CS 189/289A Introduction to Machine Learning
Spring 2022 Jonathan Shewchuk Midterm

- Please do not open the exam before you are instructed to do so. Fill out the blanks below now.
- **Electronic devices are forbidden on your person**, including phones, laptops, tablet computers, headphones, and calculators. Turn your cell phone off and **leave all electronics at the front of the room**, or **risk getting a zero** on the exam. Exceptions are made for car keys and devices needed because of disabilities.
- When you start, the **first thing you should do** is **check that you have all 7 pages and all 4 questions**. The second thing is to please **write your initials at the top right of every page after this one** (e.g., write "JS" if you are Jonathan Shewchuk).
- The exam is closed book, closed notes except your one cheat sheet.
- You have **80 minutes**. (If you are in the Disabled Students' Program and have an allowance of 150% or 200% time, that comes to 120 minutes or 160 minutes, respectively.)
- Mark your answers on the exam itself in the space provided. Do **not** attach any extra sheets. If you run out of space for an answer, write a note that your answer is continued on the back of the page.
- The total number of points is 100. There are 12 multiple choice questions worth 4 points each, and 3 written questions worth a total of 52 points.
- For multiple answer questions, fill in the bubbles for **ALL correct choices:** there may be more than one correct choice, but there is always at least one correct choice. **NO partial credit** on multiple answer questions: the set of all correct answers must be checked.

First name	
Last name	
SID	
First and last name of student to your left	
First and last name of student to your right	

Q1. [48 pts] Multiple Answer

Fill in the bubbles for **ALL correct choices**: there may be more than one correct choice, but there is always at least one correct choice. **NO partial credit**: the set of all correct answers must be checked.

(a) [4 pts] Select the true statements about Bayes decision theory.

○ A: The risk for a decision rule is the average loss over the training points that are in class C.

○ B: The Bayes decision boundary between two classes, if you're using the 0-1 loss, is the set of points x where $P(X = x|Y = 0) = P(X = x|Y = 1)$.

○ C: If the Bayes risk is nonzero in a two-class classification problem, then the distributions for each class (i.e., $P(X|Y = C)$ and $P(X|Y \neq C)$) must overlap.

○ D: There exists a loss function for which the Bayes decision rule might select the class with lower posterior probability.

(b) [4 pts] Select the true statements about least-squares linear regression.

○ A: The problem of minimizing $\|Xw - y\|_1$ often yields a "sparse" solution, where some of the components of w are exactly zero.

○ B: There is always at least one solution to the normal equations.

○ C: There are problems for which the normal equations have exactly two distinct solutions.

○ D: When the normal equations have multiple solutions, all the solutions have the same loss on test points.

(c) [4 pts] Select the true statements about ROC curves.

○ A: The horizontal axis represents posterior probability thresholds and the vertical axis represents test set accuracy.

○ B: The ROC curve is a better guide for choosing a threshold (separating negative from positive classifications) on real-world data than the threshold suggested by decision theory.

○ C: A ROC curve closer to the diagonal line $y = x$ implies that your classifier's risk is closer to Bayes optimal.

○ D: There are (at least) two points on a ROC curve that are not affected by changes in the model. (Note: we are not counting the specific choice of threshold between positive and negative as part of the model).

(d) [4 pts] Ridge regression is

○ A: a way to perform feature selection, as ridge regression encourages weights to be exactly zero.

○ B: a method in which bias tends to increase, and variance tends to decrase, as we increase the regularization parameter λ.

○ C: motivated by imposing a Gaussian prior probability on the weight vector.

○ D: a method whose cost function has a unique minimum (assuming $\lambda > 0$).

(e) [4 pts] Select the statements that are true for **every** real symmetric matrix $X \in \mathbb{R}^{n \times n}$.

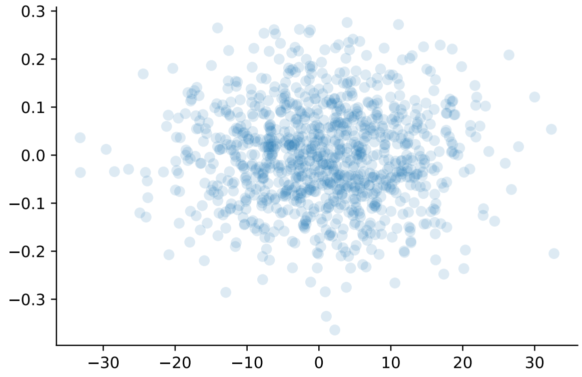
○ A: X can be factored as $X = UDU^\top$, where U is a orthogonal matrix and D is a diagonal matrix.

