CS 189:289A Introduction to Machine Learning
Fall 2021 Jennifer Listgarten and Jitendra Malik

HW3

Due 10/11 at 11:59pm

- 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.

Deliverables:

1. 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. 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 next to it. (Mac Preview and FoxIt PDF Reader, among others, have tools to let you sign a PDF file.) 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."*

HW3, ©UCB CS 189:289A, Fall 2021. All Rights Reserved. This may not be publicly shared without explicit permission.

1 Poisson Classification

Recall that the PDF of a Poisson random variable is

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

The PDF is defined for non-negative integral values.

You are given two classes ω_1, ω_2 of Poisson data with parameters λ_1 and λ_2. This means that $x|\omega_1 \sim \text{Poisson}(\lambda_1)$ and $x|\omega_2 \sim \text{Poisson}(\lambda_2)$. Assume that $P(\omega_1) = P(\omega_2) = \frac{1}{2}$.

(a) Find $P(\omega_1|x)$ in terms of λ_1 and λ_2. What type of function is the posterior?

(b) Find the optimal rule (decision boundary) for allocating an observation x to a particular class. 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, probability of correct classification for each class, and total error rate.

(c) Suppose instead of one, we can obtain two independent measurements x_1 and x_2 for the object to be classified. How does the allocation rule change? In the case where $P(\omega_1) = P(\omega_2) = \frac{1}{2}$ and $\lambda_1 = 10$ and $\lambda_2 = 15$, calculate the new total error.

2 Logistic posterior with exponential class conditionals

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.

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

Show that the posterior distribution of the class label given X is also a logistic function, however with a linear argument in X. What is the decision boundary?

3 Gaussian Classification

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

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

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

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

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

4 Bias Variance for Ridge Regression

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, $\mathbf{z} = [z_1, \ldots, z_n] \sim \mathcal{N}(0, \sigma^2 I_n)$, and $Y = [y_1, \ldots, y_n]$ 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) Write the solution for $\mathbf{w}_{\text{ols}}, \mathbf{w}_{\text{ridge}}$. No need to derive it.

(b) 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 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

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

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

Hint: The inner product of two vectors is the trace of their outer product. Also, expectation and trace commute, so $\mathbb{E}[\text{Tr}(A)] = \text{Tr}(\mathbb{E}[A])$ for any square matrix A.

(c) 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, $\mathbf{w}_{\text{ols}}$ is an *unbiased* estimator of $\mathbf{w}^*$, whereas $\mathbf{w}_{\text{ridge}}$ is a *biased* estimator of $\mathbf{w}^*$.

(d) Let $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_d$ denote the d eigenvalues of the matrix $\mathbf{X}^\top \mathbf{X}$ arranged in non-increasing order. Show that

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

Finally, use these formulas to conclude that

$$\mathrm{Tr}(\mathrm{Cov}(\mathbf{w}_{\mathrm{ridge}})) < \mathrm{Tr}(\mathrm{Cov}(\mathbf{w}_{\mathrm{ols}})) \,.$$

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 $U\mathbf{\Sigma}U^\top$.