CS 189/289A Introduction to Machine Learning
Spring 2023 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 9 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	
Name and SID of student to your left	
Name and SID 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] Suppose we are doing ridge regression with a design matrix $X \in \mathbb{R}^{n \times d}$ and a vector $y \in \mathbb{R}^n$ of labels. Recall that we find the weight vector w that minimizes the cost function $J(w) = \|Xw - y\|_2^2 + \lambda \|w\|_2^2$, where $\lambda > 0$ is a positive scalar. Suppose we solve this problem with Newton's method. Recall that its update rule is

$$w^{(k+1)} = w^{(k)} - (\nabla^2 J(w^{(k)}))^{-1} \nabla J(w^{(k)}).$$

We start the method from some arbitrary initial weight vector $w^{(0)}$. Suppose we have a magic computer that does exact arithmetic on real numbers; it can do arithmetic without ever rounding the numbers. In **how many iterations** does Newton's method converge to the true ridge regression solution?

○ A: 1

○ B: d

○ C: It might never get close to the true solution.

○ D: It might never get *exactly* to the true solution, but it gets closer with every iteration.

(b) [4 pts] Suppose we are doing classification or regression with a design matrix $X \in \mathbb{R}^{n \times d}$ and a vector $y \in \mathbb{R}^n$ of labels, where X has full rank. Below we specify pairs of the form (transformation of X, learning algorithm). Select the pairs where the transformation **might change the label** returned by the learning algorithm for some test point $z \in \mathbb{R}^d$. (Note: assume the implementation isn't stupid; it applies the same transformation to each test point z that was applied to the training points.)

○ A: (centering X, Gaussian discriminant analysis)

○ B: (whitening X, soft-margin support vector machine)

○ C: (invertible affine transformation of X, linear regression)

○ D: (decorrelating X, Lasso)

(c) [4 pts] Which of the following statements are true about Gaussian discriminant analysis?

○ A: If we whiten the design matrix X, then our class-conditional sample covariance matrices in quadratic discriminant analysis are positive definite.

○ B: Assuming our sample is drawn from a normal distribution, quadratic discriminant analysis gives us the true optimal Bayes classifier.

○ C: Linear discriminant analysis can have higher bias than quadratic discriminant analysis.

○ D: If we replace the sample covariance matrix Σ with $\Sigma + \lambda I$ for some $\lambda > 0$, we guarantee that the latter matrix is positive definite.

(d) [4 pts] Which of the following statements are true about logistic regression?

○ A: If the points are linearly separable (with a positive margin), then logistic regression (with suitable optimization software) can find a decision boundary with zero training error.

○ B: The logistic cost function can be directly applied to three-class classification.

○ C: If we don't use regularization, the logistic regression cost function might have infinitely many minimizers.

○ D: Logistic regression with quadratic features produces the same decision boundary as a quadratic discriminant analysis (QDA) model.

(e) [4 pts] Which of the following statements are true about Lasso and ridge regression?

○ A: Both ridge regression and Lasso are methods used to reduce overfitting that might occur in standard linear regression.

○ B: The ℓ_1-norm regularization used in Lasso has a tendency to induce sparsity in the weight vector.

○ C: Both ridge regression and Lasso have a cost function with a minimizing weight vector w^* that we can write as a closed-form algebraic expression.

○ D: Ridge regression shrinks the weight vector but rarely drives its components to exactly zero.

(f) [4 pts] You are given a set of sample points $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ sampled independently from a multivariate Gaussian distribution $\mathcal{N}(\mu, \sigma^2 I)$, where $\mu \in \mathbb{R}^d$, $\sigma > 0$, and I is the $d \times d$ identity matrix. We know the exact true value of σ, but we don't know μ. We want to use maximum likelihood estimation to estimate μ from the sample points. Which of the following optimization problems gives us a suitable estimate? (Select all that apply.) We choose $\hat{\mu} \in \mathbb{R}^d$ that . . .

