradescope by 11:59 PM Eastern U.S. time on Tuesday, April 20. No late days may be used on the final. For every minute past the deadline it is late, one percentage point will be deducted from the grade.

Exam Format: There are eight problems, each worth 1/7 of the exam. You may choose one problem to count as extra credit, worth 1/2 of a regular problem. To indicate your choice, mark the blank line where indicated for that problem. If you do not make a choice, or your choice is otherwise unclear, the last problem will be treated as extra credit. We will not adjust a student's selection to optimize their score.

Submission Format: You may submit a PDF using the provided Latex, or you may submit handwritten answers. You are encouraged to use Latex if possible, as any illegible parts of answers will be marked incorrect.

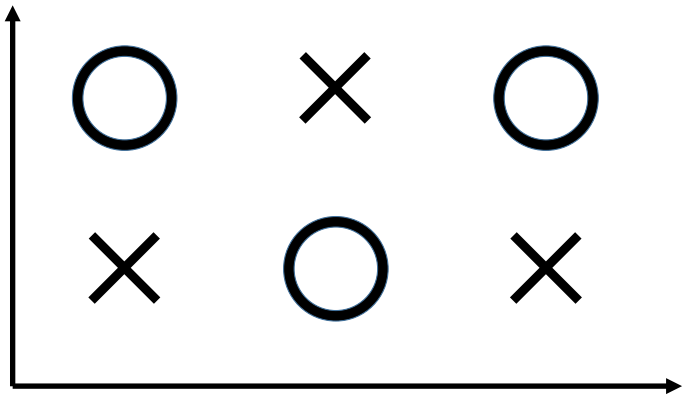
Academic Integrity: *The course collaboration policy does not apply to this exam.* Instead, you may not communicate with anyone other than course staff about the exam in any way. You may consult the course textbook, course notes, slides, homeworks, recorded lectures and discussion sessions, or existing messages on Piazza. You may not post anything new that is public to Piazza during the exam period. (See below.) Violating these instructions will be considered academic dishonesty.

Getting Help: If you have any questions about the content of the exam or technical difficulties, please first consult the "Official FAQ" that will be pinned on Piazza. If your question is not answered there, please email the HTA list: `cs1420headtas@lists.brown.edu` . We will respond as soon as possible, but please keep in mind that latency of up to 20 minutes is reasonable. In addition, the mailing list is only monitored from 9 AM to midnight Eastern U.S. time, so please plan accordingly.

If you have any other issues or concerns, such as challenging or unexpected circumstances, please contact Steve directly: `stephen_bach@brown.edu` .

Problem 1 (Representations)

Consider the following training data set with two classes in $\mathbb{R}^2$.



For each of the following hypothesis classes, state whether it can perfectly fit the above training data. Explain your answer.

a. Logistic Regression:

Solution: No, the data is not linearly separable. For example, the bottom left and top middle examples cannot be classified correctly without misclassifying the top left and/or bottom middle examples.

b. Greedy split decision tree with an arbitrary number of layers:

Solution: Yes, such a decision tree can divide the input space into an arbitrary number of rectangles with axis-aligned boundaries, because there is no limit on the number of layers.

c. Support Vector Machine with Linear Kernel:

Solution: No, the data is not linearly separable, the same as part a.

Problem 2 (Loss Functions)

Consider the 1-dimension hinge loss for a single training example $x = 1$ and $y = 1$ and a homogeneous halfspace with one weight $w \in \mathbb{R}$:

$$\ell^{\text{hinge}}(w) = \begin{cases} 0 & \text{if } w \geq 1 \\ 1 - w & \text{if } w < 1 \end{cases} = \max\{0, 1 - w\}$$

$\ell^{\text{hinge}}(w)$ is a convex function that upper bounds another loss function called the ramp loss:

$$\ell^{\text{ramp}}(w) = \begin{cases} 0 & \text{if } w \geq 1 \\ 1 - w & \text{if } 0 < w < 1 \\ 1 & \text{if } w \leq 0 \end{cases}$$

a. Show that $\ell^{\text{ramp}}(w)$ is a non-convex function.

