CS 189 Summer 2019 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 two-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 3 hours.
- Please write your initials at the top right of each page after this one (e.g., write "MK" if you are Marc Khoury). Finish this by the end of your 3 hours.
- Mark your answers on the exam itself in the space provided. Do **not** attach any extra sheets.
- The total number of points is 150. There are 26 multiple choice questions worth 3 points each, and 5 written questions worth a total of 72 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] Let $X \sim Bernoulli(\frac{1}{1+\exp\theta})$ for some $\theta \in \mathbb{R}$. What is the MLE estimator of θ?

○ X

○ 1

○ 0

● Does not exist.

$l(\theta; 1) = \frac{1}{1+\exp\theta}$ which has no maximizer in $\mathbb{R}$. $l(\theta; 0) = 1 - \frac{1}{1+\exp\theta}$ which has no maximizer in $\mathbb{R}$.

(b) [3 pts] Let $Y \sim \mathcal{N}(X\theta, I_n)$ for some unknown $\theta \in \mathbb{R}^d$ and some known $X \in \mathbb{R}^{n\times d}$ that has full column rank and $d < n$. What is the MLE estimator of θ?

● $(X^\top X)^{-1}X^\top Y$

○ $X^\top(XX^\top)^{-1}Y$

○ $Y + Z \quad \forall Z \in Null(X)$

○ Does not exist.

Maximizing the likelihood function is equivalent to minimizing $\|Y - X\theta\|_2^2$, which we can do with linear least squares. Interestingly, this means that the MLE estimator of θ when $y = X\theta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, I_n)$ is the standard least sqaures solution.

(c) [3 pts] Let $f(x) = -\sum_{i=1}^n x_i \log x_i$. For some x such that $\sum_{i=1}^n x_i = 1$ and $x_i > 0$, the Hessian of f is:

○ positive definite

● negative definite

○ positive semidefinite

● negative semidefinite

○ indefinite (neither positive semidefinite nor negative semidefinite)

● invertible

○ nonexistent

○ None of the above.

$$\nabla_x f(x) = -\vec{1} - \log(x)$$
$$\nabla_x^2 f(x) = diag(-\frac{1}{x_1}, \ldots, -\frac{1}{x_n})$$

(d) [3 pts] Which of the following statements about optimization algorithms are correct?

○ Newton's method always requires fewer iterations than gradient descent.

○ Stochastic gradient descent always requires fewer iterations than gradient descent.

● Stochastic gradient descent, even with small step size, sometimes increases the loss in some iteration for convex problems.

○ Gradient descent, regardless the step size, decreases the loss in every iteration for convex problems.

Arguments like one reasonable optimization algorithm dominates another for all problems are in general wrong. Gradient descent can work for some loss that Newton does not even converge.

SGD due to stochasticity does not necessarily decrease the loss in each iteration. One can construct case where there could be a data point which is sort of contradicting with all others, so optimizing this particular data point may increase the overall loss.

Gradient descent with extremely large step size may not be able to reach the small neighborhood of the minimum, may just jump around outside the neighborhood.

(e) [3 pts] Assume we run the hard-margin SVM algorithm on 100 d-dimensional points from 2 different classes. The algorithm outputs a solution. After which transformation to the training data would the algorithm still output a solution?

● Centering the data points

○ Transforming each data point from x to Ax for some matrix $A \in \mathbb{R}^{dxd}$

● Dividing all entries of each data point by some negative constant c

● Adding an additional feature

If all entries an A are 0, the data points all land on the origin. If a set of data points are linearly separable in d dimensions, it is linearly separable in $d + 1$ dimensions regardless of what the $d + 1$ feature is.

(f) [3 pts] Which of the following holds true when running an SVM algorithm?

● Increasing or decreasing α value only allows the decision boundary to translate.

● Given n-dimensional points, the SVM algorithm finds a hyperplane passing through the origin in the $(n + 1)$-dimensional space that separates the points by their class.

○ Decision boundary rotates if we change the constraint to $w^T x + \alpha \geq 3$.

● The set of weights that fulfill the constraints of the SVM algorithm is convex.

(g) [3 pts] Consider the set $\{x \in \mathbb{R}^d : (x - \mu)^\top \Sigma (x - \mu) = 1\}$ given some vector $\mu \in \mathbb{R}^d$ and matrix $\Sigma \in \mathbb{R}^{dxd}$. Which of the following are true?

● If Σ is the identity matrix scaled by some constant c, then the set is isotropic.

○ Increasing the eigenvalues of Σ increases the radii of the ellipsoid.

● Increasing the eigenvalues of Σ decreases the radii of the ellipsoid.

● A singular Σ produces an ellipsoid with an infinite radius.

