CS 189/289A Introduction to Machine Learning
Spring 2024 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	
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] We seek to find $w \in \mathbb{R}^d$ that minimizes a real-valued cost function $J(w)$. We know that J is continuous and smooth, and it has one and only one global minimum. (There are no other constraints on J.) Select the true statements about gradient descent on J.

○ A: A step of gradient descent is $w \leftarrow w + \epsilon \nabla J(w)$, where $\epsilon > 0$ is the step size.

○ B: The gradient descent algorithm will always converge to the global minimum of J if the step size ϵ is sufficiently small.

● C: If the global minimum of J is at the vector w^*, steps of gradient descent on J starting from $w = w^*$ will never change w.

○ D: A step of gradient descent never causes $J(w)$ to increase.

A is ascent, not descent. (The sign is wrong.) B is wrong because we might land in a local minimum that is not the global minimum. C is correct because $\nabla J(w^*) = 0$, so a step of gradient descent does not change w. D is wrong because an excessively large step size can cause J to increase; see Lecture 5 for an example.

(b) [4 pts] Which statements are true for every symmetric, real matrix $S \in \mathbb{R}^{n \times n}$?

● A: All the eigenvalues of S are real.

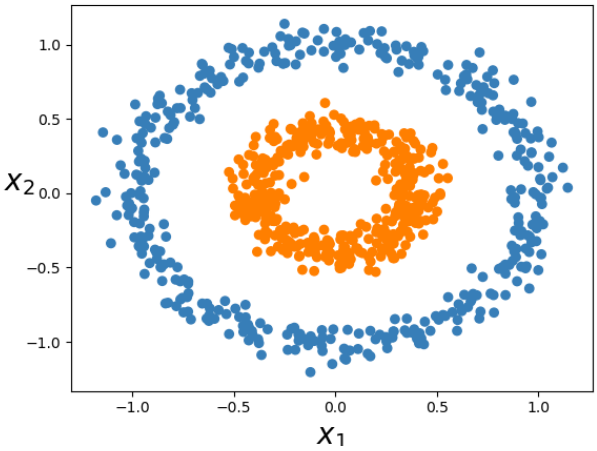
○ B: S can be written as $S = A^2$, where A is symmetric and belongs to $\mathbb{R}^{n \times n}$.

○ C: If S is positive semidefinite, then S is invertible.

● D: If all the eigenvalues of S are strictly less than zero, then S is invertible.

A is a standard result, part of the spectral theorem. For a counterexample to B, consider $-I$. In general, no matrix with a negative eigenvalue can be written as A^2. For a counterexample to C, consider $0_{n \times n}$. D is correct, because a matrix is invertible if and only if zero is not an eigenvalue of it. (To see that S has an inverse, consider the eigendecomposition $S = U \Lambda U^\top$. Then $S^{-1} = U \Lambda^{-1} U^\top$. As there are no zeros on the diagonal of Λ, Λ^{-1} exists; just take the reciprocal of each diagonal element of Λ.)

(c) [4 pts] You are given a two-class classification problem with the training points below. For each feature below, select it if adding it as a third feature (alongside x_1 and x_2) would make the two classes linearly separable.



○ A: x_2^2

● B: $\|x\|_2$

● C: $\|x\|_2^3$

○ D: $x_1 + x_2$

Choice A is incorrect since the points close to the x_1-axis (where x_2 is close to zero) still won't be linearly separable. B and C are correct because they permit the decision boundary to be a circle, centered at the origin, separating the orange and blue points. Choice D gives us no additional power at all.

(d) [4 pts] Which statements are true of **Gaussian discriminant analysis for two-class classification**, specifically quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)? (Assume that there are no added features.)

○ A: QDA for isotropic Gaussians (i.e., with the same variance in all directions) becomes the centroid method when the prior probabilities of the two classes are equal.