Solution: It is sufficient to show there exists $w_1, w_2 \in \mathbb{R}$ and $\alpha \in [0, 1]$ such that

$$\ell^{\text{ramp}}(\alpha w_1 + (1 - \alpha)w_2) > \alpha \ell^{\text{ramp}}(w_1) + (1 - \alpha)\ell^{\text{hinge}}(w_2)$$

Let $w_1 = -1$, $w_2 = 1$, and $\alpha = 0.5$. Then $\ell^{\text{ramp}}(\alpha w_1 + (1 - \alpha)w_2) = 1$ and $\alpha \ell^{\text{ramp}}(w_1) + (1 - \alpha)\ell^{\text{ramp}}(w_2) = 0.5$, completing the proof.

Arguing that the ramp loss has multiple local minima or that its epigraph is a nonconvex set are also acceptable.

b. When learning a halfspace over many training examples, what could be an advantage of using the hinge loss over the ramp loss, and vice versa, what could be an advantage of using the ramp loss over the hinge loss?

Solution: An advantage of the hinge loss is that it is convex, meaning that there will be one globally minimal value of the empirical risk over all weights.

An advantage of the ramp loss is that it bounds how much loss can be contributed to the total by a single example. It is therefore less sensitive to outlying data points that are far from the correct side of the decision boundary.

Other answers are also possible.

Problem 3 (Optimizers)

Consider the hypothesis class of two-dimensional thresholds, $H = \{h_{a,b} : a \in \mathbb{R} \text{ and } b \in \mathbb{R}\}$ where:

$$h_{a,b}(\mathbf{x}) = \begin{cases} 1 & \text{if } x_1 \leq a \text{ and } x_2 \leq b \\ -1 & \text{otherwise} \end{cases}$$

and $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \{-1, 1\}$.

Describe an algorithm for computing the ERM for this class in the realizable case. (You can assume 0-1 loss, although the solution will be the same for any reasonable loss function.) State the computational complexity of the algorithm in the context of a training data set of size m.

> **Solution:** Let a be the largest first element (x_1) of an example $\mathbf{x}$ with a label of 1, and let b be the largest second element (x_2) of an example with a label of 1. (If no examples have a label of 1, set a and b to minimum value that can be represented.) $\mathcal{O}(m)$ to find the maximum values.

Problem 4 (Empirical and Expected Risk)

For this problem, we are looking for responses that both indicate your assessment as to a possible accuracy change and your understanding of the algorithm that led to this assessment. Answers should be two or three sentences long and focus on the relevant and important issue.

a. We have trained a logistic regression model (binary vector input, binary label, no regularization) on a data set. Then, we create a new data set that is identical to the original but it includes a new feature that is set uniformly at random, with no strong correlation to any of the other features or the label, and run the same learning algorithm again. What would you expect to happen to the training ***and*** testing losses of the new learned model?

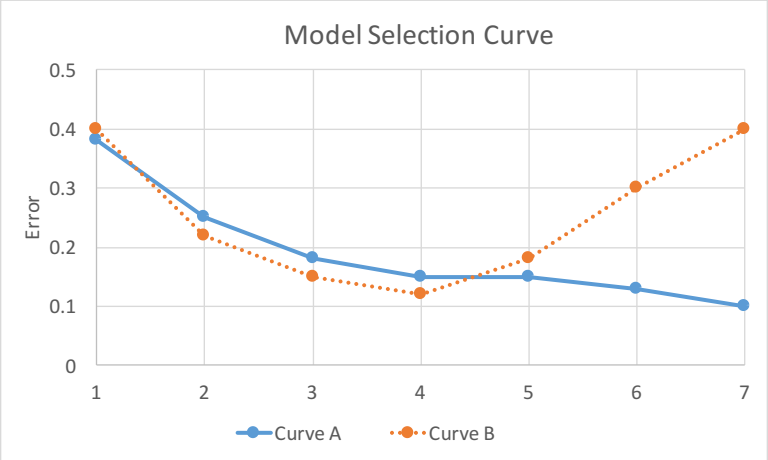
> **Solution:** We expect the training loss to decrease, because we have increased model complexity. Even though the new feature is not strongly correlated with the label, the additional feature corresponds to an additional parameter that the learning algorithm can use to fit the data.
>
> However, we expect the testing loss to increase, because the learning algorithm will assign a non-zero weight to a feature that is random noise, i.e., it overfits at least slightly to that feature.

b. We have trained a logistic regression model (binary vector input, binary label, no regularization) on a data set. Then, we create a new data set that is identical to the original but includes a new attribute that is the Boolean negation of the label and run the same learning algorithm again. What would you expect to happen to the training ***and*** testing losses of the new learned model?