(h) [3 pts] Consider the linear regression problem with full rank design matrix, which of the following regularization in general encourage more sparsity than non-regularized objective:

● L_0 regularization (number of the non-zero coordinates)

● L_1 regularization

○ L_2 (Tikhonov) regularization

○ L_3 regularization

○ L_4 regularization

○ L_∞ regularization (the maximum absolute value across all coordinates)

Think about the corresponding equivalent constrained optimization problem.

(i) [3 pts] Which of the following statements are correct?

○ In ridge regression, the regularization parameter λ is usually set as 0.1.

● SVM in general does not enforce sparsity over the parameters w and α.

○ In binary linear classification, the support vectors of SVM might contain samples from only one class even if training data has both classes.

○ In binary linear classification, suppose $\mathbf{1}\{w^\top x + \alpha \geq 0\}$ is one maximum margin linear classifier, then the margin only depends on w but not α.

Basic definitions.

(j) [3 pts] In binary classification (+1 and −1), suppose our data is linearly separable and the data matrix has full column rank ($n > d$). Which of the following formulation can guarantee to find a linear classifier that achieves 0 training error? Note that in the regression options, the prediction rule would still be $\mathbf{1}\{w^\top x + \alpha \geq 0\}$.

● Logistic regression

○ Linear regression with square loss

● SVM

● Perceptron

○ Lasso

○ None of the above

For logistic regression, note that for linearly separable data, the optimal w has infinite magnitude, which implies no tolerance of non-zero training error.

As for linear regression, consider the following case: infinite number of training points at both $(-1,-1)$ and $(1,1)$, one training point at $(1,-1)$. The $(1,-1)$ point will be misclassified.

(k) [3 pts] Analogous to positive semi-definiteness, an $n \times n$ real symmetric matrix B is called *negative semi-definite* if $x^\top Bx \leq 0$ for all vectors $x \in \mathbb{R}^n$. Which of the following conditions guarantee B is negative semi-definite?

○ B has all negative entries

○ $B = A^{-1}$, where A is positive semi-definite

● The largest eigenvalue of B is ≤ 0

● $B = -A^T A$ for some matrix A

Notice that $x^\top Bx \leq 0$ if and only if $-x^\top Bx \geq 0$. Therefore, a necessary and sufficient condition for B to be negative semi-definite is for $-B$ to be positive definite. The second and fourth options follow immediately from this observation. For the first option consider the following counterexample,

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix}^\top \begin{bmatrix} -1 & -2 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 2.$$

Also notice that if A is positive semi-definite, then so is A^{-1}, since its eigenvalues are simply the reciprocals of those of A.

(l) [3 pts] Consider two classes whose class conditionals are the scalar normal distributions $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$ respectively, where $\mu_1 < \mu_2$. Given some non-zero priors π_1 and π_2, recall the Bayes' optimal decision boundary will be a single point, x^*. Which of the following changes, holding everything else constant, would cause x^* to increase?

○ Decreasing μ_1

● Increasing μ_2

○ Increasing σ

● Increasing π_1 while decreasing π_2

Although this problem can be solved intuitively, in one dimension it is also easy to directly write out the decision boundary as the root of the following increasing linear equation.

$$Q_2'(x) - Q_1'(x) = \frac{1}{\sigma^2}(\mu_2 - \mu_1)\,(x - 0.5(\mu_2 + \mu_1)) + \ln(\pi_2) - \ln(\pi_1)$$

Immediately, notice that changing σ will not affect x^*. Meanwhile, increasing μ_2 or increasing π_1 makes $Q_2'(x^*) - Q_1'(x^*) < 0$, which shifts the root to the right. On the other hand, decreasing μ_1 will push x^* to the left.

(m) [3 pts] Let Σ be a positive definite matrix with eigenvalues $\lambda_1, \ldots, \lambda_d$. Consider the quadratic function $g(x) = x^\top (c\Sigma^{-2})x$, for some constant $c > 0$. What are the lengths of the radii of the ellipsoid at which $g(x) = 1$?

● $c^{-1/2} \cdot \lambda_i$

○ $c \cdot \lambda_i^{-1}$

○ $c^{1/2} \cdot \lambda_i$

○ $c^{1/2} \cdot \lambda_i^{-1/2}$

The axes must lie along the eigenvectors of Σ. Let v_i be one such unit eigenvalue corresponding to λ_i, and consider $g(kv_i)$. Setting this equal to 1, we get the radius of the ellipsoid in the direction of v_i in terms of c and λ_i.

$$\begin{aligned} g(kv_i) &= (kv_i)^\top (c\Sigma^{-2})(kv_i) \\ &= ck^2 v_i^\top \Sigma^{-2} v_i \\ &= ck^2 \frac{1}{\lambda_i^2} v_i^\top v_i \\ &= \frac{ck^2}{\lambda_i^2} \end{aligned}$$

Setting this equal to 1, we find that $k = c^{-1/2} \cdot \lambda_i$, which gives the length of the radius.

