Chennai Mathematical Institute

Predictive Analytics: Regression and Classification

Final Exam | 2024-11-23

Total Marks: 50
Total Time: 2 hours

Instructions:
1. Attempt all questions.
2. Write clearly and provide detailed reasoning for all answers.
3. For application questions, interpret the provided R output carefully and justify your conclusions.

Question 1: (10 points)

a) (3 points) Explain the advantages and limitations of the logit link function in logistic regression. Provide examples where this function is particularly useful.

b) (3 points) Derive the relationship between the log-likelihood function and the cross-entropy loss for a binary classification problem.

c) (4 points) Define the concept of basis expansion in regression. Explain its role in addressing non-linearity in data.

Question 2: (10 points)

a) (5 points) Explain the assumptions of linear discriminant analysis (LDA) and discuss why these assumptions may or may not hold in real-world datasets.

b) (5 points) Describe the Gaussian process regression model and explain its limitations when applied to high-dimensional data.

Question 3: (10 points)

a) (3 points) Explain how regularisation techniques (e.g., L1 and L2) influence the performance of regression models.

b) (3 points) Briefly discuss the differences between logistic regression and neural networks in terms of their mathematical formulation and practical application.

c) (4 points) Prove or disprove: The solution to ridge regression minimises the sum of squared errors plus a penalty term on the norm of the coefficients.

Question 4: (10 points)

An auto manufacturer is analysing factors that influence whether customers purchase an extended warranty. The logistic regression model output in R is provided below:

```
Call:
glm(formula = Purchase ~ Age + Income + Gender, family = binomial, data = auto_data)

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -2.456     0.345       -7.115   < 0.001 ***
Age          0.058     0.012        4.833   < 0.001 ***
Income       0.005     0.001        5.000   < 0.001 ***
GenderMale  -0.742     0.215       -3.451   < 0.001 ***
---
Signif. codes:  0 '***' 0.001
```

Answer the following:

a) (3 points) Interpret the coefficients for `Age`, `Income`, and `GenderMale`.

b) (3 points) If a 35-year-old male with an income of $50,000 is evaluated, what is the log-odds of purchasing an extended warranty?

c) (4 points) Calculate the probability of purchase for this individual and explain its implications for the company's marketing strategy.

Question 5: (10 points)

A defense manufacturer is modeling the relationship between production costs (in millions) and factors such as the number of units produced, the complexity of design (measured on a scale of 1 to 10), and the experience of the workforce (in years). The regression output in R is provided below:

```
Call:
lm(formula = Cost ~ Units + Complexity + Experience, data = defense_data)

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) 15.236     2.045        7.448   < 0.001 ***
Units        0.053     0.012        4.417   < 0.001 ***
Complexity   1.672     0.320        5.225   < 0.001 ***
Experience  -0.428     0.134       -3.194   0.002 **
---
Residual standard error: 1.45 on 47 degrees of freedom
Multiple R-squared: 0.785, Adjusted R-squared: 0.765
F-statistic: 39.11 on 3 and 47 DF,  p-value: < 0.001
```

Answer the following:

a) (3 points) Interpret the coefficients for `Units`, `Complexity`, and `Experience`.

b) (3 points) Predict the production cost for a project involving 500 units, complexity of 8, and a workforce with 10 years of experience. c) (2 points) Provide 95% predictive interval for your prediction for the same project. d) (2 points) Discuss the implications of the adjusted R-squared value and the significance of the predictors for decision-making.

End of Exam