CS 189 / 289A Introduction to Machine Learning
Fall 2022 Jennifer Listgarten, Jitendra Malik HW3

Due 10/14 at 11:59pm

- Homework 3 consists only of written questions.
- We prefer that you typeset your answers using LaTeX or other word processing software. If you haven't yet learned LaTeX, one of the crown jewels of computer science, now is a good time! Neatly handwritten and scanned solutions will also be accepted for the written questions.
- In all of the questions, **show your work**, not just the final answer.
- This homework is due a bit earlier than usual to give you time to prepare for Midterm 1.

Deliverables:

Submit a PDF of your homework to the Gradescope assignment entitled "HW3 Write-Up". **Please start each question on a new page.** If there are graphs, include those graphs in the correct sections. **Do not** put them in an appendix. We need each solution to be self-contained on pages of its own.

- In your write-up, please state with whom you worked on the homework. If you worked by yourself, state you worked by yourself. This should be on its own page and should be the first page that you submit.
- In your write-up, please copy the following statement and sign your signature underneath. If you are using LaTeX, you must type your full name underneath instead. We want to make it *extra* clear so that no one inadvertently cheats. *"I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."*

1 Logistic posterior with exponential class conditionals (6 points)

We have seen in class that Gaussian class conditionals can lead to a logistic posterior that is linear in X. Now, suppose the class conditionals are exponentially distributed with parameters λ_i, i.e.

$$p(x|Y = i) = \lambda_i \exp(-\lambda_i x), \quad \text{where } i \in \{0, 1\}, \qquad Y \sim \text{Bernoulli}(\pi)$$

(a) (4 points) Show that the posterior distribution of the class label given X is also a logistic function, however with a linear (affine) argument in X. That is, $p(Y = 1|x) = \frac{1}{1+\exp(-h(x))}$ where h is an affine function of x involving λ_0, λ_1, and π, that you should find.

(b) (2 points) Assuming $\lambda_0 < \lambda_1$, what is the decision boundary and decision rule of the Bayes classifier?

2 Gaussian Classification (6 points)

Let $P(x \mid C_i) \sim \mathcal{N}(\mu_i, \sigma^2)$ for a two-class, one-dimensional classification problem with classes C_1 and C_2 with prior probabilities $P(C_1) = P(C_2) = 1/2$, and $\mu_2 > \mu_1$.

(a) (3 points) Find the Bayes optimal decision boundary and the corresponding Bayes decision rule.

(b) (3 points) The Bayes error is the probability of misclassification,

$$P_e = P(\text{classified as } C_1 \mid C_2)\, P(C_2) + P(\text{classified as } C_2 \mid C_1)\, P(C_1).$$

Show that the Bayes error associated with this decision rule is

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-z^2/2} dz$$

where $a = \dfrac{\mu_2 - \mu_1}{2\sigma}$.

3 Poisson Classification (12 points)

Recall that the probability mass function (PMF) of a Poisson random variable is

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad x \in \{0, 1, \ldots, \infty\}$$

The PMF is defined for non-negative integral values.

You are given a sample containing a radioactive isotope. You know it is one of two potential isotopes: ω_1 and ω_2, but you do not know which one. A priori, it is equally likely to be either one, i.e. $P(\omega_1) = P(\omega_2) = \frac{1}{2}$.

Using a Geiger counter, you measure x, the number of decay events from the sample within a fixed time interval. This follows a Poisson distribution with mean dependent on the isotope. Thus, $x|\omega_1 \sim \text{Poisson}(\lambda_1)$ and $x|\omega_2 \sim \text{Poisson}(\lambda_2)$, where λ_1 and λ_2 are the expected number of decay events for each isotope considered. λ_1 and λ_2 can be treated as known; one can compute them by looking up the decay rate and atomic weight of each isotope and weighing the sample.

(a) (2 points) Find $P(\omega_1|x)$ in terms of λ_1 and λ_2. Show that the posterior $P(\omega_1|x)$ is a logistic function $\frac{1}{1+e^{-h(x)}}$ and write $h(x)$ in terms of λ_1, λ_2, and x.

