ECE 592 – Topics in Data Science

Test 4: Machine Learning – Fall 2022

November 14, 2022

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.

Solution. Find the k nearest neighbors. Identify their k corresponding labels. Our classifier (predicted class) is the majority among those k.

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

Solution. Very similar to part (a), except that this is a "plurality vote" instead of majority vote.

(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.
Solution. Similar to part (a), but output the average of the k real valued numbers.

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.

Solution. The joint pdf for the red class is

$$\Pr(\text{red}) f(X_1 = x_1, X_2 = x_2|\text{red}) = 0.4 \frac{1}{2\pi} e^{-\frac{(x_1-1)^2+(x_2)^2}{2}},$$

and for the blue class,

$$\Pr(\text{blue}) f(X_1 = x_1, X_2 = x_2|\text{red}) = 0.6 \frac{1}{2\pi} e^{-\frac{(x_1)^2+(x_2)^2}{2}}.$$

These 2 joint pdfs must be equal,

$$0.4 \frac{1}{2\pi} e^{-\frac{(x_1-1)^2+(x_2)^2}{2}} = 0.6 \frac{1}{2\pi} e^{-\frac{(x_1)^2+(x_2)^2}{2}}.$$

Performing some algebraic manipulations,

$$e^{-\frac{-2x_1+1}{2}} = 1.5.$$

Taking the log of this expression, $(2x_1 - 1)/2 = \ln(1.5)$. This is a stright line, because x_1 is constant and x_2 could take on any value.

(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.

Solution. The solution is similar to before, except that

$$\Pr(\text{red}) f(X_1 = x_1, X_2 = x_2|\text{red}) = 0.4 \frac{1}{8\pi} e^{-\frac{(x_1-1)^2 + (x_2)^2}{8}}.$$

The joint pdfs are equal,

$$0.4 \frac{1}{8\pi} e^{-\frac{(x_1-1)^2 + (x_2)^2}{8}} = 0.6 \frac{1}{2\pi} e^{-\frac{(x_1)^2 + (x_2)^2}{2}}.$$

Simplifying this expression,

$$e^{-\frac{(x_1-1)^2 + (x_2)^2}{8}} = 6 e^{-\frac{(x_1)^2 + (x_2)^2}{2}}.$$

Taking the natural logarithm and multiplying both sides of the expression by 8,

$$-[(x_1 - 1)^2 + (x_2)^2] = 8 \ln(6) - 4[(x_1)^2 + (x_2)^2].$$

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.

Solution. We can define an error function Ψ, which depends on a and b,

$$\Psi(a, b) = b^2 + (a + b - 2)^2 + (2a + b - 1)^2.$$

Taking the derivative w.r.t. a,

$$\frac{\partial}{\partial a}\Psi(a, b) = 2(a + b - 2) + 2(2a + b - 1)2 = 2(5a + 3b - 4).$$

Taking the derivative w.r.t. b,

$$\frac{\partial}{\partial b}\Psi(a, b) = 2b + 2(a + b - 2) + 2(2a + b - 1) = 2(3a + 3b - 3).$$

Both derivatives must equal 0, hence $3a + 3b = 3$. Substituting this into $5a + 3b = 4$, we see that $2a + (3a + 3b) = 2a + 3 = 4$, hence $a = \frac{1}{2}$. Because $3a + 3b = 3$, b also equals $\frac{1}{2}$.

(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 $b = 0$.)

Solution. Using $a = b = \frac{1}{2}$, meaning that $g(x) = \frac{x+1}{2}$,

$$\sum_{i=1}^{3} |y_i - g(x_i)| = |0 - \frac{1}{2}| + |2 - 1| + |1 - \frac{3}{2}| = 2.$$

Consider the perturbation $\widetilde{b} = b - \epsilon$, and $\widetilde{g}(x) = \frac{1}{2}x + \frac{1}{2} - \epsilon$. The ℓ_1 error is now

$$|0 - (\frac{1}{2} - \epsilon)| + |2 - (1 - \epsilon)| + |1 - (\frac{3}{2} - \epsilon)| = 2 - \epsilon.$$

For a small positive perturbation ϵ, the ℓ_1 error becomes smaller.
For the "you may assume" part, $g(x) = ax + b = x$. The ℓ_1 error is $|0 - 0| + |2 - 1| +$

$|1-2| = 2$. A perturbation could be $\widetilde{g}(x) = (1-\epsilon)x$, in which case the ℓ_1 error declines to $|0-0| + |2-1+\epsilon| + |1-2-2\epsilon| = 2-\epsilon$.

(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?

Solution. For $\epsilon < \frac{1}{2}$, the error keeps decreasing. However, for $\epsilon = \frac{1}{2}$, meaning that $b = \frac{1}{2} - \epsilon = 0$ and $\widetilde{g}(x) = \frac{1}{2}x$, the function becomes 0 at $x = 0$, and any further increase in ϵ will start increasing the ℓ_1 error. Similarly, the function becomes 1 at $x = 2$, and any further increase in ϵ will start increasing the ℓ_1 error. Therefore, the ℓ_1 error seems to be minimal for $\epsilon = \frac{1}{2}$. The ℓ_1 error for $\widetilde{g}(x) = \frac{1}{2}x$ is

$$|0-0| + |2-\frac{1}{2}| + |1-1| = 1\frac{1}{2}.$$

For the "you may assume" part, $\widetilde{g}(x) = (1-\epsilon)x$. Similar to what we have just seen, $\epsilon = \frac{1}{2}$ gives us an ℓ_1 error of $1\frac{1}{2}$.