Final Exam

CSCI 1420 - Spring 2021

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.

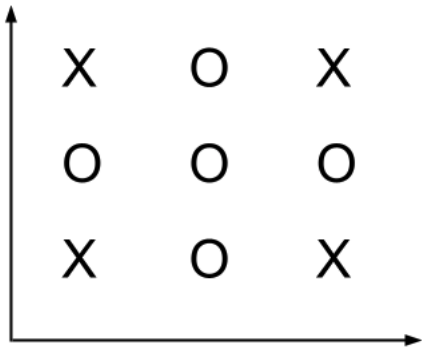
Getting Help: If you have any questions about the content of the exam or technical difficulties, please first consult the "Official FAQ" that will be pinned on Piazza. If your question is not answered there, please email the HTA list: `cs1420headtas@lists.brown.edu` . We will respond as soon as possible, 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)

_____ Mark here to count this problem as extra credit.

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. An ensemble of 4 equally-weighted decision stumps:

b. 3-Nearest Neighbors:

c. Neural network with step activation functions, 2 layers, and an arbitrary number of hidden neurons:

Problem 2 (Loss Functions)

_____ Mark here to count this problem as extra credit.

a. Suppose we learned a linear regression model using the absolute error loss (or l1 loss). Let d_1 be the maximum distance between the learned hyperplane and any point $\mathbf{x}$ in the training data. In other words, d_1 is the distance between the hyperplane of best fit and the farthest outlier. Suppose we learn another linear regression model on the same data using the squared loss (or l2 loss). Let d_2 be defined in an analogous way, the distance between the hyperplane of best fit and the farthest outlier. How do you expect d_1 and d_2 to compare? Why?

b. What is an advantage of using the 0-1 loss for binary classification, versus the log loss? Vice versa, what is an advantage of using the log loss for binary classification, versus the 0-1 loss?

Problem 3 (Optimizers)

_____ Mark here to count this problem as extra credit.

Consider the hypothesis class of d-dimensional segmentations into three sections, $H = \{h_{a,b,\mathbf{s}} : a \in \mathbb{R}, b \in \mathbb{R}, \text{ and } \mathbf{s} \in \{-1, 1\}^3\}$, where:

$$h_{a,b,\mathbf{s}}(\mathbf{x}) = \begin{cases} s_1 & \text{if } ||\mathbf{x}||_2 \leq a \\ s_2 & \text{if } ||\mathbf{x}||_2 > a \text{ and } ||\mathbf{x}||_2 \leq b \\ s_3 & \text{otherwise} \end{cases}$$

and $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$, and $a < b$.

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. Your algorithm should run in at most $\mathcal{O}(m \log m)$ time.

Problem 4 (Empirical and Expected Risk)

_____ Mark here to count this problem as extra credit.

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 one or two sentences long and focus on the relevant and important issue.

a. Suppose we train an AdaBoost ensemble of K halfspaces and achieve a non-zero empirical risk. We then continue to run one more iteration of the algorithm to add one more member to the ensemble. What will likely happen to the empirical risk of this new classifier (the one with the new member), relative to the old one (the one without the new member)? Why?

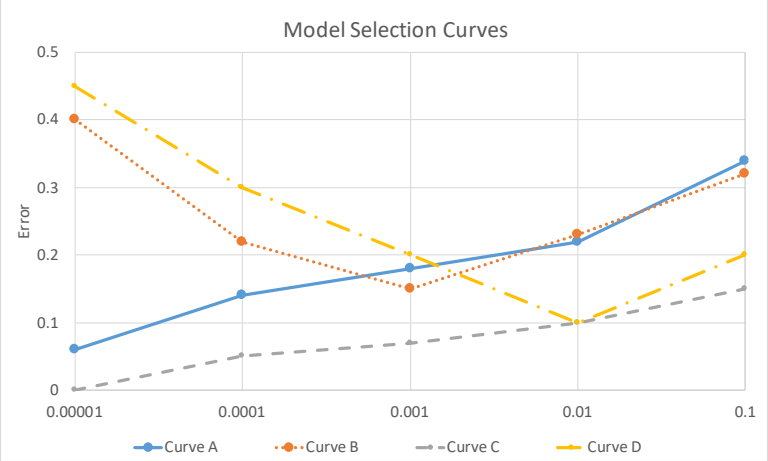
b. Suppose we train a two-layer neural network to perform classification and achieve a non-zero empirical risk. We then remove the hidden layer and keep all of the other parameters of the classifier the same. What will likely happen to the empirical risk of this new classifier, relative to the old one? Why?