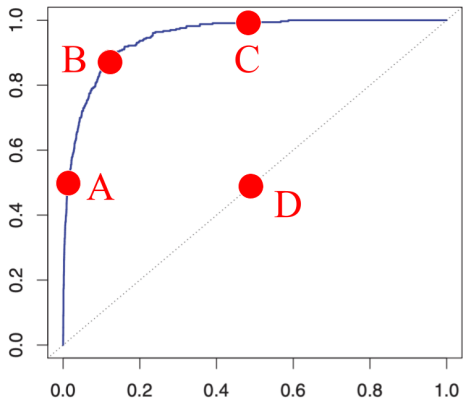
● B: QDA is more likely to overfit than LDA when the number of training points is small.

● C: LDA for isotropic Gaussians (i.e., with the same variance in all directions) becomes the centroid method when the prior probabilities of the two classes are equal.

○ D: LDA for anisotropic Gaussians can produce nonlinear decision boundaries.

- A: No, but LDA does. QDA doesn't because the covariance matrices usually differ.
- B: Yes. QDA is more likely to overfit than LDA because it has $\frac{d(d+3)}{2}$ parameters, while LDA only has d+1 parameters.
- C: Yes; see Lecture 7.
- D: No. LDA can only give nonlinear boundaries with added features.

(e) [4 pts] You trained a classifier whose **ROC curve** appears below. You would like to correctly classify points that are in-class, but you don't care what labels you assign to points that are out-of-class. In other words, your (asymmetric) loss function is 1 when your prediction is $z = -1$ and the true label is $y = 1$; otherwise, the loss is zero. The four red points below signify four classifiers. Of the four classifiers, which one has the **lowest empirical risk on the test points**?



○ A: Classifier A

○ B: Classifier B

● C: Classifier C

○ D: Classifier D

Since there is no loss for a false positive, you simply want to maximize the true positive rate to incur the least expected loss. That corresponds to the highest point on the vertical axis, which is C.

(f) [4 pts] Consider a **regression algorithm** whose cost function $J(w)$ is twice differentiable (continuous and smooth) for all $w \in \mathbb{R}^d$. (There are no other constraints on J). Consider also a second cost function $K(w) = J(w) + \lambda\|w\|^2$, which adds ℓ_2 **regularization** to J, where $\lambda > 0$ is the regularization parameter. Which statements are certain to be true?

○ A: The Hessian of $K(w)$ is positive definite for all $w \in \mathbb{R}^d$.

○ B: A step of gradient descent on K will move as far or farther than a step of gradient descent on J (with the same starting point w and learning rate ϵ for both).

● C: ℓ_2 regularization reduces the variance of the regression method if λ is sufficiently large.

● D: If the origin (i.e., the point $w^* = 0$) is a global minimum of J, then it is the only global minimum of K.

A would be true if we assumed J is convex, but we don't, and it's not. B: No; the gradient of J and the gradient of $\lambda\|w\|^2$ can easily point in opposite directions. C: Yes, because in the limit as $\lambda \to \infty$, w and the variance both approach zero. D: Yes, because for every point $v \neq 0$, $K(v) > J(v) \geq J(0) = K(0)$.

(g) [4 pts] We are given a set of linearly separable training points of two classes, with at least one point in each class. We find the **maximum margin classifier** for these points—that is, we successfully train a hard-margin support vector machine (with the usual constrants, $y_i(X_i \cdot w + \alpha) \geq 1$). Which statements are true?

○ A: In the maximum margin classifier, at least one class has at least two support vectors.

● B: The maximum margin classifier is always unique.

○ C: The weight vector w for the maximum margin classifier always has Euclidean length (magnitude) 1.

● D: Finding a linear classifier that correctly classifies all the points can be done by solving a linear program, but finding the maximum margin classifier involves solving a quadratic program whose cost function is not linear.

- A: No, just one support vector in each class is sufficient for uniqueness.
- B: Yes; so long as there is at least one point of each class, the maximum margin classifier is the unique hyperplane that bisects the shortest line segment connecting the convex hull of the points in the first class to the convex hull of the points in the second class.
- C: No; the length of the weight vector is the reciprocal of the width of the maximum margin.
- D: Yes; see Lectures 4 and 5.

(h) [4 pts] Given an $n \times d$ design matrix X and a vector of labels $y \in \mathbb{R}^n$, we perform **least-squares linear regression**—that is, we find a w that minimizes $\|Xw - y\|^2$. Which statements are true of every minimizer w^*?