○ A: minimizes $\displaystyle\sum_{i=1}^{n} \left(\frac{\|X_i - \hat{\mu}\|_2^2}{2\sigma^2} + d \ln \sqrt{2\pi} + d \ln \sigma \right)$

○ B: minimizes $\displaystyle\sum_{i=1}^{n} \left(\frac{\|X_i - \hat{\mu}\|_1^2}{2\sigma^2} + d \ln \sqrt{2\pi} + d \ln \sigma \right)$

○ C: maximizes $\displaystyle\prod_{i=1}^{n} \exp\left(-\frac{\|X_i - \hat{\mu}\|_2^2}{2\sigma^2} \right)$

○ D: minimizes $\displaystyle\sum_{i=1}^{n} \|X_i - \hat{\mu}\|_2^2$

(g) [4 pts] Select the true statements about the decision boundaries obtained by linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA).

○ A: If we add extra features, LDA could produce a decision boundary that is nonlinear relative to the original input features.

○ B: QDA never produces a linear decision boundary.

○ C: Sometimes the QDA decision boundary is an ellipsoid.

○ D: Consider using LDA for a two-class classification problem in which both classes have prior probability 0.5. LDA computes estimated means $\hat{\mu}_C$ and $\hat{\mu}_D$ for the two classes; suppose they are not equal. The midpoint halfway between $\hat{\mu}_C$ and $\hat{\mu}_D$ lies on the decision boundary obtained from LDA.

(h) [4 pts] Which of the following statements are true about ROC curves for binary classification?

○ A: They are monotonically non-decreasing.

○ B: The area under a ROC curve is always at least 0.5.

○ C: The ideal classifier has a ROC curve that passes through the top left corner at coordinate (0, 1).

○ D: ROC curves are a practical aid to choosing a loss function.

(i) [4 pts] Consider a generative two-class classifier where we estimate the distribution of each class and the prior probability of each class, but we use an asymmetrical loss function. (Assume the distributions are continuous.) What must be true **at every point $x \in \mathbb{R}^d$ that lies on the decision boundary**?

○ A: The class conditional probabilities must be equal: $P(X = x|Y = 0) = P(X = x|Y = 1)$.

○ B: The prior probabilities must be equal.

○ C: The posterior probabilities must be equal.

○ D: The risk of predicting either class (evaluated at x only) must be equal.

(j) [4 pts] Consider a design matrix $X \in \mathbb{R}^{n \times d}$. Which of the following is true of its sample covariance matrix Σ?

○ A: If the same sample point appears twice in X, then the sample covariance matrix Σ must be singular.

○ B: The PDF of the multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ has isocontours whose axes are aligned with the eigenvectors of Σ.

○ C: If some column of X is all 2's, then the sample covariance matrix Σ must be singular.

○ D: The PDF of the multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ has elliptical isocontours whose axis lengths are linearly proportional to the eigenvalues of Σ.

(k) [4 pts] We are solving a least-squares linear regression problem without regularization. Suppose that the following two weight matrices both have the same cost: $w_1 = \begin{bmatrix}1\\0\\6\end{bmatrix}$ and $w_2 = \begin{bmatrix}-2\\3\\3\end{bmatrix}$. That is, $\text{RSS}(w_1) = \text{RSS}(w_2)$. Which of the following statements become true if regularization is incorporated into the regression?

○ A: If l_2 regularization is added with sufficiently high λ, w_1 will be preferred over w_2.

○ B: If l_∞ regularization is added with sufficiently high λ, w_1 will be preferred over w_2. (Recall that $\|w\|_\infty = \max_i |w_i|$.)

○ C: If l_1 regularization is added with sufficiently high λ, w_1 will be preferred over w_2.

○ D: If l_2 regularization is added with $\lambda > 0$, there might be infinitely many weight vectors that minimize $\text{RSS}(w)$.

(l) [4 pts] Suppose we are doing ridge regression with a design matrix $X \in \mathbb{R}^{n \times d}$ and a vector $y \in \mathbb{R}^n$ of labels. The ridge regression cost function is $J(w) = \|Xw - y\|_2^2 + \lambda \|w\|_2^2$, where $\lambda \geq 0$. Which of the following conditions are **sufficient** for the Hessian matrix $\nabla^2 J$ to be positive definite?

