CS 189 Spring 2016 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 odd-numbered page (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 3 written questions worth a total of 40 points.
- For multiple-choice 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-choice 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 Choice

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] Which of the following learning algorithms will return a classifier if the training data is not linearly separable?

- ○ Hard-margin SVM
- ● Soft-margin SVM
- ○ Perceptron
- ● Linear Discriminant Analysis (LDA)

(b) [3 pts] With a soft-margin SVM, which samples will have non-zero slack variables ξ_i?

- ● All misclassified samples
- ● All samples inside the margin
- ○ All samples lying on the margin boundary
- ○ All samples outside the margin

(c) [3 pts] Recall the soft-margin SVM objective function $|\mathbf{w}|^2 + C\sum_i \xi_i$. Which value of C is most likely to overfit the training data?

- ○ $C = 0.01$
- ○ $C = 1$
- ○ $C = 0.00001$
- ● $C = 100$

(d) [3 pts] There are several ways to formulate the hard-margin SVM. Consider a formulation in which we try to directly maximize the margin β. The training samples are $X_1, X_2, \ldots, X_n$ and their labels are $y_1, y_2, \ldots, y_n$. Which constraints should we impose to get a correct SVM? (*Hint*: Recall the formula for the distance from a point to a hyperplane.) Maximize β subject to ...

- ○ $y_i X_i^T \mathbf{w} \leq \beta \quad \forall i \in [1, n]$.
- ● $y_i X_i^T \mathbf{w} \geq \beta \quad \forall i \in [1, n]$.
- ○ $|\mathbf{w}| \geq 1$.
- ● $|\mathbf{w}| = 1$.

(e) [3 pts] In the homework, you trained classifiers on the digits dataset. The features were the pixels in each image. What features could you add that would improve the performance of your classifier?

- ○ Maximum pixel intensity
- ● Average pixel intensity
- ● Number of enclosed regions
- ● Presence of a long horizontal line

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

- ● the class distributions $P(X|Y)$ do not overlap.
- ○ the training data is linearly separable.
- ○ the loss function $L(z, y)$ is symmetrical.
- ○ the Bayes decision rule perfectly classifies the training data.

(g) [3 pts] Let $L(z, y)$ be a loss function (where y is the true class and z is the predicted class). Which of the following loss functions will *always* lead to the same Bayes decision rule as L?

- ● $L_1(z, y) = aL(z, y)$, $a > 0$
- ○ $L_2(z, y) = aL(z, y)$, $a < 0$
- ● $L_3(z, y) = L(z, y) + b$, $b > 0$
- ● $L_4(z, y) = L(z, y) + b$, $b < 0$

(h) [3 pts] Gaussian discriminant analysis

- ○ models $P(Y = y|X)$ as a Gaussian.
- ● is an example of a generative model.
- ● models $P(Y = y|X)$ as a logistic function.
- ● can be used to classify points without ever computing an exponential.

(i) [3 pts] Which of the following are valid covariance matrices?

- ○ $A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$
- ○ $C = \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix}$
- ● $B = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$
- ● $D = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

(j) [3 pts] Consider a d-dimensional multivariate normal distribution that is isotropic (i.e., its isosurfaces are spheres). Let Σ be its $d \times d$ covariance matrix. Let I be the $d \times d$ identity matrix. Let σ be the standard deviation of any one component (feature). Then

- ○ $\Sigma = \sigma I$.
- ○ $\Sigma = \frac{1}{\sigma} I$.
- ● $\Sigma = \sigma^2 I$.
- ○ $\Sigma = \frac{1}{\sigma^2} I$.
- ○ None of the above.

(k) [3 pts] In least-squares linear regression, imposing a Gaussian prior on the weights is equivalent to

- ○ logistic regression
- ● L_2 regularization
- ○ adding a Laplace-distributed penalty term
- ○ L_1 regularization

(l) [3 pts] Logistic regression

- ○ assumes that we impose a Gaussian prior on the weights.
- ○ has a closed-form solution.
- ● minimizes a convex cost function.
- ● can be used with a polynomial kernel.

(m) [3 pts] Ridge regression

- ○ is more sensitive to outliers than ordinary least-squares.
- ○ adds an L_1-norm penalty to the cost function.
- ● reduces variance at the expense of higher bias.
- ○ often sets several of the weights to zero.

