School of Computing
UNIVERSITY OF GEORGIA

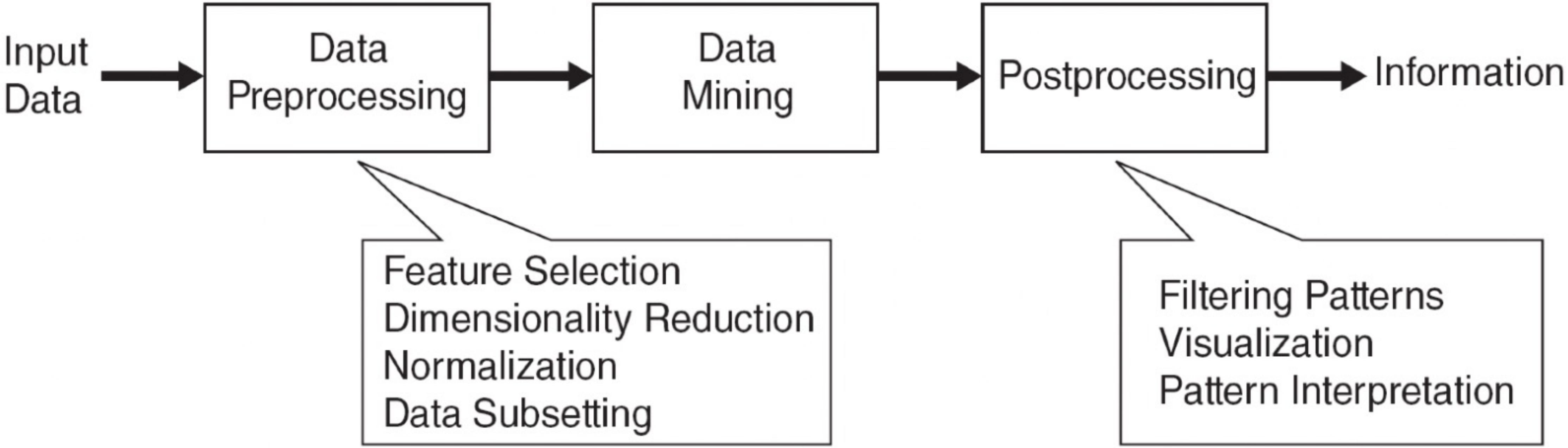
# CSCI 4380/6380 Data Mining


Fei Dou

Assistant Professor
School of Computing
University of Georgia


October 18, 2023

# Recap: Data Mining Process



# Test 2

- Time: 75 mins on October 26 during class
- Materials: Course slides
- Covering Topics:
  - Classification: Decision Tree, kNN, Naïve Bayes, Logistic Regression
  - Evaluation, CV, Loss Function, Gradient Descent
  - Clustering: K-Means

- Format:
  - True or False, Yes or No
  - Multiple choice questions
  - Short answer problems

# Imbalanced Classification Problem

# Class Imbalance Problem

- Lots of classification problems where the classes are skewed (more records from one class than another)
  - Credit card fraud
  - Intrusion detection
  - Defective products in manufacturing assembly line
  - COVID-19 test results on a random sample

- **Key Challenge**:
  - Evaluation measures such as accuracy are not well-suited for imbalanced class

# Recap: Confusion Matrix

| | PREDICTED CLASS | | |
|---|---|---|---|
| ACTUAL CLASS | | Class=Yes | Class=No |
| | Class=Yes | a | b |
| | Class=No | c | d |

- a: TP (true positive)
- b: FN (false negative)
- c: FP (false positive)
- d: TN (true negative)

# Recap: Accuracy

| | PREDICTED CLASS | |
|---|---|---|
| ACTUAL CLASS | Class=Yes | Class=No |
| Class=Yes | a (TP) | b (FN) |
| Class=No | c (FP) | d (TN) |

Most widely-used metric:

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

# Problem with Accuracy

- Consider a 2-class problem
  - Number of Class NO examples = 990
  - Number of Class YES examples = 10
- If a model predicts everything to be class NO, accuracy is 990/1000 = 99 %
  - This is misleading because this trivial model does not detect any class YES example
  - Detecting the rare class is usually more interesting (e.g., frauds, intrusions, defects, etc)

| | | PREDICTED CLASS | |
|---|---|---|---|
| ACTUAL CLASS | | Class=Yes | Class=No |
| | Class=Yes | 0 | 10 |
| | Class=No | 0 | 990 |

