Practice Questions for the Final Exam of ELEC292

Provided by: Mahzabeen Emu & Mojtaba Kolahdouzi

1. Given the 2 tables below, create the output of the following query.

students table		
student_id	**first_name**	**last_name**
1	Trevor	Mcdonald
2	Ali	Gerretsen
3	James	Halpert
4	Pam	Beesly

majors table	
student_id	**major_code**
1	101
1	120
2	5080
5	002

SELECT students.first_name, students.last_name, majors.major_code
FROM students
RIGHT JOIN majors
ON students.student_id = majors.student_id;

Answer:

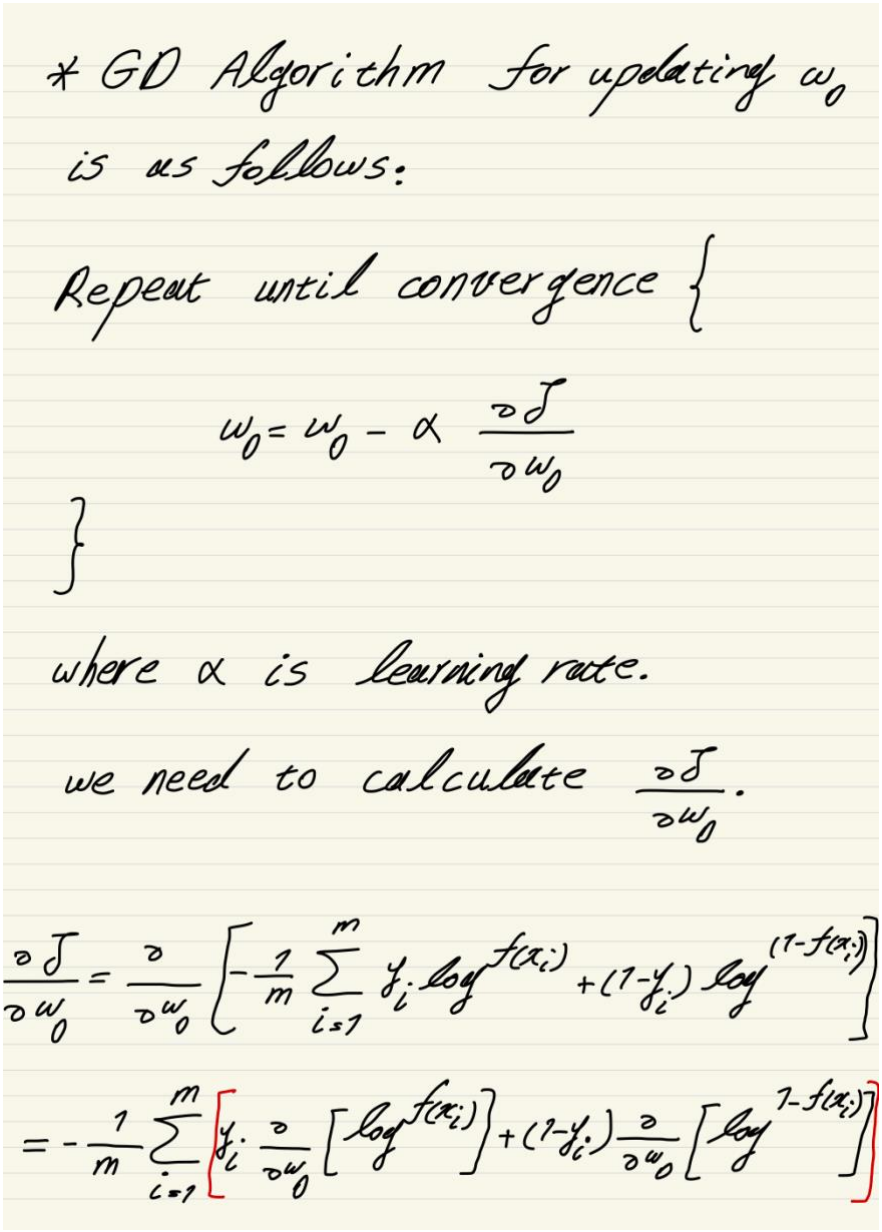
first_name	last_name	major_code
Trevor	Mcdonald	101
Trevor	Mcdonald	120
Ali	Gerretsen	5080
NULL	NULL	002

2. Given the following logistic regression with 1-D input $x_i \in R$, implement the gradient descent algorithm for updating w_0.

$$f(x_i) = \frac{1}{1 + e^{-(w_0 + w_1 x_i)}}$$

$$J(w_0, w_1) = \frac{-1}{m} \sum_{i=0}^{m} [\, y_i log^{f(x_i)} + (1 - y_i)\, log^{(1-f(x_i))}]$$

Answer:



