# 2-6 Evaluation of Binary Classification Models

Zhonglei Wang

WISE and SOE, XMU, 2025

# Contents

# Recall

1. Binary classification models
   - Logistic regression model
   - Support vector machine
   - Random forest
   - Neural network models
2. For most models,
   - A certain score is estimated
   - A certain threshold is used to determine the estimated class
3. We consider the case that a model has estimated class labels for each example in the test dataset.

# Confusion matrix

1. Among the binary responses in the test dataset,
   - True positive number: number of true positive responses
   - True negative number: number of true negative responses
2. A certain binary classification model predicts responses for the test dataset
   - Predicted positive number: number of predicted positive responses
   - Predicted negative number: number of predicted negative responses

# Confusion matrix

1. There are four possibilities
   - $TP$ (True Positive): number of correctly estimated positives
   - $FN$ (False Negative): number of falsely estimated negatives
     (should be positives but estimated as negatives)
   - $TN$ (True Negative): number of correctly estimated negatives
   - $FP$ (False Positive): number of falsely estimated positives
     (should be negative but estimated as positives)

# Confusion matrix

1. Confusion matrix

| | Predicted positive | Predicted negative |
|---|---|---|
| True positive | $TP$ | $FN$ |
| True negative | $FP$ | $TN$ |

# Measures

| | Predicted positive | Predicted negative |
|---|---|---|
| True positive | $TP$ | $FN$ |
| True negative | $FP$ | $TN$ |

1. Accuracy

$$\begin{aligned}\text{Accuracy} &= \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \\ &= \frac{TP + TN}{TP + TN + FP + FN}\end{aligned}$$

2. Precision

$$\begin{aligned}\text{Precision} &= \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \\ &= \frac{TP}{TP + FP}\end{aligned}$$

# Measures

| | Predicted positive | Predicted negative |
|---|---|---|
| True positive | $TP$ | $FN$ |
| True negative | $FP$ | $TN$ |

3. Recall or True Positive Rate, TPR

$$\begin{aligned}\text{Recall (TPR)} &= \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \\ &= \frac{TP}{TP + FN}\end{aligned}$$

4. False Positive Rate (FPR)

$$\begin{aligned}\text{FPR} &= \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} \\ &= \frac{FP}{FP + TN}\end{aligned}$$

# Measures

5. F1 score

$$\text{F1 score} = 2 \times \frac{\text{Precission} \times \text{Recall}}{\text{Precission} + \text{Recall}}$$

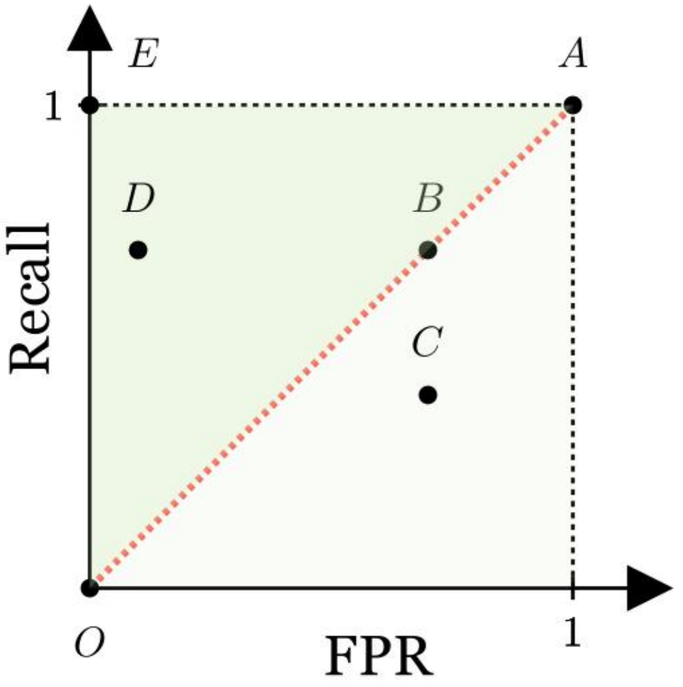
6. For a certain model,
   - It only has one set of these measures if it estimates the classes directly
   - It corresponds to different set of measures if it estimates the "scores" first
     - Different thresholds correspond to different estimated labels.

# ROC curve

1. Assume the availability of estimated classes for the test dataset
   - Obtain the Recall (sensitivity)
   - Obtain the FPR (1-specificity)
2. ROC curve plots Recall against FPR
3. Two scenarios for the ROC curve:
   - For a model predicting LABELS directly, it corresponds to a POINT
   - For a model predicting SCORES, it corresponds to a CURVE

