CIS 520: Machine Learning Spring 2020

Final Exam
Version A

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 answer questions worth 2 points each (total 18 points).
   - Problem 2 contains 9 short answer questions worth 3 points each (total 27 points).
   - Problems 3–5 are long answer questions (10 points each).
3. **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).****
4. For Problems 1 and 2, we expect you to provide only the final answer and not your calculations; we will grade only the final answer.
5. For Problems 3–5, you must provide clear explanations and detailed calculations.
6. 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 Answer Questions [$9 \times 2 = 18$ **points**]

(a) [2 points] Consider a binary classification task with 60 features: $\mathcal{X} = \mathbb{R}^{60}$, $\mathcal{Y} = \{\pm 1\}$. Suppose you want to learn a linear logistic regression classifier. How many parameters do you need to learn? For simplicity, ignore bias/threshold terms.

(b) [2 points] Consider a regression problem with 10 features: $\mathcal{X} = \mathbb{R}^{10}$, $\mathcal{Y} = \mathbb{R}$. Suppose you want to learn a one-hidden-layer neural network model with 15 hidden units, each with a ReLU 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:

i	$\mathbf{x}_i$	y_i
1	$(0, 1)$	$+1$
2	$(2, 2)$	-1
3	$(1, 0)$	$+1$
4	$(3, 1)$	-1
5	$(1, 5)$	-1
6	$(1, 1)$	$+1$

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 $\mathbf{x} = (0, 0)$?

(d) [2 points] Consider a binary classification task with 10 features: $\mathcal{X} = \mathbb{R}^{10}$, $\mathcal{Y} = \{\pm 1\}$. You are given a training sample containing 1000 labeled examples. Suppose you train 4 support vector machine (SVM) classifiers on the same training sample: h_1 is trained using a linear kernel, h_2 using a degree-2 kernel, h_3 using a degree-3 kernel, and h_4 using a degree-4 kernel. Surprisingly, they all have the same training error of 0.114. You decide to compute VC-dimension based upper bounds on their generalization errors using confidence parameter $\delta = 0.05$. List the 4 classifiers (h_1, h_2, h_3, and h_4) in increasing order of these generalization error bounds (i.e. list the classifier with smallest generalization error bound first, and so on).

(e) [2 points] Consider a binary classification task with d features: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{\pm 1\}$. You are given a training sample containing m labeled examples. You train a kernel support vector machine (SVM) classifier on this sample using a radial basis function (RBF) kernel with width parameter σ: $K(\mathbf{x}, \mathbf{x}') = \exp\left(\frac{-\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\sigma^2}\right)$. Your learned model has k support vectors. What is the time complexity in big-O notation of predicting the class of *one* new test point $\mathbf{x}$? (Assume that division by σ^2 takes unit time.) Express your answer in terms of m, k, d, σ.

(f) [2 points] Consider a collaborative filtering problem with 5000 users and 1000 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 20 latent factors, what is the total number of parameters to be learned (ignore any bias terms)?

(g) [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, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$, 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:

$$\widehat{\eta}(\mathbf{x}_1) = 0.33\,; \quad \widehat{\eta}(\mathbf{x}_2) = 0.42\,; \quad \widehat{\eta}(\mathbf{x}_3) = 0.56\,; \quad \widehat{\eta}(\mathbf{x}_4) = 0.74\,.$$

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

(h) [2 points] Suppose you are given a (fully specified) Markov decision process with state space $\mathcal{S} = \{1, 2, 3\}$ and action space $\mathcal{A} = \{\mathsf{a}, \mathsf{b}, \mathsf{c}, \mathsf{d}\}$. 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	3.6	2.6	3.9	2.6
2	3.1	4.5	2.8	4.0
3	2.0	1.8	2.3	1.9

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

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

$$\mathbf{w}_t = (1,\, 2)^\top .$$

The instance received on this round is

$$\mathbf{x}_t = (2,\, 0)^\top .$$

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

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

(a) [3 points] Consider a binary classification task on the unit square: $\mathcal{X} = [0,1]^2$, $\mathcal{Y} = \{\pm 1\}$. Let D be a probability distribution on $\mathcal{X} \times \mathcal{Y}$ under which an example $(\mathbf{x}, y)$ is generated as follows:

- First an instance $\mathbf{x}$ is drawn uniformly at random from $\mathcal{X}$;
- Given $\mathbf{x}$, a binary label in $\{\pm 1\}$ is drawn randomly as follows:

$$\mathbf{P}(Y = +1 \,|\, X = \mathbf{x}) = \begin{cases} 0.7 & \text{if } x_1 < 0.5 \\ 0.1 & \text{otherwise.} \end{cases}$$

What is the Bayes 0-1 error, $\text{er}_D^{\text{0-1},*}$?

(b) [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 $\epsilon \sim \mathcal{N}(0,1)$.

Now consider the specific instance $\mathbf{x} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$. For this instance, what is the predicted label $f^*(\mathbf{x})$ according to the optimal regression model f^* (under squared loss)?

