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.

A: False, it is the expected loss over all values of x and y, and it has nothing to do with training points.

B: False; it's the set of points x where $P(Y = 0|X = x) = P(Y = 1|X = x)$.

C: True, when the class distributions are non-overlapping, each x will uniquely lie in one classes distribution, which means there exists a (optimal) classifier which perfectly classifies all points, and has Bayes risk of zero. Similarly, if Bayes risk is nonzero, there must be uncertainty for an x, and thus the class distributions must overlap.

D: True, for a nonsymmetric loss function.

(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.

A is false. Minimizing $\|w\|_1$ encourages weights to be exactly zero, but minimizing $\|Xw - y\|_1$ does not.
B is true. (Proved in discussion section.)
C is false. (See discussion section.)
D is false. The different solutions give different values when you move off the affine subspace that all the training points lie on. Those different values will give different losses 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).

A is false because both axes of the ROC curve show classification rates. The possible classifier thresholds are implicitly-not explicitly- shown on the ROC curve. B is true (see lecture note 11) C is nearly the opposite of the truth, because the line $y = x$ represents a random classifier. D is true (the points (0,0) and (1,1) are always present since they arise when we only classify as one class).

(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 $\lambda > 0$).

A: Lasso does that, but ridge regression rarely produces weights of exactly zero.

B: True. As $\lambda \to \infty$, the weights approach zero, and so does the variance, whereas the bias increases.

C: True. Another way to derive ridge regression is via *maximum a posteriori* with a Gaussian prior on the weights.

D: True. Ridge regression's regularization term makes the objective function strictly convex.

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

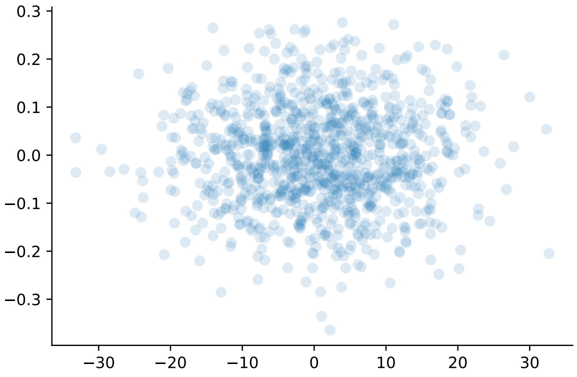
A: True, this is a direct statement of spectral theorem.

B: Not true, as this entails X being PSD.

C: Not true, X can be $-I_n$.

D: True. Proof can be found in HW2.

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

A is valid. B would be spherical. C would not have the right coordinates. Remember, the widths of the Gaussian are the standard deviations in each eigevector direction; and the standard deviations are the square roots of the variances, which are the eigenvalues of Σ. D cannot be a covariance matrix.

(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.

A is true as training on more data reduces variance and decreases overfitting in general. B is false since polynomial features increase the variance and the risk of overfitting. C is false because increasing C enforces a harder margin constraint, leading to more overfitting. D is true because decreasing C allows for more slack, decreasing the risk of overfitting.

(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.

A is False; if all classes have the same anisotropic covariance matrix, the decision boundary is linear. B is true, and can be shown by Bayes' rule. C is true since QDA has a higher number of parameters than LDA. D is false, since the decision boundary can split the feature space into three regions if the class covariances are not equal.

(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.

A is false: Newton's method doesn't always converge, and when it does, the critical point it finds isn't always a minimum.
B is true as the globally minimum solution is b and $\nabla f(x) = 2\delta(x-b)$ and $\nabla^2 f(x) = 2\delta I$ and $x_1 = x_0 - \frac{2\delta(x_0-b)}{2\delta} = x_0-(x_0-b) = b$
C is false as gradient descent will not converge if the learning rate is too high.
D is true because the perceptron risk function has a zero Hessian matrix everywhere, except at points where the Hessian isn't even defined.

(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.

A is false since increasing the number of samples generally reduces the variance.
B is generally true, since you typically underfit if you have insufficient model capacity to even perform well on the training set.
C is not always true, as increasing the number of parameters can lead to overfitting.
D is true, regularization effectively reduces model complexity.

(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.

A is not true in general. Lasso *tends* to produce sparser solutions, but there are cases where it may produce equivalently sparse solutions as other forms of regression.
B is false because Lasso's objective function is nondifferentiable.
C is true. This is due to the square-shaped isocontours introduced by the ℓ_1 norm being more likely to intersect other isocontours on axes than the circular isocontours introduced by the ℓ_2 norm.
D is true — intuitively, the smaller the p we select for our p-norm, the pointier the penalty's isocontours become, making us more likely to choose a sparse weight vector.

(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

A: This does not work because $d > n$, meaning $X^\top X$ will be rank-deficient and thus not invertible.

B: Logistic regression is limited to predicting values between 0 and 1, which is not appropriate for our use case (predicting time in days).

C: Ridge regression avoids the problem with (A) because the matrix $X^\top X + \lambda I$ will always be invertible.

D: As we saw in discussion, the Moore-Penrose pseudoinverse allows us to find the unique least norm least squares solution even if $X^\top X$ is non-invertible.

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.

