CS589 Checkpoint Quiz 1

September 9, 2020

- This quiz covers material from Lectures 1, 2 and 3. It is meant as a self-check after you've watched the lectures.
- It is highly encouraged that you also watch Discussions 1-4 if the lecture material is still not clear.
- The answers to this quiz should be submitted via Moodle.
- The deadline is the 14t h of September 2020, 11:59pm.
- There are 4 multiple choice questions in this quiz.
- Each question is worth 0.5 points. Total: 2 points for the quiz.
- The test is multiple choice, so only correct answers will be counted.
- You may use any resources to answer the questions.
- You may discuss the questions with your peers, but please do not ask for the answers outright.

Good Luck!

1 K-Nearest Neighbors

You are tasked with building a case-based screening tool for COVID-19 and given the data in Table 1. Due to the small number of samples, you decide to apply the K-NN classifier with $k = 3$. As the features are categorical, with no particular ordering, you use the Manhattan distance.

Sample	Fever (X_1)	Sneezing (X_2)	Cough (X_3)	Shortness of breath (X_4)	COVID? (Y)
1	No	Yes	None	Yes	1
2	Yes	Yes	Dry	Yes	1
3	Yes	No	Wet	Yes	0
4	No	No	Wet	No	0
5	Yes	No	Dry	No	1
6	Yes	Yes	None	Yes	0

Table 1: Training data for the COVID-19 classifier.

A new patient comes in with a fever, no sneezing, dry cough and shortness of breath. Does the subject have COVID-19, according to your classifier?

(a) Yes

(b) No

2 Decision Trees

You decide to train a decision tree using the training data in Table 1 using accuracy as the splitting metric. What should you select as the root node of the tree out of the following options?

(a) $X_1 = No$

(b) $X_2 = No$

(c) $X_3 = Wet$

(d) $X_3 = None$

3 Probabilistic classification

You are given additional information about the COVID symptoms, as follows:

- X_1 has no information about Y
- X_2 is independent of the other features given Y
- X_3 and X_4 are independent of the other features given Y

In light of this information, what is the form of the Bayes optimal classifier?

(a) $\arg \max_{c \in \{0,1\}} P(X_1, X_2|Y = c)P(X_3|Y = c)P(X_4|Y = c)P(Y = c)$

(b) $\arg \max_{c \in \{0,1\}} P(X_2|Y = c)P(X_3, X_2|Y = c)P(X_4|Y = c)P(Y = c)$

(c) $\arg \max_{c \in \{0,1\}} P(X_2|Y = c)P(X_3, X_4|Y = c)P(Y = c)$

(d) $\arg \max_{c \in \{0,1\}} P(X_2)P(X_3, X_4|Y = c)P(Y = c)$

4 Naive Bayes

Assume you are training a Naive Bayes classifier for the above problem, and are trying to estimate $\phi_{3,0}(x_3)$ using MAP estimation. What prior should we use and how will the MAP estimate look like?

(a) Gaussian prior $\mathcal{N}(\mu, \sigma)$;

$$\phi_{3,0}^{MAP}(x_3) \sim \mathcal{N}(\mu + \frac{\#(x_3 = None)}{n}, \sigma)$$

where the symbol # counts the number of instances that match the condition and n is the total number of samples.

(b) Dirichlet prior (3, α_1, α_2, α_3)

$$\phi_{3,0}^{MAP}(x_3) \sim Dirichlet(3, \#(x_3 = v_1) + \alpha_1, \#(x_3 = v_2) + \alpha_2, \#(x_3 = v_3) + \alpha_3)$$

where v_1 =None, v_2 =Wet, v_3 =Dry and the symbol # counts the number of instances.

(c) Beta prior, β_1, β_2

$$\phi_{3,0}^{MAP}(x_3) \sim Beta(\#(x_3 = Wet) + \beta_1, \#(x_3 = Dry) + \beta_2)$$