# Which model is better?

A

| | | PREDICTED | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL | Class=Yes | 0 | 10 |
| | Class=No | 0 | 990 |

**Accuracy: 99%**

B

| | | PREDICTED | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL | Class=Yes | 10 | 0 |
| | Class=No | 500 | 490 |

**Accuracy: 50%**

# Which model is better?

A

| | PREDICTED | | |
|---|---|---|---|
| ACTUAL | | Class=Yes | Class=No |
| | Class=Yes | 5 | 5 |
| | Class=No | 0 | 990 |

B

| | PREDICTED | | |
|---|---|---|---|
| ACTUAL | | Class=Yes | Class=No |
| | Class=Yes | 10 | 0 |
| | Class=No | 500 | 490 |

# Recap: Alternative Measures

| | PREDICTED CLASS | |
|---|---|---|
| **ACTUAL CLASS** | Class=Yes | Class=No |
| Class=Yes | a | b |
| Class=No | c | d |

- Precision(p) $= \frac{a}{a+c}$
- Recall(r) $= \frac{a}{a+b}$
- F-measure(F) $= \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$

# Alternative Measures

| | PREDICTED CLASS | | |
|---|---|---|---|
| ACTUAL CLASS | | Class=Yes | Class=No |
| | Class=Yes | 10 | 0 |
| | Class=No | 10 | 980 |

- $\text{Precision}(p) = \frac{10}{10+10} = 0.5$
- $\text{Recall}(r) = \frac{10}{10+0} = 1$
- $\text{F-measure}(F) = \frac{2*1*0.5}{1+0.5} = 0.62$
- $\text{Accuracy} = \frac{990}{1000} = 0.99$

# Alternative Measures

| | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 10 | 0 |
| | Class=No | 10 | 980 |

$$\text{Precision (p)} = \frac{10}{10+10} = 0.5$$

$$\text{Recall (r)} = \frac{10}{10+0} = 1$$

$$\text{F - measure (F)} = \frac{2*1*0.5}{1+0.5} = 0.62$$

$$\text{Accuracy} = \frac{990}{1000} = 0.99$$

| | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 1 | 9 |
| | Class=No | 0 | 990 |

$$\text{Precision (p)} = \frac{1}{1+0} = 1$$

$$\text{Recall (r)} = \frac{1}{1+9} = 0.1$$

$$\text{F - measure (F)} = \frac{2*0.1*1}{1+0.1} = 0.18$$

$$\text{Accuracy} = \frac{991}{1000} = 0.991$$

# Which of these classifiers is better?

**A**

| | PREDICTED CLASS | | |
|---|---|---|---|
| ACTUAL CLASS | | Class=Yes | Class=No |
| | Class=Yes | 40 | 10 |
| | Class=No | 10 | 40 |

**B**

| | PREDICTED CLASS | | |
|---|---|---|---|
| ACTUAL CLASS | | Class=Yes | Class=No |
| | Class=Yes | 40 | 10 |
| | Class=No | 1000 | 4000 |

# Which of these classifiers is better?

**A**

| | PREDICTED CLASS | |
|---|---|---|
| ACTUAL CLASS | Class=Yes | Class=No |
| Class=Yes | 40 | 10 |
| Class=No | 10 | 40 |

$$\text{Precision (p)} = 0.8$$
$$\text{Recall (r)} = 0.8$$
$$\text{F - measure (F)} = 0.8$$
$$\text{Accuracy} = 0.8$$

**B**

| | PREDICTED CLASS | |
|---|---|---|
| ACTUAL CLASS | Class=Yes | Class=No |
| Class=Yes | 40 | 10 |
| Class=No | 1000 | 4000 |

$$\text{Precision (p)} =\sim 0.04$$
$$\text{Recall (r)} = 0.8$$
$$\text{F - measure (F)} =\sim 0.08$$
$$\text{Accuracy} =\sim 0.8$$

# Measures of Classification Performance

| | PREDICTED CLASS | |
|---|---|---|
| ACTUAL CLASS | Yes | No |
| Yes | TP | FN |
| No | FP | TN |

α is the probability that we reject the null hypothesis when it is true. This is a Type I error or a false positive (FP).

β is the probability that we accept the null hypothesis when it is false. This is a Type II error or a false negative (FN).

