CS 189 Spring 2020

Introduction to Machine Learning

Midterm A

- Please do not open the exam before you are instructed to do so.
- The exam is closed book, closed notes except your cheat sheets.
- Please write your name at the top of each page of the Answer Sheet. (You may do this before the exam.)
- You have 80 minutes to complete the midterm exam (6:40–8:00 PM). (If you are in the DSP program and have an allowance of 150% or 200% time, that comes to 120 minutes or 160 minutes, respectively.)
- When the exam ends (8:00 PM), **stop writing**. You have 15 minutes to scan the exam and turn it into Gradescope. You must remain visible on camera while you scan your exam and turn it in (unless the scanning device is your only self-monitoring device). Most of you will use your cellphone and a third-party scanning app. If you have a physical scanner in your workspace that you can make visible from your camera, you may use that. Late exams will be penalized at a rate of 10 points per minute after 8:15 PM. (The midterm has 100 points total.) Continuing to work on the exam after 8:00 PM (or not being visible prior to submission) **may incur a score of zero**.
- Mark your answers on the Answer Sheet. If you absolutely must use overflow space for a written question, use the space for "Written Question #5" (but please try hard not to overflow). Do **not** attach any extra sheets.
- The total number of points is 100. There are 10 multiple choice questions worth 4 points each, and three written questions worth 20 points each.
- 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.
- For written questions, **please write your full answer in the space provided** and **clearly label all subparts of each written question**. Again, do not attach extra sheets.

First name	
Last name	
SID	

Q1. [40 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] Let X be an $m \times n$ matrix. Which of the following are always equal to rank(X)?

● A: rank(X^T)

○ C: $m -$ dimension(nullspace(X))

● B: rank($X^T X$)

● D: dimension(rowspace(X))

Option C is not equal because by Rank-Nullity Theorem: $n-$dim(nullspace(X)) =rank(X)
Options A and D are equal, since dim(rowspace(X)) =dim(columnspace(X^T)) =rank(X^T) =rank(X)

(b) [4 pts] Which of the following types of square matrices can have negative eigenvalues?

● A: a symmetric matrix

● C: an orthonormal matrix (M such that $M^\top M = I$)

○ B: $I - uu^T$ where u is a unit vector

● D: $\nabla^2 f(x)$ where $f(x)$ is a Gaussian PDF

Top left: A symmetric matrix can have negative eigenvalues, they just have to be real. Bottom left: u is a unit vector that can be expressed as a linear combination of the standard vectors $\sum_{i=1}^{n} c_i * e_i$ where $c_i < 1$. $(\sum_{i=1}^{n} c_i * e_i)(\sum_{i=1}^{n} c_i * e_i)^T = (\sum_{i=1}^{n} c_i * e_i)(\sum_{i=1}^{n} c_i * e_i)^T$ Top right: An orthogonal matrix has eigenvalues of 1 and -1. Bottom right: The gaussian of a PDF is a concave function, thus the hessian must have negative eigenvalues.

(c) [4 pts] Choose the correct statement(s) about Support Vector Machines (SVMs).

● A: if a finite set of training points from two classes is linearly separable, a **hard-margin** SVM will always find a decision boundary correctly classifying every training point

○ B: if a finite set of training points from two classes is linearly separable, a **soft-margin** SVM will always find a decision boundary correctly classifying every training point

● C: every trained two-class **hard-margin** SVM model has at least one point of each class at a distance of exactly $1/\|w\|$ (the margin width) from the decision boundary

○ D: every trained two-class **soft-margin** SVM model has at least one point of each class at a distance of exactly $1/\|w\|$ (the margin width) from the decision boundary

Option A is correct: fundamental material about SVMs from lectures.

(d) [4 pts] Suppose we perform least-squares linear regression, but we don't assume that all weight vectors are equally reasonable; instead, we use the maximum *a posteriori* method to impose a normally-distributed prior probability on the weights. Then we are doing

● A: L_2 regularization

○ C: logistic regression

○ B: Lasso regression

● D: ridge regression

As shown in Lecture 13, the Bayesian justification for ridge regression is derived by applying MAP to the posterior probability with a Gaussian prior on the weights.

(e) [4 pts] Which of the following statements regarding ROC curves are true?

● A: the ROC curve is monotonically increasing

○ B: for a logistic regression classifier, the ROC curve's horizontal axis is the posterior probability used as a threshold for the decision rule

○ C: the ROC curve is concave

● D: if the ROC curve passes through $(0, 1)$, the classifier is always correct (on the test data used to make the ROC curve)

The axes of an ROC curve do not correspond to the "knob" we're turning when we plot the curve.

