CS 189 Spring 2017 Introduction to Machine Learning Midterm

- Please do not open the exam before you are instructed to do so.
- The exam is closed book, closed notes except your one-page cheat sheet.
- **Electronic devices are forbidden on your person**, including cell phones, iPods, headphones, and laptops. Turn your cell phone off and **leave all electronics at the front of the room**, or **risk getting a zero** on the exam.
- You have 1 hour and 20 minutes.
- Please write your initials at the top right of each page after this one (e.g., write "JS" if you are Jonathan Shewchuk). Finish this by the end of your 1 hour and 20 minutes.
- Mark your answers on the exam itself in the space provided. Do **not** attach any extra sheets.
- The total number of points is 100. There are 20 multiple choice questions worth 3 points each, and 4 written questions worth a total of 40 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. [60 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) [3 pts] For a *nonconvex* cost function J, which of the following step sizes guarantee that batch gradient descent will converge to the global optimum? Let i denote the ith iteration.

○ $\epsilon = 10^{-2}$

○ $\epsilon = \frac{1}{\nabla^2 J}$

○ $\epsilon = 10^{-i}$

● None of the above

(b) [3 pts] Which of the following optimization algorithms attains the optimum of an unconstrained, quadratic, convex cost function in the fewest iterations?

○ Batch gradient descent

● Newton's method

○ Stochastic gradient descent

○ The simplex method

(c) [3 pts] You train a linear classifier on 10,000 training points and discover that the training accuracy is only 67%. Which of the following, done in isolation, has a good chance of improving your training accuracy?

● Add novel features

○ Use linear regression

○ Train on more data

● Train on less data

(d) [3 pts] You train a classifier on 10,000 training points and obtain a training accuracy of 99%. However, when you submit to Kaggle, your accuracy is only 67%. Which of the following, done in isolation, has a good chance of improving your performance on Kaggle?

○ Set your regularization value (λ) to 0

● Use validation to tune your hyperparameters

● Train on more data

○ Train on less data

(e) [3 pts] You are trying to improve your Kaggle score for the spam dataset, but you must use logistic regression with no regularization. So, you decide to extract some additional features from the emails, but you forget to normalize your new features. You find that your Kaggle score goes down. Why might this happen?

○ The new features make the sample points linearly separable

● The new features have significantly more noise and larger variances than the old features

● The new features are uncorrelated with the emails being HAM or SPAM

● The new features are linear combinations of the old features

(f) [3 pts] In a soft-margin support vector machine, if we increase C, which of the following are likely to happen?

○ The margin will grow wider

● Most nonzero slack variables will shrink

○ There will be more points inside the margin

● The norm $|w|$ will grow larger

(g) [3 pts] If a hard-margin support vector machine tries to minimize $|w|^2$ subject to $y_i(X_i \cdot w + \alpha) \geq 2$ instead, what will be the width of the slab (the point-free region bracketing the decision boundary)?

○ $\frac{1}{\|w\|}$

● $\frac{4}{\|w\|}$

○ $\frac{2}{\|w\|}$

○ $\frac{1}{2\|w\|}$

(h) [3 pts] There is a 50% chance of rain on Saturday and a 30% chance of rain on Sunday. However, it is twice as likely to rain on Sunday if it rains on Saturday than if it does not rain on Saturday. What is the probability it rains on neither of the days?

○ 15%

● 40%

○ 25%

○ 45%

(i) [3 pts] The Bayes risk for a decision problem is zero when

○ the training data is linearly separable after lifting it to a higher-dimensional space.

○ the Bayes decision rule perfectly classifies the training data.

● the class distributions $P(X|Y)$ do not overlap.

● the prior probability for one class is 1.

(j) [3 pts] Consider using a Bayes decision rule classifier in a preliminary screen for cancer patients, as in Lecture 6. We want to reduce the probability that someone is classified as cancer-free when they do, in fact, have cancer. On the ROC curve for the classifier, an asymmetric loss function that implements this strategy

● Picks a point on the curve with higher *sensitivity* than the 0-1 loss function.

○ Picks a point on the curve that's closer to the y-axis than the 0-1 loss function.

○ Picks a point on the curve with higher *specificity* than the 0-1 loss function.