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

$$ErrorRate = 1 - accuracy$$

$$Precision = Positive\ Predictive\ Value = \frac{TP}{TP + FP}$$

$$Recall = Sensitivity = TP\ Rate = \frac{TP}{TP + FN}$$

$$Specificity = TN\ Rate = \frac{TN}{TN + FP}$$

$$FP\ Rate = \alpha = \frac{FP}{TN + FP} = 1 - specificity$$

$$FN\ Rate = \beta = \frac{FN}{FN + TP} = 1 - sensitivity$$

$$Power = sensitivity = 1 - \beta$$

# Alternative Measures

| A | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 40 | 10 |
| | Class=No | 10 | 40 |

$$\text{Precision (p)} = 0.8$$
$$\text{TPR} = \text{Recall (r)} = 0.8$$
$$\text{FPR} = 0.2$$
$$\text{F−measure (F)} = 0.8$$
$$\text{Accuracy} = 0.8$$
$$\frac{\text{TPR}}{\text{FPR}} = 4$$

| B | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 40 | 10 |
| | Class=No | 1000 | 4000 |

$$\text{Precision (p)} = 0.038$$
$$\text{TPR} = \text{Recall (r)} = 0.8$$
$$\text{FPR} = 0.2$$
$$\text{F−measure (F)} = 0.07$$
$$\text{Accuracy} = 0.8$$
$$\frac{\text{TPR}}{\text{FPR}} = 4$$

# ROC (Receiver Operating Characteristic)

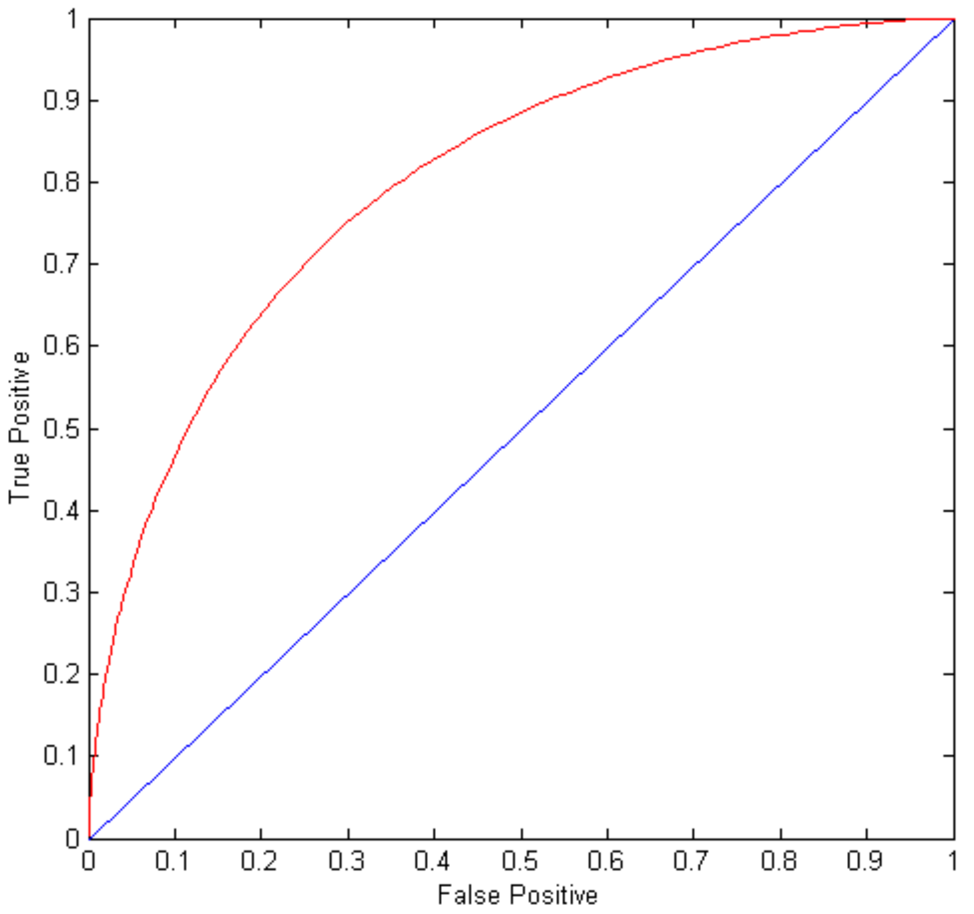
- A graphical approach for displaying trade-off between detection rate and false alarm rate
- Developed in 1950s for signal detection theory to analyze noisy signals
- ROC curve plots TPR against FPR
  - Performance of a model represented as a point in an ROC curve

