Quiz for Summer Analytics Week 3

Total points 37/43

Hope that you've gone through the course content for week-3.

- This form accepts the solution only once, so make sure you don't press the submit button accidentally. No requests will be entertained.
- **Use the SAME email ID which you used for registering for Summer Analytics 2025**.
- Please follow the honor code, which otherwise may lead to harsh actions being taken.

All the best :)

0 of 0 points

Name *

Sakshi Rastogi

Email ID *

sakshi97182@gmail.com

Are you from IIT Guwahati? *

- ○ Yes
- ◉ No

If you are from IIT Guwahati, provide your roll number

Did you explore and star the repositories by our friends at Pathway (needed for the course and final hackathon) Star 🌟 so that it's easier for you to navigate as and when required! Pathway Main Repo with all Updates: https://github.com/pathwaycom/pathway Pathway LLM App Templates: https://github.com/pathwaycom/llm-app *

- (•) Yes
- ( ) No

Quiz Starts from here. 37 of 43 points

✓ On the same dataset, compare three regressors: 3/3

1. Without any regularization , 2. Ridge with very large λ, 3. Lasso with moderate λ.
Which ordering is correct for variance (highest → lowest)?

- ( ) A. 2 → 1 → 3
- (•) B. 1 → 3 → 2 ✓
- ( ) C. 3 → 1 → 2
- ( ) D. 1 → 2 → 3

✓ You plot train- and validation-MSE vs. λ for Ridge. Both curves start high at λ≈0, validation dips then rises. Which region indicates underfitting, and which overfitting? 2/2

- ( ) A. Underfitting at low λ; overfitting at high λ
- (•) B. Underfitting at high λ; overfitting at low λ ✓
- ( ) C. Both under- and overfitting at low λ
- ( ) D. Neither; this shape is inconclusive

Given a classifier with TP, FP, TN, FN, which metric remains unchanged if you swap the positive and negative labels? 2/2

- A. Precision
- B. Recall (Sensitivity)
- C. Accuracy
- D. Specificity

As you raise the decision threshold t for calling "positive": 2/2

- A. Precision ↑, Recall ↓
- B. Precision ↓, Recall ↑
- C. Both Precision & Recall ↑
- D. Both Precision & Recall ↓

A binary classifier outputs prediction probabilities for class 1 uniformly distributed on [0, 2/3] and class 0 uniformly distributed on [1/3, 1]. What is the AUC-ROC value for this classifier? 0/3

- A) 0.5 (random classifier performance)
- B) 0.75 (good discriminative ability)
- C) 0.875 (excellent performance)
- D) Cannot be determined without threshold information

Correct answer

- C) 0.875 (excellent performance)

Consider a regularized logistic regression model for medical diagnosis where the cost function is: $J(\theta) = -\sum[y^{(i)}\log(h\theta(x^{(i)})) + (1-y^{(i)})\log(1-h\theta(x^{(i)}))] + \lambda\sum\theta_j^2$. If increasing λ from 0.01 to 1.0 changes the decision boundary from highly curved to nearly linear, this indicates:

2/2

- A) The model is transitioning from overfitting to optimal fit
- B) L2 regularization is forcing the model toward higher bias
- C) The features are becoming less correlated with the target
- D) The regularization is improving feature selection

Consider two regularized logistic regression models for email spam detection: Model 1 uses L1 with $\lambda_1=0.1$, Model 2 uses L2 with $\lambda_2=0.01$. Both achieve similar validation accuracy. In a production environment with limited computational resources, which model characteristic would favor Model 1?

3/3

- A) Better handling of correlated email features
- B) Lower memory requirements due to sparse coefficient vector
- C) More robust predictions for new email types
- D) Higher interpretability of spam indicators

✓ The Minkowski distance metric is defined as: $d(x,y) = (\Sigma_i|x_i-y_i|^p)^\wedge(1/p)$. What is the mathematical relationship between Minkowski distance and other common distance metrics? 2/2

- A) p=1 gives Euclidean, p=2 gives Manhattan
- B) p=2 gives Euclidean, p=1 gives Manhattan ✓
- C) p=∞ gives Euclidean, p=1 gives Chebyshev
- D) p=0 gives Manhattan, p=1 gives Euclidean

