ECE 592 – Topics in Data Science

Test 4: Machine Learning – Fall 2022

November 14, 2022

Please remember to justify your answers carefully.

Last name: ____________________ First name: ____________________

Please recall the course academic integrity policy (from the syllabus): When working on tests, no cooperation or "collaboration" between students is allowed. While it could be tempting to text or email a friend during a test that is administered electronically, this is not allowed. You will be allowed to use your notes, books, a browser, and software such as Matlab and/or Python.[1] However, while working on the test you should not text, email, or communicate with other people (certainly not other students) in any way, unless you are consulting with the course staff. By submitting the test, you will be acknowledging that you completed the work on your own without the help of others in any capacity. Any such aid would be unauthorized and a violation of the academic integrity policy.

[1] You can use the browser to access Moodle, the course webpage, and look up technical topics. Similar to a normal test, you must not communicate with other people.

Question 1 (Nearest neighbors regression.)
This question deals with k nearest neighbors classification, and extends it to regression.

(a) For binary classes, explain in words how k nearest neighbors classification works. You should not use code or mathematical equations.

(b) How would your answer in part (a) change if there are more than 2 classes?

(c) While classification is a machine learning (ML) approach that attempts to output which among a discrete set of possible classes our data is believed to belong to, in regression our output is typically a real valued number. Again, explain in words how you could perform k nearest neighbors regression.

Question 2 (Linear and quadratic discriminant analysis.)
Based on two classes, red and blue, a 2-dimensional (2D) random vector $[X_1 \quad X_2] \in \mathbb{R}^2$ is generated in different ways. The probabilities of the two classes satisfy $\Pr(\text{blue}) = 0.6$ and $\Pr(\text{red}) = 0.4$. The distributions of the random vector for each class are 2D Gaussian densities centered around $\mu_{\text{blue}} = [0 \quad 0]$ and $\mu_{\text{red}} = [1 \quad 0]$, the Gaussian noise for each coordinate has variance 1, and all Gaussian random variables in this question are assumed to be independent. Recall that a scalar Gaussian random variable X with mean μ and variance σ^2 has probability density function (pdf) given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Therefore, for the red class, the 2 corresponding random variables in our random vector, $[X_1 \quad X_2]$, follow the joint pdf,

$$f(X_1 = x_1, X_2 = x_2|\text{red}) = \frac{1}{\sqrt{2\pi \cdot 1}} e^{-\frac{(x_1-1)^2}{2\cdot 1}} \frac{1}{\sqrt{2\pi \cdot 1}} e^{-\frac{(x_2-0)^2}{2\cdot 1}},$$

where x_1 and x_2 are the numerical values that X_1 and X_2 take, we remind the reader that X_1 and X_2 are statistically independent, and their expected values are 1 and 0, respectively. The joint pdf for the blue class has a similar form, where the expected value of X_1 is 0.

(a) This part involves linear discriminant analysis (LDA). The pdfs for the blue and red classes have the same variance, hence the decision boundary will be a straight line.
Compute the decision boundary for deciding between red and blue given a 2D vector whose values are $[x_1 \quad x_2]$. (In other words, what condition must x_1 and x_2 satisfy such that $\Pr(\text{red}|x_1, x_2) = 0.5$?) Explain why this is a straight line.

(b) This part involves quadratic discriminant analysis (QDA). The pdfs for the blue and red classes have different variances, and the decision boundary will be conic (either a straight line, circle, parabola, hyperbola, or ellipse).
In part (b), the variance of the red class is 4 instead of 1, meaning that

$$f(X_1 = x_1, X_2 = x_2|\text{red}) = \frac{1}{\sqrt{2\pi \cdot 4}} e^{-\frac{(x_1-1)^2}{2\cdot 4}} \frac{1}{\sqrt{2\pi \cdot 4}} e^{-\frac{(x_2-0)^2}{2\cdot 4}},$$

and $f(X_1 = x_1, X_2 = x_2|\text{blue})$ is the same as in part (a). For this modified variance, compute the decision boundary for deciding between red and blue given a 2D vector whose values are $[x_1 \quad x_2]$. Your boundary should have some conic shape, although describing it is beyond the scope of our course.

Question 3 (Linear regression with an ℓ_1 penalty.)
Consider a function $y = f(x)$, where we are given the following 3 data points:

$$(x_1 = 0, y_1 = 0), \quad (x_2 = 1, y_2 = 2), \quad (x_3 = 2, y_3 = 1). \tag{1}$$

In this question, you wil perform linear regression to fit affine functions (lines) to these 3 data points using ℓ_1 and ℓ_2 penalties. While ℓ_2 regression (minimization of squared error) is standard, the ℓ_1 part of the question should be more interesting.

(a) Find the coefficients a and b such that the affine function,

$$g(x) = ax + b,$$

minimizes its ℓ_2 error with respect to (w.r.t.) our 3 data points (1). To be specific, find a, b that minimize $\sum_{i=1}^{3}(y_i - g(x_i))^2$. Make sure to justify your answer.

(b) We now want to find coefficients $\widetilde{a}$ and $\widetilde{b}$ such that $\widetilde{g}(x)$, which is defined in a manner analogous to $g(x)$ in part (a), minimizes its ℓ_1 error w.r.t. our 3 data points. To be specific, we want to find $\widetilde{a}$, $\widetilde{b}$ such that $\sum_{i=1}^{3} |y_i - \widetilde{g}(x_i)|$ is minimized.
Compute the ℓ_1 error obtained using g from part (a). Next, perturb a and/or b, for example replace a by $a + \epsilon$ or b by $b - \epsilon$, where $\epsilon > 0$ is some small value. Design a perturbation that reduces the ℓ_1 error. What is the new ℓ_1 error as a function of ϵ? (Hint: if you are unsure about your answer for part (a), you may assume that $a = 1$ and $b = 0$.)

(c) Using the perturbation style from part (b), increase ϵ in order to minimize the ℓ_1 error. What ϵ minimizes the ℓ_1 error? What is your ℓ_1 error?