

**Subject:**  
**Instructor:**  
**Place:**  
**Data & Time:**

Statistical Inference EE2102575  
Suwichaya Suwanwimolkul  
Eng. Building 3, 205  
21/02/2024 Wed. 11:00 – 12:30

---

### Exam Instructions

Total 75 marks to be collected as Homework3 ( $\approx 10\%$ ) of the scores.

1. **Time Limit:** you have **1.5 hours** to complete this examination.
2. **Materials:** Open books.
3. **Electronic Devices:** switch off and stow away all electronic devices, including mobile phones, smartwatches, and any other electronic gadgets, and **leave all of them at the provided locations**.
4. **Seating:** Maintain a one-seat gap between you and other candidates. Do not communicate or share any materials during the exam.
5. **Answering format:** use **a pen with blue or black ink** for the exam. Don't forget to fill in **your name and student ID** at the top of the pages.
6. **Instructions:** read all questions carefully!!! Ensure you understand the directions and requirements for each section.
7. **Early Submission:** if you finish early, quietly leave the examination hall, ensuring not to disturb others.

"I hereby acknowledge that my signature constitutes my understanding and agreement to comply with the conditions stipulated above."

Signature \_\_\_\_\_

Name \_\_\_\_\_

Student ID \_\_\_\_\_

Section	1	2	3.1	3.2
Your Scores				
Total Scores	5	30	30	10

## 1 Quizzes (5 scores)

### 1.1 True/False Questions (5 scores)

For each statement, answer True or False. (Put one "X" in each.)	True	False
a) KNN is a supervised leawrning.		
b) The parameters of LDA is estimated from the mean and covariance of training data		
c) LDA assumes that the equal covariance between $X Y_j$ for all $j \in \{1, 2, \dots, K\}$		
d) Unlike LDA, QDA did not assume the equal covariance		
e) Unlike KNN, LDA and QDA estimate a set of parameters to draw the boundaries which depend on the data distributions. Then, estimates the posterior probability of the classes using the estimated parameters.		

## 2 Logistic Regression: loan amount dataset (30 scores)

Data.

- **Input features:** Loan\_ID, Gender, Married, Dependents, Education, Self\_Employed

ApplicantIncome, CoapplicantIncome, LoanAmount, Loan\_Amount\_Term, Credit\_History, and  
Property\_Area

- **Target:** Loan\_status

### 2.1 Logistic regressoin (intro)

In this section, let's analyze the results from logistic regression. The logistic regressor will try to give you the answer of whether or not you you give the loan to a person, based on his/her information ?

Here, we ran the logistic regressor given **2 different sets of input features**.

- **Input features:** LoanAmount

```
Optimization terminated successfully.
Current function value: 0.607423
Iterations 4
```

```

                        Logit Regression Results
=====
Dep. Variable:          y      No. Observations:          246
Model:                  Logit  Df Residuals:              245
Method:                  MLE   Df Model:                0
Date:                   Sun, 18 Feb 2024  Pseudo R-squ.:      0.001087
Time:                   21:59:14          Log-Likelihood:     -149.43
converged:               True    LL-Null:             -149.59
Covariance Type:         nonrobust  LLR p-value:         nan
=====
               coef      std err          z      P>|z|      [0.025      0.975]
-----
x1              0.0081      0.001       6.142     0.000      0.005      0.011
=====
```

- **Input features:** LoanAmount, Credit\_History, Gender, Education

Optimization terminated successfully.  
Current function value: 0.476945  
Iterations 6

```

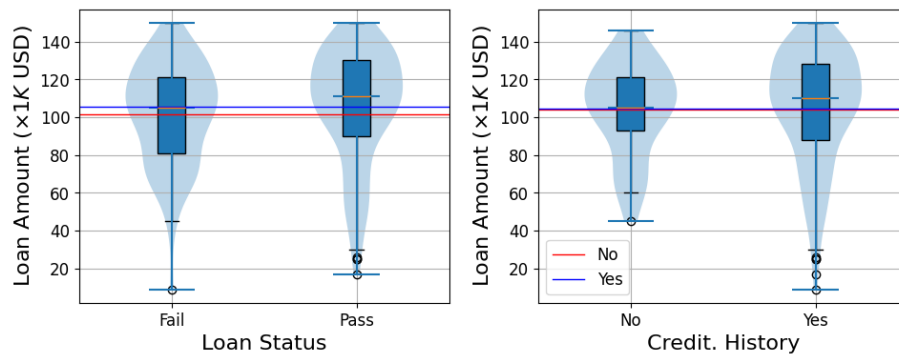
Logit Regression Results
=====
Dep. Variable:          y      No. Observations:      246
Model:                Logit   Df Residuals:        242
Method:               MLE     Df Model:           3
Date:                Sun, 18 Feb 2024   Pseudo R-squ.:      0.2157
Time:                21:59:14   Log-Likelihood:     -117.33
Converged:            True     LL-Null:            -149.59
Covariance Type:      nonrobust   LLR p-value:        6.353e-14
=====
               coef    std err          z      P>|z|     [0.025     0.975]
-----
x1            -0.0122     0.004     -2.966     0.003     -0.020     -0.004
x2             2.9862     0.452      6.603     0.000      2.100      3.873
x3            -0.1059     0.397     -0.267     0.790     -0.884     0.672
x4            -0.5153     0.361     -1.426     0.154     -1.224     0.193
=====