# ROC Curve

(TPR,FPR):

- (0,0): declare everything to be negative class
- (1,1): declare everything to be positive class
- (1,0): ideal

- Diagonal line:
  - Random guessing
  - Below diagonal line:
    - prediction is opposite of the true class



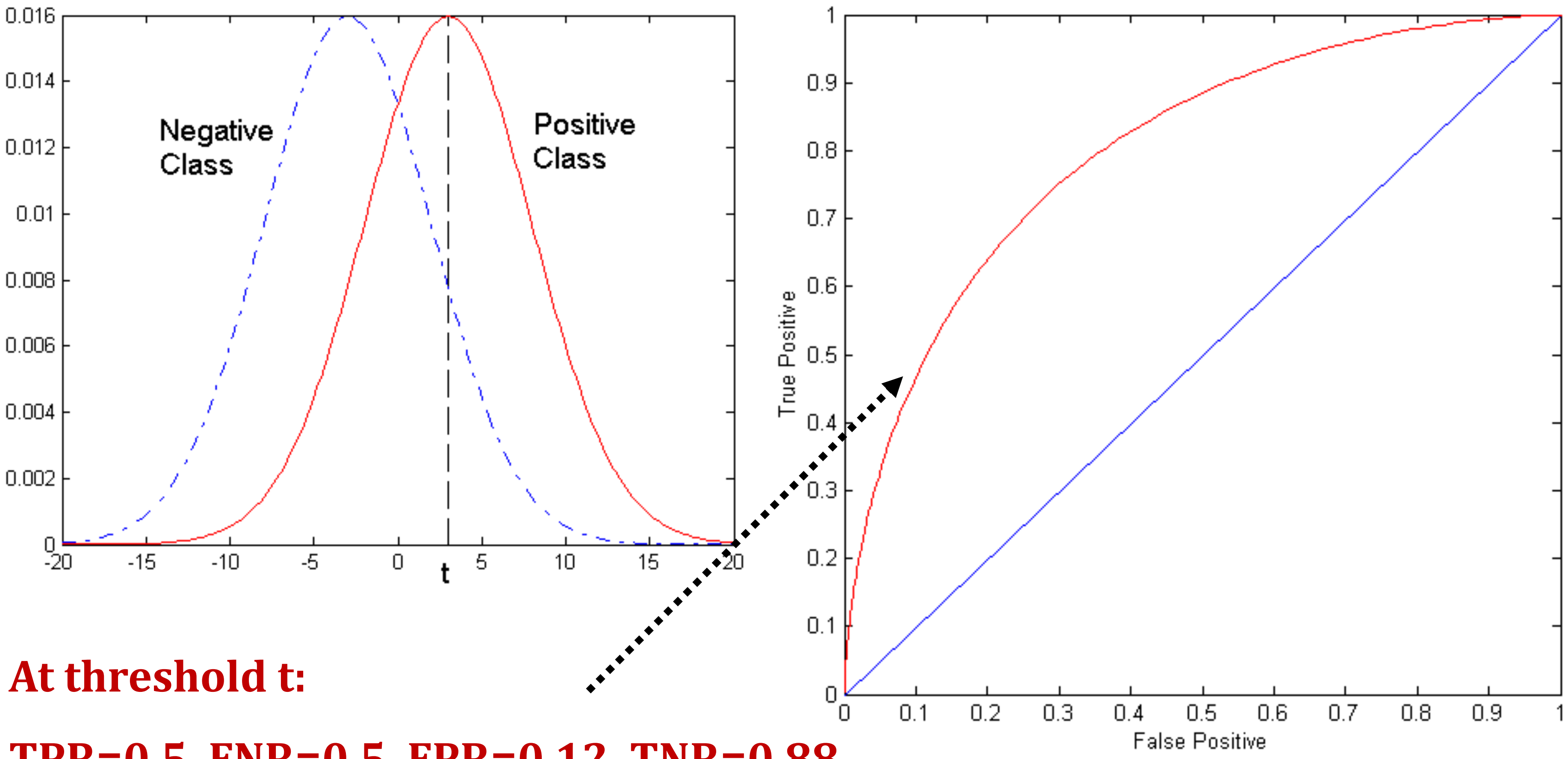
# ROC (Receiver Operating Characteristic)