(n) [3 pts] Given a design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and labels $\mathbf{y} \in \mathbb{R}^n$, which of the following techniques could potentially decrease the empirical risk on the training data (assuming the loss is the squared error)?

- ● Adding the feature "1" to each data point.
- ● Centering the vector $\mathbf{y}$ by subtracting the mean $\bar{y}$ from each component y_i.
- ● Adding polynomial features to each data point.
- ○ Penalizing the model weights with L_2 regularization.

(o) [3 pts] In terms of the bias-variance trade-off, which of the following is/are substantially more harmful to the test error than the training error?

- ◯ Bias
- ◯ Loss
- ● Variance
- ◯ Risk

(p) [3 pts] Consider the bias-variance trade-off in fitting least-squares surfaces to two data sets. The first is US census data, in which we want to estimate household income from the other variables. The second is synthetic data we generated by writing a program that randomly creates samples from a known normal distribution, and assigns them y-values on a known smooth surface $y = f(\mathbf{x})$ plus noise drawn from a known normal distribution. We can compute or estimate with high accuracy

- ◯ the bias component of the empirical risk for the US census data.
- ● the bias component of the empirical risk for the synthetic data.
- ◯ the variance component of the empirical risk for the US census data.
- ● the variance component of the empirical risk for the synthetic data.

(q) [3 pts] The kernel trick

- ◯ is necessary if we want to add polynomial features to a learning algorithm.
- ● can improve the speed of high-degree polynomial regression.
- ◯ can be applied to any learning algorithm.
- ◯ can improve the speed of learning algorithms when the number of samples is very large.

(r) [3 pts] In the usual formulation of soft-margin SVMs, each training sample has a slack variable $\xi_i \geq 0$ and we impose a regularization cost $C \sum_i \xi_i$. Consider an alternative formulation where we impose the additional constraints $\xi_i = \xi_j$ for all i, j. How does the minimum objective value $|\mathbf{w}|^2 + C \sum_i \xi_i$ obtained by the new method compare to the one obtained by the original soft-margin SVM?

- ◯ They are always equal.
- ◯ Original SVM minimum $\geq$ new minimum.
- ● New minimum $\geq$ original SVM minimum.
- ◯ New minimum is sometimes larger and sometimes smaller.

(s) [3 pts] In Gaussian discriminant analysis, if two classes come from Gaussian distributions that have different means, may or may not have different covariance matrices, and may or may not have different priors, which decision boundary shapes are possible?

- ● a hyperplane
- ◯ a surface that is not a quadric
- ● a nonlinear quadric surface (quadric = the isosurface of a quadratic function)
- ● the empty set (the classifier always returns the same class)

(t) [3 pts] Let the class conditionals be given by $P(X|Y = i) \sim \mathcal{N}(\mathbf{0}, \Sigma_i)$, where $i \in \{0, 1\}$ and $\Sigma_0 = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$ and $\Sigma_1 = \begin{bmatrix} b & 0 \\ 0 & a \end{bmatrix}$ with $a, b > 0$, $a \neq b$. Both conditionals have mean zero, and both classes have the prior probability $P(Y = 0) = P(Y = 1) = 0.5$. What is the shape of the decision boundary?

- ◯ a line
- ● multiple lines
- ◯ a nonlinear quadratic curve
- ◯ not defined

Q2. [15 pts] Quadratics and Gaussian Isocontours

(a) [4 pts] Write the 2×2 matrix Σ whose unit eigenvectors are $\begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$ with eigenvalue 1 and $\begin{bmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}$ with eigenvalue 4. Write out **both** the eigendecomposition of Σ and the final 2×2 matrix Σ.

$$\Sigma = \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} = \begin{bmatrix} 17/5 & -6/5 \\ -6/5 & 8/5 \end{bmatrix}.$$

