**140.644: Statistical Machine Learning** **Name:** _____

**Instructor: Yiqun Chen TA: Junwei (Ivy) Sun**

# Quiz 5 — Practice — Non-Linear Methods, Trees, and Ensembles

**Not graded. Bring questions to class.**

This is a shortened sample so you know what to expect. The real quiz has **4 multiple choice** and **2 questions in Part B**, and is worth 100 points; this one has 2 and 2, and the matching question uses two panels rather than four. The style and the difficulty are the same. No calculator is needed.

## Part A: Multiple Choice [20 points]

---

Circle the single best answer. **(10 points)** each.

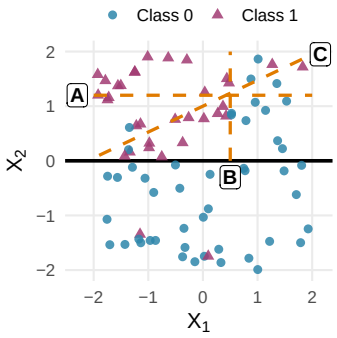
1. A classification tree has made its first split at $X_2 = 0$ (solid line). Which dashed line is the most likely next split?

   (A) Line A

   (B) Line B

   (C) Line C

   (D) None of the above




2. In a classification tree for 30-day readmission, one leaf holds seven training patients: 5 were readmitted and 2 were not. Across the whole training set, 20% of patients were readmitted. What does the tree predict for a new patient who falls in this leaf?

   (A) Readmitted, the leaf's majority class

   (B) Not readmitted, the overall majority

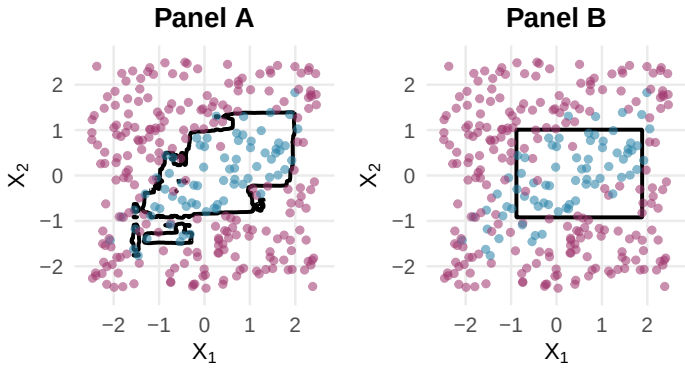
   (C) The class of the nearest of the seven

   (D) It depends on where in the leaf they fall

**Part B: Decision Boundaries** **[40 points]**

---

The panels below show two classifiers fitted to the *same simulated* two-predictor data. The black curve is the estimated decision boundary. One is a single classification tree and the other is a random forest of 500 trees.




1. **(10 points)** Match each panel to its model.

| Panel A | Panel B |
|---|---|
| _____ | _____ |

2. **(30 points)** The table gives the training error and the test error (on 10,000 new points) of the same two models.

   (a) **(10 points)** Fill in the model for each line of the table.

| Model | Training error (%) | Test error (%) |
|---|---|---|
| _____ | 11.3 | 15.5 |
| _____ | 2.3 | 12.1 |

   (b) **(10 points)** Why is the random forest's test error lower than the single tree's?

   _____

   _____

   (c) **(10 points)** Without a test set, how would we choose the number of leaves of the single tree?

   _____

   _____