ML Final 2019 Answer (V1.0) Dr\ Hanaa Bayoumi

Solved by **Ahmed Sallam** if you find any mistakes, please contact me.

لا تنسونا من صالح دعائكم

Question 1 : same as Question 1 in 2020 Final Exam

Question 2 :

We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied.

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

Remember The equations:

Entropy: $H(S) = - p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$

$$Gain(S,A) = H(S) - \sum_{V \in Values(A)} \frac{|S_V|}{|S|} H(S_V)$$

V … possible values of A
S … set of examples {X}
S_v … subset where $X_A = V$

a) What is the entropy H(Passed)?

$$\text{H(Passed)} = -(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}) = 0.92$$

b) What is the entropy H(Passed | GPA)?

هنطبق قانون اسمه **conditional entropy**

طب ايه فكرته ؟

دلوقتي احنا عارفين ايه الناتج بتاع ال GPA بس هل ال GPA ده passed ولا لا

بمعني اخر دلوقتي واحد جايب GPA=3 هل ده ناجح ولا لا

$$H(Y|X) = -\sum_{ij} p(y_j, x_i) log[p(y_j|x_i)] = -\sum_{ij} p(y_j, x_i) log\frac{p(y_j, x_i)}{p(x_i)}.$$

$$\text{H(Passed | GPA)} = -\frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) - \frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) - -\frac{1}{3}(1\log_2 1) = \frac{2}{3}$$

C) What is the entropy H(Passed | Studied)

$$H(Y|X) = -\sum_{ij} p(y_j, x_i) log[p(y_j|x_i)] = -\sum_{ij} p(y_j, x_i) log \frac{p(y_j, x_i)}{p(x_i)}.$$

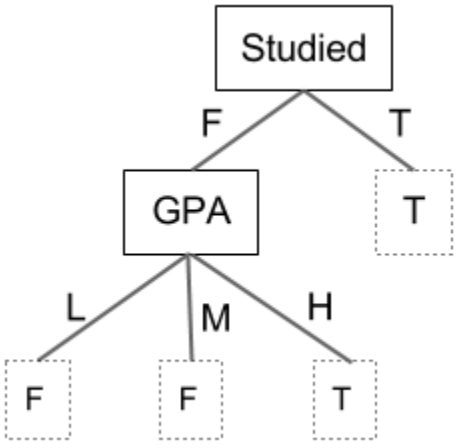
H(Passed | Studied)= $-\frac{1}{2}(\frac{2}{3}\log_2\frac{2}{3}+\frac{1}{3}\log_2\frac{1}{3})-\frac{1}{2}(1\log_2 1)] = 0.46$

d)Draw the full decision tree that would be learned for this dataset. You do not need to show any calculations.

We want to split first on the variable which maximizes the information.

gain H (Passed) − H (Passed|A).

This is equivalent to minimizing H (Passed|A), so we should split on "Studied?" first.



Question 3: Same as Question 3 in 2020 Final Exam

Question 4 :

1) Describe the difference between parametric methods and nonparametric methods.

Parametric model:

Definition: Parametric methods make assumptions about the underlying functional form of the relationship between the input features and the output. These assumptions are usually based on a specific mathematical model with a fixed number of parameters.

الملخص :

- Model fit the data exactly

- this models have a parameters that model are try to find and calculate them exactly

Ex. linear regression, logistic regression, *and linear SVM (Support Vector Machines).*

Non- Parametric model:

Definition: Nonparametric methods do not make explicit assumptions about the functional form of the underlying relationship. Instead, they aim to learn the structure from the data itself, often adapting to the complexity of the data.

الملخص:

- The data tell you what the fit method look like.

- they have a parameters but we don't know how many of them, the data will tell the model how many of them.

Ex. k-nearest neighbors (KNN), decision trees, random forests, and *support vector machines with non-linear kernels.*

2) we need re-estimate probabilities (smoothing) in Naive Bayes classifier.

• to prevent zero probabilities and improve the model's Accuracy by avoiding situations where a feature in the test data was not observed in the training data, leading

3) What is the similarity and difference between feature selection and dimensionality reduction?

Feature selection involves choosing a subset of relevant features from the original feature set, while dimensionality reduction aims to transform the data into a lower-dimensional space, preserving essential information by combining or projecting the original features.

4) True/ false, a single perceptron can only compute linear variation AND, OR, and XOR? Explain in one sentence

False. A single perceptron can only compute linearly separable functions like AND and OR, but not non-linearly separable functions like XOR.

5) When it is possible to run Gradient Descent algorithm, what is guaranteed by the algorithm (1 sentence)? And what isn't guaranteed by the algorithm (1 sentence)?

Guaranteed by Gradient Descent:

The Gradient Descent algorithm guarantees convergence to a local minimum of the cost function when applied to convex optimization problems.