(n) [3 pts] Let $X \sim \mathcal{N}(\mu, \Sigma)$ be a multivariate normal random variable. Which the the following statements of about linear functions of X are always true, where A is some square matrix and b a vector?

○ $\mathrm{Var}(AX) = A^2\Sigma$

○ AX is isotropic if $A = \Sigma^{-1}$

● $AX + b$ is also multivariate normal

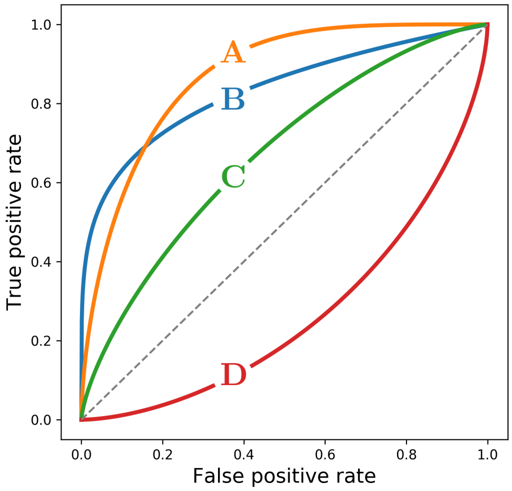
● $\mathbb{E}[AX + b] = A\mu + b$

As mentioned in the class, $AX + b$ is multivariate normal. From linearity of expectation, we can directly compute the mean and variance.

$$\begin{aligned} \mathbb{E}[AX + b] &= A\mathbb{E}[X] + b \\ &= A\mu + b \\ \mathrm{var}(AX) &= E[(AX)(AX)^\top] \\ &= E[AXX^\top A^\top] \\ &= AE[XX^\top]A^\top \\ &= A\Sigma A^\top \end{aligned}$$

Letting $A = \Sigma^{-1}$ in the formula above tell us the covariance becomes Σ^{-1}, which does not necessarily have to be diagonal.

(o) [3 pts] You have trained four binary classifiers A, B, C, and D, observing the following ROC curves when evaluating them:



We say that a classifier G *strictly dominates* a classifier H if G's true positive rate is always greater than H's true positive rate for all possible false positive rates in $(0, 1)$.

Mark *all* of the below relations between (A, B, C, and D) which are true under this definition.

● C strictly dominates D ○ A strictly dominates B

○ D strictly dominates C ● B strictly dominates C

○ B strictly dominates A ○ D strictly dominates A

(p) [3 pts] We are doing binary classification on classes $\{1, 2\}$. We have a single dataset of size N of which a fraction α of the elements are in class 1. To construct a test set, we randomly choose a fraction β of the dataset to put in the test set, keeping the remaining elements in a training set.

We would like to avoid the situation where in the training or test sets, either class appears less than $0.1N$ times. In which of the following situations does this occur, in expectation?

○ $\alpha = 50\%, \beta = 50\%$ ● $\alpha = 70\%, \beta = 20\%$

● $\alpha = 20\%, \beta = 70\%$ ○ $\alpha = 30\%, \beta = 60\%$

○ $\alpha = 40\%, \beta = 60\%$ ● $\alpha = 60\%, \beta = 80\%$

The expected proportions we are looking at in each case are $\alpha\beta$, $\alpha(1-\beta)$, $(1-\alpha)\beta$, $(1-\alpha)(1-\beta)$. The smallest of these is $\min(\alpha, 1-\alpha) \cdot \min(\beta, 1-\beta)$, so if $\min(\alpha, 1-\alpha) \cdot \min(\beta, 1-\beta) > 0.1$ then the underrepresentation will not occur.

(q) [3 pts] Assume that for a k-class problem, all classes have the same prior probability, i.e. $\pi = [\frac{1}{k}, \ldots, \frac{1}{k}]$. You build two different models:

- (Model A) You train QDA once for all k classes, and to classify a data point you return the class with the highest posterior probability.
- (Model B) You train QDA pairwise $\binom{k}{2}$ times, restricting the training data each time to only the data points from two of the k classes. To classify a test point, you return the class that has the higher posterior probability most often from the $\binom{k}{2}$ independent models.