○ A: $\lambda > 0$

○ B: $XX^\top$ has full rank

○ C: $X^\top X$ has full rank

○ D: $n < d$

Q2. [20 pts] Convergence of Gradient Descent

Consider the problem of finding a weight vector $w \in \mathbb{R}^n$ that minimizes the cost function $J(w) = w^\top Aw + b^\top w + \alpha$, where $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix, $b \in \mathbb{R}^n$ is a vector, and $\alpha \in \mathbb{R}$ is a scalar.

(a) [5 pts] Write **an expression for** $\nabla J(w)$ and **a closed-form expression for a critical point** w^* **of** $J(w)$. For full points, your expression for w^* should be a critical point even when A is singular, assuming that J has a critical point. (Hint: you can simply use the notation you learned in Discussion Section 6.)

(b) [5 pts] Write **the gradient descent update rule** for this optimization problem, with step size ϵ, in the form "$w^{(k+1)} \leftarrow$ some function of $w^{(k)}$". (Here, $w^{(k)}$ denotes the kth iterate of the weight vector, given some fixed starting value $w^{(0)}$.)

(c) [5 pts] We define the "error" in an iterate $w^{(k)}$ to be $e^{(k)} = w^{(k)} - w^*$. **Rewrite your update rule** in the form "$e^{(k+1)} \leftarrow$ some function of $e^{(k)}$", so that $e^{(k)}$ appears **only once** and the letters w and b **do not appear** at all. (Hint: remember the equation you solved for part (a).) **Simplify** the expression as much as you can.

(d) [5 pts] Show that if every eigenvalue λ_i of A satisfies $0 < \lambda_i < \frac{1}{\epsilon}$, then **each successive error** $e^{(k+1)}$ **is shorter than the previous error** $e^{(k)}$, so the iterations cause the error to converge to zero. Hint: use the eigendecomposition of A and the fact that an orthonormal matrix does not change the length of a vector it is multiplied by.

Q3. [12 pts] Gaussian Discriminant Analysis

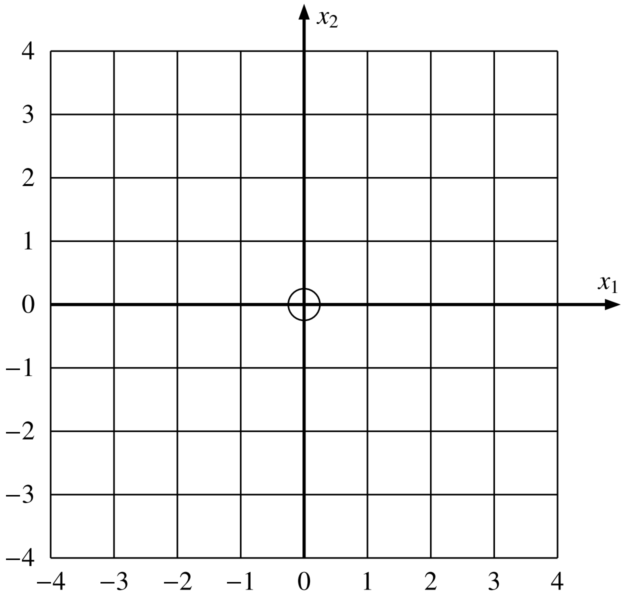
Suppose we know the true distributions of two classes, $f(X = x|Y = 1) = \mathcal{N}(\mu_1, \Sigma_1)$ and $f(X = x|Y = 2) = \mathcal{N}(\mu_2, \Sigma_2)$, where $\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\mu_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$. Σ_1 has eigenvector $\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ with eigenvalue 1 and eigenvector $\begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ with eigenvalue 4, and Σ_2 has the same eigenvectors with the eigenvalues swapped: eigenvector $\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ with eigenvalue 4 and eigenvector $\begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ with eigenvalue 1.