Not Guaranteed by Gradient Descent:

Gradient Descent does not guarantee convergence to the global minimum in the case of non-convex optimization problems, as the algorithm can get stuck in local minima.

6)

7) In the K-nearest neighbor classifier, which of the following statement(s) are true?

a) A KNN is supervised classifier	(T)
b) The hyper parameter K in KNN is typically set to an odd number	(T)
c) When K is set to an extremely large number, it is more likely that the classifier will overfit than underfit.	(F)
d) Both KNN and K-means are unsupervised learning techniques	(F)
e) Increase k in a k-nearest neighbor classifier increase bias	(T)

if the value of k is too high, then it can underfit the data

if k value is too small then it can overfit the data

The k, represent the number of closest neighbors that you are comparing, right? So, no matter if you have 2 or n classes, if you choose an even k, there is a risk of a tie in the decision of which class you should set a new instance. This is why the k is usually odd

7) 9) For the methods below, indicate whether the method is parametric or nonparametric using a "P" for parametric and a "N" for nonparametric:

Method	P/N
Linear Regression	parametric
k-Nearest Neighbor	non-parametric
Support Vector Machines	
Multivariate Linear Regression	parametric
Logistic Regression	parametric
Perceptron	parametric
Multilayer Feed-Forward Neural Network	parametric
K-Means	non-parametric

Linear SVM = parametric
Non-Linear SVM = non-parametric

Question 5: Consider the neural network architecture shown above for a 2-class (0, 1) classification problem. The values for weights and biases are shown in the figure. We define

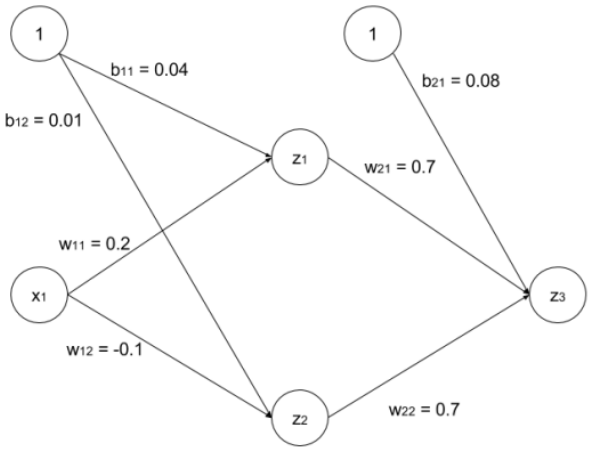
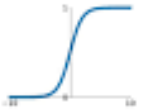


Figure 1: neural network

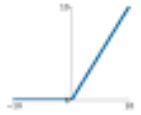
$$a_1 = w_{11}x_1 + b_{11}$$
$$a_2 = w_{12}x_1 + b_{12}$$
$$a_3 = w_{21}z_1 + w_{22}z_2 + b_{21}$$
$$z_1 = \text{relu}(a_1)$$
$$z_2 = \text{relu}(a_2)$$
$$z_3 = \sigma(a_3),\ \sigma(x) = \frac{1}{1+e^{-x}}$$

Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$

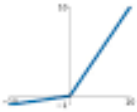


tanh
$\tanh(x)$

ReLU
$\max(0, x)$

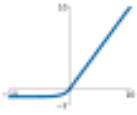


Leaky ReLU
$\max(0.1x, x)$



Maxout
$\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU
$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$



(1) for x1 = 0.3, compute z3 , in terms of e. Show all work.

$$Z_1 = X_1 W_{11} + b_{11} = 0.3 * 0.2 + 0.04 = 0.1$$
$$Z_1 = \text{relu}(0.1) = \max(0, 0.1) = 0.1$$

$$Z_2 = X_1 W_{12} + b_{12} = 0.3 * -0.1 + 0.01 = -0.02$$
$$Z_2 = \text{relu}(-0.02) = \max(0, -0.02) = 0$$

$$Z_3 = Z_1 * W_{21} + Z2 * W_{22} + b_{21}$$
$$= 0.1 * 0.7 + 0.08 = 0.15$$

$$Z_3 = \frac{1}{1 + e^{0.15}}$$

(ii) [**2 pts**] To which class does the network predict the given data point ($x_1 = 0.3$), i.e., $\hat{y} = ?$ Note that $\hat{y} = 1$ if $z_3 > \frac{1}{2}$, else $\hat{y} = 0$.

Circle one: 0 1

$\hat{y}(x_1 = 0.3) = 1$

Question 6: CNN

Question 7:

الصراحه معرفتش اشرح الإجابات في ملف word فعذرا بقي 😂

Solution: (d).Single-linkage combines clusters with the smallest pairwise distance. Complete-linkage combines clusters with the smallest (across cluster pairs) of the largest pairwise distance (across word pairs between two clusters).