- To draw ROC curve, classifier must produce continuous-valued output
  - Outputs are used to rank test records, from the most likely positive class record to the least likely positive class record
  - By using different thresholds on this value, we can create different variations of the classifier with TPR/FPR tradeoffs

- Many classifiers produce only discrete outputs (i.e., predicted class)
  - How to get continuous-valued outputs?
    - Decision trees, neural networks, Bayesian classifiers, k-nearest neighbors

# ROC Curve Example

- 1-dimensional data set containing 2 classes (positive and negative)
- Any points located at x > t is classified as positive




**At threshold t:**

**TPR=0.5, FNR=0.5, FPR=0.12, TNR=0.88**

# How to Construct an ROC curve

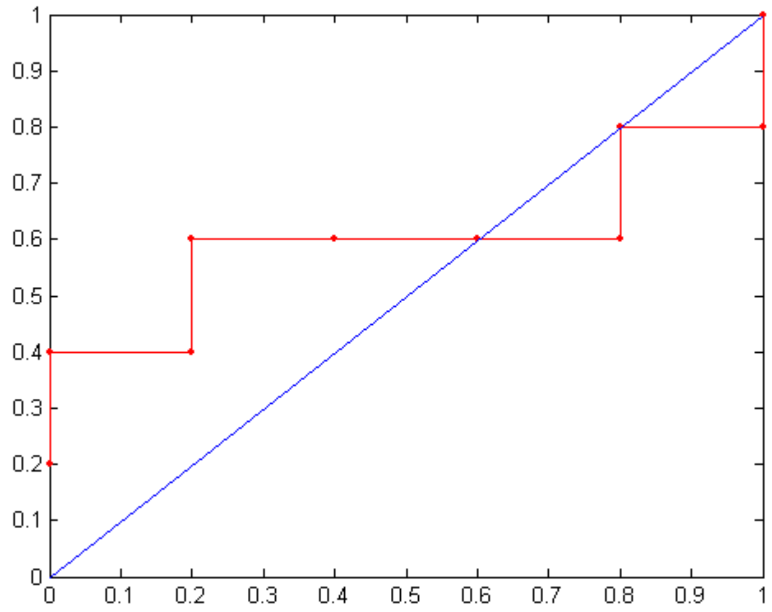
| Instance | Score | True Class |
|---|---|---|
| 1 | 0.95 | + |
| 2 | 0.93 | + |
| 3 | 0.87 | - |
| 4 | 0.85 | - |
| 5 | 0.85 | - |
| 6 | 0.85 | + |
| 7 | 0.76 | - |
| 8 | 0.53 | + |
| 9 | 0.43 | - |
| 10 | 0.25 | + |

- Use a classifier that produces a continuous-valued score for each instance
  - The more likely it is for the instance to be in the + class, the higher the score
- Sort the instances in decreasing order according to the score
- Apply a threshold at each unique value of the score
- Count the number of TP, FP, TN, FN at each threshold
  - TPR = TP/(TP+FN)
  - FPR = FP/(FP + TN)