○ B: X can be factored as $X = UU^\top$, where U is a orthogonal matrix.

○ C: $\lambda_{\max}(X) \geq 0$, where $\lambda_{\max}(X)$ denotes the greatest eigenvalue of X.

○ D: $a^\top X a \leq \lambda_{\max}(X)\, \|a\|_2^2$ for all $a \in \mathbb{R}^n$.

(f) [4 pts] Below are 1,000 sample points drawn from a two-dimensional multivariate normal distribution. Which of the following matrices could (without extreme improbability) be the covariance matrix of the distribution? (Pay attention to the numbers on the axes!)



○ A: $\Sigma = \begin{bmatrix} 100 & 0 \\ 0 & 0.01 \end{bmatrix}$

○ B: $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

○ C: $\Sigma = \begin{bmatrix} 10 & 0 \\ 0 & 0.1 \end{bmatrix}$

○ D: $\Sigma = \begin{bmatrix} -10 & 0 \\ 0 & -0.1 \end{bmatrix}$

(g) [4 pts] You are training a soft-margin SVM on a binary classification problem. You find that your model's training accuracy is very high, while your validation accuracy is very low. Which of the following are likely to improve your model's performance on the validation data?

○ A: Training your model on more data.

○ B: Adding a quadratic feature to each sample point.

○ C: Increasing the hyperparameter C.

○ D: Decreasing the hyperparameter C.

(h) [4 pts] Select the true statements about Gaussian Discriminant Analysis.

○ A: If a class-conditional covariance matrix is anisotropic (the eigenvalues are not equal), the decision boundary is guaranteed to be nonlinear.

○ B: The QDA posterior probability is a logistic function composed with (applied to) a quadratic function of the feature space.

○ C: QDA is more prone to overfitting than LDA.

○ D: The Bayes decision boundary arising from two normally distributed classes can split the feature space into *at most* two regions.

(i) [4 pts] Select the true statements about finding a minimum of a cost function $f(x)$.

○ A: Newton's method always converges to a globally minimum solution for any twice-differentiable function f.

○ B: For the cost function $f(x) = \delta\|x - b\|^2 + \gamma$ with $\delta > 0$, Newton's method always converges to a globally minimum solution.

○ C: If f is convex, is differentiable, and has exactly one local minimum, then (batch) gradient descent always converges to that minimum for any choice of learning rate.

○ D: It is not possible to execute an iteration of Newton's method on the perceptron risk function.

(j) [4 pts] In the following statements, the word “bias” is referring to the bias-variance decomposition. Select the true ones.

○ A: A model trained with n training points is likely to have lower variance than a model trained with $2n$ training points.

○ B: If my model is underfitting, it is more likely to have high bias than high variance.

○ C: Increasing the number of parameters (weights) in a model usually improves the test set accuracy.

○ D: Adding ℓ_2-regularization usually reduces variance in linear regression.

(k) [4 pts] Which of the following statements are true regarding Lasso regression?

○ A: Lasso’s optimization problem can be stated as a quadratic program.

○ B: The cost function minimized by Lasso has points where its gradient is not well-defined, and the solution (minimum) is often at such a point.

○ C: Lasso often produces sparser results (more zero weights) than ridge regression.

○ D: A version of Lasso using a penalty term of $\lambda\|w\|_{\ell_{0.5}}$ (that is, the $\ell_{0.5}$-norm) will be more inclined to produce sparse solutions than Lasso.

(l) [4 pts] Let X be an $n \times d$ design matrix where $n = 10$ and $d = 12$, representing information about various loan borrowers. Let $y \in \mathbb{R}^n$ be a vector of labels such that y_i represents the time (in days) between when borrower i took a loan and when it was fully repaid. We would like to train a regression model on this data. Which of the following methods would be reasonable choices for this task?