Thus, we need to calculate 2 terms:

$$\frac{\partial}{\partial w_0}\left[\log^{f(x_i)}\right] \quad , \quad \frac{\partial}{\partial w_0}\left[\log^{1-f(x_i)}\right]$$

$$\frac{\partial}{\partial w_0}\left[\log^{\overset{u}{f(x_i)}}\right] \overset{\text{chain rule}}{=\!=\!=} \frac{\partial}{\partial w_0}\left(\log^{u}\right) =$$

$$\frac{\partial}{\partial u}\log^{u} \cdot \frac{\partial u}{\partial w_0} = \frac{1}{u} \cdot \frac{e^{-(w_0+w_1x_i)}}{\left(1+e^{-(w_0+w_1x_i)}\right)^2}$$

$$= \frac{1}{\frac{1}{1+e^{-(w_0+w_1x_i)}}} \cdot \frac{e^{-(w_0+w_1x_i)}}{\left(1+e^{-(w_0+w_1x_i)}\right)^2} =$$

$$\frac{e^{-(w_0+w_1x_i)}}{1+e^{-(w_0+w_1x_i)}}$$

$$\frac{\partial}{\partial w_0}\left[\log\left(1-f(x_i)\right)\right] = \frac{\partial}{\partial u}\log^{1-u}\cdot\frac{\partial u}{\partial w_0}$$

$$= \frac{-1}{1-u}\cdot\frac{e^{-(w_0+w_1x_i)}}{\left(1+e^{-(w_0+w_1x_i)}\right)^2} =$$

$$\frac{-\left(1+e^{-(w_0+w_1x_i)}\right)}{e^{-(w_0+w_1x_i)}}\cdot\frac{e^{-(w_0+w_1x_i)}}{\left(1+e^{-(w_0+w_1x_i)}\right)^2} =$$

$$\frac{-1}{1+e^{-(w_0+w_1x_i)}}$$

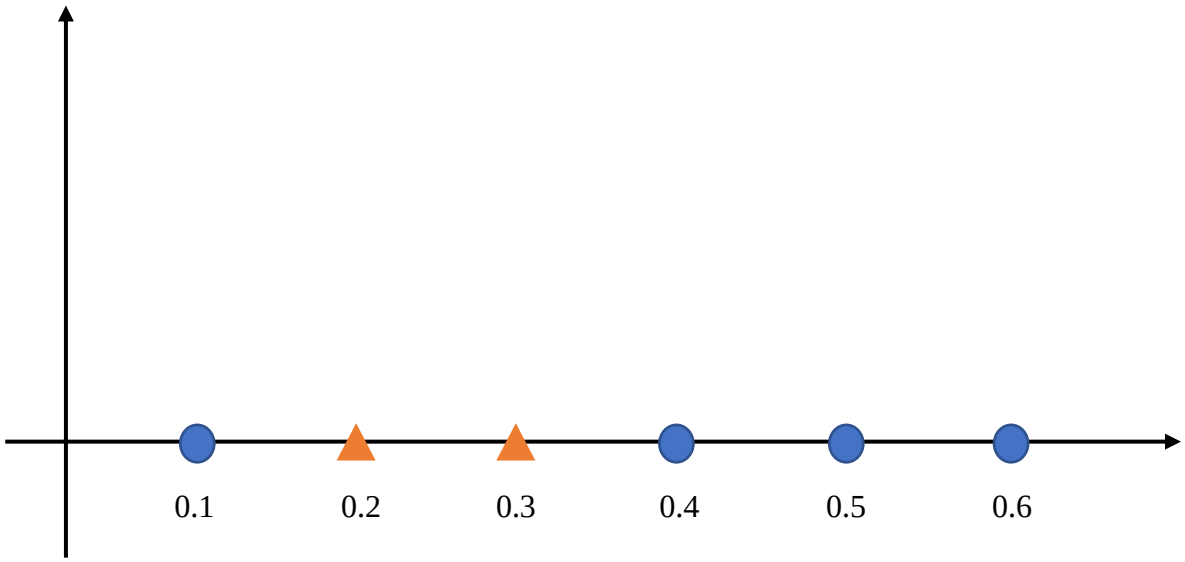
final eq:

$$w_0 = w_0 - \alpha\left[\frac{-1}{m}\sum_{i=1}^{m}\left[\frac{y_i\, e^{-(w_0+w_1x_i)}}{1+e^{-(w_0+w_1x_i)}} + \frac{y_i-1}{1+e^{-(w_0+w_1x_i)}}\right]\right]$$

No need for further simplification!