Always predicting positive will give us 100

Does not have to be concave, just needs to be increasing.

Since it's increasing, the curve is a horizontal line at y=1. So, we have no false positives nor false negatives.

(f) [4 pts] One way to understand regularization is to ask which vectors minimize the regularization term. Consider the set of unit vectors in the plane: $\{x \in \mathbb{R}^2 : \|x\|_2^2 = 1\}$. Which of the following regularization terms are minimized solely by the four unit vectors $\{(0, 1), (1, 0), (-1, 0), (0, -1)\}$ and no other unit vector?

● A: $f(x) = \|x\|_0 =$ the # of nonzero entries of x

● B: $f(x) = \|x\|_1$

○ C: $f(x) = \|x\|_2^2$

○ D: $f(x) = \|x\|_\infty = \max\{|x_1|, |x_2|\}$

The first option is almost true by definition: these are the sparsest unit vectors. The second option follows Cauchy–Schwartz. Intuitively, however, we know also that the ℓ_1-norm promotes sparsity, so we should expect this to be true. Finally, notice that $\|x\|_2^2$ always equals 1 and that $\max(x_1, x_2)$ is minimized when $x_1 = x_2$, so both these options are incorrect.

(g) [4 pts] Suppose we train a soft-margin SVM classifier on data with d-dimensional features and binary labels. Below we have written four pairs of the form "modification → effect." For which ones would a model trained on the modified data **always** have the corresponding effect relative to the original model?

● A: augment the data with polynomial features → optimal value of the objective function (on the training points) decreases or stays the same

○ B: multiply each data point by a fixed invertible $d \times d$ matrix A; i.e., $X_i \leftarrow AX_i$ → all training points are classified the same as before

● C: multiply each data point by a fixed orthonormal $d \times d$ matrix U and add a fixed vector $z \in \mathbb{R}^d$; i.e., $X_i \leftarrow UX_i + z$ → all training points are classified the same as before.

○ D: normalize each feature so that its mean is 0 and variance is 1 → all training points are classified the same as before

1) The original optimal values are possible in the modified setup, since we can use the original w on the original features and set $w_i = 0$ for all w_i corresponding to the newly added features. Thus, the new optimum must be at least as good as the old one.

2) An invertible matrix could still send the data points to a transformed space in which the relative scale of different components is very different from the original, for instance, affecting the optimization of the new w.

3) Multiplying by an orthogonal matrix is an isometry, meaning that it preserves the norm of all vectors, so the relative scale stays the same. Adding z to every data point just shifts them. Concretely, take $w' = Uw$ and $\alpha' = \alpha - z^T w'$ to recover the original optimum, since $\|w'\| = \|Uw\| = \|w\|$.

4) Normalizing each feature changes the relative scale of different components; see 2).

(h) [4 pts] A real-valued $n \times n$ matrix P is called a projection matrix if $P^2 = P$. Select all the true statements about eigenvalues of P.

● A: P can have an eigenvalue of 0

● B: P can have an eigenvalue of 1

○ C: P can have an eigenvalue of -1

○ D: P can have an eigenvalue that isn't 0, 1, or -1

(i) [4 pts] Let X be a real-valued $n \times d$ matrix. Let Ω be a diagonal, real-valued $n \times n$ matrix whose diagonal entries are all positive. Which of the following are true of the matrix product $M = X^T \Omega X$?

○ A: M could have negative eigenvalues

● B: M could have eigenvalues equal to zero

● C: M could have positive eigenvalues

○ D: the eigenvalues of M are the values on the diagonal of Ω

The matrix M is PSD:

$$\forall v \in R^d, v^T X^T \Omega X v = \sum_{i=1}^{n} \omega_{ii} (x_i^T v)^2 \geq 0$$

It therefore cannot have negative eigenvalues. It can have positive eigenvalues, or eigenvalues equal to 0 (for example, when all entries of X are 0). Note that the eigenvalues are not necessarily the values along the diagonal of Ω. That would be the case when the matrix X is orthogonal, but the matrix X isn't even necessarily square.

(j) [4 pts] Which of the following regression methods always have just one unique optimum, regardless of the data?

○ A: least Squares Regression

● B: ridge Regression

○ C: Lasso Regression

○ D: logistic Regression

Only ridge regression is guaranteed to converge to have a unique optimum as the regularization term makes the objective function strictly convex. Other options may also have convex objectives, but there may be multiple points in weight space at which the same loss value is obtained.