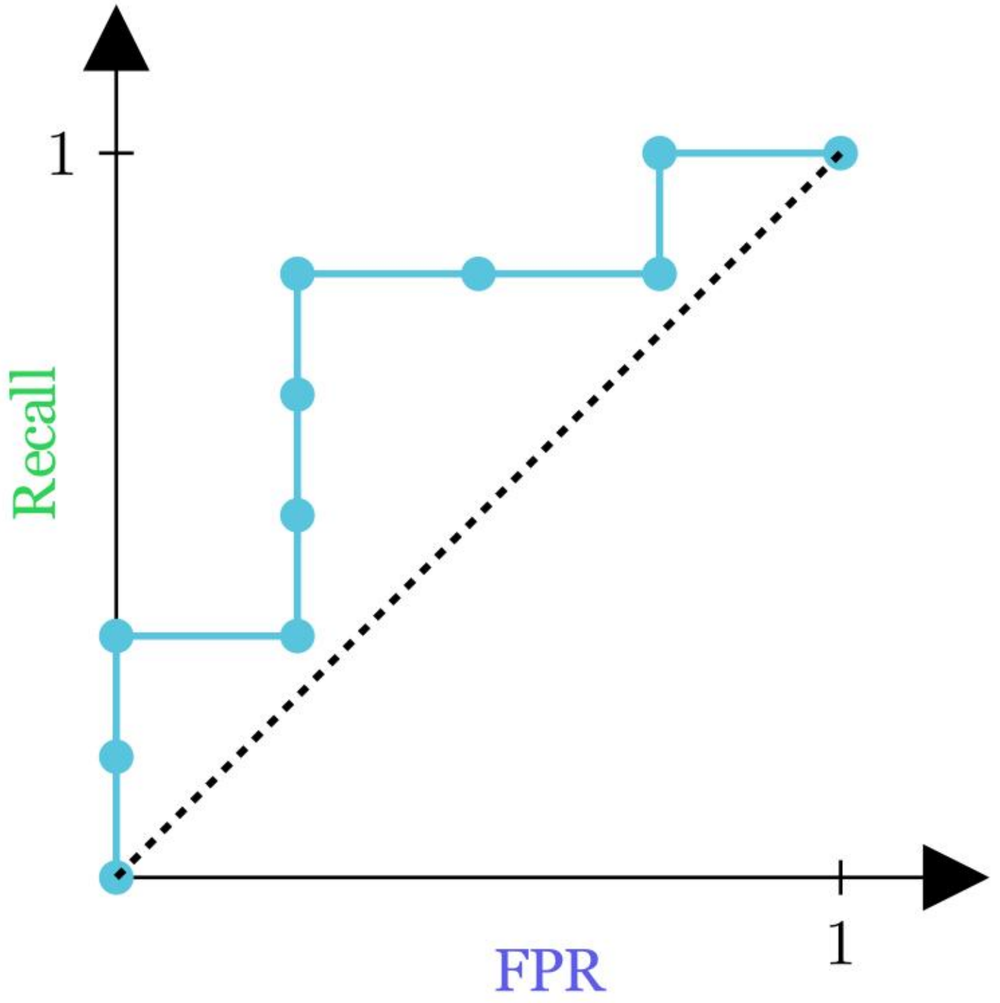
# ROC curve

1. Different estimation corresponds to different points
   - Points on the diagonal: Random guess with different probabilities
     - $O\,(0,0)$ : Assign all test examples to the negative class.
     - $B\,(0.7,0.7)$ : Assign a test example to the positive class with probability 0.7
     - $A\,(1,1)$ : Assign all test examples to the positive class
   - Points above the diagonal: Good models
     - $D\,(0.1,0.7)$ : It can correctly identify 70% positives with low FPR
     - $E\,(0,1)$ : Perfect model
     - The more up-left, the better!
   - Points below the diagonal: Actually, not that bad!
     - $C\,(0.7,0.4)$: Although recall is low but FPR is high, the model messes up the two labels!



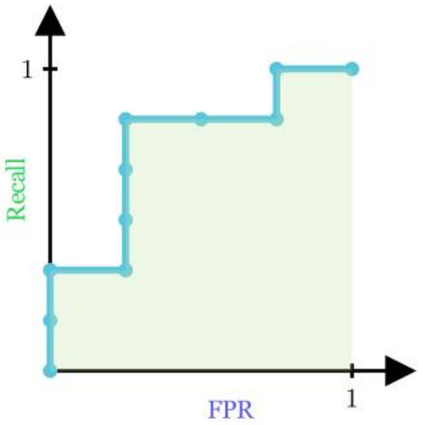
# Draw an ROC curve

1. For models predicting "scores",
   - Different thresholds correspond to different points
   - Join those points produces an ROC curve
2. We use a toy example to show how to obtain an ROC curve




| Sam. Ind. | True Lab. | Est. Lab. | Est. Sco. |
|---|---|---|---|
| 1 | 1 | | 0.9 |
| 2 | 1 | | 0.8 |
| 3 | 0 | | 0.7 |
| 4 | 1 | | 0.6 |
| 5 | 1 | | 0.55 |
| 6 | 1 | | 0.54 |
| 7 | 0 | | 0.53 |
| 8 | 0 | | 0.52 |
| 9 | 1 | | 0.51 |
| 10 | 0 | | 0.505 |

# AUC




1. AUC is short for "area under the ROC curve"
   - It calculates the area under this curve
2. A good classification model corresponds to high AUC
   - If a model "ranks" positive examples higher than negative ones, its AUC is high
   - A perfect model is the one with AUC being 1, where the scores for positives are uniformly larger than those for the negatives.
3. If the AUC is very small, the associated model is not so bad, since it messes up positives and negatives