c. Suppose we sample a large amount of data from an arbitrary (i.e., not necessarily Naive Bayes) generative model $\mathcal{P}(\mathbf{x}, y)$. We then train a logistic regression classifier on the data. We also copy the generative model, randomly re-initialize its parameters, and re-estimate them using maximum likelihood estimation on the same data. We then use the resulting posterior $\mathcal{P}(y|\mathbf{x})$ as a second classifier. How will the expected risk of the two classifiers likely compare? Why?

Problem 5 (Model Selection)

_____ Mark here to count this problem as extra credit.

Consider the following (partially labeled) model selection curves for two ℓ_2-regularized logistic regression classifiers:



The training of the two classifiers differed only in the input data. The first model ("Model 1") was trained by regularized risk minimization using the log loss. The second model ("Model 2") was trained in an identical fashion, but had some features *removed* from each example.

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 10^{-5} through 10^{-1}):

Curve A (solid line):

Curve B (dotted line):

Curve C (dashed line):

Curve D (dashed/dotted line):

Problem 6 (Generative Models)

_____ Mark here to count this problem as extra credit.

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
1	0	1
0	1	0
1	1	0
0	0	0
1	1	1

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

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

Problem 7 (Unsupervised Learning)

_____ Mark here to count this problem as extra credit.

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.

Suppose we applied principal component analysis (PCA) to a data set in $\mathbb{R}^5$ before learning a halfspace classifier on the reduced data set.

a. After decomposing the matrix A, we obtain the following eigenvalues: 4, 3.8, 0.05, 0.01, and 0.005. Which value of K, the dimensionality of the linear subspace to which we will reduce the data, would you pick? Why?

b. Suppose you *increased* the value of K beyond what you picked in part a. What would you expect to happen to the training error of the halfspace classifier? Why?

c. Suppose you would rather not spend your time writing the PCA code unless it can potentially help reduce the testing error. Is there a scenario where you would want to implement and apply PCA to your data? If yes, what is it? Why?

d. Suppose that one of the original five attributes is a legally protected attribute such as a person's race or gender. Will training a halfspace classifier on the dimensionality-reduced data produced by PCA prevent the classifier from discriminating based on that attribute? Why? (For this question, discrimination is defined as an ϵ difference in the probability of outputing 1 conditioned on any two different values of the protected attribute x, i.e., $|p(y = 1|x = a) - p(y = 1|x = b)| > \epsilon$.)

Problem 8 (VC Dimension)

_____ Mark here to count this problem as extra credit.

Consider (again, see problem 3) the hypothesis class of d-dimensional segmentations into three sections, $H = \{h_{a,b,\mathbf{s}} : a \in \mathbb{R}, b \in \mathbb{R}, \text{ and } \mathbf{s} \in \{-1,1\}^3\}$, where:

$$h_{a,b,\mathbf{s}}(\mathbf{x}) = \begin{cases} s_1 & \text{if } ||\mathbf{x}||_2 \leq a \\ s_2 & \text{if } ||\mathbf{x}||_2 > a \text{ and } ||\mathbf{x}||_2 \leq b \\ s_3 & \text{otherwise} \end{cases}$$

and $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$, and $a < b$.

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