(c) [3 points] You have just learned a linear logistic regression classifier which, given an instance $\mathbf{x}$, estimates the probability of a positive label to be

$$\widehat{\eta}(\mathbf{x}) = \frac{1}{1 + e^{-\widehat{\mathbf{w}}^\top \mathbf{x}}}\,.$$

You are now told that the cost of a false negative (mis-predicting a positive example as negative) will be $\frac{1}{5}$, and that of a false positive will be $\frac{4}{5}$. In order to classify a new instance $\mathbf{x}$ as positive or negative, the plug-in decision rule should be of the form $h(\mathbf{x}) = \text{sign}(\widehat{\mathbf{w}}^\top \mathbf{x} - c)$ for some threshold $c \in \mathbb{R}$. What is the correct value of c?

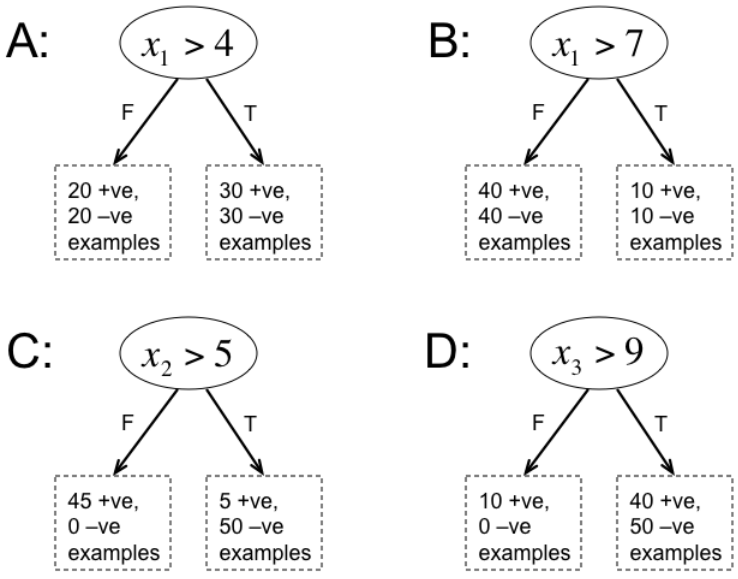
(d) [3 points] Consider a binary classification problem in which the class +1 is rare, and you wish to 'detect' instances from this class. You have learned a binary classifier $h : \mathcal{X} \to \{\pm 1\}$ from some training data. On a test set of 100 data points, you observe the following confusion matrix:

		$h(x)$	
		-1	$+1$
y	-1	45	25
	$+1$	13	17

What is the precision of the classifier h on the above test set?

(e) [3 points] Consider a 10-class classification task (such as handwritten digit recognition) involving 25 binary features: $\mathcal{X} = \{0,1\}^{25}$, $\mathcal{Y} = \{1, 2, \ldots, 10\}$. How many parameters are needed for a Naïve Bayes classifier?

(f) [3 points] Suppose you have a binary classification problem with 3-dimensional feature vectors $\mathbf{x} \in \mathbb{R}^3$. You are given 50 positive and 50 negative training examples, and want to build a decision tree classifier. Consider 4 possible splits at the root node:



Which of the above splits would give the highest information gain: A, B, C, or D? (You should be able to answer this by visual inspection; you do not need a calculator.)

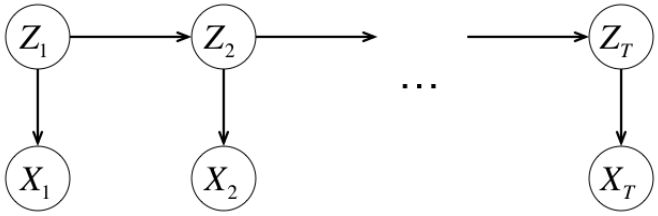
(g) [3 points] Consider an unlabeled data set containing m 2-dimensional points $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^2$, with sample mean $\overline{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i$. You form the mean-centered data matrix $\widetilde{\mathbf{X}} \in \mathbb{R}^{m \times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

$$\begin{bmatrix} 3 & \sqrt{2} \\ \sqrt{2} & 2 \end{bmatrix}.$$

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

(h) [3 points] Suppose you want to model data points in 3 dimensions, $\mathcal{X} = \mathbb{R}^3$, using a Gaussian mixture model with 2 mixture components. How many (independent) parameters are needed?

(i) [3 points] Consider a hidden Markov model in which each hidden state Z_t takes one of 2 possible values, and each observation X_t takes one of 10 possible values.



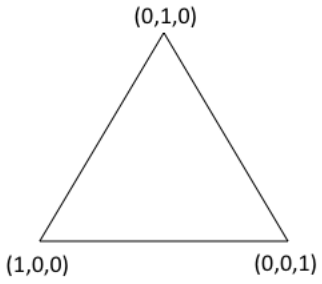
Assume the model is homogeneous, so that transition and emission probabilities are the same for all t. How many (independent) parameters are needed in this model?