```

## 2.2 True/False Questions (10 scores)

You should use the given codes to recheck your answer.

For each statement, answer True or False. (Put one "X" in each.)	True	False
a) LoanAmount alone has a positive relationship with 'Loan_status == True'		
b) Credit_History alone has a positive relationship with 'Loan_status == True'		
c) At a Credit_History, the LoanAmount alone has a negative relationship with 'Loan_status == True'		
d) At a fixed values of Credit_History, Gender, Education, the LoanAmount alone has a negative relationship with 'Loan_status == True'		
e) P-value indicates that the coefficients of Gender and Education are likely to have zero values.		
f) From the P-values of Credit_History, the parameter coefficient associated with Credit_History is likely to be non-zero, so it could give a crucial information.		
g) The correlation between LoanAmount and Credit_History can cause the confusing conclusion between LoanAmount and Loan_status.		
h) The correlation between LoanAmount and Education can cause the confusing conclusion between LoanAmount and Loan_status.		
i) Credit_History alone has a negative correlation with LoanAmount		
j) Loan_amount_term is one of the least correlated features with Loan_status		



### 2.3 Questions (20 scores)

**Questions, 2.3.1:** Compare the relationship between `Loan_Status` and `LoanAmount` ... versus the relationship between `Credit_History` and `LoanAmount` ? Are they somewhat in the opposite direction? You can explain this by using the minimum (Q1), maximum (Q3), and the median values of the loan amount in each case.

---



---



---

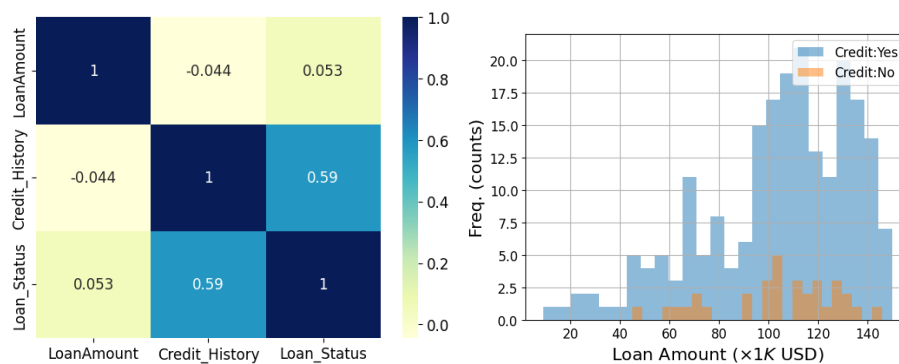


---



---

**Questions, 2.3.2:** Explain why `Credit_History` might have a negative relationship with `LoanAmount` ?




---



---



---



---



---

**Questions, 2.3.3:** Do you think are there any other feature correlation that can cause the problem similar to Credit\_History and LoanAmount ? If there is,

- confirm what are these features ?
- explain their relationship (you can provide the box plots/histogram similar to what I have done).
- please draw the correlation matrix as the evidence.



---

---

---

---


### 3 KNN, LDA, and QDA Classifiers (40 scores)

#### 3.1 Two features (30 scores)

- Use **two features** Credit\_History LoanAmount for the classification of Loan\_status.
- Find the precision-recall trade-off of each k of the KNN, LDA and QDA on the validated dataset.

**Two features, 3.1.1:** Based on the precision-recall trade-off, which value of  $k$  of the KNN algorithm gives the best trade-off (you can also use other evaluations, *e.g.*, specificity to help)...(5 scores)

- confirm by plotting the graph and explain.