○ A: Every minimizer can be written as $w^* = X^+y + z$ for some $z \in \text{Row}\, X$.

○ B: Every minimizer can be written as $w^* = X^+y + z$ for some $z \in \text{Col}\, X$.

● C: Every minimizer can be written as $w^* = X^+y + z$ for some $z \in \text{Null}\, X$.

○ D: Every minimizer can be written as $w^* = X^+y + z$ for some $z \in \text{Null}\, X^\top$.

We showed in Discussion 6 that every solution to $X^\top Xw = X^\top y$ can be written in the form $w = w_0 + z$ where $w_0 = X^+y$ is the unique solution in the row space of X and z is some component in the null space of X.

(i) [4 pts] Which statements are true about **ridge regression and Lasso** (with $\lambda > 0$ for both).

○ A: Ridge regression has a unique solution if and only if the design matrix has full rank.

○ C: Ridge regression can be formulated as a linear programming problem.

● B: There are points in feature space where the gradient of Lasso's cost function is not defined.

● D: One of Lasso's virtues is its tendency to set some weights to zero.

- A: Incorrect. Ridge regression always has a unique solution no matter the rank of X.

- B: Correct. At any point where some coordinate x_i is zero, the directional derivative of Lasso's cost function is undefined in the direction of the x_i-axis.
- C: Incorrect, the objective function of ridge regression is quadratic.
- D: Of course.

(j) [4 pts] Two classes of observations are drawn from two **univariate normal** distributions: $D_1 = \mathcal{N}(\mu_1, \sigma_1^2)$ for class 1 and $D_2 = \mathcal{N}(\mu_2, \sigma_2^2)$ for class 2. We know the parameters and prior probabilities of each class (the priors may or may not be equal), and we construct their Bayes classifier (with a 0-1 loss function). Which statements are true of the **Bayes optimal decision boundary**?

- ● A: It might be $\emptyset$ (no points).
- ○ C: It might have exactly three points.
- ● B: It might have exactly two points.
- ○ D: It might have exactly ten points.

The Bayes optimal decision boundary consists of the real roots of a quadratic equation, so it cannot have more than two points. It is easy to exhibit the zero-point and two-point cases by choosing the parameters accordingly. Note that the zero-point case requires the two distributions to have different prior probabilities. The two-point case is easier to see if you give the two distributions different variances.

(k) [4 pts] Which of the following classifiers are guaranteed to assign the same classes to the test data if we apply to all points (training and test points) an invertible linear transformation that **whitens** the training points? (By "the same classes," we mean the same predictions as if we didn't whiten the data.)

- ○ A: Soft-margin support vector machine
- ● C: Quadratic discriminant analysis
- ○ B: k-nearest neighbor classifier
- ● D: Linear discriminant analysis

- A: The soft-margin SVM cost function is influenced by the magnitudes of the features, so rescaling features can change their influence on the decision boundary. Whitening can stretch space in some directions and shrink it in others.
- B: Whitening changes distances between points in an anisotropic fashion, so a point's nearest neighbor can change.
- C: Whitening the data also whitens the sample means and covariances, thus any point in the whitened space will still have the same posterior probability as the corresponding un-whitened point in the original space.
- D: The sample covariance is shared across all classes in LDA, but whitening will also whiten it proportionally, so the same argument as QDA applies.

(l) [4 pts] Consider a **continuous uniform distribution** $\mathcal{U}[0, b]$, from which we draw a random real number between 0 and b. The probability density function (PDF) of $\mathcal{U}[0, b]$ is

$$f(x) = \begin{cases} 1/b & 0 \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$$

We want to use **maximum likelihood estimation** to estimate the parameter b. We draw three points at random from $\mathcal{U}[0, b]$ and obtain $x_1 = 44.4$, $x_2 = 8$, and $x_3 = 41.2$. What is the maximum likelihood estimate $\hat{b}$ of b?

- ○ A: $\hat{b} = 44.4 + 8 + 41.2$.
- ○ C: $\hat{b} = (44.4 + 8 + 41.2)/3$.
- ● B: $\hat{b} = 44.4$.
- ○ D: $\hat{b} = \frac{4}{3} \cdot 44.4$.