(b) (4 points) Find the Bayes optimal rule (decision boundary) for classifying the observation x as either isotope. In the case where $P(\omega_1) = P(\omega_2) = \frac{1}{2}$ and $\lambda_1 = 10$ and $\lambda_2 = 15$, calculate the decision boundary and the 2×2 confusion matrix $C \in [0, 1]^{2\times 2}$:

$$C = \begin{bmatrix} P(\hat{\omega}(x) = \omega_1, \omega_1) & P(\hat{\omega}(x) = \omega_1, \omega_2) \\ P(\hat{\omega}(x) = \omega_2, \omega_1) & P(\hat{\omega}(x) = \omega_2, \omega_2) \end{bmatrix}$$

where $\hat{\omega}(x)$ denotes the class predicted by the Bayes optimal rule. For reference, $P(\hat{\omega}(x) = \omega_1, \omega_1)$ denotes the joint probability that $\hat{\omega}(x) = \omega_1$ and the true class is ω_1. *Hint: first, compute the diagonal elements.*

Using the confusion matrix, compute the total error rate, precision, recall, and F_1 score. F_1 score is the harmonic mean of precision and recall, and has the formula

$$F_1 = \frac{2TP}{2TP + FP + FN} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

For the latter three metrics, treat ω_1 as the positive class and ω_2 as the negative class (perhaps ω_1 is a radioactive isotope you are particularly interested in).

(c) (4 points) Suppose instead of one, we can obtain two independent measurements x_1 and x_2 from the radioactive substance. How does the classification rule change? In the case where $P(\omega_1) = P(\omega_2) = \frac{1}{2}$ and $\lambda_1 = 10$ and $\lambda_2 = 15$, recalculate the confusion matrix and metrics from the previous part.

(d) (2 points) Comment on how your answers to part (c) compare to your answers to part (b). Give an explanation for what you observe.

4 Exploring Bias Variance between Ridge and Least Squares Regression (18 points)

Recall the statistical model for ridge regression from lecture. We have a design matrix $\mathbf{X}$, where the rows of $\mathbf{X} \in \mathbb{R}^{n \times d}$ are our data points $\mathbf{x}_i \in \mathbb{R}^d$. We assume a linear regression model

$$Y = \mathbf{X}\mathbf{w}^* + \mathbf{z}$$

Where $\mathbf{w}^* \in \mathbb{R}^d$ is the true parameter we are trying to estimate, $\mathbf{z} = [z_1, \ldots, z_n]^\top \sim \mathcal{N}(0, \sigma^2 I_n)$, and $Y = [y_1, \ldots, y_n]^\top$ is the random variable representing our labels.

Throughout this problem, you may assume $\mathbf{X}^\top\mathbf{X}$ is invertible. Given a realization of the labels $Y = \mathbf{y}$, recall these two estimators we have studied:

$$\mathbf{w}_{\text{ols}} = \min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$$
$$\mathbf{w}_{\text{ridge}} = \min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

(a) (1 point) Write the solution for $\mathbf{w}_{\text{ols}}, \mathbf{w}_{\text{ridge}}$. No need to derive it.

(b) (2 points) Let $\widehat{\mathbf{w}} \in \mathbb{R}^d$ denote any estimator of $\mathbf{w}^*$. In the context of this problem, an estimator $\widehat{\mathbf{w}} = \widehat{\mathbf{w}}(Y)$ is any function which takes the data $\mathbf{X}$ and a realization of Y, and computes a guess of $\mathbf{w}^*$.

Define the MSE (mean squared error) of the estimator $\widehat{\mathbf{w}}$ as

$$\text{MSE}(\widehat{\mathbf{w}}) := \mathbb{E}\left[\left\|\widehat{\mathbf{w}} - \mathbf{w}^*\right\|_2^2\right].$$

Above, the expectation is taken w.r.t. the randomness inherent in $\mathbf{z}$. Note that this is a multivariate generalization of the mean squared error we have seen previously.

Define $\widehat{\boldsymbol{\mu}} := \mathbb{E}[\widehat{\mathbf{w}}]$. Show that the MSE decomposes as such

$$\text{MSE}(\widehat{\mathbf{w}}) = \underbrace{\left\|\widehat{\boldsymbol{\mu}} - \mathbf{w}^*\right\|_2^2}_{\text{Bias}(\widehat{\mathbf{w}})} + \underbrace{\text{Tr}(\text{Cov}(\widehat{\mathbf{w}}))}_{\text{Var}(\widehat{\mathbf{w}})}$$

Note that this is a multivariate generalization of the bias-variance decomposition we have seen previously.

Hint: Given vectors x and y, the following equality holds: $\langle x, y\rangle = \text{Tr}(xy^\top)$. Also, expectation and trace commute, so $\mathbb{E}[\text{Tr}(A)] = \text{Tr}(\mathbb{E}[A])$ for any square matrix A.

(c) (2 points) Show that

$$\mathbb{E}[\mathbf{w}_{\text{ols}}] = \mathbf{w}^*, \qquad \mathbb{E}[\mathbf{w}_{\text{ridge}}] = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{w}^*.$$

That is, $\text{Bias}(\mathbf{w}_{\text{ols}}) = 0$, and hence $\mathbf{w}_{\text{ols}}$ is an *unbiased* estimator of $\mathbf{w}^*$, whereas $\mathbf{w}_{\text{ridge}}$ is a *biased* estimator of $\mathbf{w}^*$.

