CIS 520: Machine Learning Spring 2021

Final Exam
Version D

University of Pennsylvania

Total Time: 120 minutes **Total Points:** 75

Read the following instructions carefully before starting.

1. The exam is **open-notes, open-book, closed-electronics, and closed-communication**. This means that you can refer to your notes or books, but you are not allowed to use a calculator, phone, computer, internet etc during the exam, and you are not allowed to talk to anyone else about the exam (through phone, text, chat, email, personal conversation, or any other means) until the 24-hour Gradescope submission period is over.*
2. This exam contains 5 problems:
   - Problem 1 contains 9 short questions worth 2 points each (total 18 points).
   - Problem 2 contains 9 short questions worth 3 points each (total 27 points).
   - Problems 3–5 are long answer questions (10 points each).
3. For all problems, you must provide clear explanations and detailed calculations. You must also highlight your final answers.
4. **For each of the 5 problems, you must write your solution on a new sheet of paper, with your name, 8-digit Penn ID, exam version (A/B/C/D), and problem number marked clearly at the top of each sheet. Your solutions must be handwritten and must be submitted as a single PDF file on Gradescope before the announced submission deadline. You must clearly indicate the page number for each problem when submitting on Gradescope in order for our team to be able to grade your solutions (and your writing must be clearly legible).****
5. Good luck!

---

* If your notes are stored on a tablet or computer and you want to refer to them during the exam, or if you want to write your solutions on a tablet, then you must **disconnect the device from the internet throughout the duration of the exam** and use it only for the purposes of referring to your notes or writing your solutions.

** If you are writing your solutions on physical paper, then after finishing the exam, you can either scan your answer sheets or take clear photos of them, and then upload everything as a single PDF file.

PROBLEM 1 – Short Questions [$9 \times 2 = 18$ **points**]

Please include explanations for your answers.

(a) [2 points] Consider a binary classification task with 150 features: $\mathcal{X} = \mathbb{R}^{150}$, $\mathcal{Y} = \{\pm 1\}$. You are given a training sample containing 75 labeled examples. You are considering two possible models to use: linear logistic regression and kernel logistic regression (using RBF kernel with parameter $\sigma = 1$). Which model learns fewer parameters (ignore bias/threshold terms)? How many fewer?

(b) [2 points] Consider a regression problem with 15 features: $\mathcal{X} = \mathbb{R}^{15}, \mathcal{Y} = \mathbb{R}$. Suppose you want to learn a one-hidden-layer neural network model with 5 hidden units, each with a sigmoid activation function. How many parameters (weights) do you need to learn? For simplicity, ignore bias/threshold terms.

(c) [2 points] Consider a binary classification problem with 2 features: $\mathcal{X} = \mathbb{R}^2, \mathcal{Y} = \{\pm 1\}$. You are given a training set with the following features and labels:

$\mathbf{x}$	y
$(-1, 0)$	$+1$
$(2, 0)$	-1
$(1, 1)$	-1
$(-2, 2)$	-1
$(0, -1)$	-1
$(0, 3)$	$+1$

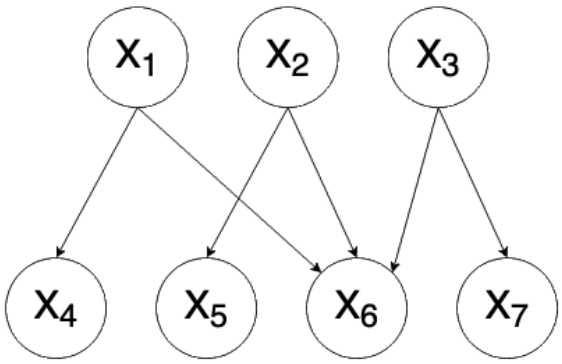
What is the estimated class probability $\widehat{\eta}(\mathbf{x}) = \widehat{\mathbf{P}}(Y = +1 | X = \mathbf{x})$ returned by a 3-nearest neighbor classifier (using Euclidean distance), for the point $\mathbf{x} = (1, 0)$?

(d) [2 points] Suppose you are running AdaBoost on a training set $((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \{\pm 1\})^m$, and your third weak classifier h_3 makes the following predictions on the first five training examples:

x_i	y_i	$h_3(x_i)$
x_1	-1	-1
x_2	-1	-1
x_3	$+1$	$+1$
x_4	$+1$	-1
x_5	-1	$+1$

Which of these training examples will have higher weight in the fourth iteration than they had in the third iteration?