● Picks a point on the curve that's further from the x-axis than the 0-1 loss function.

(k) [3 pts] For which of the following cases are the scalar random variables X_1 and X_2 guaranteed to be independent?

○ $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 \sim \mathcal{N}(0, 1)$.

○ $\mathbb{E}\left[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])\right] = -1$

● $\text{Cov}(X_1, X_2) = 0$ and $[X_1 \quad X_2]^\top$ has a multivariate normal distribution.

● $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 & 0 \\ 0 & 7 \end{bmatrix}\right)$

(l) [3 pts] Given $X \sim \mathcal{N}(0, \Sigma)$ where the precision matrix Σ^{-1} has eigenvalues λ_i for $i = 1, \ldots, d$, the isocontours of the probability density function for X are ellipsoids whose relative axis lengths are

○ λ_i

○ $\sqrt{\lambda_i}$

○ $1/\lambda_i$

● $1/\sqrt{\lambda_i}$

(m) [3 pts] In LDA/QDA, what are the effects of modifying the sample covariance matrix as $\tilde{\Sigma} = (1 - \lambda)\Sigma + \lambda I$, where $0 < \lambda < 1$?

● $\tilde{\Sigma}$ is positive definite

● $\tilde{\Sigma}$ is invertible

○ Increases the eigenvalues of Σ by λ

● The isocontours of the quadratic form of $\tilde{\Sigma}$ are closer to spherical

(n) [3 pts] Let w^* be the solution you obtain in standard least-squares linear regression. What solution do you obtain if you scale all the input features (but not the labels y) by a factor of c before doing the regression?

● $\frac{1}{c}w^*$

○ cw^*

○ $\frac{1}{c^2}w^*$

○ c^2w^*

(o) [3 pts] In least-squares linear regression, adding a regularization term can

● increase training error.

● increase validation error.

○ decrease training error.

● decrease validation error.

(p) [3 pts] You have a design matrix $X \in \mathbb{R}^{n\times d}$ with $d = 100{,}000$ features and and vector $y \in \mathbb{R}^n$ of binary 0-1 labels. When you fit a logistic regression model to your design matrix, your test error is much worse than your training error. You suspect that many of the features are useless and are therefore causing overfitting. What are some ways to eliminate the useless features?

● Use ℓ_1 regularization.

○ Use ℓ_2 regularization.

● Iterate over features; check if removing feature i increases validation error; remove it if not.

○ If the ith eigenvalue λ_i of the sample covariance matrix is 0, remove the ith feature/column.

(q) [3 pts] Recall the data model, $y_i = f(X_i) + \epsilon_i$, that justifies the least-squares cost function in regression. The statistical assumptions of this model are, for all i,

● ϵ_i comes from a Gaussian distribution.

● all ϵ_i have the same mean

○ all y_i have the same mean

● all y_i have the same variance

(r) [3 pts] How does ridge regression compare to linear regression with respect to the bias-variance tradeoff?

● Ridge regression usually has higher bias.

○ Ridge regression usually has higher variance.

○ Ridge regression usually has higher irreducible error.

● Ridge regression's variance approaches zero as the regularization parameter $\lambda \to \infty$.

(s) [3 pts] Which of the following quantities affect the bias-variance tradeoff?

● λ, the regularization coefficient in ridge regression

○ ϵ, the learning rate in gradient descent

● C, the slack parameter in soft-margin SVM

● d, the polynomial degree in least-squares regression

(t) [3 pts] Which of the following statements about maximum likelihood estimation are true?

● MLE, applied to estimate the mean parameter μ of a normal distribution $\mathcal{N}(\mu, \Sigma)$ with a known covariance matrix Σ, returns the mean of the sample points

○ For a sample drawn from a normal distribution, the likelihood $\mathcal{L}(\mu, \sigma; X_1, \ldots, X_n)$ is equal to the probability of drawing exactly the points $X_1, \ldots, X_n$ (in that order) when you draw n random points from $\mathcal{N}(\mu, \sigma)$

○ MLE, applied to estimate the covariance parameter Σ of a normal distribution $\mathcal{N}(\mu, \Sigma)$, returns $\hat{\Sigma} = \frac{1}{n}X^T X$, where X is the design matrix