○ A: Least squares linear regression with the solution $w^* = (X^\top X)^{-1} X^\top y$

○ B: Logistic regression

○ C: Least squares linear regression using the Moore–Penrose pseudoinverse, $w^* = X^\dagger y$

○ D: Ridge regression

Q2. [17 pts] Gaussian Discriminant Analysis

You want to create a model to predict student performance on the CS 189/289A Midterm. You survey several past students and record how many hours they studied for the exam, and whether or not they passed, yielding the two classes.

Passed: [4, 5, 5.5, 6.5, 7, 8]
Failed: [0, 1, 2, 3, 4]

The hours spent studying is the only feature we have for each student ($d = 1$). Assume that the number of hours is normally distributed for both the passing and failing students. Consider two ways of modeling this data: Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Use the 0-1 loss function to define risk.

(a) [8 pts] Calculate the sample means μ_p, μ_f and the variances σ_p^2, σ_f^2 computed for **QDA**. (The subscripts mean "pass" and "fail.") Express your answers as the simplest fractions (not decimals) possible.

(b) [4 pts] Calculate the sample means and variances used by **LDA**. Express your answers as the simplest fractions (not decimals) possible.

(c) [5 pts] Calculate the decision boundary for **LDA**. Use fractions, not decimals, and express the answer in as simple a form as possible (but expect it to have a logarithm in it).

Q3. [15 pts] Symmetric Matrices

(a) [6 pts] Derive the 2×2 symmetric matrix whose eigenvalues are 5 and 2, such that $(2, -1)$ is an eigenvector with eigenvalue 5.

(b) [6 pts] Consider the two-dimensional bivariate normal distribution $\mathcal{N}(0, \Sigma)$ where the covariance matrix Σ is the matrix you derived in part (a) and the mean is $\mu = 0$. Let $f(x)$ be the PDF of that normal distribution, where $x \in \mathbb{R}^2$. What are the lengths of the major and minor axes of the ellipse

$$f(x) = \frac{1}{4\pi\sqrt{10}}?$$

Justify your answer.

(c) [3 pts] Consider a cost function $J(w)$ over a weight vector w, and suppose that at every point $w \in \mathbb{R}^d$, the Hessian matrix $\nabla^2 J$ is positive definite. Is it always true that $J(w)$ has exactly one unique local minimum $w^* \in \mathbb{R}^d$? Why or why not?

Q4. [20 pts] Linear Regression with Laplacian Noise

In lecture, we saw how least-squares regression is motivated by maximum likelihood estimation if we think our data obeys a linear relationship but has added noise that is normally distributed. But what if the noise is better modeled by the Laplace distribution (which you reviewed in Homework 4)?

Let $\epsilon \sim \text{Laplace}(\mu, \beta)$ indicate a random variable ϵ drawn from a univariate Laplace distribution with mean μ and scale parameter β. The PDF of this distribution is

$$f(\epsilon; \mu, \beta) = \frac{1}{2\beta} \exp\left(\frac{-|\epsilon - \mu|}{\beta}\right).$$

Following our customary notation, the input is an $n \times d$ design matrix X and a vector y such that y_i is the label for sample point X_i, where $X_i^\top$ is the ith row of X. To keep things simple, we will do linear regression through the origin (no bias term α), so the regression function is $h(x) = w \cdot x$. Our model is that each label y_i comes from a linear relationship perturbed by Laplacian noise,

$$y_i \sim \text{Laplace}(w \cdot X_i, \beta),$$

where $w \in \mathbb{R}^d$ is the true linear relationship. We will use maximum likelihood estimation to try to estimate w.

(a) [5 pts] Write the likelihood function $\mathcal{L}(w; X, y)$ for the parameter w, given the fixed data X and y.

(b) [3 pts] Write the log likelihood function $\ell(w; X, y)$ for the parameter w, given the fixed data X and y, in as simple a form as you can. (Make sure your logarithms have the correct base.)

(c) [3 pts] What is the simplest cost function we can minimize that gives us the same value of w as maximizing the likelihood?

(d) [4 pts] How is the cost function you just derived different from standard least-squares regression? Is it more or less sensitive to outliers? Why?

(e) [5 pts] Write the batch gradient descent rule for minimizing your cost function, using η for the step size (aka learning rate). You may omit training points whose losses have undefined gradients. *Hint:* Recall that $\frac{d}{d\alpha}|\alpha|$ is 1 for $\alpha > 0$, -1 for $\alpha < 0$, and undefined for $\alpha = 0$.