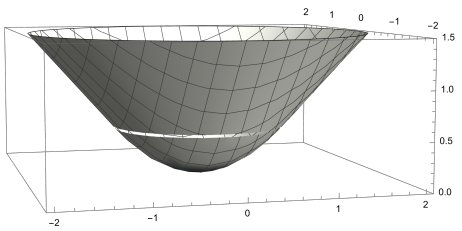
MLE chooses $\hat{b}$ to maximize $\mathcal{L}(b) = f(x_1)f(x_2)f(x_3)$, which equals $1/b^3$ if $b \ge 44.4$ and zero otherwise. Hence, $\hat{b} = 44.4$.

Extra space: if you need extra space for your answer to a written problem on pages 5–7, you may write here. **Be sure to write "see page 4" under the unfinished answer!**

Q2. [18 pts] Optimizing Huber Loss

Given a vector prediction $z \in \mathbb{R}^k$, a vector true label $y \in \mathbb{R}^k$, a fixed constant $\delta > 0$, and the ℓ_2-norm $\|v\| = \sqrt{v^\top v}$, the Huber ℓ_2-loss function is

$$L_\delta(z, y) = \begin{cases} \frac{1}{2}\|z - y\|^2, & \|z - y\| \le \delta, \\ \delta \cdot (\|z - y\| - \frac{1}{2}\delta), & \text{otherwise.} \end{cases}$$



At right is a plot of $L_1(z, 0)$ for $k = 2$. This loss can be used for regression where the regression function returns a k-dimensional vector.

The Huber ℓ_2-loss is designed to be similar to the loss function $\|z - y\|$ (i.e., the Euclidean distance, which in 1D we call the absolute loss); but unlike the Euclidean distance, it is smooth at the minimum, $z = y$. For a fixed y, the Huber ℓ_2-loss is quadratic in z in a small region near y, but it is shaped like a cone farther away from y. The Huber ℓ_2-loss is continuous and convex.

(a) [7 pts] Compute the gradient $\nabla_z L_\delta(z, y)$ of the ℓ_2-Huber loss (for a fixed δ and y).

$$\nabla_z L_\delta(z, y) = \begin{cases} z - y, & \|z - y\| \le \delta, \\ \frac{\delta}{\|z-y\|}(z - y), & \text{otherwise.} \end{cases}$$

(The second case is the tricky one. Recall that $\|z - y\| = (\|z - y\|^2)^{1/2}$, so by the chain rule, $\frac{\partial}{\partial z}\|z - y\| = \frac{1}{2}(\|z - y\|^2)^{-1/2} 2(z - y)$.)

(b) [4 pts] If we optimize z with gradient descent on L_δ (for a fixed y), **what learning rate (step size)** ϵ guarantees that we will eventually reach the exact minimum (rather than just inching closer and closer forever)? **Why?**

$\epsilon = 1$, because it ensures that gradient descent from any point z in the parabolic regions jumps directly to y. (In that region, the gradient descent rule is $z \leftarrow z - \epsilon(z - y)$.)

(c) [2 pts] Suppose we use Newton's method to find a z that minimizes the ℓ_2-Huber loss. (Technically, the Hessian of L_δ with respect to z is not defined where $\|z - y\| = \delta$, but we fix that by simply using the Hessian of $\frac{1}{2}\|z - y\|^2$ at those points.)

If we start at a point $z = z_0$ that satisfies $\|z_0 - y\| < \delta$, what will the value of z be after one step of Newton's method? **Why**?

One step of Newton's method will set $z = y$, because Newton's method approximates the function $\frac{1}{2}\|z - y\|^2$ with the paraboloid $\frac{1}{2}\|z - y\|^2$ (i.e., the same function) and jumps directly to the bottom of the parabola.

(d) [3 pts] If we start at a point $z = z_0$ that satisfies $\|z_0 - y\| > \delta$, what will one step of Newton's method do? **Why**? (Note: for this question and the next one, we want a qualitative answer; you don't need to calculate a Hessian.)