---

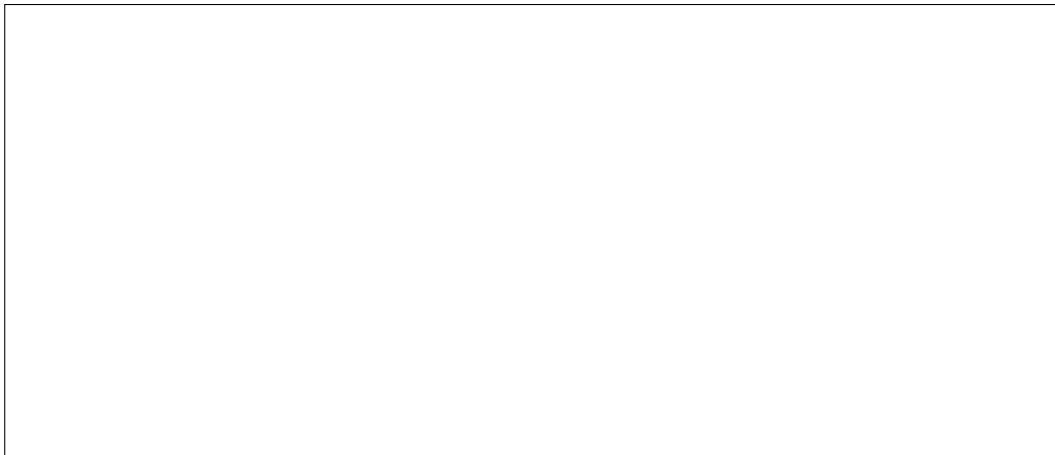
---

---

---

**Two features, 3.1.2:** Compare the results of LDA and QDA with the previous precision-recall trade-off (you can also use other evaluations, *e.g.*, specificity to help)...(5 scores)

- Between LDA, QDA, and KNN, which is the best ? confirm by plotting the graph and explain.



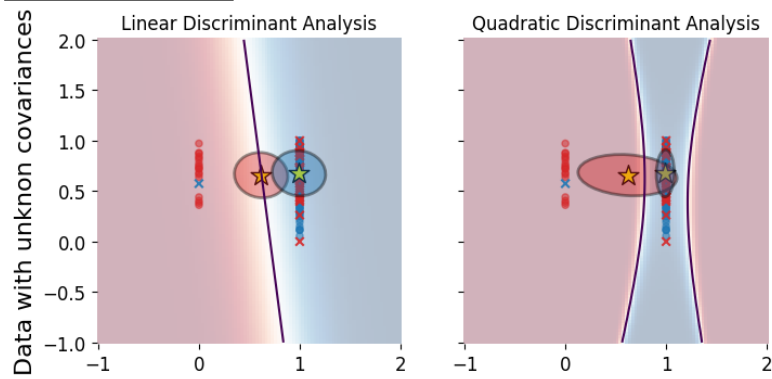
Name

ID

**Two features, 3.1.3:** Compare the confusion matrix of KNN with the best-k, LDA and QDA ...(5 scores)

- Between LDA, QDA, and KNN, which is the best ? confirm by drawing the confusion matrix of KNN, LDA and QDA. Your analysis should be supported by the quantity derived from the confusion matrix, *e.g.*, true positive, false positive, true negative, and false negative.



**Two features, 3.1.4:** Visualize the boundary of LDA and QDA. (5 scores)

- From the above figure can you explain how LDA and QDA decide the boundary.
- Which of the two features that LDA and QDA rely on the most ?

---

---

---

---

---

---

---

---

**Two features, 3.1.5:** Suggestions from the bounary. Which of the two features that LDA and QDA rely on the most when deciding the boundary? (10 scores)

- Check the performance of LDA and QDA when using **only** Credit\_History / LoanAmount .
- Please compare the performance using the confusion matrix.



Name	ID

### 3.2 7 features (10 scores)

- **7 features:** Gender, Married, LoanAmount, Dependents, Self\_Employed, ApplicantIncome, CoapplicantIncome, Credit\_History for the classification of Loan\_status.

7 features, 3.2.1: On validation set:

- Between LDA, QDA, and KNN, which is the best ? confirm by plotting the graph and explain.
- Do you get different best value of k for KNN compared to when using only two features?

Name

ID

**7 features, 3.2.2:** which is the best ? The confusion matrix of KNN with the best-k vs LDA vs QDA...

- Do you get a different testing performance from using two features? Why do you think?