● Maximizing the log likelihood is equivalent to maximizing the likelihood

Q2. [10 pts] Logistic Posterior for Poisson Distributions

Consider two classes C and D whose class conditionals are discrete Poisson distributions with means $\lambda_{\mathrm{C}} > 0$ and $\lambda_{\mathrm{D}} > 0$. Their probability mass functions are

$$P(K = k|Y = \mathrm{C}) = \frac{\lambda_{\mathrm{C}}^k\, e^{-\lambda_{\mathrm{C}}}}{k!}, \quad P(K = k|Y = \mathrm{D}) = \frac{\lambda_{\mathrm{D}}^k\, e^{-\lambda_{\mathrm{D}}}}{k!}, \quad k \in \{0, 1, 2, \ldots\}.$$

Their prior probabilities are $P(Y = \mathrm{C}) = \pi_{\mathrm{C}}$ and $P(Y = \mathrm{D}) = \pi_{\mathrm{D}} = 1 - \pi_{\mathrm{C}}$. We use the standard 0-1 loss function.

(a) [7 pts] Derive the posterior probability and show that it can be written in the form $P(Y = \mathrm{C}|K = k) = s(f(k, \lambda_{\mathrm{C}}, \lambda_{\mathrm{D}}, \pi_{\mathrm{C}}))$, where s is the logistic function and f is another function.

By Bayes' Theorem, we have

$$\begin{aligned}
P(Y = \mathrm{C}|K = k) &= \frac{P(K = k|Y = \mathrm{C})\,\pi_{\mathrm{C}}}{P(K = k|Y = \mathrm{C})\,\pi_{\mathrm{C}} + P(K = k|Y = \mathrm{D})\,\pi_{\mathrm{D}}} \\
&= \frac{1}{1 + \frac{P(K=k|Y=\mathrm{D})\,\pi_{\mathrm{D}}}{P(K=k|Y=\mathrm{C})\,\pi_{\mathrm{C}}}} \\
&= \frac{1}{1 + \frac{\lambda_{\mathrm{D}}^k e^{-\lambda_{\mathrm{D}}}(1-\pi_{\mathrm{C}})}{\lambda_{\mathrm{C}}^k e^{-\lambda_{\mathrm{C}}}\pi_{\mathrm{C}}}} \\
&= \frac{1}{1 + \exp\Big(- \big(k \ln \frac{\lambda_{\mathrm{C}}}{\lambda_{\mathrm{D}}} + \lambda_{\mathrm{D}} - \lambda_{\mathrm{C}} - \ln \frac{1-\pi_{\mathrm{C}}}{\pi_{\mathrm{C}}}\big)\Big)} \\
&= \frac{1}{1 + \exp(-f(k, \lambda_{\mathrm{C}}, \lambda_{\mathrm{D}}, \pi_{\mathrm{C}}))} \\
&= s(f(k, \lambda_{\mathrm{C}}, \lambda_{\mathrm{D}}, \pi_{\mathrm{C}}))
\end{aligned}$$

(b) [3 pts] What is the maximum number of points in the Bayes optimal decision boundary? (Note: as the distribution is discrete, we are really asking for the maximum number of integral values of k where the classifier makes a transition from predicting one class to the other.)

As f is linear in k, there is only one root, and the decision boundary is a single point.

Q3. [10 pts] Error-Prone Sensors

We want to perform linear regression on the outputs of d building sensors measured at n different times, to predict the building's energy use. Unfortunately, some of the sensors are inaccurate and prone to large errors and, occasionally, complete failure. Fortunately, we have some knowledge of the relative accuracy and magnitudes of the sensors.

Let X be a $n \times (d+1)$ design matrix whose first d columns represent the sensor measurements and whose last column is all 1's. (Each sensor column has been normalized to have variance 1.) Let y be a vector of n target values, and let w be a vector of $d+1$ weights (the last being a bias term α). We decide to minimize the cost function

$$J(w) = \|Xw - y\|_1 + \lambda w^\top D w,$$

where D is a diagonal matrix with diagonal elements D_{ii} (with $D_{d+1,d+1} = 0$ so we don't penalize the bias term).