> **Solution:** We expect the training loss of our algorithm to approach 0 as we train. This is because there is one feature that exactly corresponds to the opposite label, so its partial derivative is always negative (rigorous proof not needed). Therefore, its weight will decrease on each gradient update.
>
> At the end of training, we expect our learned model to have loss approaching zero on both the training and test data, because it always predicts the opposite value of the feature as the label, and a very large, negative (possibly overflowing) weight for that feature.

Problem 5 (Model Selection)

Consider the following (partially labeled) model selection curve for boosted halfspace classifiers learned with AdaBoost:



a. Describe the likely interpretation of the following parts of the above figure, based on the bias-complexity tradeoff. Include a specific statement of what that part of the figure represents, and provide a brief explanation justifying your interpretation.

The horizontal axis (with values 1 through 7):

Solution: The number of iterations of AdaBoost determines the complexity of the learned hypothesis. It is likely that this axis is the number of iterations because a model selection curve is used to choose a hyperparameter that trades off between bias and complexity.

Curve A (solid line):

Solution: It is likely the training error, since it continues to decrease as model complexity increases. Making a model more complex decreases, but does not increase, the training error.

Curve B (dotted line):

Solution: It is likely the validation error, because it starts to decrease as model complexity increases, but eventually grows again. Making a model more complex can increase the estimation error. If it increases more than the approximation error decreases, the error on the held-out validation set will grow.

b. If you were using the above model selection curve to choose a specific value on the horizontal axis to use for the corresponding task, which would you choose? Why?

Solution: 4 is a good choice because it minimizes the validation error.

Problem 6 (Generative Models)

Consider the following training data for binary classification of two-bit vectors, i.e., $\mathcal{X} = \{0,1\}^2$ and $\mathcal{Y} = \{0,1\}$:

x_1	x_2	y
0	0	1
0	1	0
1	0	0
1	1	1
1	0	0

x_1	x_2	y
1	0	1
1	0	0
0	1	0
0	1	0
0	0	0

Using a maximum likelihood Naive Bayes model with Laplace smoothing of 1, what is the probability $\mathcal{P}(y = 1|x_1 = 1, x_2 = 0))$, i.e., the probability that a test example $(1, 0)$ has the label 1? (Assume that the distributions $\mathcal{P}(y)$ and all $\mathcal{P}(x_i|y)$ are Bernoulli, i.e., binary.)

Solution:

First compute

$$\mathcal{P}(y) = \frac{4}{12} = 0.333$$

$$\mathcal{P}(x_1 = 1|y = 1) = \frac{3}{5} = 0.600$$

$$\mathcal{P}(x_2 = 0|y = 1) = \frac{3}{5} = 0.600$$

$$\mathcal{P}(x_1 = 1|y = 0) = \frac{4}{9} = 0.444$$

$$\mathcal{P}(x_2 = 0|y = 0) = \frac{5}{9} = 0.556$$

Next, compute

$$\mathcal{P}(x_1 = 1, x_2 = 0, y = 1) = 0.600 \cdot 0.600 \cdot 0.333 = 0.120$$

$$\mathcal{P}(x_1 = 1, x_2 = 0, y = 0) = 0.444 \cdot 0.556 \cdot 0.667 = 0.165$$

By normalizing, obtain

$$\mathcal{P}(y|x_1 = 1, x_2 = 0) = \frac{0.120}{0.120 + 0.164} = 0.421$$

Problem 7 (Unsupervised Learning)

A mixture of K Gaussians learns a set of K normal distributions as the generative model for a given set of data. What if we wanted to learn a generative model that uses *uniform* distributions over intervals in $\mathbb{R}$ instead? In this setup, there is still a multinomial distribution over mixture components, but instead of each component being a Gaussian distribution, it is a uniform distribution with a density defined by the parameters a, b where $a < b$:

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Suppose we want to learn such a K uniform distributions mixture model where $K = 3$ using expectation maximization. Further suppose our data comprises four points in $\mathbb{R}$: 0.13, 0.21, 0.57, 0.82.

a. The E step, given a set of parameters and the data, should compute a probability that each data point came from each interval. Fill in the conditional probability $p(\cdot|x)$ for each of the data points in the column corresponding to each mixture component A, B, and C, with the specified current estimates of the parameters (a, b). Assume that the current estimates of the prior are $p(A) = p(B) = p(C) = \frac{1}{3}$.