Q2. [20 pts] Gradient Descent

Let's use gradient descent to solve the optimization problem of finding the value of $x \in \mathbb{R}^2$ that minimizes the objective function

$$J(x) = \frac{1}{2} x^T A x, \qquad A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}.$$

(a) [7 pts] Let $x^{(t)}$ represent the value of x after t iterations of gradient descent from some arbitrary starting point $x^{(0)}$. Write the standard gradient descent update equation in the form $x^{(t+1)} \leftarrow f(x^{(t)})$ (you tell us what the function f is) with a step size of $\epsilon = \frac{1}{4}$. Then manipulate it into the form $x^{(t+1)} = Bx^{(t)}$ where B is a matrix (you tell us what B is). Show your work.

$$\begin{aligned} x^{(t+1)} &= x^{(t)} - \epsilon \nabla J(x^{(t)}); \\ \nabla J(x^{(t)}) &= Ax^{(t)}; \\ x^{(t+1)} &= (I - \epsilon A)x^{(t)} \\ &= \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \frac{1}{4} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \right) x^{(t)} \\ &= \begin{bmatrix} \frac{3}{4} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} x^{(t)}. \\ B &= \begin{bmatrix} \frac{3}{4} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}. \end{aligned}$$

(b) [4 pts] The minimum of $J(x)$ is at $x^* = 0$, so we hope that our algorithm will converge: that is, $\lim_{t\to\infty} x^{(t)} = 0$. Show that for any starting point $x^{(0)}$, your gradient descent algorithm converges to x^*.

Since B is diagonal, $x_0^{(t)} = \left(\frac{3}{4}\right)^t x_0^{(0)}$ and $x_1^{(t)} = \left(\frac{1}{2}\right)^t x_1^{(0)}$. Therefore, $\lim_{t\to\infty} x_0^{(t)} = 0$ and $\lim_{t\to\infty} x_1^{(t)} = 0$.

(c) [3 pts] Suppose we change the step size to $\epsilon = 1$. What is B? How does gradient descent behave with this step size?

Now $B = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}$. So the component x_0 is reduced to zero immediately, but the component x_1 repeatedly flips its sign and never reaches the critical point.

(d) [3 pts] Suppose we replace A with another diagonal matrix with positive diagonal entries. What is the optimal step size for fastest convergence, expressed in terms of the diagonal entries A_{11} and A_{22}?

We achieve the fastest convergence by minimizing $\max\{|1-\epsilon A_{11}|, |1-\epsilon A_{22}|\}$, which happens when $1-\epsilon A_{11} = -(1-\epsilon A_{22})$, which implies that

$$\epsilon = \frac{2}{A_{11} + A_{22}}.$$

(e) [3 pts] Your argument in part (b) can be adapted to prove convergence for *any* diagonal A with positive diagonal entries, so long as we choose a suitably small step size ϵ as derived in part (d). Suppose we replace A with another matrix that is symmetric and positive definite but *not* diagonal. Suppose we choose a suitably small step size ϵ. Without writing any equations, give a mathematical explanation (in English) why your argument in part (b) applies here and gives us confidence that gradient descent will converge to the minimum, even though A is not diagonal. *Hint: One approach is to change the coordinate system.*

If we find the eigenvectors of A and change the coordinate system to one whose primary axes are the eigenvector directions, then once again we can express our minimization problem with a diagonal A in that new coordinate system. Then the same arguments apply.

Q3. [20 pts] Gaussians and Linear Discriminant Analysis

Suppose that the training and test points for a class come from an anisotropic multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ is symmetric and positive definite. Recall that (for $x \in \mathbb{R}^d$) its probability density function (PDF) is

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

(a) [7 pts] In lecture, I claimed that if Σ is diagonal, you can write this PDF as a product of d univariate Gaussian PDFs, one for each feature. What if Σ is not diagonal? Show that if you substitute Σ's **eigendecomposition** for Σ, you can write the PDF above as a **product of d univariate Gaussian PDFs**, one aligned with each *eigenvector* of Σ. For simplicity, please set $\mu = 0$ (prove it just for the mean-zero case).

Hints: Use the shorthand $\tau = 1/\left((\sqrt{2\pi})^d \sqrt{|\Sigma|}\right)$. Write the eigendecomposition as a summation with one term per eigenvalue/vector. The determinant $|\Sigma|$ is the product of Σ's eigenvalues (all d of them).