The Hessian is singular, so Newton's method will fail. [The Hessian has an eigenvalue of zero corresponding to the eigenvector direction $z - y$, the direction with no curvature.]

(e) [2 pts] Suppose we add an ℓ_2 regularization term $\lambda\|z\|^2$ to the ℓ_2-Huber loss, with $\lambda > 0$, and perform one step of Newton's method on the ℓ_2-regularized ℓ_2-Huber loss. How do your answers to (c) and (d) change (qualitatively), and **why**?

The answer to (c) will no longer be y, unless y happens to be the origin.

The answer to (d) changes because the Hessian will always be positive definite, so Newton's method will always be able to take a step. [It won't reach the minimum in one step, though.]

Q3. [17 pts] Quadratic Discriminant Analysis

We want to predict whether a person prefers vanilla or chocolate ice cream based on a single feature: their age. We suspect that the ages of vanilla-lovers are normally distributed, and so are the ages of chocolate-lovers, so we build a classifier with **quadratic discriminant analysis (QDA)** and **a 0-1 loss function**. Our survey of 13 random people turns up 8 vanilla lovers and 5 chocolate lovers of the following ages.

Vanilla: [21, 26, 27, 28, 30, 30, 31, 31] Chocolate: [15, 18, 21, 22, 24]

(a) [17 pts] Please do QDA. Determine the **distribution parameters and prior probabilities** of vanilla lovers and chocolate lovers (as exact, simplified integers or fractions). Then determine the **probability that a person of age x prefers vanilla over chocolate** (substituting the numbers so your answer is an exact, simplified function of x, which can include logistic functions $s(\cdot)$ or exponentials). Also, determine the **decision boundary** (as one or more numbers written as simplified expressions, possibly with logarithms and fractions). Show all your work! (Hint: as MNIST taught us, $28^2 = 784$.)

$$\begin{aligned}
\hat{\mu}_{\text{vanilla}} &= \frac{1}{8}(21+26+27+28+30+30+31+31) = \frac{1}{8}(224) = 28. \\
\hat{\sigma}^2_{\text{vanilla}} &= \frac{1}{8}((21-28)^2+(26-28)^2+(27-28)^2+(28-28)^2+(30-28)^2+(30-28)^2+(31-28)^2+(31-28)^2) \\
&= \frac{1}{8}(49+4+1+0+4+4+9+9) = \frac{1}{8}(80) = 10. \\
\hat{\pi}_{\text{vanilla}} &= 8/13. \\
\hat{\mu}_{\text{chocolate}} &= \frac{1}{5}(15+18+21+22+24) = \frac{1}{5}(100) = 20. \\
\hat{\sigma}^2_{\text{chocolate}} &= \frac{1}{5}((15-20)^2+(18-20)^2+(21-20)^2+(22-20)^2+(24-20)^2) = \frac{1}{5}(25+4+1+4+16) = \frac{1}{5}(50) = 10. \\
\hat{\pi}_{\text{chocolate}} &= 5/13.
\end{aligned}$$

[1 point for each estimate above, summing to 6 points.]

From here, there are two ways to proceed: you could work out the quadratic discriminant functions, or you could do it the hard way: by equating the posterior probabilities and plugging in the normal PDFs. [We'll need a different scoring system for each.]

The quadratic discriminant functions are

$$\begin{aligned}
Q_{\text{vanilla}}(x) &= -\frac{\|x-\mu_{\text{vanilla}}\|^2}{2\sigma^2_{\text{vanilla}}} - d\ln\sigma_{\text{vanilla}} + \ln\pi_{\text{vanilla}} = -\frac{\|x-28\|^2}{20} - d\ln\sqrt{10} + \ln\frac{8}{13}, \\
Q_{\text{chocolate}}(x) &= -\frac{\|x-20\|^2}{20} - d\ln\sqrt{10} + \ln\frac{5}{13}.
\end{aligned}$$

The QDA decision function is $Q_{\text{vanilla}}(x) - Q_{\text{chocolate}}(x) = -\frac{\|x-28\|^2}{20} + \frac{\|x-20\|^2}{20} + \ln\frac{8}{13} - \ln\frac{5}{13} = \frac{4}{5}x - \frac{96}{5} + \ln\frac{8}{5}$.

