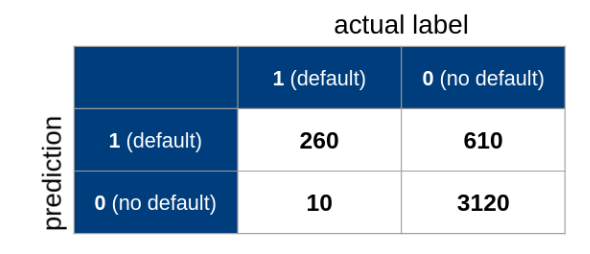
**CS5228 Tutorial 3 – Classification & Regression**

**Q1: Classification Metrics**

Assume that you have trained a binary classifier that aims to predict if a bank customer will default on his/her credit. Your test data contained 4,000 samples and your final model yields the following confusion matrix:



1. Calculate the Specificity, Sensitivity, Recall, Precision, and F1 score.

1. For each metric, provide a verbal interpretation of the resulting value in the context of predicting a customer’s likelihood to default on his or her credit! Discuss if we can be happy with the result, or what result might cause problems in practice!

1. In imbalanced datasets, a *majority* class contains most of the samples whereas a *minority* class contains only a fraction of samples (assuming a binary classification task). List example applications where you would expect a very imbalanced dataset.

1. While not covered in the lecture, what do you think can be done to correct imbalance (beyond picking the right metric)?

**Q2: Assessing Classification Errors**

In binary classification, we can make 2 types of errors:

* **False Positives** (FP), also called Type I Error
* **False Negatives** (FN), also called Type II Error

In many cases these two types of errors are not equally problematic.

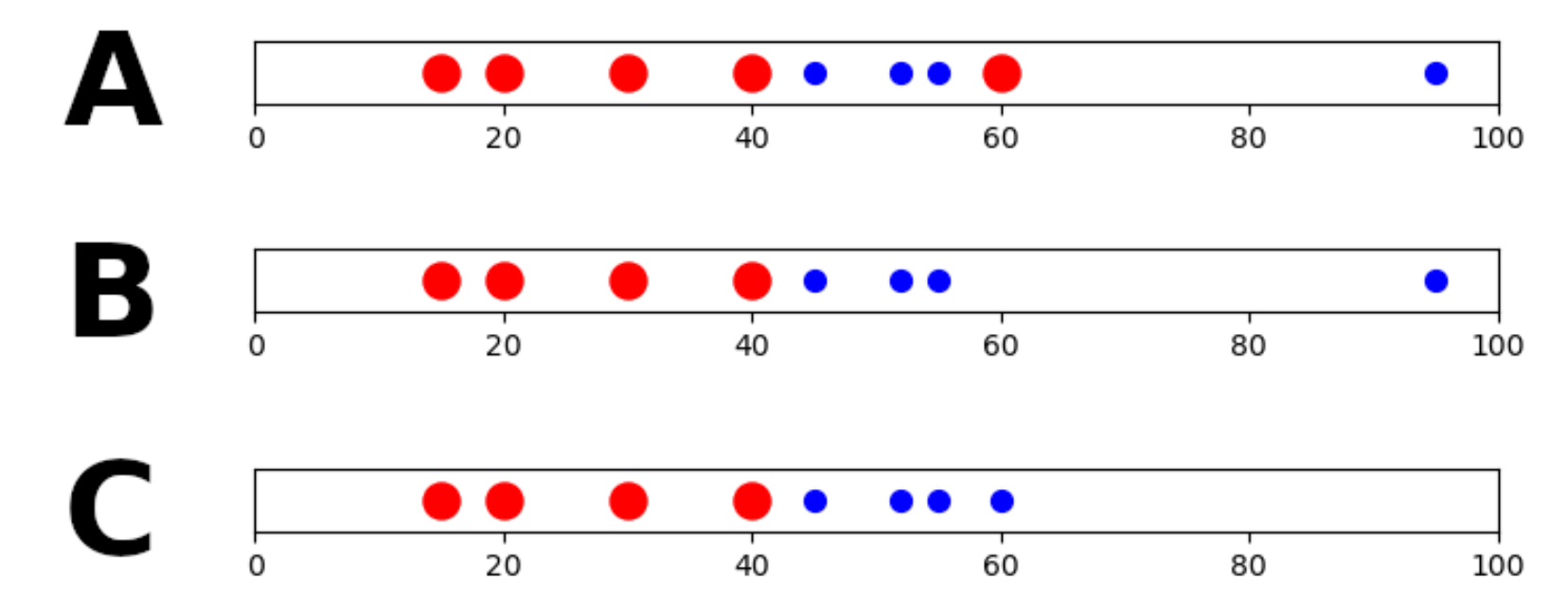
1. List 1 example application where False Positives are more problematic than False Negatives, and vice versa. Provide a brief explanation!
2. Say you trained a binary sentiment classifier that classifies social media posts (e.g., tweets) into ”negative” and ”positive”. Your datasets for training and testing were balanced and sufficiently large. Let’s assume that the F1 score of your classifier is 0.85. How would you assess if this is a good result?