Solution:

	A	B	C
	$(0.1, 0.3)$	$(0.2, 0.7)$	$(0.8, 0.9)$
0.13	1	0	0
0.21	$\frac{5}{7}$	$\frac{2}{7}$	0
0.57	0	1	0
0.82	0	0	1

b. The M step should find the maximum likelihood intervals for each component with respect to the conditional probabilities computed in the E step. Suppose after a different E step from the one above, the following probabilities were obtained. Compute the result of the M step using this table. Your answer should be estimates for 9 parameters: $a_A, b_A, a_B, b_B, a_C, b_C, p(A), p(B)$, and $p(C)$.

	A	B	C
0.13	0.0	0.0	1.0
0.21	0.2	0.8	0.0
0.57	0.7	0.3	0.0
0.82	0.1	0.0	0.9

Solution:

$$a_A = .21, b_A = .82$$
$$a_B = .21, b_B = .57$$
$$a_C = .13, b_C = .82$$
$$p(A) = \frac{1}{4}, \;\; p(B) = \frac{1.1}{4}, \;\; p(C) = \frac{1.9}{4}$$

Problem 8 (VC Dimension)

Consider (again, see problem 3) the hypothesis class of two-dimensional thresholds, $H = \{h_{a,b} : a \in \mathbb{R} \text{ and } b \in \mathbb{R}\}$ where:

$$h_{a,b}(\mathbf{x}) = \begin{cases} 1 & \text{if } x_1 \leq a \text{ and } x_2 \leq b \\ -1 & \text{otherwise} \end{cases}$$

and $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \{-1, 1\}$.

What is the VC dimension of this hypothesis class? Provide a complete proof.

Solution: Recall that to show that the VC dimension of this hypothesis is k, we need to show that

1. H shatters some set of size k.
2. Every set of size $k+1$ cannot be shattered by H.

We will show that the VC dimension of H is 2:

1. H shatters some set of size 2. Consider the set of points $X = \{(1,0),(0,1)\}$. H shatters X because for every mapping from X to $\mathcal{Y}$, we can find appropriate a and b values to correctly classify the points according to that mapping.

$$\begin{array}{c|cc} \mathbf{x} & (1,0) & (0,1) \\ \hline y & -1 & -1 \end{array} \implies \begin{array}{cc} a & b \\ \hline 0 & 0 \end{array} \tag{1}$$

$$\begin{array}{c|cc} \mathbf{x} & (1,0) & (0,1) \\ \hline y & 1 & -1 \end{array} \implies \begin{array}{cc} a & b \\ \hline 1 & 0 \end{array} \tag{2}$$

$$\begin{array}{c|cc} \mathbf{x} & (1,0) & (0,1) \\ \hline y & -1 & 1 \end{array} \implies \begin{array}{cc} a & b \\ \hline 0 & 1 \end{array} \tag{3}$$

$$\begin{array}{c|cc} \mathbf{x} & (1,0) & (0,1) \\ \hline y & 1 & 1 \end{array} \implies \begin{array}{cc} a & b \\ \hline 1 & 1 \end{array} \tag{4}$$

2. Every set of size 3 cannot be shattered by H. Consider any three points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ all in $\mathbb{R}^2$. Without loss of generality assume that they are partially ordered along each dimension, so that $x_{11} \leq x_{21}$ or $x_{11} \leq x_{31}$, as well as $x_{12} \leq x_{22}$ or $x_{12} \leq x_{32}$.

   Now, let $\mathbf{x}_1$ be labeled -1, and let $\mathbf{x}_2$ and $\mathbf{x}_3$ be labeled 1. To correctly classify $\mathbf{x}_2$ and $\mathbf{x}_3$, we must choose a and b such that $x_{21} \leq a$ and $x_{31} \leq a$, as well as $x_{22} \leq b$ and $x_{32} \leq b$. However, any such a and b will misclassify $\mathbf{x}_1$, so we have shown that no set of size 3 can be shattered by H.

Thus, we have shown that the VC dimension of H is 2.