Hence, the posterior probability that a person of age x prefers vanilla is

$$P(Y=\text{vanilla}|X=x) = s(Q_{\text{vanilla}}(x) - Q_{\text{chocolate}}(x)) = s\left(\frac{4}{5}x - \frac{96}{5} + \ln\frac{8}{5}\right).$$

You could stop here and get all the points for the posterior probability, or you could simplify a bit more. For example,

$$P(Y=\text{vanilla}|X=x) = \frac{1}{1+e^{-4x/5+96/5-\ln(8/5)}} = \frac{1}{1+5e^{-4x/5+96/5}/8}.$$

[6 points for writing the correct posterior probability in a form as simple as these, with a sliding scale for increasing less simplified or correct variants as well as variants where not all the substitutions were made. 5 points of partial credit for writing the decision function in a simplified, correct form with all substitutions. If that is not earned, 2 points of partial credit for expressing $Q_{\text{vanilla}}(x)$ in the simplest possible form and 2 more for $Q_{\text{chocolate}}(x)$.]

The decision boundary is the set of points where the decision function is zero. That is,

$$x = 24 - \frac{5}{4}\ln\frac{8}{5} \text{ years old.}$$

(FYI, this is approximately 23.41. Note that this can also be written $x = 24 + \frac{5}{4}\ln\frac{5}{8}$.)

[5 points for writing the correct decision boundary in the simplest possible form, with a sliding scale for increasing less simplified or correct variants as well as variants where not all the substitutions were made. 2 points of partial credit for simply setting the decision function to zero, or setting $Q_{\text{vanilla}}(x) = Q_{\text{chocolate}}(x)$.]

Here's the more tedious alternative solution, which equates the posterior probabilities. First we find the decision boundary.

$$\begin{aligned}
P(Y = \text{chocolate}|X = x) &= P(Y = \text{vanilla}|X = x) \\
\frac{f(X|Y = \text{chocolate})\,\hat{\pi}_{\text{chocolate}}}{f(X = x)} &= \frac{f(X|Y = \text{vanilla})\,\hat{\pi}_{\text{vanilla}}}{f(X = x)} \\
\frac{5}{13\sqrt{2\pi\cdot 10}}\exp\left(-\frac{(x-20)^2}{2\cdot 10}\right) &= \frac{8}{13\sqrt{2\pi\cdot 10}}\exp\left(-\frac{(x-28)^2}{2\cdot 10}\right) \\
\ln 5 - \frac{(x-20)^2}{20} &= \ln 8 - \frac{(x-28)^2}{20} \\
(x^2 - 56x + 784) - (x^2 - 40x + 400) &= 20\ln\frac{8}{5} \\
384 - 20\ln\frac{5}{8} &= 16x \\
x &= 24 - \frac{5}{4}\ln\frac{8}{5} \text{ years old.}
\end{aligned}$$

[5 points for writing the correct decision boundary in the simplest possible form, with a sliding scale, as above. 2 points of partial credit for simply getting to the second line (equating posteriors + applying Bayes' Theorem).]

The posterior probability that a person of age x prefers vanilla is

$$\begin{aligned}
P(Y = \text{vanilla}|X = x) &= \frac{f(X|Y = \text{vanilla})\,\hat{\pi}_{\text{vanilla}}}{f(X = x)} \\
&= \frac{f(X|Y = \text{vanilla})\,\hat{\pi}_{\text{vanilla}}}{f(X|Y = \text{vanilla})\,\hat{\pi}_{\text{vanilla}} + f(X|Y = \text{chocolate})\,\hat{\pi}_{\text{chocolate}}} \\
&= \frac{1}{1 + \frac{f(X|Y=\text{chocolate})\,\hat{\pi}_{\text{chocolate}}}{f(X|Y=\text{vanilla})\,\hat{\pi}_{\text{vanilla}}}} \\
&= \frac{1}{1 + \frac{\sqrt{2\pi\cdot 10}}{\sqrt{2\pi\cdot 10}}\exp\left(-\frac{(x-20)^2}{2\cdot 10}\right)\exp\left(\frac{(x-28)^2}{2\cdot 10}\right)\frac{5}{8}} \\
&= \frac{1}{1 + 5\exp\left(\frac{(x^2-56x+784)-(x^2-40x+400)}{20}\right)/8} \\
&= \frac{1}{1 + 5\exp\left(\frac{-16x+384}{20}\right)/8} \\
&= \frac{1}{1 + 5\exp\left(\frac{-4x+96}{5}\right)/8}.
\end{aligned}$$

