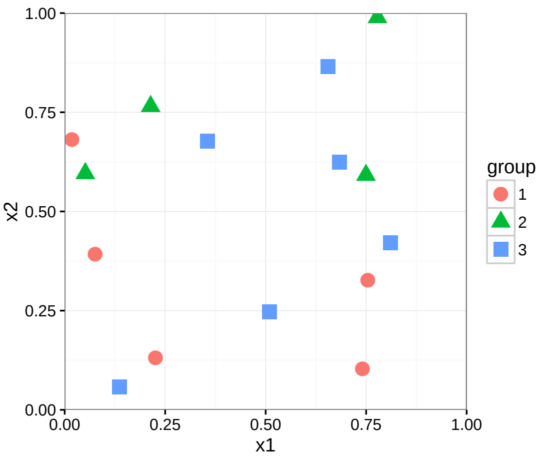
Math 243: Statistical Learning

1. The plot below represents the predictor space (on X_1 and X_2) with a training data set plotted and the class of their response variable indicated by the color.



(a) If we consider this a classification tree without any splits yet (i.e. only one region), what would be the prediction for *every* new observation?

(b) What is the (training) misclassification rate?

(c) What is the GINI index?

(d) What is the cross-entropy?

2. Add a straight line, parallel to one of the axes, that splits the predictor space into two regions. Choose the split in a way that you think will lead to the best overall improvement in the metrics above. Label the new regions R_1 and R_2 and calculate the metrics for each.

R_1

(a) What is the predicted class?

(b) What is the misclassification rate?

(c) What is the GINI index?

(d) What is the cross-entropy?

R_2

(a) What is the predicted class?

(b) What is the misclassification rate?

(c) What is the GINI index?

(d) What is the cross-entropy?

3. To decide if the split in Q2 was optimal, we need to evaluate how much the metrics in Q1 have improved. This requires combining the metrics across R_1 and R_2 in Q2. Please do so in a sensible way so that you can answer: what was the overall decrease each metric going from one region/node to two?

Misclassification: GINI: Cross-entropy:

4. On the back of this page, please draw the (very simple) tree corresponding to your partition.