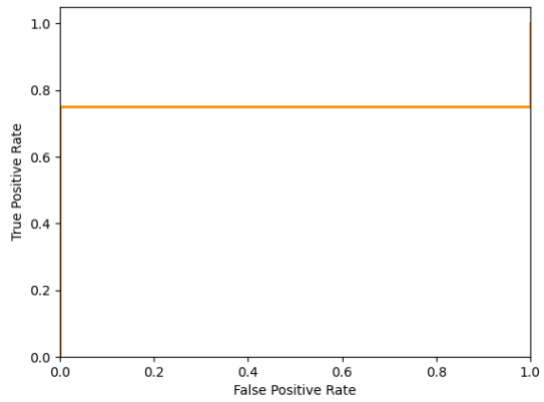
What if we change the cross-entropy loss to mean squared error?

3. Given the data below, calculate true positive rate (tpr) and false positive rate (fpr) at different thresholds. Draw the resulting receiver operating characteristic (ROC) curve and calculate the area under curve (AUC). (labels: **1**: ● **0**: ▲)



ANSWER:

Threshold	tpr	fpr
inf	0	0
0.6	1/4	0
0.5	2/4	0
0.4	3/4	0
0.3	3/4	1/2
0.2	3/4	1
0.1	1	1



AUC is 0.75.

4. Suppose that you are assigned a task to apply principal component analysis (PCA) on a 2d dataset for dimension reduction. You do the calculations and find out that the first principal component is positioned at an angle of 5 degrees contraclockwise from the x axis. What can you infer about the variance of the dataset along x and y axes?

ANSWER:

Since the first principal component is aligned only 5 degrees from x axis, it provides insight into the variance of the dataset. The first principal component is the direction of maximum variance in the dataset. Since it's only slightly tilted from the x-axis, this suggests that a significant portion of the variance in the data occurs along the x-axis, or close to it, with less variance along the y-axis.

5. Imagine that you're spearheading an initiative to develop an augmented reality (AR) application that uses a smartphone's camera to identify and provide educational content about various animal species encountered in nature. The goal is to make zoology engaging and accessible for both students and wildlife enthusiasts, offering instant information about animals through image recognition. Your team decides to compile and utilize a comprehensive dataset consisting of both professionally captured images in zoos and wildlife reserves and user-contributed photographs from their natural encounters. Given the assortment of data sources, including user submissions, your team opts to employ a hybrid approach for data annotation and validation. Considering this scenario, answer the following questions:

a) How would your team execute a hybrid/mixed approach for data labelling?

b) What challenges could emerge from integrating professional images with user-contributed wildlife photos, and how would you tackle these challenges to ensure the AR application's success?

Answer:

a) To implement a hybrid method for data labeling and checking, the team would use automated labelling tools to quickly sort and tag animal images, then let wildlife experts manually review and correct any unclear or mistaken cases. This ensures the information is accurate, even for rare animals. Also, by letting app users report errors or confirm data, the team can continuously improve the app's accuracy with help from its community.

b) The wildlife identification app combines professional and user-taken photos, facing challenges like background noise and poor lighting that can lower recognition accuracy. To improve this, the team can clean up photos with advanced preprocessing techniques to remove noise and identify animals, no matter the photo quality.

6. As a data scientist working for a nutrition analysis startup, you are tasked with developing a machine learning model to predict the likelihood of vitamin deficiencies in individuals based on their dietary intake and blood test results. Among the features in your dataset are Vitamin D level, measured in nanograms per milliliter (ng/mL) with values typically ranging from 20-50 ng/mL, and Daily Caloric Intake, measured in calories, with a typical range from 1,200 to 3,500 calories per day. Considering the substantial difference in scale between these measurements, what preprocessing step would you deem essential to perform on these features before using them in your predictive model, and why?

Answer:
Given the significant disparity in scale between Vitamin D levels and Daily Caloric Intake, it is essential to apply normalization as a preprocessing step before incorporating these variables into the machine learning model. Normalization ensures that each feature contributes proportionately to the model's prediction, preventing features with larger numerical ranges from dominating those with smaller ranges. Two such common scaling methods are normalization (rescaling the data to a [0, 1] range) and standardization (rescaling the data to have a mean of 0 and a standard deviation of 1). By scaling the Vitamin D levels and Daily Caloric Intake to a similar scale, the predictive model can more effectively learn patterns and relationships between the features and the likelihood of vitamin deficiencies.

7. The TechInspect team is developing a system for diagnosing problems in industrial machinery by analyzing acoustic signals. They encounter a challenge with low-frequency background noise interfering with their signal data. Which type of filter (High-pass filter/Low-pass filter) would be most effective for isolating the relevant frequencies indicative of machinery faults? Explain your reasoning.

Answer:
To address the issue of low-frequency background noise in the acoustic signals used for diagnosing industrial machinery problems, employing a High-pass filter would be the most effective strategy. A High-pass filter allows frequencies higher than its cutoff frequency to pass through while attenuating frequencies lower than the cutoff frequency. Given that the noise interfering with the signal data is of low frequency, the High-pass filter would effectively remove or reduce this unwanted noise, thereby allowing the system to focus on the higher frequency components that are more indicative of machinery faults.