(a) [2 pts] Why might we choose to minimize the ℓ_1-norm $\|Xw - y\|_1$ as opposed to the ℓ_2-norm $|Xw - y|^2$ in this scenario?

Least-squares regression gives too much power to outliers, which is inappropriate for inaccurate or failing sensors. The ℓ_1-normalized cost function does not try as hard to fit the outliers.

(b) [2 pts] Why might we choose to minimize $w^\top D w$ as opposed to $|w'|^2$? What could the values D_{ii} in D represent?

We might want to more heavily penalize the weights associated with the less accurate sensors. Each D_{ii} can be thought of how much we don't trust sensor i.

(c) [6 pts] Derive the batch gradient descent rule to minimize our cost function. Hint: let p be a vector with components $p_i = \text{sign}(X_i^\top w - y_i)$, and observe that $\|Xw - y\|_1 = (Xw - y)^\top p$. For simplicity, assume that no $X_i^\top w - y_i$ is ever exactly zero.

$$\begin{aligned}
\nabla_w(\|Xw - y\|_1 + \lambda w^\top D w) &= \nabla_w((Xw - y)^\top p + \lambda w^\top D w) \\
&= \nabla_w(w^\top X^\top p - y^\top p + \lambda w^\top D w) \\
&= X^\top p + 2\lambda D w
\end{aligned}$$

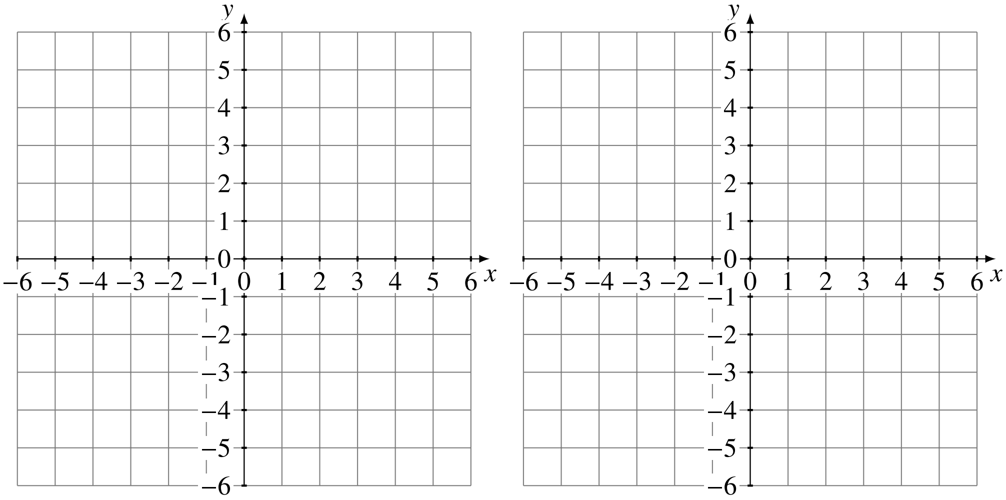
Therefore, the update rule is $w^{(t+1)} \leftarrow w^{(t)} - \epsilon(X^\top p + 2\lambda D w)$.

Q4. [10 pts] Gaussian Discriminant Analysis

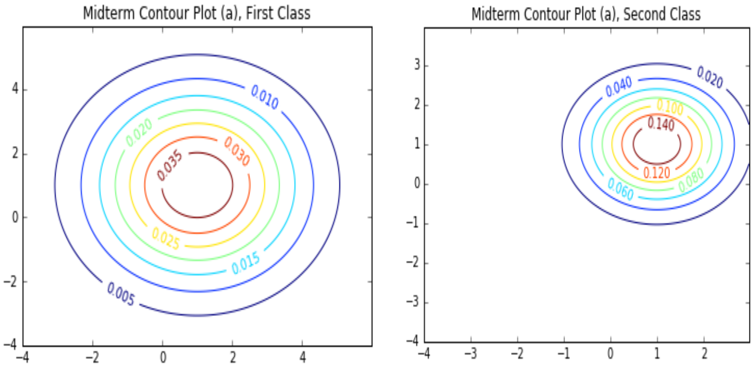
Consider a two-class classification problem in $d = 2$ dimensions. Points from these classes come from multivariate Gaussian distributions with a common mean but different covariance matrices.