The sample mean for a class is

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

By plugging in the data, we have [4 pts]

$$\hat{\mu}_p = 6, \quad \hat{\mu}_f = 2.$$

The sample variance for a class is

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{\mu})^2$$

By plugging in the data, we have [4 pts]

$$\hat{\sigma}_p^2 = \frac{7}{4}, \quad \hat{\sigma}_f^2 = 2.$$

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

The sample means used by LDA are the same as the sample means we computed above for QDA. [1 pt]

For LDA, we calculate just one variance, an average of the class variances weighted by the priors [3 pts]:

$$\hat{\sigma}^2 = \frac{7}{4} \times \frac{6}{11} + 2 \times \frac{5}{11} = \frac{41}{22}.$$

(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).

We can find the decision boundary by solving for the points at which $P(Y = \text{pass}|X) = P(Y = \text{fail}|X)$.

$$\begin{aligned}
6f(X|Y = \text{pass}) &= 5f(X|Y = \text{fail}) \\
\frac{6}{\sqrt{2\pi}\hat{\sigma}}\exp\left(-\frac{1}{2}\frac{(x-6)^2}{\hat{\sigma}^2}\right) &= \frac{5}{\sqrt{2\pi}\hat{\sigma}}\exp\left(-\frac{1}{2}\frac{(x-2)^2}{\hat{\sigma}^2}\right) \\
\ln 6 - \frac{1}{2\hat{\sigma}^2}(x-6)^2 &= \ln 5 - \frac{1}{2\hat{\sigma}^2}(x-2)^2 \\
2\hat{\sigma}^2 \ln\frac{6}{5} + (x-2)^2 - (x-6)^2 &= 0 \\
8x - 32 &= -2\hat{\sigma}^2 \ln\frac{6}{5} \\
x &= 4 - \frac{41}{88}\ln\frac{6}{5}
\end{aligned}$$

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.

From the eigendecomposition, we have

$$\frac{1}{\sqrt{5}} \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix} \frac{1}{\sqrt{5}} \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 22/5 & -6/5 \\ -6/5 & 13/5 \end{bmatrix}.$$

(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.

The determinant of Σ is 10 (the product of the eigenvalues, or you can compute it the hard way). The normal PDF is

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)\right) = \frac{1}{2\pi\sqrt{10}} \exp\left(-\frac{1}{2} x^\top \Sigma^{-1} x\right).$$

Setting that to $1/(4\pi\sqrt{10})$ gives

$$\begin{aligned} -\frac{1}{2} x^\top \Sigma^{-1} x &= \ln \frac{1}{2} \\ \|\Sigma^{-1/2} x\|^2 &= 2 \ln 2 \\ \|\Sigma^{-1/2} x\| &= \sqrt{2 \ln 2}. \end{aligned}$$

The eigenvalues of $\Sigma^{-1/2}$ are $1/\sqrt{5}$ and $1/\sqrt{2}$. The major axis of the ellipse is the eigenvector x that solves this equation and has eigenvalue $1/\sqrt{5}$ (or eigenvalue 5 for Σ); thus the major axis has length $\sqrt{10 \ln 2}$. The minor axis is the eigenvector x that solves this equation and has eigenvalue $1/\sqrt{2}$ (or eigenvalue 2 for Σ); thus the minor axis has length $2\sqrt{\ln 2}$.

(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?

No, because $J(w)$ doesn't have to have a minimum. For example, the function $J(w) = \sum_{i=1}^{n} e_i^{w}$ has a Hessian that is positive definite everywhere, but J approaches zero only in the limit as all the weights approach $-\infty$.

Recall another example: the logistic regression cost function doesn't have a minimum when the points are linearly separable.

Grading note: full marks will be given to answers that simply point out that some cost functions don't have a minimum because you can get ever smaller costs by moving w toward infinity, yet they can still have positive definite Hessians everywhere. But the answer does have to make clear how a function can have no minimum, either by giving an example or saying something like "it goes downhill forever" or "it keeps decreasing as w goes to infinity."

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.

$$\mathcal{L}(w; X, y) = \prod_{i=1}^{n} f(y_i; g(X_i), \beta) = \frac{1}{(2\beta)^n} \exp\left(-\frac{1}{\beta}\sum_{i=1}^{n} |y_i - w \cdot X_i|\right).$$

(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.)

$$\ell(w; X, y) = -n \ln(2\beta) - \frac{1}{\beta}\sum_{i=1}^{n} |y_i - w \cdot X_i|.$$

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

$$J(w) = \sum_{i=1}^{n} |y_i - w \cdot X_i|.$$

(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?

The solution minimizes the mean absolute error instead of the mean squared error. It is less sensitive to outliers because the error is not squared, so labels with large errors do not dominate as much.

(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 $\alpha = 0$.

$$\nabla_w J = \sum_{w \cdot X_i > y_i} X_i - \sum_{w \cdot X_i < y_i} X_i.$$

So the batch update rule is

$$w \leftarrow w + \eta \sum_{w \cdot X_i < y_i} X_i - \eta \sum_{w \cdot X_i > y_i} X_i.$$

Grading note: no harm is done if we sum over the points such that $w \cdot X_i \leq y_i$ or $w \cdot X_i \geq y_i$. Although the gradient is not technically defined at some points, it does no harm to include subgradients for them anyway.