(e) [2 points] Given the Bayesian network below, write down the joint probability distribution $p(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ as a product of 7 factors.



(f) [2 points] Consider an active learning setup for binary classification with labels $\{\pm 1\}$ and 0-1 loss. You are given a small labeled training set, from which you learn a logistic regression model. You are also given four more unlabeled data points, $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$, and are allowed to query the label of one of these. Your logistic regression model predicts the probabilities of each of these instances having label $+1$ as follows:

$$\widehat{\eta}(\mathbf{x}_1) = 0.31\,; \quad \widehat{\eta}(\mathbf{x}_2) = 0.7\,; \quad \widehat{\eta}(\mathbf{x}_3) = 0.68\,; \quad \widehat{\eta}(\mathbf{x}_4) = 0.29\,.$$

If you use an uncertainty sampling approach, which of the above instances would be chosen to query a label for?

(g) [2 points] Suppose you are given a (fully specified) Markov decision process with state space $\mathcal{S} = \{1, 2, 3\}$ and action space $\mathcal{A} = \{\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}\}$. You calculate the optimal state-action value $Q^*(s, a)$ for each state-action pair (s, a) to be as follows:

	a	b	c	d
1	1.6	2.6	2.3	2.8
2	3.1	4.5	2.8	4.0
3	2.0	1.8	2.3	3.9

Find an optimal deterministic policy $\pi^* : \mathcal{S} \to \mathcal{A}$.

(h) [2 points] Consider a collaborative filtering problem with 3,000 users and 2,000 movies, where a subset of user-movie ratings are observed. Consider learning a latent factor model via matrix factorization methods in order to predict the unobserved user-movie ratings. If you use 5 latent factors, what is the total number of parameters to be learned (ignore any bias terms)?

(i) [2 points] You want to learn a Latent Dirichlet Allocation (LDA) model using a training set of 200 documents. Each of the documents has 1000 words. The vocabulary is of size 500. You decide that the target number of topics is 50. How many parameters are needed for an LDA model?

PROBLEM 2 – Short Questions [$9 \times 3 = 27$ **points**]

Please include explanations for your answers.

(a) [3 points] Consider a binary classification problem with instance space $\mathcal{X} = \mathbb{R}$ and label space $\mathcal{Y} = \{+1, -1\}$. Suppose that $\mathbf{P}(Y = +1) = 0.75$, and that for each $y \in \mathcal{Y}$, the class-conditional density of x given $Y = y$ is $\mathcal{N}(2y, 1)$ (Normal distribution with mean $2y$ and variance 1). What is the Bayes optimal classifier $h^*(x)$?

(b) [3 points] Consider a 5-class classification task involving 100 features: $\mathcal{X} = \mathbb{R}^{100}$, $\mathcal{Y} = \{1, 2, \ldots, 5\}$. How many parameters are needed for a Gaussian Naïve Bayes classifier?

(c) [3 points] Consider a regression problem with $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \mathbb{R}$. Let D be a probability distribution on $\mathcal{X} \times \mathcal{Y}$ under which labeled examples $(\mathbf{x}, y)$ are generated as follows:

- First an instance $\mathbf{x}$ is drawn by drawing each of the two features x_1, x_2 randomly and independently from a standard normal distribution, $\mathcal{N}(0, 1)$;
- Given $\mathbf{x} = (x_1, x_2)^\top$, a label $y \in \mathbb{R}$ is generated randomly as follows:

$$y = 3x_1 - 2x_2 + \epsilon\,,$$

  where $\epsilon \sim \text{Uniform}(1, 2)$, a uniform distribution.

Now consider the specific instance $\mathbf{x} = \begin{pmatrix} 2 \\ -3 \end{pmatrix}$. For this instance, what is the predicted label $f^*(\mathbf{x})$ according to the optimal regression model f^* (under squared loss)?

(d) [3 points] You have fitted a soft-margin SVM to a dataset $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)) \in (\mathcal{X} \times \{\pm 1\})^m$, producing a weight vector $\widehat{\mathbf{w}}$ and a bias term $\widehat{b}$. Suppose you have the following values for $\widehat{\mathbf{w}}^T \mathbf{x}_i + \widehat{b}$ for the first five training examples:

$\mathbf{x}_i$	y_i	$\widehat{\mathbf{w}}^T \mathbf{x}_i + \widehat{b}$
$\mathbf{x}_1$	-1	-3
$\mathbf{x}_2$	-1	0
$\mathbf{x}_3$	$+1$	1
$\mathbf{x}_4$	$+1$	2
$\mathbf{x}_5$	$+1$	-2

Which of these training examples are support vectors?

