CS301 - Data Science Midterm Spring 2023

PLEASE WRITE CLEARLY AND DO NOT WRITE ON THESE PAGES. USE YOUR OWN PAGES AND SUBMIT THEM ONLY.

Question Set 1 (20 points)

QS1.A (5 points)

Figure 1: dog

My dog is an English cream golden retriever. If the probability of barking (b) given that its night (n) is

$$p(b|n) = 5\%$$

write the expression and calculate the number of nights per year **(365 nights)** that I will be waken up.

QS1.B (5 points)

What is the maximum Area Under the Curve (AUC) of a Receiver Operating Characteristic (ROC) that you can have in a classification system (4 points) and why (6 points)?

QS1.C (10 points)

Which of the expressions below, if any, correspond to (a) True Positive Rate and (b) False Positive Rate (4 points). Explain your answer (6 points).

(a) $p(\hat{y} = 1|y = 0, x)$ **FP**

(b) $p(\hat{y} = 1|y = 0, x) + p(\hat{y} = 0|y = 1, x)$

(c) $p(\hat{y} = 1|y = 1, x) + p(\hat{y} = 0|y = 0, x)$

(d) $p(\hat{y} = 0|y = 0, x)$ **TN**

(e) $p(y = 1|\hat{y} = 0, x)$

(f) $p(y = 1|\hat{y} = 0, x) + p(y = 0|\hat{y} = 1, x)$

(g) $p(y = 1|\hat{y} = 1, x) + p(y = 0|\hat{y} = 0, x)$

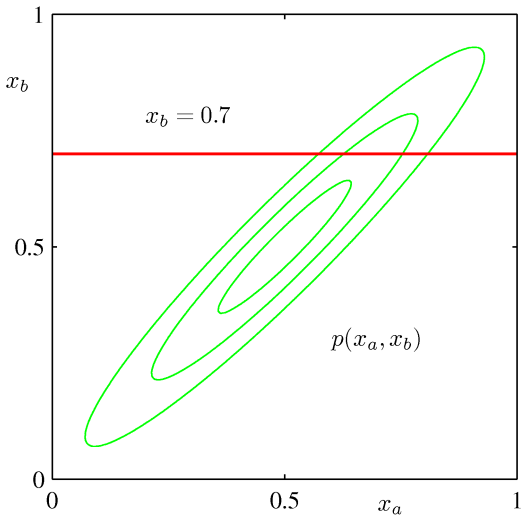
(h) $p(y = 0|\hat{y} = 0, x)$

Question Set 2 (15 points)

QS2.A (5 points)

Your manager is telling you that a dataset they have been using came from a bivariate joint distribution shown below. They need your help to plot

- the marginal $p(x_a)$ (5 points) and
- the conditional $p(x_a|x_b = 0.7)$ (5 points)



QS2.B (10 points)

In the figure below you are given the data points $x_n, n = 1, ..., 7$ and you are asked to fit a Gaussian model on these points with mean μ and variance σ^2.

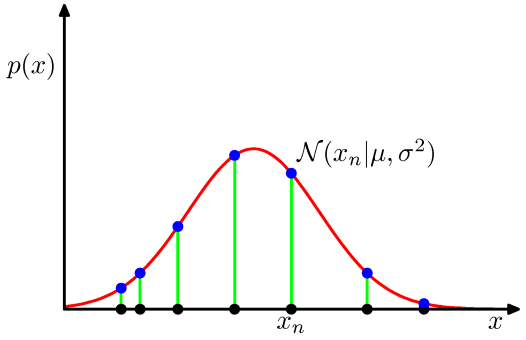


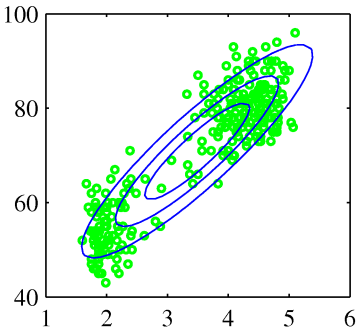
Figure 2: likelihood

Write the equation for the negative log likelihood function and describe how it can be minimized with SGD. To answer the `how` you need to draw a block diagram of the process quoting with equations all the steps.

Question Set 3 (25 points)

A colleague of yours used the model of a bivariate Gaussian distribution ($p_{model}(\mathbf{x})$ to fit the data shown below.

- (5 points) Explain why this was not a good idea.
- (5 points) Draw the contour plot of the ($p_{model}(y|\mathbf{x}, \mathbf{w})$ you would used instead. To do so, you will need to recreate in your paper the plot you see with the data only (approximately) and plot the contour lines of the desired function $p_{model}(y|\mathbf{x}, \mathbf{w})$.
- (5 points) Write the likelihood function equation that includes two Gaussian components and specify the number of parameters that need to be estimated.
- (10 points) Write the SGD update equation for the parameters of the model.



QUESTION SET 4 (10 points)

QS4.A (5 points)

Explain how correlation of features will affect the prediction performance with linear predictor (performing regression or classification tasks). Provide this explanation using a dataset that has binary features e.g. Male and Female canines and employs one-hot encoding.

QS4.B (5 points)

Describe a method on how to detect strong correlations between features that are numerical in nature.

Question Set 5 (30 points)

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

6 true
6 false

Someone is giving you the following dataset of 12 examples and is asking you to help them design a stump, a decision tree of depth 1 (i.e. only one split). The stump will ultimately determine if a patron will wait for a table to be freed in a restaurant or not (last column).

Out of all the features (attributes) you are given, you are asked to compare between:

Pat (Patrons): how many people are in the restaurant (values are None, Some, and Full).

Type: the restaurant's cuisine (French, Thai, Burger, Italian)

Show all your calculations that lead to the determination of the root note between these two choices. If just the root node is quoted you will be automatically granted the grade of 0. Make sure that your answer is clearly handwritten.

HINT: You need to use the Information Gain from your notes to select between the two features. The Information Gain is the difference between the entropy of the parent node and the weighted average of the entropy of the children nodes.

$$IG = H(D) - \sum_j \frac{|D^j|}{|D|} H(D^j)$$

where $|.|$ is the cardinality operator i.e. the number of elements in the corresponding set.

GOOD LUCK !