Let Σ's eigendecomposition be $\Sigma = V\Lambda V^\top$, where Λ is diagonal. Then $\Sigma^{-1} = V\Lambda^{-1}V^\top = \sum_{i=1}^d \frac{1}{\Lambda_{ii}} v_i v_i^\top$ where v_i is the unit eigenvector in column i of V, and

$$f(x) = \tau \exp\left(-\frac{1}{2}x^\top \left(\sum_{i=1}^d \frac{1}{\Lambda_{ii}} v_i v_i^\top\right) x\right) = \tau \exp\left(\sum_{i=1}^d -\frac{1}{2\Lambda_{ii}} x^\top\, v_i\, v_i^\top\, x\right) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi\Lambda_{ii}}} \exp\left(-\frac{(x^\top\, v_i)^2}{2\Lambda_{ii}}\right).$$

(b) [2 pts] When you express the multivariate PDF as a product of univariate PDFs, what is the **variance** of the univariate distribution along the direction of the ith eigenvector v_i?

Λ_{ii}.

(c) [7 pts] Consider performing **linear discriminant analysis** (LDA) with two classes. Class C has the class-conditional distribution $\mathcal{N}(\mu_\text{C}, \Sigma)$, and class D has the class-conditional distribution $\mathcal{N}(\mu_\text{D}, \Sigma)$. Note that they both have the same covariance matrix but different means. Recall that we define a quadratic function

$$Q_\text{C}(x) = \ln\left((\sqrt{2\pi})^d\, f_\text{C}(x)\, \pi_\text{C}\right),$$

where $f_\text{C}(x)$ is the PDF for class C and π_C is the prior probability for class C. For class D, we define $Q_\text{D}(x)$ likewise. For simplicity, assume $\pi_\text{C} = \pi_\text{D} = \frac{1}{2}$.

Write down the LDA decision boundary as an equation in terms of $Q_\text{C}(x)$ and $Q_\text{D}(x)$. Then substitute the definition above and show that the decision boundary has the form $\{x : w \cdot x + \alpha = 0\}$ for some $w \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$. **What is the value of w?**

The LDA decision boundary is $Q_\text{C}(x) - Q_\text{D}(x) = 0$. (More pedantically, it's $\{x : Q_\text{C}(x) - Q_\text{D}(x) = 0\}$.)

$$\begin{aligned}
Q_\text{C}(x) &= -\frac{1}{2}(x-\mu_\text{C})^\top \Sigma^{-1} (x-\mu_\text{C}) - \frac{1}{2}\ln|\Sigma| + \ln \pi_\text{C}. \\
Q_\text{C}(x) - Q_\text{D}(x) &= -\frac{1}{2}(x-\mu_\text{C})^\top \Sigma^{-1} (x-\mu_\text{C}) + \frac{1}{2}(x-\mu_\text{D})^\top \Sigma^{-1} (x-\mu_\text{D}) \\
&= -\frac{1}{2}x\,\Sigma^{-1}\,x + \mu_\text{C}^\top \Sigma^{-1}\,x - \frac{1}{2}\mu_\text{C}^\top \Sigma^{-1}\,\mu_\text{C} + \frac{1}{2}x\,\Sigma^{-1}\,x - \mu_\text{D}^\top \Sigma^{-1}\,x + \frac{1}{2}\mu_\text{D}^\top \Sigma^{-1}\,\mu_\text{D} \\
&= (\mu_\text{C} - \mu_\text{D})^\top \Sigma^{-1}\,x - \frac{1}{2}(\mu_\text{C}^\top \Sigma^{-1}\,\mu_\text{C} - \mu_\text{D}^\top \Sigma^{-1}\,\mu_\text{D}). \\
w &= \Sigma^{-1}(\mu_\text{C} - \mu_\text{D}).
\end{aligned}$$

(d) [2 pts] What is the relationship between w and the decision boundary?

w is normal (orthogonal) to the decision boundary.

(e) [2 pts] **Is w always an eigenvector of Σ?** (That is, is it always true that $w = \omega v_i$ for some scalar ω and unit eigenvector v_i of Σ?) **Why or why not?**

No. The means μ_C and μ_D are arbitrary, so $w = \Sigma^{-1}(\mu_\text{C} - \mu_\text{D})$ does not have to be an eigenvector of Σ.

Q4. [20 pts] Double Regression

Let's work out a two-way least-squares linear regression method. The input is n observations, recorded in two vectors $s, t \in \mathbb{R}^n$. The ith observation is the ordered pair (s_i, t_i). We're going to view these data in two ways: (1) s_i is a sample point in one dimension with label t_i; or (2) t_i is a sample point in one dimension with label s_i. We will use least-squares linear regression to (1) take a test point $s_T \in \mathbb{R}$ and predict its label t_T, with a hypothesis $\hat{t}(s_T) = \beta s_T$, and (2) take a test point $t_T \in \mathbb{R}$ and predict its label s_T, with a hypothesis $\hat{s}(t_T) = \gamma t_T$.