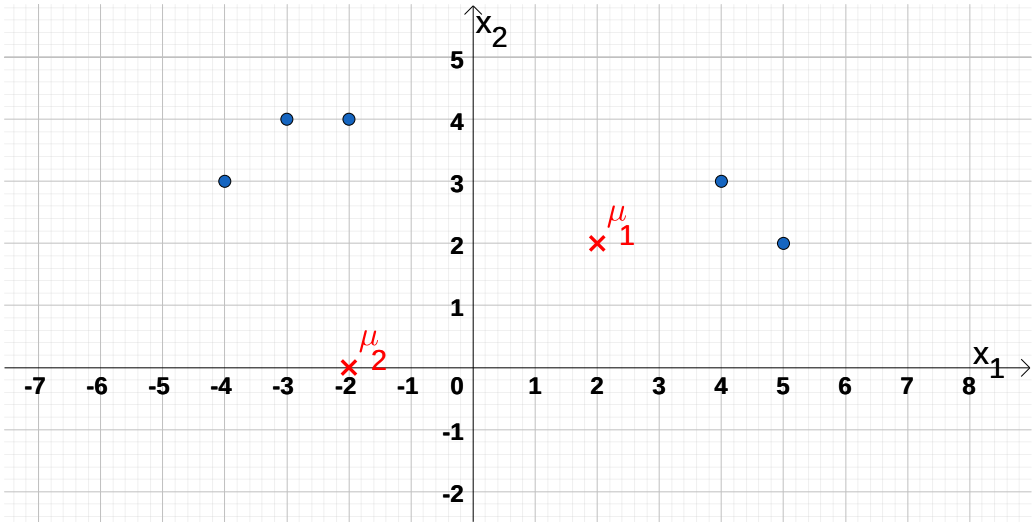
(e) [3 points] A group of students created a machine learning model that predicts whether a person has pets or not. Then they tested the model on 1000 volunteers and predicted that only 500 of them have pets. Assume the precision and accuracy of their model are 0.6 and 0.6. How many people actually have pets but were predicted wrongly?

(f) [3 points] Consider an unlabeled data set containing m 2-dimensional points $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^2$, with sample mean $\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i$. You form the mean-centered data matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{m \times 2}$ containing $\tilde{\mathbf{x}}_i^\top = (\mathbf{x}_i - \bar{\mathbf{x}})^\top$ in row i, and calculate the matrix $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$ to be

$$\begin{bmatrix} 6 & \sqrt{6} \\ \sqrt{6} & 5 \end{bmatrix}.$$

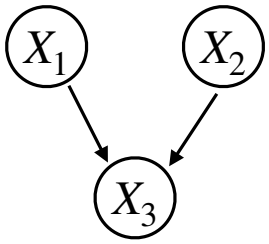
What fraction of the total variance in the data is explained by the first principal component?

(g) [3 points] Consider a clustering problem with 2 features: $\mathcal{X} = \mathbb{R}^2$. You are given a training set consisting of the following data points (shown as blue dots):



You want to run the K-Means algorithm with $K = 2$ and the initialization μ_1, μ_2 (shown as red crosses in the figure). What is the value of μ_1 after the first iteration of the algorithm?

(h) [3 points] Consider three random variables X_1, X_2, X_3, each of which takes one of 4 distinct values. Suppose their joint probability distribution is known to factor according to the Bayesian network structure below:



Given this information, how many parameters are needed to specify the joint probability distribution?

(i) [3 points] Consider running the least mean squares algorithm with parameter $\eta = 1$ for an online regression task where instances have two real-valued features: $\mathcal{X} = \mathbb{R}^2$, $Y = \mathbb{R}$. Suppose that at the beginning of round t, the current weight vector is

$$\mathbf{w}_t = (-3, 2)^\top$$

The instance received on this round is

$$\mathbf{x}_t = (1, -1)^\top$$

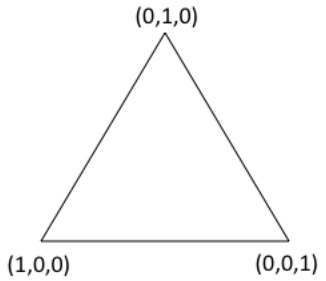
After the algorithm makes its prediction $\widehat{y}_t = -5$, the true label received is $y_t = -4$. What is the weight vector $\mathbf{w}_{t+1}$ in the next round?