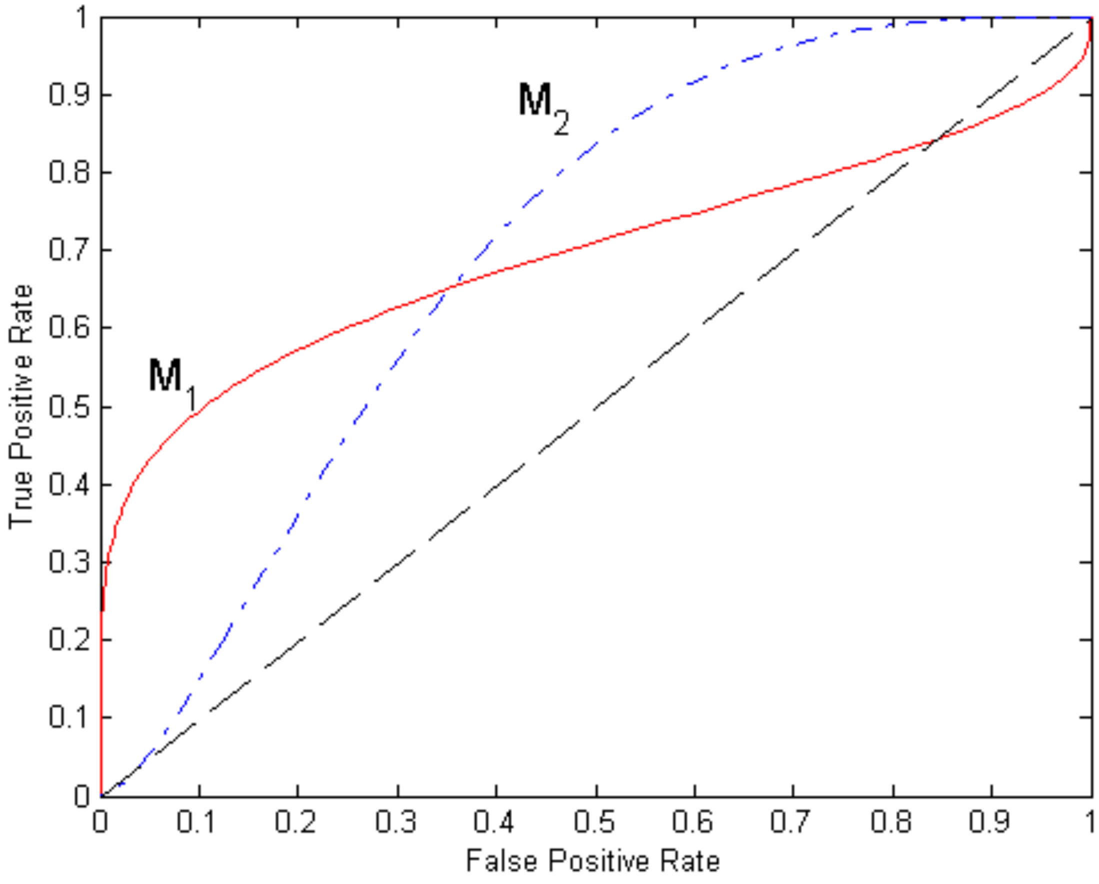
# How to construct an ROC curve

| Class | + | - | + | - | - | - | + | - | + | + | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Threshold >= | 0.25 | 0.43 | 0.53 | 0.76 | 0.85 | 0.85 | 0.85 | 0.87 | 0.93 | 0.95 | 1.00 |
| TP | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | 1 | 0 |
| FP | 5 | 5 | 4 | 4 | 3 | 2 | 1 | 1 | 0 | 0 | 0 |
| TN | 0 | 0 | 1 | 1 | 2 | 3 | 4 | 4 | 5 | 5 | 5 |
| FN | 0 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 4 | 5 |
| → TPR | 1 | 0.8 | 0.8 | 0.6 | 0.6 | 0.6 | 0.6 | 0.4 | 0.4 | 0.2 | 0 |
| → FPR | 1 | 1 | 0.8 | 0.8 | 0.6 | 0.4 | 0.2 | 0.2 | 0 | 0 | 0 |

**ROC Curve:**



# Using ROC for Model Comparison




- No model consistently outperforms the other
  - $M_1$ is better for small FPR
  - $M_2$ is better for large FPR
- Area Under the ROC curve (AUC)
  - Ideal:
    - Area = 1
  - Random guess:
    - Area = 0.5

# Dealing with Imbalanced Classes - Summary

- Many measures exists, but none of them may be ideal in all situations
  - Random classifiers can have high value for many of these measures
  - TPR/FPR provides important information but may not be sufficient by itself in many practical scenarios
  - Given two classifiers, sometimes you can tell that one of them is strictly better than the other
    - C1 is strictly better than C2 if C1 has strictly better TPR and FPR relative to C2 (or same TPR and better FPR, and vice versa)
  - Even if C1 is strictly better than C2, C1's F-value can be worse than C2's if they are evaluated on data sets with different imbalances
  - Classifier C1 can be better or worse than C2 depending on the scenario at hand (class imbalance, importance of TP vs FP, cost/time tradeoffs)