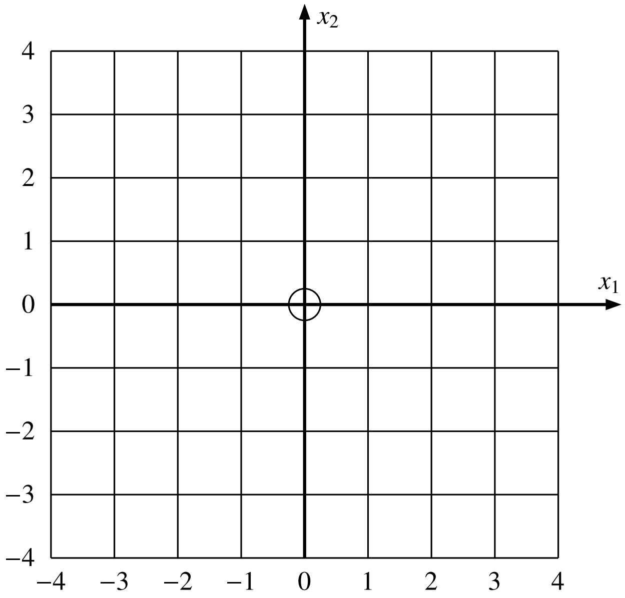
(a) [6 pts] What is the value of the matrix Σ_1, with the eigenvectors and eigenvalues listed above?

(b) [6 pts] Draw some isocontours of the quadratic functions $Q_1(x) = (x - \mu_1)^\top \Sigma_1^{-1} (x - \mu_1)$ and $Q_2(x) = (x - \mu_2)^\top \Sigma_2^{-1} (x - \mu_2)$, using the same isovalues for both functions (so you can easily compare them). You may choose any isovalues you like, but draw at least three isocontours for each function. Show clearly which contours are for Q_1 and which are for Q_2.

Then draw and label the Bayes optimal decision boundary. (Hint: Where do the isocontours intersect each other? The Bayes optimal decision boundary is not a straight line; try to get it curving in the right direction.)

(Note: there are two copies of the grid below so if you mess it up, you can start over. But you only need to draw on one.)





Q4. [20 pts] Coins, Clones, and Maximum Likelihood

As described in class, imagine that we repeatedly flip a biased coin that comes up heads with probability p and tails with probability $1 - p$. If we flip the coin n times, the number of times it comes up heads is a random variable $X \sim \mathcal{B}(n, p)$ drawn from a binomial distribution. Note that $X \in [0, n]$.

In this problem, we grow 100 clones of you in a whiskey vat. The clones are numbered, from 1 to 100. Clone number n flips the biased coin n times. All coin flips are independent of each other. The number of heads obtained by clone number n is a random variable $X_n \sim \mathcal{B}(n, p)$. Note that n is different for every clone, but p is always the same. We will use maximum likelihood estimation to estimate p from these counts.

(a) [2 pts] Let's start with clone number 3, who flips the coin $n = 3$ times. What is the probability of getting exactly two heads (as a function of p)?

(b) [3 pts] For an arbitrary $n \in [1, 100]$, what is the probability $P(X_n = x)$ that clone number n will get exactly x heads? (Hint: You can use the expression $\binom{n}{x}$ to denote the number of ways to choose x items from a total of n items. If you can't figure out the right constant for the binomial distribution, we'll give partial credit for getting the dependence on p right.)

(c) [4 pts] We want to use maximum likelihood estimation to estimate the parameter p, knowing that clone number n got x heads, but having no data yet about the other clones. Write the likelihood function $\mathcal{L}(p; x)$ and the log likelihood function $\ell(p; x)$. Simplify the log likelihood so it has **no exponents**.

(d) [5 pts] Suppose we collect data from all the clones, and we know that clone n reported x_n heads. Write the likelihood function $\mathcal{L}(p; x_1, x_2, \ldots, x_{100})$ and the log likelihood function $\ell(p; x_1, x_2, \ldots, x_{100})$. Simplify the log likelihood so it has **a summation but no exponents**. (Note: feel free to just write "const" for any constant that will not affect the optimal value of p.)

(e) [6 pts] What is the estimate $\hat{p}$ of the parameter p obtained by maximum likelihood estimation as a function of $x_1, x_2, \ldots, x_{100}$? Please simplify it as much as you can.