[6 points for writing the correct posterior probability in a form as simple as this, with a sliding scale, as above. 2 points of partial credit for simply getting to the second line. If that is not earned, 1 point for simply getting to the first line (Bayes' Theorem).]

Q4. [17 pts] Estimating the Noise in Linear Regression

In Lecture 12 we suggested a model of reality in which we want to determine a linear natural law $g(z) = v^\top z$ (g = "ground truth") mapping each data point $z \in \mathbb{R}^d$ to a label in $\mathbb{R}$. (For simplicity, we don't use a bias term α in this question; assume our natural law satisfies $g(0) = 0$.) But the measurements are noisy, so what we get is an $n \times d$ design matrix X and a vector $y \in \mathbb{R}^n$ of labels such that $y_i = v^\top X_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is random noise (and each ϵ_i is independent of the others). In class we applied maximum likelihood estimation (MLE) to justify using the mean of the squared losses as the cost function for linear regression to compute a weight vector $w \in \mathbb{R}^d$ that is an estimate of v. Now we will use MLE to estimate σ^2, the variance of the measurement noise.

(a) [5 pts] Write the **likelihood function** $\mathcal{L}(\sigma; y, X, v)$ for obtaining the labels y_i given fixed values of X and v. (Note: for the purposes of this problem, X and v are **not** random. There should be no μ or other unlisted parameters in your answer.)

$$\mathcal{L}(\sigma; y, X, v) = \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi}\sigma)} \exp\left(-\frac{(y_i - v^\top X_i)^2}{2\sigma^2}\right).$$

(b) [6 pts] Write the **log likelihood function** $\ell(\sigma; y, X, v)$ and **find the value of σ^2 that maximizes** ℓ. Show your work. (Note: you do not need to prove it's a maximum.)

$$\begin{aligned} \ell(\sigma; y, X, v) &= -n \ln \sqrt{2\pi} - n \ln \sigma - \sum_{i=1}^{n} \frac{(y_i - v^\top X_i)^2}{2\sigma^2}, \\ \frac{d}{d\sigma}\ell(\sigma; y, X, v) &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - v^\top X_i)^2. \end{aligned}$$

Setting $\frac{d\ell}{d\sigma} = 0$ gives the maximizer

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - v^\top X_i)^2.$$

(c) [2 pts] What formula that you're familiar with does your optimal value of σ^2 look like? ("The mean variance of the labels y_i" doesn't count. It's something else too.)

It happens to be the cost function for linear least-squares regression, if we interpret v as a vector of weights.

(d) [4 pts] Unfortunately, we don't know the value of v. How should we estimate σ^2, given that we cannot obtain v? **Write an estimate of your estimate for σ^2 expressed solely in terms of X, y, and n, with no v.** (This is your maximum likelihood estimator $\hat{\sigma}^2$ for the true σ^2.) For full points, write your final answer in matrix notation with no summation. You may assume that X has rank d.

The weights w computed by least-squares linear regression are our estimate of v. These weights are $w = X^+ y$, so we have

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (X^+ y)^\top X_i)^2 = \frac{1}{n} \|y^\top - (X^+ y)^\top X^\top\|^2 = \frac{1}{n} \|y - XX^+ y\|^2.$$

The first formula is worth three points and either of the last two is worth four. As X has rank d, $X^\top X$ is invertible and you can write $w = (X^\top X)^{-1} X^\top y$, so we will also accept

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - ((X^\top X)^{-1} X^\top y)^\top X_i)^2 = \frac{1}{n} \|y^\top - ((X^\top X)^{-1} X^\top y)^\top X^\top\|^2 = \frac{1}{n} \|y - X(X^\top X)^{-1} X^\top y\|^2.$$