$$X_C \sim \mathcal{N}\left(\mu = \begin{bmatrix}1\\1\end{bmatrix}, \Sigma_C = \begin{bmatrix}4 & 0\\0 & 4\end{bmatrix}\right), \quad X_D \sim \mathcal{N}\left(\mu = \begin{bmatrix}1\\1\end{bmatrix}, \Sigma_D = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right).$$

(a) [5 pts] Plot some isocontours of the probability distribution function $P(\mu, \Sigma_C)$ of X_C on the left graph. (The particular isovalues don't matter much, so long as we get a sense of the isocontour shapes.) Plot the isocontours of $P(\mu, \Sigma_D)$ *for the same isovalues* (so we can compare the relative spacing) on the right graph.



As there are no covariance terms, the isocontours are axis-aligned. As the variances are equal, the isocontours are circles. The student should make an attempt to demonstrate that they understand that higher standard deviations results in larger "gaps" between significant isocontours, as below.


Midterm Contour Plot (a), First Class

Midterm Contour Plot (a), Second Class

(b) [5 pts] Suppose that the priors for the two classes are $\pi_C = \pi_D = \frac{1}{2}$ and we use the 0-1 loss function. Derive an equation for the points x in the Bayes optimal decision boundary and simplify it as much as possible. What is the geometric shape of this boundary? (Hint: try to get your equations to include the term $|x - \mu|^2$ early, then keep it that way.) (Hint 2: you can get half of these points by guessing the geometric shape.)

$$
\begin{aligned}
P(Y = 1|X) &= P(Y = 2|X) \\
P(X|Y = 1)\,\pi_{\mathrm{C}} &= P(X|Y = 2)\,\pi_{\mathrm{D}} \\
\frac{1}{2\pi\sqrt{|\Sigma_{\mathrm{C}}|}} \exp\left(-\frac{1}{2}(x-\mu)^{\top}\Sigma_{\mathrm{C}}^{-1}(x-\mu)\right) &= \frac{1}{2\pi\sqrt{|\Sigma_{\mathrm{D}}|}} \exp\left(-\frac{1}{2}(x-\mu)^{\top}\Sigma_{\mathrm{D}}^{-1}(x-\mu)\right) \\
\frac{1}{4}\exp\left(-\frac{1}{2}\frac{1}{4}|x-\mu|^2\right) &= \exp\left(-\frac{1}{2}|x-\mu|^2\right) \\
-\frac{1}{8}|x-\mu|^2 - \ln 4 &= -\frac{1}{2}|x-\mu|^2 \\
|x-\mu|^2 &= \frac{8}{3\ln 4}
\end{aligned}
$$

This is a circle with center (1, 1) and radius $\sqrt{\frac{8\ln 4}{3}} \approx 1.92$.

Q5. [10 pts] Quadratic Functions

(a) [4 pts] Derive the 2×2 symmetric matrix whose eigenvalues are 7 and 1, such that $(1, 1)$ is an eigenvector with eigenvalue 7.

From the eigendecomposition, we have

$$\frac{1}{\sqrt{2}}\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\begin{bmatrix}7 & 0\\ 0 & 1\end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix}=\begin{bmatrix}4 & 3\\ 3 & 4\end{bmatrix}.$$

(b) [4 pts] Is the function $f(x_1, x_2) = x_1^4 + 2x_1^2 + 3x_1x_2 + 2x_2^2 - 7x_1 - 12x_2 - 18$ convex? Justify your answer.

Yes. The Hessian of f is

$$\begin{bmatrix}4 + 12x_1^2 & 3\\ 3 & 4\end{bmatrix},$$

which, it follows from part (a), is positive definite for all values of x_1.

(c) [2 pts] Consider the cost function $J(w)$ for least-squares linear regression. Can $J(w)$ ever be unbounded below? In other words, is there a set of input sample points X and labels y such that we can walk along a path in weight space for which the cost function $J(w)$ approaches $-\infty$? Explain your answer.

No. $J(w)$ is a sum of squares, so it never drops below zero.