PROBLEM 3 [10 points]

Consider a 3-class classification problem with instance space $\mathcal{X}$ and label and prediction spaces $\mathcal{Y} = \widehat{\mathcal{Y}} = \{1, 2, 3\}$, in which the loss on predicting $\widehat{y}$ when the true label is y, $\ell(y, \widehat{y})$, is given by the following loss matrix:

	$\widehat{y} = 1$	$\widehat{y} = 2$	$\widehat{y} = 3$
$y = 1$	0	2	2
$y = 2$	1	0	1
$y = 3$	2	2	0

Derive a Bayes optimal classifier $h^* : \mathcal{X} \to \mathcal{Y}$ for this loss. Express your answer in terms of the class probabilities $\eta_1(x) = \mathbf{P}(Y = 1|X = x)$, $\eta_2(x) = \mathbf{P}(Y = 2|X = x)$, and $\eta_3(x) = \mathbf{P}(Y = 3|X = x)$, giving clear conditions on these class probabilities under which the classifier should predict class 1, 2 or 3. Explain your reasoning and show your calculations. Also draw "decision regions" in the probability simplex below indicating the sets of class probability vectors $(\eta_1(x), \eta_2(x), \eta_3(x))$ for which h^* should predict 1, 2, and 3.



PROBLEM 4 [10 points]
Assume that real DNA sequences can be described by a hidden Markov model with hidden states representing different types of nucleotide composition. Consider an HMM that includes two hidden states H and L for higher and lower nucleotide composition content, respectively. There are 4 kinds of nucleotides which can be viewed as being emitted from the hidden states:

T, C, A, G

Your model has been trained on DNA sequences. The initial state probabilities $\boldsymbol{\pi}$, hidden state transition probabilities $\mathbf{A}$ and observation/emission probabilities ϕ in your model are as follows:

$\boldsymbol{\pi}$:	H	L
	0.3	0.7

A:	H	L
H	0.4	0.6
L	0.3	0.7

ϕ :	T	C	A	G
H	0.2	0.1	0.4	0.3
L	0.3	0.2	0.3	0.2

For all problem parts below, use the above model to compute the associated probabilities. Show your calculations for all parts.

(a) [2 points] Compute the probability of the hidden state sequence (L, H, L).

(b) [3 points] Given the initial observation of nucleotide is T, what is the probability that the initial nucleotide composition is H?

(c) [5 points] Compute the probability of the observation DNA sequence (A,C).

PROBLEM 5 [10 points]

Suppose you run the Perceptron algorithm on the following dataset without a bias term:

$$(\mathbf{x}_1, y_1) = \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, -1 \right), \quad (\mathbf{x}_2, y_2) = \left(\begin{bmatrix} -1 \\ -1/2 \end{bmatrix}, +1 \right), \quad (\mathbf{x}_3, y_3) = \left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}, -1 \right),$$
$$(\mathbf{x}_4, y_4) = \left(\begin{bmatrix} 2 \\ -1/2 \end{bmatrix}, +1 \right), \quad (\mathbf{x}_5, y_5) = \left(\begin{bmatrix} 3 \\ 3 \end{bmatrix}, -1 \right).$$

(a) [2 points] Is this dataset linearly separable? If so, find a linear separator $\mathbf{u} \in \mathbb{R}^2$ with $\|\mathbf{u}\|_2 = 1$.

(b) [3 points] Using the vector $\mathbf{u}$ you found in part (a), find the margin of separation, defined as the largest number γ such that $y_i \mathbf{u}^T \mathbf{x}_i \geq \gamma$ for all pairs $(\mathbf{x}_i, y_i)$ in the dataset. (If the dataset is not linearly separable and you did not find a vector $\mathbf{u}$, then choose $\mathbf{u}$ arbitrarily and compute its margin. You will get a negative number.)

(c) [5 points] Is there an upper bound for how many times the perceptron algorithm needs to iterate over this dataset to find a linear separator? If so, find an upper bound. Explain your answer.