Mark all of the following which are true in general, in the comparison of bias and variance between models A and B:

○ B has higher bias than A ○ B has higher variance than A

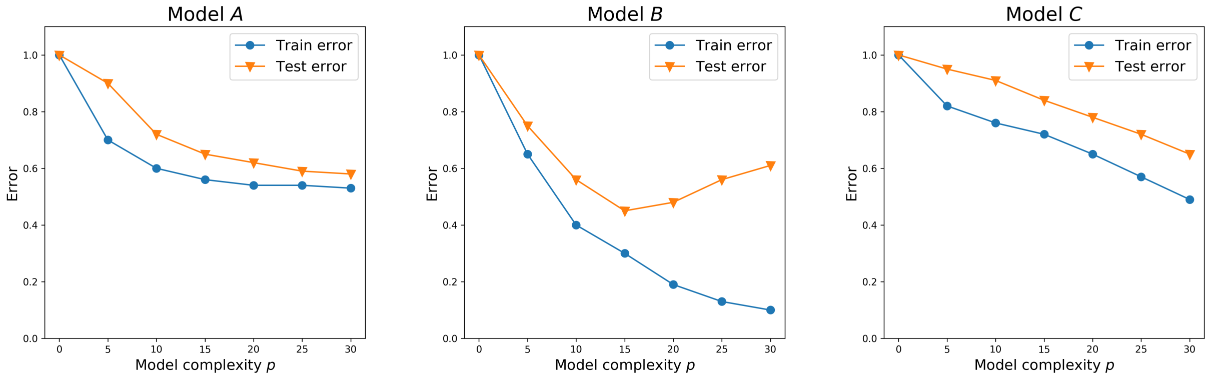
● B has the same bias as A ● B has the same variance as A

○ B has lower bias as A ○ B has lower variance as A

The covariance matrices and means, and therefore posterior probabilities will be exactly the same between Model A and the $\binom{k}{2}$ models contained in the ensemble of B. Therefore, the class that has the highest posterior probability in A will win the most elections for B, every time.

This means that the predictions are exactly the same for the two models, i.e. they *are* the same model. So they have the same bias and variance.

(r) [3 pts] You observe the following train and test error as a function of model complexity p for three different models:



Mark the values of p and models where the test and train error indicate overfitting.

○ model A at $p = 10$ ○ model A at $p = 20$ ○ model A at $p = 30$

○ model B at $p = 10$ ● model B at $p = 20$ ● model B at $p = 30$

○ model C at $p = 10$ ○ model C at $p = 20$ ○ model C at $p = 30$

A divergence in the test and train error indicates overfitting. This only happens for model B at $p = 20$ and $p = 30$

(s) [3 pts] Which models, if any, appear to be underfit for all settings of p?

○ A ○ B ● C

The test and train error are still decreasing for model C, so it is possible it is underfit

(t) [3 pts] Consider the minimum possible bias for each model over all settings of p for $0 \leq p \leq 30$. Which of the following are true in comparing the minimum bias between the three models?

● Model A has higher minimum bias than B ○ Model B has higher minimum bias than C

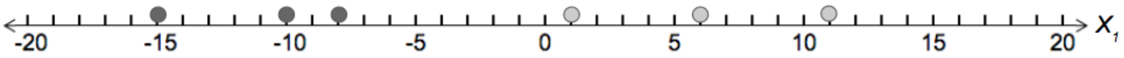
○ Model A has the same minimum bias as B ○ Model B has the same minimum bias as C

○ Model A has lower minimum bias than C ● Model B has lower minimum bias than C

The train error is a good indicator of the bias. *B* has the lowest possible among the three models, and *A* and *C* have around the same.

Q2. [10 pts] Comparing Classification Algorithms

Find the decision boundary given by the following algorithms. Provide a range of values if the algorithm allows for multiple feasible decision boundaries. If there exists no feasible decision boundary, state ”None.”

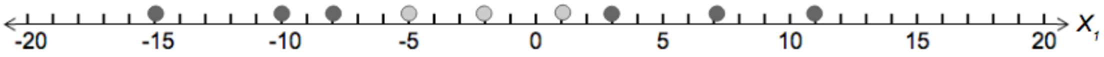


(a) [1 pt] Perceptron: X_1 = ________________

(b) [2 pts] Hard-Margin SVM: X_1 = ________________

(c) [2 pts] Linear Discriminant Analysis: X_1 = ________________

Perceptron: [-8, 1]
Hard-Margin SVM: -3.5
Linear Discriminant Analysis: -2.5



(d) [1 pt] Perceptron: X_1 = ________________

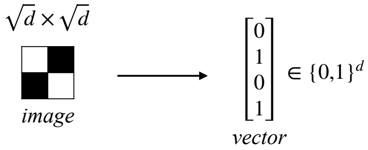
(e) [2 pts] Hard-Margin SVM: X_1 = ________________

(f) [2 pts] Linear Discriminant Analysis: X_1 = ________________

Perceptron: None
Hard-Margin SVM: None
Linear Discriminant Analysis: None

Q3. [15 pts] Binary Image Classification

A *binary image* is a digital image where each pixel has only possibles values: zero (white) or one (black). A binary image, which consists of a grid of pixels, can therefore naturally be represented as a vector with entries in $\{0, 1\}$.