✓ In multinomial Naive Bayes with Laplace smoothing, the probability P(word|class) is calculated as: $P(w_i|c) = (\text{count}(w_i,c) + \alpha)/(\text{count}(c) + \alpha\times|V|)$. If a vocabulary has 1000 words, $\alpha=1$, and class C has 500 word occurrences with word "excellent" appearing 5 times, what is P("excellent"|C)? 5/5

- A) 5/500
- B) 6/501
- C) 6/1500 ✓
- D) 5/1000

✓ In Gaussian Naive Bayes, each feature follows a normal distribution: $P(x_i|c) = (1/\sqrt{(2\pi\sigma^2c)})\exp(-(x_i-\mu c)^2/2\sigma^2c)$. If feature values for class C have $\mu c=10$, $\sigma^2c=4$, what is the relative likelihood of observing $x_i=12$ versus $x_i=8$? 4/4

- A) They have equal likelihood (symmetric around mean) ✓
- B) $x_i=12$ is twice as likely as $x_i=8$
- C) $x_i=8$ is twice as likely as $x_i=12$
- D) The ratio depends on other features

A medical diagnostic system uses Gaussian Naive Bayes to classify diseases from continuous biomarker measurements. The ROC curve shows AUC=0.92, but the confusion matrix reveals 15% false negative rate for a critical disease. From a mathematical perspective, what does this suggest about the optimal threshold selection?

3/3

- A) The current threshold maximizes overall accuracy
- B) The threshold should be lowered to increase sensitivity ✓
- C) The AUC value is inconsistent with the confusion matrix
- D) The model suffers from severe class imbalance

Spotify's music recommendation system processes 70 million songs with 13 audio features (danceability, energy, speechiness, acousticness, etc.). If they use weighted KNN with inverse distance weighting w(d) = 1/d, what happens mathematically when two songs have identical feature vectors (d = 0)?

3/3

- A) The weight becomes undefined, requiring regularization to w(d) = 1/(d + ε) ✓
- B) The algorithm automatically excludes the duplicate song
- C) The weight is set to the maximum possible value in the system
- D) Standard KNN voting is used instead of weighted voting

Amazon's "Customers who bought this item also bought" feature uses item-based collaborative filtering with KNN. For a specific smartphone case, the k=5 nearest products in the recommendation space are: [wireless charger: 0.85, screen protector: 0.82, car mount: 0.78, headphones: 0.71, tablet case: 0.65] where numbers represent cosine similarity scores. If Amazon uses weighted voting with similarity-based weights, what is the relative influence of the wireless charger compared to the tablet case?

3/3

- (x) A) 1.31 times more influential (0.85/0.65)
- ( ) B) 0.20 times more influential (difference of 0.20)
- ( ) C) 1.31 times more influential, but only if k > 3
- ( ) D) Equal influence since both are in the k-nearest neighbors

The linear regression pipeline in the Pathway template computes aggregates like sum_x, sum_y, sum_x_y, and sum_x_square. Suppose the incoming Kafka stream has missing or malformed rows like "x, " or " ,5". What would be the best way to handle these errors in a streaming setting while ensuring your regression continues producing meaningful results?

3/3

- ( ) Ignore all rows with missing values silently
- ( ) Drop the stream and restart from the latest offset
- (x) Add preprocessing to validate input, and log or route errors to a dead-letter topic
- ( ) Fill missing values with the median of the seen data so far

You're building a real-time system that continuously estimates the slope (a) and intercept (b) of a linear relationship between variables x and y using streaming summary statistics: …/3

count: number of observations

sum_x, sum_y: sums of x and y values

sum_x_y: sum of x * y

sum_x_square: sum of x^2

Write a robust Python function that computes a and b using only these aggregate values. Your function should **safely handle all edge cases.** The return should be a tuple (a, b) or (None, None) if the result is undefined.

```
def linear_regression_params(count, sum_x, sum_y, sum_x_y, sum_x_square):
    if count <= 1:
        return (None, None)

    denominator = (count * sum_x_square) - (sum_x ** 2)
    if denominator == 0:
        return (None, None)

    a = ((count * sum_x_y) - (sum_x * sum_y)) / denominator
    b = (sum_y - a * sum_x) / count

    return (a, b)
```