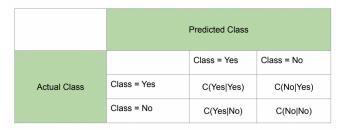
## Lecture 12 Model Evaluation

- Confusion Matrix
  - $\circ \quad Accuracy = (a+d)/(a+b+c+d)$

	Predicted Class		
Actual Class		Class = Yes	Class = No
	Class = Yes	a (TP)	b (FN)
	Class = No	c (FP)	d (TN)

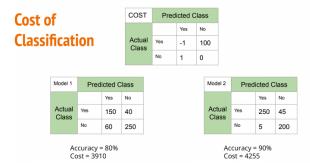
- $\circ$   $\circ$   $\mathsf{TP} = \mathsf{true}\;\mathsf{positive}$
- $\circ$  FP = false positive
- $\circ$  FN = false negative
- $\circ$  TN = True Negative
- o Accuracy can be misleading
  - If one particular class has a majority of the data set, a predictor that only predicts that class will yield a very high accuracy
- Cost Matrix



0

COST	Predicted Class		
Actual Class		Yes	No
	Yes	а	b
	No	С	d

- $\begin{array}{ccc}
  \circ & \\
  \circ & \text{Precision} = \frac{a}{a} & \text{(a+c)}
  \end{array}$
- $\circ$  Recall = a/(a+b)
- $\circ$  F-Measure = 2RP/(R+P)



0

- Methods of Estimation
  - Goal: get a reliable estimate of performance of the model on unseen data
  - Holdout:
    - Split data into two sets a testing and a training set
    - Use \( \frac{1}{4} \) of the dataset for testing and use \( \frac{3}{4} \) for training
  - Cross Validaton
    - Split data into K equal sized folds (subsets)
    - Train the model on K-1 fols
    - Test the moden on the remaining 1 fold
      - o Repeat this K times so each fold serves at a test set once
      - o Average the performance metrics over all k runs
    - Partition into K disjoint subsets
    - K-fold: train on K-1 partitions, test on the remaining
    - K=n leave one out
      - Special case where K=n and n is the number of data points
      - For each example
        - Train on all other n-1 samples
        - Test on the one left out
          - This is very accurate but computationally expensive for large datasets

Method	Accuracy	Speed	<b>Overfitting Risk</b>
Holdout	Medium	Fast	Higher
K-Fold CV	High	Moderate	Lower
Leave-One-Out	Very High	Very Slow	Very Low

- Ensemble Methods
  - Ensembling in machine learning is the technique of combining multiple models (often called "learners" or "classifiers") to improve overall performance usually in terms of accuracy, robustness, or generalization.
  - Reduces error
  - Increases stability
- Suppose you have 17 independent classifier each with an error rate of e = 0.20 and you take a majority vote to make a final decision
  - o The majority needs to make a mistake and get at least 9/17 wrong
  - What is the change the ensemble gets it wrong
    - Binomial probability problem

$$P(X \ge 9) = \sum_{k=9}^{17} {17 \choose k} (.2)^k (1 - .2)^{17-k} = 0.002581463$$

- This value is much smaller than 0.20
- This shows how combining multiple weak learners can drastically reduce error if they are independent and better than random

## Bagging

- o Goal: reduce varible by training classifiers on different subsets of the data
- Generate bootstrap samples by randomly sampling from the dataset to create multiple training sets
- o Train a separate model on each bootstrap sample
- o Combine their predictions (majority vote or average)
- o Good for unstable learners like decision trees
- o Random forests is bagging applied to decision trees

## Boosting

- o Reduce bias by focusing on errors made by previous models
  - Train the first classifier
  - Increase the weights of misclassified points
  - Train the next classifier to focus more on those errors
    - Repeat.
  - Combine all classifiers using a weight vote
- o Learners are not independent but work in sequence to correct each other