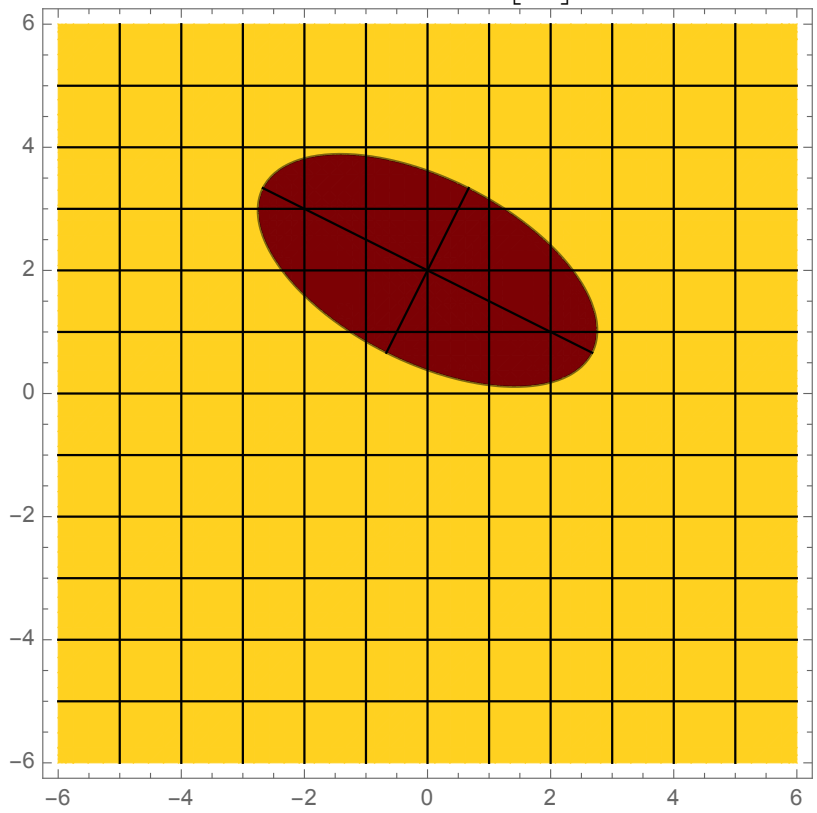
(b) [3 pts] Write the symmetric square root $\Sigma^{1/2}$ of Σ. (The eigendecomposition is optional, but it might earn you partial credit if you get $\Sigma^{1/2}$ wrong.)

$$\Sigma^{1/2} = \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} = \begin{bmatrix} 9/5 & -2/5 \\ -2/5 & 6/5 \end{bmatrix}.$$

(c) [3 pts] Consider the bivariate Gaussian distribution $X \sim \mathcal{N}(\mu, \Sigma)$. Let $P(X = \mathbf{x})$ be its probability distribution function (PDF). Write the formula for the isocontour $P(\mathbf{x}) = e^{-\sqrt{5}/2}/(4\pi)$, substitute in the value of the determinant $|\Sigma|$ from part (a) (but leave μ and Σ^{-1} as variables), and simplify the formula as much as you can.

$$\begin{aligned} \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)}{2}\right) &= \frac{e^{-\sqrt{5}/2}}{4\pi} \\ (\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu) &= \sqrt{5} \end{aligned}$$

(d) [5 pts] Draw the isocontour $P(\mathbf{x}) = e^{-\sqrt{5}/2}/(4\pi)$ where $\mu = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$ and Σ is given in part (a).



Q3. [15 pts] Linear Regression

Recall that if we model our input data as linear plus Gaussian noise in the y-values, $Y \mid \mathbf{x} \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2)$, then the maximum likelihood estimator is the $\mathbf{w}$ that minimizes the residual sum of squares $\sum_{i=1}^n (X_i^\top \mathbf{w} - y_i)^2$, where the training samples are $X_1, X_2, \ldots, X_n$ and their labels are $y_1, y_2, \ldots, y_n$.

Let's model noise with a Laplace distribution instead of a normal distribution. The probability density function (PDF) of Laplace(μ, b) is

$$P(y) = \frac{1}{2b} \exp\left(-\frac{|y-\mu|}{b}\right).$$

(a) [6 pts] Show that if we model our input data as a line plus Laplacian noise in the y-values, i.e.

$$Y \mid \mathbf{x} \sim \text{Laplace}(\mathbf{w}^\top \mathbf{x}, b),$$

then the maximum likelihood estimator is the $\mathbf{w}$ that minimizes the *sum of absolute residuals*

$$\sum_{i=1}^n |X_i^\top \mathbf{w} - y_i|.$$

We wish to maximize the log-likelihood

$$\begin{aligned} \ln \prod_{i=1}^n P(y_i|X_i) &= \sum_{i=1}^n \ln\left(\frac{1}{2b} e^{-|y_i - X_i^\top \mathbf{w}|/b}\right) \\ &= -\frac{1}{b}\sum_{i=1}^n |y_i - X_i^\top \mathbf{w}| - n\ln(2b), \end{aligned}$$

which is equivalent to minimizing $\sum_{i=1}^n |X_i^\top \mathbf{w} - y_i|$.