(d) (4 points) Let $\{\gamma_i\}_{i=1}^d$ denote the d eigenvalues of the matrix $\mathbf{X}^\top\mathbf{X}$. Show that

$$\mathrm{Tr}(\mathrm{Cov}(\mathbf{w}_{\mathrm{ols}})) = \sigma^2 \sum_{i=1}^{d} \frac{1}{\gamma_i}, \qquad \mathrm{Tr}(\mathrm{Cov}(\mathbf{w}_{\mathrm{ridge}})) = \sigma^2 \sum_{i=1}^{d} \frac{\gamma_i}{(\gamma_i+\lambda)^2} .$$

The quantity $\sum_{i=1}^{d} \frac{\gamma_i}{(\gamma_i+\lambda)^2}$ is referred to as the degree of freedom in ridge regression.

Finally, use these formulas to conclude that

$$\mathrm{Var}(\mathbf{w}_{\mathrm{ridge}}) < \mathrm{Var}(\mathbf{w}_{\mathrm{ols}}) .$$

Note that this is opposite of the relationship between the bias terms. In the next parts, we will explore this trade-off further.

Hint: Recall that for a random vector v, its covariance matrix is defined as $\mathrm{Cov}(v) = \mathbb{E}[(v - \mathbb{E}[v])(v - \mathbb{E}[v])^\top]$. Also, recall that if A and B are constant matrices and M is a random matrix, then $\mathbb{E}[AMB] = A\mathbb{E}[M]B$.

Hint: Remember the relationship between the trace and the eigenvalues of a matrix. Also, for the ridge variance, consider writing $\mathbf{X}^\top\mathbf{X}$ in terms of its eigen-decomposition $U\Sigma U^\top$. Note that $\mathbf{X}^\top\mathbf{X} + \lambda I_d$ has the eigendecomposition $\mathrm{U}(\Sigma + \lambda I_d)U^\top$.

(e) (6 points) Let's use the bias and variance trade-off to study the choice between OLS and Ridge Regression.

(i) Using the bias-variance decomposition, show that

$$\mathrm{MSE}(\mathbf{w}_{\mathrm{ols}}) = \sigma^2 \sum_{i=1}^{d} \frac{1}{\gamma_i}.$$

(ii) Using the bias-variance decomposition, show that $\mathrm{MSE}(\mathbf{w}_{\mathrm{ridge}})$ is bounded above as,

$$\mathrm{MSE}(\mathbf{w}_{\mathrm{ridge}}) \le \sigma^2 \sum_{i=1}^{d} \frac{\gamma_i}{(\gamma_i+\lambda)^2} + \left\|\mathbf{w}^*\right\|_2^2 \sum_{i=1}^{d} \frac{\lambda^2}{(\gamma_i+\lambda)^2} .$$

Hint: Use following Cauchy-Schwarz matrix vector product inequality in your upper bound: $\|Ax\|_2^2 \le \|A\|_F^2\|x\|_2^2 = \|x\|_2^2 \sum_{i=1}^{d} \lambda_i(A^\top A)$, where $\lambda_i(A^\top A)$ is the ith eigenvalue of $A^\top A$. (In fact, a tighter bound holds, where $\|A\|_F^2$ is replaced with $\|A\|_2^2 = \lambda_{\max}(A^\top A)$, but the looser bound will be easier to minimize.)

(iii) Show that the upper-bound is tight for $\lambda = 0$, meaning that when $\lambda = 0$, the upper-bound equals $\mathrm{MSE}(\mathbf{w}_{\mathrm{ridge}})$.

(iv) Optimize the upper-bound, showing that the lowest value is achieved at

$$\lambda_* = \frac{\sigma^2}{\|\mathbf{w}^*\|_2^2}.$$

which is related to the inverse of the widely used signal to noise ratio statistic (SNR) in signal processing and machine learning. You may use without proof the fact that the upper-bound is convex as a function of λ, implying a stationary point is a global minimum.

(f) (2 points) In practice, both $\|\mathbf{w}^*\|_2$ and σ are unknown. Show that at any fixed $\|\mathbf{w}^*\|_2$ and σ, there exists a $\lambda > 0$ such that, $\text{MSE}(\mathbf{w}_{\text{ridge}}) < \text{MSE}(\mathbf{w}_{\text{OLS}})$. How would you find such a λ in practice?

(g) (1 point) Suppose, a sensor suddenly degrades multiplying the noise $\mathbf{z}$ in recorded observation by a factor of 3. Guided by the form of the minimizer of the $\text{MSE}_{\text{ridge}}$ upper-bound, how would you adjust the regularization λ from its old value of 21 to the new setting?