Introduction to Statistical Learning, Fall 2025

# Homework exercise 3

Due date: 25 December 2025 — submission via Moodle, in groups of 1 or 2

1. **LDA and QDA**

   (a) In the generative 2-class classification models LDA and QDA, what type of distribution does $P(Y|X = x)$ have?

      i. Unknown — can be anything
      ii. Gaussian
      iii. Bernoulli

   (b) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

   (c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

2. **Classification evaluation measures and class distribution**

   (a) In class we explained that if only class distribution changes ($\pi_0, \pi_1$ in the generative description), but the class-conditional distributions are unchanged ($f_0(x), f_1(x)$ in the generative description), then AUC of the population for a given model is unchanged, and for a specific random test sample it is almost unchanged. This means that class sampling aspects like case-control sampling do not affect AUC. For each of the following measures, state whether its expected behavior is like AUC (not sensitive to class proportions) or if it is in fact sensitive and expected to change drastically due to changes in class proportions. Explain each answer briefly, try to make your arguments concise, specific and accurate:

      i. Average Bernoulli loss
      ii. Misclassification error
      iii. FPR and TPR
      iv. Precision and Recall

   (b) Assume we have two competing models, mod1 and mod2, for the same 2-class classification problem, evaluated on the same test set. Denote their respective AUCs by AUC1, AUC2, their precisions by Precision1, Precision2, etc. Using the definitions we gave in class, which of the following guarantees that mod1 has higher accuracy on the test set? Explain briefly.

      i. TP1>TP2 and FP1<FP2
      ii. Recall1>Recall2 and Precision1<Precision2
      iii. AUC1>AUC2
      iv. FP1<FP2 and FPR1<FPR2

(c) (* Extra credit) I have two prediction models for the same 2-class classification problem. Mod1 gives AUC of 0.5 on a big test set, while Mod2 gives an AUC of 1 on a different big test set (drawn from the same distribution). Now we have a third test set, also drawn from the same distribution, and I score it using the following rule: I randomly choose half of the data and score it using Mod1, and I use Mod2 on the other half. What is the AUC expected to be on this test set? Explain briefly. You can assume there are no ties in the scores.
**Hint:** Remember the interpretation we gave of AUC in class, and use it carefully with conditional probability calculations.

3. **5-fold CV and LOOCV**

In this problem we will compare 5-fold CV and LOOCV, and compare the automated implementation in R/Python to a manual implementation of each.
Will use the Netflix data we have been using in class, with the classification problem of whether y (Miss Congeniality grade) is bigger than 3. Use only the first 2000 rows of the data as your training set (no validation or test set), and do model assessment for the logistic regression model using the cross validation approaches. The assessment we do will be on the misclassification error (so prediction is by comparing predictions to 0.5). With this setting, perform the following four tasks:

(a) Leave-one-out CV implemented manually with a loop

(b) Leave-one-out CV by using the existing implementation in your software of choice (can use cv.glm() in R or cross_validate() in Python, or any other implementation you know)

(c) 5-fold CV implemented manually with a loop

(d) 5-fold CV by using the existing implementation in your software of choice (can use cv.glm in R or cross_validate in Python, or any other implementation you know)

For each of these report the CV classification error estimate, and the running time, and report:

(a) Do the two implementations of LOOCV give the exact same result? What about the two implementations of 5-fold CV? Explain your results (try to make your arguments concise, specific and accurate).

(b) Do the manual implementations take longer to run than the automated implementations?

**Note:** Make sure you properly implement the classification error in the cross validation implementation (for example, if using cv.glm() in R, read the documentation of the cost argument).