$\sqrt{d} \times \sqrt{d}$
image
$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \in \{0,1\}^d$
vector

In this problem, we consider a classification scheme based on a simple generative model. Let X be a random binary image, represented as a d-dimensional binary vector, drawn from one of two classes: P or Q. Assume every pixel X_i is an independent Bernoulli random variable with parameter p_i and q_i when drawn from classes P and Q respectively.

$$X_i \mid Y = P \sim \text{Bernoulli}(p_i) \qquad \text{independently for all } 1 \leq i \leq d$$
$$X_i \mid Y = Q \sim \text{Bernoulli}(q_i) \qquad \text{independently for all } 1 \leq i \leq d$$

(a) [1 pt] Of course, when working with real data, the true parameters p_i and q_i will be unknown and therefore must be estimated from the data. Given the following 5 2-dimensional training points from class P, find the maximum likelihood estimates of p_1 and p_2.

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$\hat{p}_1^{\text{MLE}} =$ $\qquad\qquad$ $\hat{p}_2^{\text{MLE}} =$

$$\hat{p}_1^{\text{MLE}} = \frac{0+0+0+1+0}{5} = 0.2 \qquad \hat{p}_2^{\text{MLE}} = \frac{0+1+1+1+1}{5} = 0.8$$

Important. The other parameters and priors could similarly be estimated. For the remainder of the problem, however, we focus on the ideal case, where the true values of p_i and q_i, along with priors π_p and π_q, are known.

(b) [1 pt] Fill in the blanks in the statement below.

To minimize risk with the (symmetric) 0-1 loss function, we should pick the class with the ______________ ______________ probability, which gives the Bayes' optimal classifier.

biggest, posterior

(c) [4 pts] Given an image $x \in \{0,1\}^d$, compute the probabilities $\Pr(X = x|Y = P)$ and $\Pr(X = x|Y = Q)$ in terms of the priors, image pixels and/or class parameters. Your answer must be a single expression for each probability.

Notice that $(x_i, 1 - x_i)$ is either $(0,1)$ or $(1,0)$ depending on the value of $x_i \in \{0,1\}$.

$$\Pr(X = x|Y = P) = \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1-x_i}$$

$$\Pr(X = x|Y = Q) = \prod_{i=1}^{d} q_i^{x_i}(1 - q_i)^{1-x_i}$$

(d) [2 pts] In terms of the probabilities above, write an equation which holds if and only if x is at the decision boundary of the Bayes' optimal classifier, assuming a (symmetric) 0-1 loss function. No simplification is necessary for full credit.

From Bayes' rule the posterior probability of a class is proportional to the product of its prior and likelihood. At the decision boundary, these posteriors must be exactly equal, which gives the following equation.

$$\pi_p \Pr(X = x|Y = P) = \pi_q \Pr(X = x|Y = Q)$$

(e) [7 pts] It turns out that the decision boundary derived above is actually linear in the features of x, so for some vectors w and scalar b, it can be succinctly expressed as:

$$\{x \in \{0,1\}^d : w^\top x + b = 0\}$$

Find the entries of the vector w and value of b in terms of class priors and parameters, using them to fill in the blanks on the line below.

$w_i =$ $\qquad\qquad\qquad$ $b =$

We proceed directly by writing out the equality above in full and taking the logarithm of both sides.

$$\pi_p \Pr(X = x|Y = P) = \pi_q \Pr(X = x|Y = Q)$$

$$\pi_p \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1-x_i} = \pi_q \prod_{i=1}^{d} q_i^{x_i}(1 - q_i)^{1-x_i}$$

$$\ln \pi_p + \sum_{i=1}^{d} [x_i \ln(p_i) + (1 - x_i)\ln(1 - p_i)] = \ln \pi_q + \sum_{i=1}^{d} [x_i \ln(q_i) + (1 - x_i)\ln(1 - q_i)]$$

$$\ln \frac{\pi_p}{\pi_q} + \sum_{i=1}^{d} x_i \frac{\ln p_i}{\ln q_i} + \sum_{i=1}^{d} (1 - x_i)\frac{\ln(1 - p_i)}{\ln(1 - q_i)} = 0$$

From this expression, we can peel off the coefficient of x_i and the constant term at the very end.

$$w_i = \frac{\ln p_i}{\ln q_i} - \frac{\ln(1 - p_i)}{\ln(1 - q_i)} \qquad\qquad b = \ln \frac{\pi_p}{\pi_q} + \sum_{i=1}^{d} \frac{\ln(1 - p_i)}{\ln(1 - q_i)}$$

Q4. [15 pts] Gaussian Mean Estimation

Suppose $Y \in \mathbb{R}^d$ is a random variable distributed as $\mathcal{N}(\theta, I_{d\times d})$ for some unknown $\theta \in \mathbb{R}^d$. We observe a sample $y \in \mathbb{R}^d$ of Y and want to estimate θ.

(a) [2 pts] What is the maximum likelihood estimator(MLE) of θ? Write down the answer in the box.

y

Now we are going to use ridge regression to solve this problem. Namely, solve $\min_\theta \left\{||y - \theta||^2 + \frac{1}{2}\lambda||\theta||^2\right\}$ to get an estimate of θ.