# Which classifier is better?

| T1 | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 50 | 50 |
| | Class=No | 1 | 99 |

$$\text{Precision (p)} = 0.98$$
$$\text{TPR} = \text{Recall (r)} = 0.5$$
$$\text{FPR} = 0.01$$
$$\text{TPR/FPR} = 50$$
$$\text{F} - \text{measure} = 0.66$$

| T2 | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 99 | 1 |
| | Class=No | 10 | 90 |

$$\text{Precision (p)} = 0.9$$
$$\text{TPR} = \text{Recall (r)} = 0.99$$
$$\text{FPR} = 0.1$$
$$\text{TPR/FPR} = 9.9$$
$$\text{F} - \text{measure} = 0.94$$

| T3 | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | 99 | 1 |
| | Class=No | 1 | 99 |

$$\text{Precision (p)} = 0.99$$
$$\text{TPR} = \text{Recall (r)} = 0.99$$
$$\text{FPR} = 0.01$$
$$\text{TPR/FPR} = 99$$
$$\text{F} - \text{measure} = 0.99$$

# Which classifier is better? Medium Skew case

| T1 | PREDICTED CLASS | |
|---|---|---|
| ACTUAL CLASS | | Class=Yes | Class=No |
| | Class=Yes | 50 | 50 |
| | Class=No | 10 | 990 |

$$\text{Precision (p)} = 0.83$$
$$\text{TPR} = \text{Recall (r)} = 0.5$$
$$\text{FPR} = 0.01$$
$$\text{TPR/FPR} = 50$$
$$\text{F} - \text{measure} = 0.62$$

| T2 | PREDICTED CLASS | | |
|---|---|---|---|
| ACTUAL CLASS | | Class=Yes | Class=No |
| | Class=Yes | 99 | 1 |
| | Class=No | 100 | 900 |

$$\text{Precision (p)} = 0.5$$
$$\text{TPR} = \text{Recall (r)} = 0.99$$
$$\text{FPR} = 0.1$$
$$\text{TPR/FPR} = 9.9$$
$$\text{F} - \text{measure} = 0.66$$

| T3 | PREDICTED CLASS | | |
|---|---|---|---|
| ACTUAL CLASS | | Class=Yes | Class=No |
| | Class=Yes | 99 | 1 |
| | Class=No | 10 | 990 |

$$\text{Precision (p)} = 0.9$$
$$\text{TPR} = \text{Recall (r)} = 0.99$$
$$\text{FPR} = 0.01$$
$$\text{TPR/FPR} = 99$$
$$\text{F} - \text{measure} = 0.94$$

# Which classifier is better? High Skew case

| T1 | PREDICTED CLASS | |
|---|---|---|
| | Class=Yes | Class=No |
| ACTUAL CLASS Class=Yes | 50 | 50 |
| ACTUAL CLASS Class=No | 100 | 9900 |

$$\text{Precision (p)} = 0.3$$
$$\text{TPR} = \text{Recall (r)} = 0.5$$
$$\text{FPR} = 0.01$$
$$\text{TPR/FPR} = 50$$
$$\text{F} - \text{measure} = 0.375$$

| T2 | PREDICTED CLASS | |
|---|---|---|
| | Class=Yes | Class=No |
| ACTUAL CLASS Class=Yes | 99 | 1 |
| ACTUAL CLASS Class=No | 1000 | 9000 |

$$\text{Precision (p)} = 0.09$$
$$\text{TPR} = \text{Recall (r)} = 0.99$$
$$\text{FPR} = 0.1$$
$$\text{TPR/FPR} = 9.9$$
$$\text{F} - \text{measure} = 0.165$$

| T3 | PREDICTED CLASS | |
|---|---|---|
| | Class=Yes | Class=No |
| ACTUAL CLASS Class=Yes | 99 | 1 |
| ACTUAL CLASS Class=No | 100 | 9900 |

$$\text{Precision (p)} = 0.5$$
$$\text{TPR} = \text{Recall (r)} = 0.99$$
$$\text{FPR} = 0.01$$
$$\text{TPR/FPR} = 99$$
$$\text{F} - \text{measure} = 0.66$$

# Building Classifiers with Imbalanced Training Set

- Modify the distribution of training data so that rare class is well-represented in training set
  - Undersample the majority class
  - Oversample the rare class

- Loss Functions:
  - WRAP Loss (IJCAI 2011)
  - Focal Loss (ICCV 2017)
  - r-th Root Ranking loss (KDD 2018)
  - MultiSimilarity (MS) (ECCV 2020)
  - SoftTriple (ECCV 2020)