Mock Exam Questions - Module 2

Medium Answer

1. When developing a predictive model for retail sales using hundreds of store and market indicators, what are the fundamental distinctions between applying feature selection techniques and using feature extraction methods?

2. Given the results of a diagnostic test on 150 lung disease patients where the test predicted 60 as positive and 90 as negative, and the actual condition shows 50 truly positive cases and 100 truly negative cases (with the model making 45 true positives and 15 false positives), construct the confusion matrix and then compute the test's precision and recall.

3. A sentiment analysis model labels customer reviews as either Positive ('Pos') or Negative ('Neg'). Below is a list of 12 reviews with their actual sentiments and the model's predictions:

Review	**Actual**	**Predicted**
1	Neg	Neg
2	Neg	Neg
3	Neg	Pos
4	Pos	Neg
5	Pos	Pos
6	Pos	Pos
7	Neg	Pos
8	Pos	Pos
9	Neg	Neg
10	Pos	Pos
11	Pos	Neg
12	Neg	Neg

   Using the Positive class as the class of interest, calculate:

   (i) Precision

   (ii) Recall

   (iii) Number of True Positives

   (iv) F1 score

4. Explain how you would use the bootstrap resampling technique to estimate the 95% confidence interval for the median height of trees based on a field sample of 40 measurements. Under what circumstances is the bootstrap method the preferred approach?

5. In a credit card fraud detection scenario, explain the importance of the ROC curve for evaluating model performance. How does comparing the ROC curves of two different algorithms help you understand their relative trade-offs between detecting true fraud and raising false alarms?

6. Suppose we build a system to detect red cars in street surveillance footage that also contains blue cars. The system flags 8 vehicles as red cars in a scene where there are actually 12 red cars among other blue cars. Of the 8 flagged cars, 5 are truly red cars while the other 3 are actually blue cars. What are the precision and recall of this system?

7. A scholarly database contains 120 papers on renewable energy. A search query returns 90 papers, of which 60 are pertinent to the topic. Calculate the precision and recall of the search.

8. In evaluating a fraud detection model, you test it on a dataset containing 250 fraudulent transactions and 750 legitimate ones. The ROC curve for this model passes through the point where TPR = FPR = 0.1. What is the precision at this operating point?

9. A spam detection model is tested on a dataset of 1,200 emails, of which 30% are spam (positive class) and 70% are legitimate (negative class). The classifier achieves a sensitivity (recall) of 75% and an overall accuracy of 70%.

   (a) Construct the full confusion matrix for this evaluation.

   (b) From your confusion matrix, calculate the classifier's precision, F$_1$-measure, and specificity.

10. Consider a binary classification dataset where there are 200 positive (rare) examples and 10,000 negative (common) examples. Model A predicts 250 examples as positive, of which 180 are true positives and 70 are false positives. Model B predicts 100 examples as positive, of which 85 are true positives and 15 are false positives. Compare the precision of Model A versus Model B.

11. In a supervised learning regression task, define empirical risk, true risk, bias, and variance. Then discuss what occurs when a model is too simplistic or too complex (underfitting versus overfitting), and explain how the bias–variance trade-off governs this behavior.

12. In predictive modeling, what does a cost function $C(\text{actual}, \text{predicted})$ quantify? Provide one example of such a cost function used for predicting real-valued outcomes (e.g., housing prices) and another example used for classifying data into discrete categories (e.g., spam vs. not spam).

13. In evaluating a medical diagnostic algorithm for predicting heart disease risk, explain how you would implement a 5-fold cross validation procedure and why it is used.

14. In model selection, what is the Bayesian Information Criterion (BIC)? Explain how you would apply BIC to choose the best regression model when comparing several models with different numbers of predictors. As the sample size grows without bound, which model complexity does BIC asymptotically favor?

15. Draw a graph illustrating how the training error and the validation error change as you increase the degree of a polynomial in a regression problem (i.e., model flexibility). Based on this graph, explain how you would pick the best polynomial degree.

16. For a regression model f_θ and a training set of n pairs (x_i, y_i), define the model's training error in terms of the squared-loss function $\ell(f_\theta(x_i), y_i)$. How is the overall training error expressed?

Long Answer

1. Explain the purpose of an ROC curve and how to compute the AUC. Imagine you have built a credit-risk classifier that assigns each of six loan applicants a probability of default. The true default labels for these six applicants are $y_{\text{true}} = [0, 1, 0, 1, 0, 1]$, and your model's predicted probabilities are $y_{\text{scores}} = [0.25, 0.85, 0.15, 0.65, 0.45, 0.75]$. Using decision thresholds $T = [0, 0.2, 0.4, 0.6, 0.8, 1.0]$, calculate the false positive rate and true positive rate at each threshold, draw the ROC curve, and compute the AUC.

2. Given a customer churn dataset, write pseudocode for both forward stepwise selection and backward stepwise elimination when choosing features for a predictive model. Include how you evaluate and add or drop features and the stopping criteria.

3. Explain how to compute and interpret the ROC curve and AUC for a binary classifier. Suppose you have a spam detection model whose true labels for six emails are $y = [0, 1, 0, 1, 1, 0]$ and the model's predicted probabilities are $y_{\text{pred}} = [0.12, 0.68, 0.35, 0.82, 0.27, 0.59]$. Using thresholds $[0, 0.25, 0.5, 0.75, 1]$, plot the ROC curve and calculate the AUC.

4. Explain the five-fold cross-validation procedure step by step. Then describe how you would properly integrate a dimensionality reduction step (e.g., principal component analysis) into this five-fold cross-validation workflow to ensure no data leakage before training your model.

5. In the context of assessing a binary classifier for email spam detection, when would you resort to ROC and AUC? Define what ROC and AUC mean. Additionally, explain the measures specificity, sensitivity, precision, and recall. Lastly, why is it useful to plot an ROC curve?

6. Consider the table below listing eight bank transactions with the true label ('Fraud' or 'Legit') and the risk score assigned by a fraud-detection model:

Transaction	**Risk Score**	**True Label**
T1	0.85	Fraud
T2	0.40	Legit
T3	0.60	Fraud
T4	0.20	Legit
T5	0.90	Fraud
T6	0.50	Legit
T7	0.30	Fraud
T8	0.10	Legit

   A transaction is classified as Fraud if its risk score is at least a chosen threshold, and as Legit otherwise. Using threshold values 0.7 and 0.3, compute the accuracy, precision, and recall of this fraud detector at each threshold.

7. In supervised learning, how do you choose the optimal predictive model? Define an error function J that quantifies the difference between the true outcome T and the model's prediction $q(w)$. Then specify the form of J for continuous-valued targets (regression) and for discrete class labels (classification).

8. Explain how you would perform 8-fold cross-validation to evaluate a predictive model. If you apply factor analysis to reduce input dimensions before training, how should you integrate the cross-validation steps to prevent information leakage and ensure a valid performance estimate?

9. You've created a computer vision model for classifying wildlife photos, but its accuracy stagnates. Outline the systematic procedure you would follow to isolate and fix issues in your learning pipeline. Additionally, what rule of thumb would you apply when partitioning your data into training, validation, and test sets for effective debugging? Provide a brief evaluation of your approach.