(b) [2 pts] What is the closed form of estimator $\hat{\theta}(y)$ from ridge regression with regularization parameter λ? Write down the answer in the box.

Taking derivatives we obtain that $2(\theta - y) + \lambda\theta = 0$, which implies $\hat{\theta}(y) = \frac{2}{\lambda+2}y$

(c) [4 pts] Derive the population risk $\mathbb{E}||Y - \hat{\theta}(y)||^2$ for ridge regression estimator (expectation is taken with respect to all the randomness including testing time Y and training sample y).

n

$$\begin{aligned}\mathbb{E}||Y - \hat{\theta}(y)||^2 &= ||\mathbb{E}\hat{\theta} - \theta||^2 + \mathbb{E}||\hat{\theta} - \mathbb{E}\hat{\theta}||^2 + d \\ &= \left(\frac{\lambda}{2+\lambda}\right)^2 ||\theta||^2 + \left(\frac{2}{2+\lambda}\right)^2 d + d\end{aligned}$$

(d) [3 pts] what is the population risk for the MLE estimator. Write down the answer in the box.

$d + d = 2d$

(e) [1 pt] Suppose we choose $\lambda = d$, find out the condition on θ such that the ridge regression estimator has a lower risk than the MLE estimator.

Comparing the risk of the two estimators, suffices to find θ such that

$$\frac{d^2}{(2+d)^2}||\theta||^2 + \frac{4d}{(2+d)^2} \leq d,$$

which is equivalent to $||\theta|| \leq \sqrt{d+4}$.

(f) [2 pts] This implies that MLE estimator, although seems to be the most natural estimator, does not always achieve the lowest risk. Briefly explain the reasons behind this fact.

In this particular case, if we know the magnitude of θ is small, adding regularization penalty utilizes this prior knowledge.

(As an aside, the phenomenon that MLE is not the best estimator is much more general than this context. See James-Stein estimator for more details.)

(g) [1 pt] Based on the previous parts, write down one potential advantage of ridge regression over ordinary least square regression (namely, why do we sometimes add the regularization term).

In some circumstances, ridge regression (with cross validation to choose λ) could achieve lower population risk.

Q5. [12 pts] Estimation of Linear Models

In all of the following parts, write your answer as the solution to a norm minimization problem, potentially with a regularization term. **You do not need to solve the optimization problem.** Simplify any sums using matrix notation for full credit.
Hint: Recall that the MAP estimator maximizes $P(\theta|Y)$: $\hat{\theta} = \arg\max P(Y|\theta)P(\theta)/P(Y) = \arg\max_{\theta\in\mathbb{R}^d} P(Y|\theta)P(\theta)$. The difference between MAP and MLE is the inclusion of a prior distribution on θ in the objective function.

For the following problems assume you are given $X \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$ as training data.

(a) [3 pts] Let $y = X\theta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \Sigma)$ for some positive definite, diagonal Σ. Write the MLE estimator of θ as the solution to a weighted least squares problem, potentially with a regularization term.

$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^d}$\____

Observe that $y \sim \mathcal{N}(X\theta, \Sigma)$. From the PDF of y, maximizing the likelihood the data given θ is equivalent to minimizing:

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta\in\mathbb{R}^d}(y - X\theta)^\top\Sigma^{-1}(y - X\theta)\\ &= \arg\min_{\theta\in\mathbb{R}^d} ||\Sigma^{-\frac{1}{2}}(y - X\theta)||_2^2\end{aligned}$$

Observe that this is weighted least squares.

(b) [3 pts] Let $y|\theta \sim \mathcal{N}(X\theta, \Sigma)$ for some positive definite, diagonal Σ. Let $\theta \sim \mathcal{N}(0, \lambda I_d)$ for some $\lambda > 0$ be the prior on θ. Write the MAP estimator of θ as the solution to a weighted least squares minimization problem, potentially with a regularization term.

$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^d}$\____

Observe that $(y, \theta) \sim \mathcal{N}(X\theta, \Sigma)\mathcal{N}(0, \lambda I_d)$. Finding the MAP estimator of θ is equivalent to minimizing:

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2}(y - X\theta)^\top\Sigma^{-1}(y - X\theta) + \frac{1}{2\lambda}\theta^\top\theta\\ &= \arg\min_{\theta\in\mathbb{R}^d} ||\Sigma^{-\frac{1}{2}}(y - X\theta)||_2^2 + \frac{1}{\lambda}||\theta||_2^2\end{aligned}$$

Observe that this is weighted ridge regression.