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	1	2
$y = 2$	1	0	1
$y = 3$	2	1	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]

You are trying to model the tweeting behaviors of young adults. You are given data on the number of tweets posted by m users over a 1-year period. You know the users in the study come from two communities, A and B, although you don't know which user is from which community. You decide to model the number of tweets posted by each user from Community A as a Poisson random variable with parameter λ_A, and the number of tweets posted by each user from Community B as a Poisson random variable with parameter λ_B, all assumed to be independent of each other. You further assume that each user in the study comes from Community A with probability π, independently of all other users.

For each user i, let X_i denote the number of tweets posted by user i, and let Z_i denote a random variable indicating which community user i comes from:

$$Z_i = \begin{cases} 1 & \text{if user } i \text{ comes from Community A} \\ 0 & \text{otherwise.} \end{cases}$$

Since you don't know which user comes from which community, the Z_i are unobserved or latent variables; all you get to see is the number of tweets posted by each user, i.e. the values $x_1, \ldots, x_m$ taken by the random variables $X_1, \ldots, X_m$. Based on this data, you want to estimate the parameters $\theta = (\pi, \lambda_A, \lambda_B)$.

To help with this, note that for each user i, the joint distribution of (X_i, Z_i) under parameters θ is given by

$$p(x_i, z_i;\, \theta) = \left(\pi \frac{\lambda_A^{x_i}\, e^{-\lambda_A}}{x_i!}\right)^{z_i} \left((1-\pi) \frac{\lambda_B^{x_i}\, e^{-\lambda_B}}{x_i!}\right)^{1-z_i}, \quad \text{for } x_i \in \{0, 1, 2, \ldots\},\ z_i \in \{0, 1\}.$$

(a) [2 points] Write out an expression for the complete-data log-likelihood, $\ln \mathcal{L}_c(\theta) = \sum_{i=1}^m \ln p(x_i, z_i;\, \theta)$ (in terms of $x_i, z_i, \pi, \lambda_A, \lambda_B$).

(b) [4 points] Suppose you knew the values z_i taken by the latent variables Z_i. What would be the maximum-likelihood parameter estimates $\widehat{\theta}$? Give expressions for $\widehat{\pi}$, $\widehat{\lambda}_A$, and $\widehat{\lambda}_B$ (in terms of x_i and z_i). Show your calculations.

(c) [4 points] In the absence of knowledge of z_i, one possibility for estimating θ is to use the EM algorithm. Recall that the algorithm starts with some initial parameter estimates θ^0, and then on each iteration t, performs an E-step followed by an M-step. For a given iteration t, let θ^t denote the parameter estimates at the start of the iteration, and for each user i, let γ_i^t denote the posterior probability of Z_i taking value 1 under parameters θ^t as computed in the E-step:

$$\gamma_i^t = \mathbf{P}(Z_i = 1 \,|\, X_i = x_i;\, \theta^t)\,.$$

Then the expected complete-data log-likelihood with respect to these posterior distributions is

$$\sum_{i=1}^m \Big(\gamma_i^t \cdot \ln p(x_i, 1;\, \theta) + (1 - \gamma_i^t) \cdot \ln p(x_i, 0;\, \theta)\Big)\,.$$

The M-step of the EM algorithm requires finding parameters θ^{t+1} that maximize this expected complete-data log-likelihood. Determine the updated parameters θ^{t+1}. Give expressions for π^{t+1}, λ_A^{t+1}, and λ_B^{t+1} (in terms of x_i and γ_i^t). Show your calculations.

PROBLEM 5 [10 points]

You are working with a National Geographic team that studies penguins, and have built a small hidden Markov model for sequences of penguin behaviors. In particular, your model incorporates the following 4 observed behavior elements:

$$\mathsf{rest},\ \mathsf{walk},\ \mathsf{swim},\ \mathsf{preen}$$

Your model assumes 2 hidden states, which you label as follows:

$$\mathsf{calm},\ \mathsf{excited}$$

Your model has been trained on video sequences of penguins filmed by the team; for the purposes of training your (discrete-time, homogeneous) model, time-steps correspond to 1-second segments in the video sequences. The initial state probabilities $\boldsymbol{\pi}$, hidden state transition probabilities $\mathbf{A}$, and observation/emission probabilities $\boldsymbol{\phi}$ in your model are as follows:

$\boldsymbol{\pi}$:	calm	excited
	0.7	0.3

$\mathbf{A}$:	calm	excited
calm	0.6	0.4
excited	0.2	0.8

$\boldsymbol{\phi}$:	rest	walk	swim	preen
calm	0.3	0.2	0.1	0.4
excited	0.1	0.3	0.5	0.1

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 $(\mathsf{calm}, \mathsf{excited}, \mathsf{calm})$.

(b) [3 points] Given the initial observation is swim, what is the probability that the initial hidden state is calm?

(c) [5 points] Compute the probability of the observation sequence $(\mathsf{rest}, \mathsf{swim})$.