1. Say you trained a classifier that identifies whether an image contains a Car, Boat, or Plane, and the F1-score is very high, say, 0.99. Your datasets for training and testing were balanced and sufficiently large. What might be a reason why the classifier would suddenly perform poorly in practice?

**Q3: Data Distribution and Decision Trees**

The figure below shows 3 distributions of the values for a single feature. The color and shape of the dots reflect the class label. Since we only have two colors/sizes, the example application is a binary classification task.

In A, we observe what can be called a “mislabeled point” at x=60, and an “outlier” at x=95. In B, the mislabeled point is removed, while in C, the outlier is removed.

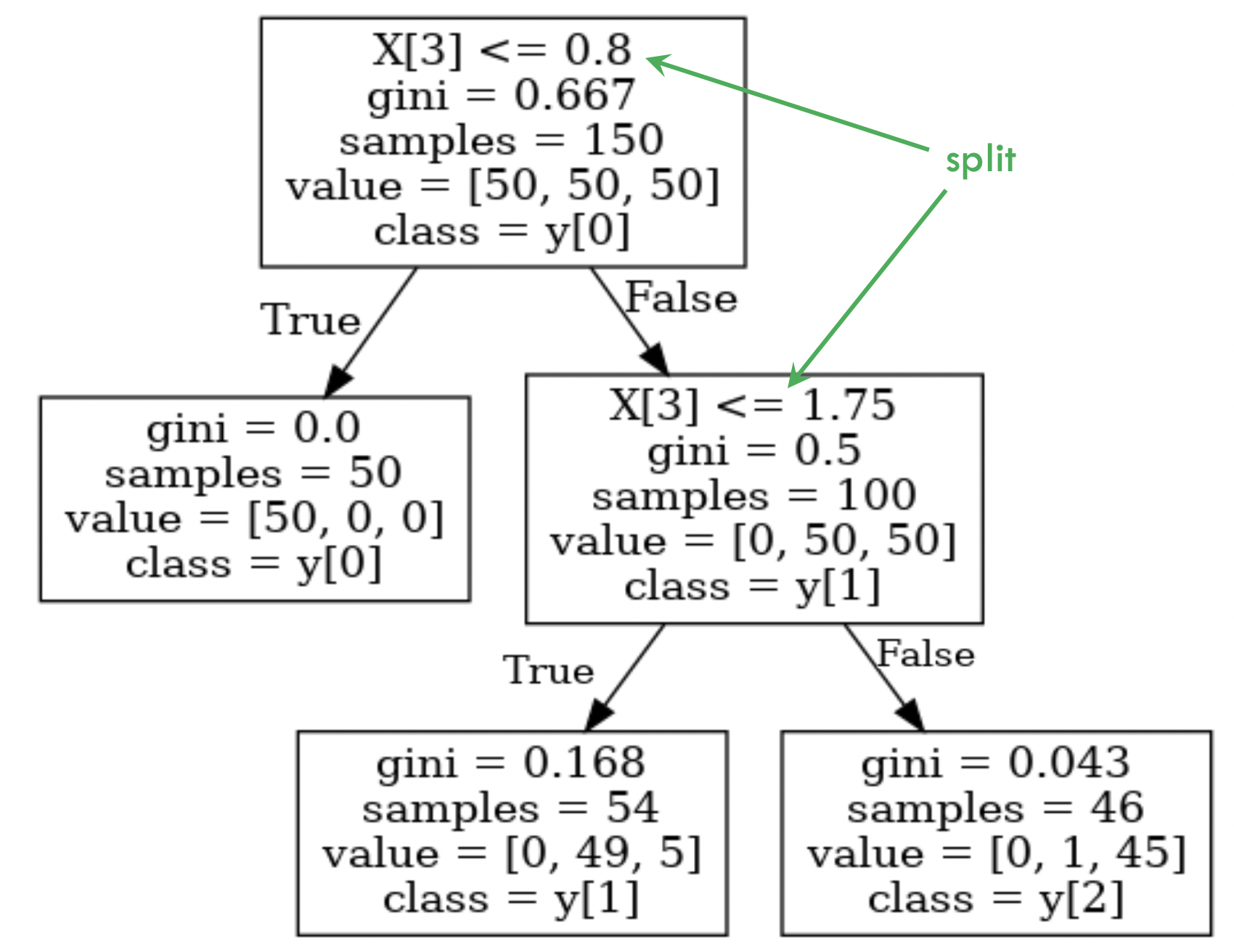
* 

1. How many splits would the decision tree fitting algorithm perform in each of these cases?

1. Given your previous answer, summarize how “mislabeled points” and “outliers” affect the learned decision tree in this case.

**Q4: Interpreting Decision Trees**

The Decision Tree shown below has been trained over the iris Dataset with a maximum depth of 2. Each data sample has 4 numerical features, and labeled with 1 out of 3 classes. Below, X[3] refers to the 3rd feature; samples refers to the number of samples in a node; value refers to the distribution of class labels of the samples in a node; and class refers to the predicted class of a node.



1. Based on the learned decision tree, how would you describe the characteristics of each class in the dataset?

1. How would the decision tree change if we additionally standardized each feature before training the decision tree?