(b) [6 pts] Derive the batch gradient descent rule for minimizing the sum of absolute residuals. (*Hint*: You will probably need "**if**" statements or equations with conditionals because of the absolute value operators in the cost function. Don't worry about points where the gradient is undefined.)

The batch gradient descent rule with learning rate ϵ:

$$w \leftarrow w - \epsilon \nabla_{\mathbf{w}} \sum_{i=1}^n |X_i^\top \mathbf{w} - y_i| = w + \epsilon \sum_{i=1}^n \begin{cases} -X_i, & X_i^\top \mathbf{w} - y_i > 0, \\ X_i, & X_i^\top \mathbf{w} - y_i < 0. \end{cases}$$

Alternatively, it can be written as pseducode:

for $i \to 1$ **to** n
  if $X_i^\top \mathbf{w} - y_i > 0$
    $w \leftarrow w - \epsilon X_i$
  else
    $w \leftarrow w + \epsilon X_i$

Students can get partial credit by deriving (or coming close to) the stochastic gradient descent rule:

$$w \leftarrow w + \begin{cases} -\epsilon X_i, & X_i^\top \mathbf{w} - y_i > 0, \\ \epsilon X_i, & X_i^\top \mathbf{w} - y_i < 0. \end{cases}$$

(c) [3 pts] Why might we prefer to minimize the sum of absolute residuals instead of the residual sum of squares for some data sets? (*Hint*: What is one of the flaws of least-squares regression?)

The sum of absolute residuals is less sensitive to outliers than the residual sum of squares.

Q4. [10 pts] Discriminant Analysis

Let's derive the decision boundary when one class is Gaussian and the other class is exponential. Our feature space is one-dimensional ($d = 1$), so the decision boundary is a small set of points.

We have two classes, named N for normal and E for exponential. For the former class ($Y = N$), the prior probability is $\pi_N = P(Y = N) = \frac{\sqrt{2\pi}}{1+\sqrt{2\pi}}$ and the class conditional $P(X|Y = N)$ has the normal distribution $\mathcal{N}(0, \sigma^2)$. For the latter, the prior probability is $\pi_E = P(Y = E) = \frac{1}{1+\sqrt{2\pi}}$ and the class conditional has the exponential distribution

$$P(X = x|Y = E) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

Write an equation in x for the decision boundary. (Only the positive solutions of your equation will be relevant; ignore all $x < 0$.) Use the 0-1 loss function. Simplify the equation until it is quadratic in x. (You don't need to solve the quadratic equation. It should contain the constants σ and λ. Ignore the fact that 0 might or might not also be a point in the decision boundary.) **Show your work**, starting from the posterior probabilities.

Ignoring the possibility of $x = 0$, the decision boundary is the set of positive solutions to

$$\begin{aligned} P(Y = N|X = x) &= P(Y = E|X = x) \\ \frac{P(X = x|Y = N)P(Y = N)}{P(X = x)} &= \frac{P(X = x|Y = E)P(Y = E)}{P(X = x)} \\ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \frac{\sqrt{2\pi}}{1+\sqrt{2\pi}} &= \lambda e^{-\lambda x} \frac{1}{1+\sqrt{2\pi}} \\ -\ln\sigma - \frac{x^2}{2\sigma^2} &= \ln\lambda - \lambda x \\ 0 &= \frac{x^2}{2\sigma^2} - \lambda x + \ln\lambda + \ln\sigma. \end{aligned}$$

Note that the last term can be abbreviated to $\ln(\lambda\sigma)$. The last line above is not necessary for full credit; the second-last line counts as a "quadratic equation." The first line of math also is not necessary for full credit, but Bayes' Theorem must implicitly be present.