(c) [3 pts] Let $y = X\theta + \epsilon$ where $\epsilon_i \overset{i.i.d.}{\sim} Laplace(0, 1)$. Recall that the pdf for $Laplace(\mu, b)$ is $p(x) = \frac{1}{2b}\exp\left(-\frac{1}{b}|x - \mu|\right)$. Write down the MLE estimator of θ as the solution to a norm minimization optimization problem.

$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^d}$\____

Observe that $y_i \sim Laplace(x_i^\top \theta, 1)$.

$$\begin{aligned}
\hat{\theta} &= \arg\max_{\theta \in \mathbb{R}^d} \prod_{i=1}^{N} \frac{1}{2} \exp\left(-|y_i - x_i^\top \theta|\right) \\
&= \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{N} |y_i - x_i^\top \theta| \\
&= \arg\min_{\theta \in \mathbb{R}^d} \|y - X\theta\|_1
\end{aligned}$$

Observe that this estimator minimizes the sum of absolute differences.

(d) [3 pts] Let $y|\theta \sim \mathcal{N}(X\theta, \Sigma)$ for some positive definite, diagonal Σ. Let $\theta_i \overset{i.i.d.}{\sim} Laplace(0, \lambda)$ for some positive scalar λ. Write the MAP estimator of θ as the solution to a weighted least squares minimization problem, potentially with a regularization term.

$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^d}$ ____

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^{d} \frac{1}{\lambda}|\theta_i| + \frac{1}{2}(y - X\theta)^\top \Sigma^{-1}(y - X\theta) \\ &= \arg\min_{\theta\in\mathbb{R}^d} \|\Sigma^{-\frac{1}{2}}(y - X\theta)\|_2^2 + \frac{2}{\lambda}\|\theta\|_1\end{aligned}$$

Observe that this is weighted LASSO regression.

Q6. [9 pts] Bias-Variance for Least Squares

For this problem, we would like to analyze the performance of linear regression on our given data $(X, \tilde{Y})$, where $X \in \mathbb{R}^{n\times d}$ and $\tilde{Y} \in \mathbb{R}^n$.

The original Y perfectly fit a line from the original data, i.e. $Y = X\beta$. However, we do not know the original data, we only know $\tilde{Y} = Y + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I_n)$, i.e. we only know the $\tilde{y}$s that are distorted from the actual ys with mean-zero, independent variance-one Gaussian noise.

Recall that via least squares, the predicted regression coefficients are $\tilde{\beta} = (X^T X)^{-1} X^T \tilde{Y}$.

(a) [1 pt] For a test data point $(z, y) \in (\mathbb{R}^d, \mathbb{R})$, call the predicted value from our model $\tilde{f}(z)$. What is $\tilde{f}(z)$?

The predicted value is $\tilde{f}(z) = z^t\tilde{\beta}$.

(b) [1 pt] Note that for our test data point z, we do not observe the true value y ever, only $\tilde{y} = y + \varepsilon_y$ where $\varepsilon_y \sim \mathcal{N}(0, 1)$. We would like to calculate the expected squared-error $\mathbb{E}[(\tilde{y} - \tilde{f}(z))^2]$. Apply the Bias-Variance decomposition to decompose this expected squared error into three pieces, you do not have to simplify further. Clearly label what the three pieces correspond to.

$$\mathbb{E}[(\tilde{y} - \tilde{f}(z))^2] = \mathrm{var}[\varepsilon_y] + (y - \mathbb{E}[\tilde{f}(z)])^2 + \mathrm{var}[\tilde{f}(z)]$$

1. $\mathrm{var}[\varepsilon_y]$ is irreducible noise
2. $(y - \mathbb{E}[\tilde{f}(z)])^2$ is bias squared
3. $\mathrm{var}[\tilde{f}(z)]$ is the variance

(c) [2 pts] Derive the bias. Show your work.

We can write $\tilde{\beta} = (X^T X)^{-1} X^T \tilde{Y} = (X^T X)^{-1} X^T Y + (X^T X)^{-1} X^T \varepsilon$. Because the noise is mean-zero, this means $\mathbb{E}[\tilde{\beta}] = \beta$, so $(\beta^t z - \beta^t z)^2 = 0$.

This is not surprising because the data perfectly fit a line, and our model is a linear regression, so it is possible for the bias to be zero.

(d) [3 pts] Show that the variance is $z^t (X^T X)^{-1} z$.

First we calculate $\mathrm{var}[\beta]$.

$$\mathrm{var}[\beta] = \mathrm{var}[(X^T X)^{-1} X^T \varepsilon] = (X^T X)^{-1} X^T I_n X (X^T X)^{-1} = (X^T X)^{-1}$$

Now, $\mathrm{var}[\tilde{f}(z)] = \mathrm{var}[z^t\tilde{\beta}] = z^t (X^T X)^{-1} z$.

(e) [2 pts] Argue why in this case the variance is always at least the bias, for any potential test data point z.