We do not use bias terms, so both regression functions will pass through the origin. Our optimization problems are

$$\boxed{\text{Find } \beta \text{ that minimizes } \sum_{i=1}^{n} (\beta s_i - t_i)^2} \quad \boxed{\text{Find } \gamma \text{ that minimizes } \sum_{i=1}^{n} (\gamma t_i - s_i)^2}$$

A natural question, which we will explore now, is whether both regressions find the same relationship between s_T and t_T.

(a) [7 pts] Derive a **closed-form expression for the optimal regression coefficient** β. Write your final answer in terms of **vector operations**, not summations. **Show all your work.**

The squared error loss is convex, hence we find β by taking the derivative and setting it equal to zero.

$$0 = \frac{\partial}{\partial \beta} \sum_{i=1}^{n} (\beta s_i - t_i)^2 = 2 \sum_{i=1}^{n} s_i(\beta s_i - t_i) = 2\beta \|s\|^2 - 2s \cdot t \implies \beta = \frac{s \cdot t}{\|s\|^2}.$$

(b) [2 pts] What is a closed-form expression for the optimal regression coefficient γ? (This follows from symmetry; you don't need to repeat the derivation, unless you want to.)

$$\gamma = \frac{s \cdot t}{\|t\|^2}.$$

(c) [4 pts] The hypotheses $t_T = \beta s_T$ and $s_T = \gamma t_T$ represent the same equation if and only if $\beta\gamma = 1$. **Prove that** $\beta\gamma \leq 1$ and **determine under what condition equality holds**. *Hint: remember the Cauchy–Schwarz inequality.*

$\beta\gamma = (s \cdot t)^2/(\|s\|^2\|t\|^2)$. By the Cauchy–Schwarz inequality, $s \cdot t \leq \|s\|\,\|t\|$, so it follows immediately that $\beta\gamma \leq 1$. This equality is tight only when s and t are parallel vectors; that is, $s = \rho t$ for some $\rho \in \mathbb{R}$.

(d) [5 pts] We might want to compute these coefficients with ℓ_1-regularization. For some regularization parameter $\lambda > 0$, consider the optimization problem

$$\boxed{\text{Find } \beta \text{ that minimizes } \lambda|\beta| + \sum_{i=1}^{n} (\beta s_i - t_i)^2}$$

In Homework 4, we analyzed this optimization problem and concluded that that there is at most one point where the derivative is zero. If such a point exists, it is the minimum; otherwise, the minimum is at the discontinuity $\beta = 0$. For simplicity, let's consider only the case where the solution happens to be positive ($\beta > 0$).

Derive a **closed-form expression for the optimal regression coefficient** β **in the case** $\beta > 0$. Write your final answer in terms of vector operations, not summations. Show all your work.

For $\beta \neq 0$,

$$\frac{\partial}{\partial \beta}\left(\lambda|\beta| + \sum_{i=1}^{n} (\beta s_i - t_i)^2\right) = \lambda \operatorname{sign}(\beta) + 2 \sum_{i=1}^{n} s_i(\beta s_i - t_i) = \lambda \operatorname{sign}(\beta) + 2\beta\|s\|^2 - 2s \cdot t.$$

If there is a critical point $\beta > 0$, it satisfies

$$\lambda + 2\beta\|s\|^2 - 2s \cdot t = 0 \implies \beta = \frac{2s \cdot t - \lambda}{2\|s\|^2}.$$

(e) [2 pts] **What necessary and sufficient condition (inequality) should** s, t, **and** λ **satisfy** to assure us that the optimal β is indeed positive?

The expression for β above is positive if and only if $2s \cdot t > \lambda$.

Addendum (irrelevant to grading). Symmetrically, if there is a critical point $\beta < 0$, it satisfies

$$\beta = \frac{2s \cdot t + \lambda}{2\|s\|^2},$$

but that expression is negative if and only if $2s \cdot t < -\lambda$.

When $2s \cdot t \in [-\lambda, \lambda]$ (neither condition is satisfied), the minimum must lie at the only remaining critical point, zero. So in all cases, we can write

$$\beta = \mathrm{sign}(s \cdot t) \max\left\{\frac{2|s \cdot t| - \lambda}{2\|s\|^2}, 0\right\}.$$