Recall that $\mathrm{var}[\beta] = (X^T X)^{-1}$. Variance-covariance matrices are PSD, so $z^t (X^T X)^{-1} z \geq 0$ for all z. The bias is 0, so the variance is always at least the bias.

Q7. [5 pts] Estimates of Variance

Assume that data points $X_1, \ldots, X_n$ are sampled i.i.d from a normal distribution $\mathcal{N}(0, \sigma^2)$. You know that the mean of this distribution is 0, but you do not know the variance σ^2.

Recall that if a variable $X \sim \mathcal{N}(\mu, \sigma^2)$, its probability density function is:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

(a) [3 pts] Write the expression for the log-likelihood $\ln \mathbb{P}(X_1, \ldots, X_n \mid \sigma^2)$. Simplify as much as possible.

$$\begin{aligned}\ln \mathbb{P}(X_1, \ldots, X_n \mid \sigma^2) &= \ln \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2}\,\pi} e^{\frac{-x_i^2}{2\sigma^2}} \\ &= -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2\right)\end{aligned}$$

(b) [2 pts] Find the $\hat{\sigma}^2$ that maximizes this expression, i.e. the MLE.

Take the derivative and set it to 0:

$$\begin{aligned}\frac{-n}{\sigma} + \frac{\sum_{i=1}^{n} x_i^2}{\sigma^3} &= 0 \\ \sigma^2 &= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right)\end{aligned}$$

Q8. [6 pts] Train and Test Error

Assume a general setting for regression with arbitrary loss function $L(y, \hat{y}) \geq 0$. We have devised a family of models $r_\theta \colon \mathbb{R}^d \to \mathbb{R}$ parameterized by $\theta \in \Theta$.

Let $\{(x_1, y_1), \dots (x_n, y_n)\}$ be a test set and $\{(\tilde{x}_1, \tilde{y}_1), \dots (\tilde{x}_m, \tilde{y}_m)\}$ be a training set, both sampled from the same joint distribution $(X, Y) \in (\mathbb{R}^d, \mathbb{R})$. Then we have $R_{tr}(\theta) = \frac{1}{n}\sum_{i=1}^{n} L(r_\theta(x_i), y_i)$ and $R_{te}^{(m)}(\theta) = \frac{1}{m}\sum_{i=1}^{m} L(r_\theta(\tilde{x}_i), \tilde{y}_i)$ as the train and test error depending on the setting of θ, respectively.

We have found the optimal $\hat{\theta} = \text{argmin}_{\theta\in\Theta} R_{tr}(\theta)$, and would like to show that

$$\mathbb{E}[R_{tr}(\hat{\theta})] \leq \mathbb{E}[R_{te}^{(m)}(\hat{\theta})]$$

(a) [2 pts] Show that $\mathbb{E}[R_{te}^{(m)}(\hat{\theta})]$ is the same regardless of the size of the test set m.

$$\mathbb{E}[R_{te}(\hat{\theta})] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} L(r_\theta(\tilde{x}_i), \tilde{y}_i)\right] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[L(r_\theta(\tilde{x}_i), \tilde{y}_i)\right]$$
$$= \left(\frac{1}{m} \cdot m\right) \mathbb{E}\left[L(r_\theta(\tilde{x}_i), \tilde{y}_i)\right]$$

All the $(\tilde{x}_i, \tilde{y}_i)$ are sampled from the same distribution, so the expected average test error is always $\mathbb{E}\left[L(r_\theta(\tilde{x}_i), \tilde{y}_i)\right]$.

(b) [2 pts] Due to the previous part, we can work with a test set that is the same size as the training set. Argue that

$$\mathbb{E}[R_{tr}(\hat{\theta})] = \mathbb{E}[\text{min}_{\theta\in\Theta} R_{te}^{(n)}(\theta)]$$

Both sides represent the same thing now, just under different names. On either side we are finding the smallest possible average error for n data points, finding the best θ for that set of data points.

(c) [1 pt] Argue that $\mathbb{E}[\text{min}_{\theta\in\Theta} R_{te}^{(n)}(\theta)] \leq \mathbb{E}[R_{te}^{(n)}(\hat{\theta})]$, completing the proof.

We have that $\text{min}_{\theta\in\Theta} R_{te}^{(n)}(\theta) \leq R_{te}^{(n)}(\hat{\theta})$ by definition, so

$$\mathbb{E}[\text{min}_{\theta\in\Theta} R_{te}^{(n)}(\theta)] \leq \mathbb{E}[R_{te}^{(n)}(\hat{\theta})]$$

(d) [1 pt] True or False: For all training and test datasets,

$$R_{tr}(\hat{\theta}) \leq R_{te}(\hat{\theta})$$

○ True ● False

We just proved that this inequality holds, but just *in expectation*. It's still possible to get a 'lucky' test set with smaller average error than the training set. We showed that we don't expect this to happen, but it can still happen.