

# CS685: DATA MINING BASICS OF CLASSIFICATION

Arnab Bhattacharya  
arnabb@cse.iitk.ac.in

Computer Science and Engineering,  
Indian Institute of Technology, Kanpur  
<http://web.cse.iitk.ac.in/~cs685/>

1<sup>st</sup> semester, 2020-21  
Mon 1030-1200 (online)

# Classification

- A dataset of  $n$  objects  $O_i, i = 1, \dots, n$
- A total of  $k$  classes  $C_j, j = 1, \dots, k$
- Each object belongs to a *single* class
- If object  $O_i$  belongs to class  $C_j$ , then  $C(O_i) = j$
- Given a new object  $O_q$ , **classification** is the problem of determining its class, i.e.,  $C(O_q)$  out of possible  $k$  choices

# Classification

- A dataset of  $n$  objects  $O_i, i = 1, \dots, n$
- A total of  $k$  classes  $C_j, j = 1, \dots, k$
- Each object belongs to a *single* class
- If object  $O_i$  belongs to class  $C_j$ , then  $C(O_i) = j$
- Given a new object  $O_q$ , **classification** is the problem of determining its class, i.e.,  $C(O_q)$  out of possible  $k$  choices
- If, instead of  $k$  discrete classes, there is a continuum of values, the problem of determining the value  $V(O_q)$  of a new object  $O_q$  is called **prediction**

- Total available data is divided *randomly* into two parts: **training set** and **testing set**
- Classification algorithm or model is built using *only* the training set
- Testing set should not be used *at all*
- Quality of method is measured using testing set
- Sometimes **validation set** is separated from training set to evaluate method

- **Stratified**

- If representation of each class in training set is proportional to the overall ratios

- **Stratified**
  - If representation of each class in training set is proportional to the overall ratios
- **$k$ -fold cross-validation**
  - Training data is divided into  $k$  random parts
  - $k - 1$  groups are used as training set and the  $k^{\text{th}}$  group as validation set
  - Training is repeated  $k$  times with a new validation set each time
- **Leave-one-out cross-validation (LOOCV):** When  $k = n$

- **Stratified**
  - If representation of each class in training set is proportional to the overall ratios
- **$k$ -fold cross-validation**
  - Training data is divided into  $k$  random parts
  - $k - 1$  groups are used as training set and the  $k^{\text{th}}$  group as validation set
  - Training is repeated  $k$  times with a new validation set each time
- **Leave-one-out cross-validation (LOOCV):** When  $k = n$
- **Stratified cross-validation**
  - When representation in each of the  $k$  random groups is proportional to the overall ratios

- **Stratified**
  - If representation of each class in training set is proportional to the overall ratios
- **$k$ -fold cross-validation**
  - Training data is divided into  $k$  random parts
  - $k - 1$  groups are used as training set and the  $k^{\text{th}}$  group as validation set
  - Training is repeated  $k$  times with a new validation set each time
- **Leave-one-out cross-validation (LOOCV):** When  $k = n$
- **Stratified cross-validation**
  - When representation in each of the  $k$  random groups is proportional to the overall ratios
- Classification is called **supervised learning**
  - Algorithm or model is “supervised” by class information



# Over-Fitting and Under-Fitting

- Over-fitting

- Algorithm or model classifies the training set too well
- It is too complex or uses too many parameters
- Generally performs poorly with testing set
- Ends up modeling noise rather than data characteristics

# Over-Fitting and Under-Fitting

- **Over-fitting**

- Algorithm or model classifies the training set too well
- It is too complex or uses too many parameters
- Generally performs poorly with testing set
- Ends up modeling noise rather than data characteristics

- **Under-fitting**

- The opposite problem
- Algorithm or model does not classify the training set well at all
- It is too simple or uses too less parameters
- Generally performs poorly with testing set
- Ends up modeling overall data characteristics instead of per class

# Over-Fitting and Under-Fitting

- **Over-fitting**

- Algorithm or model classifies the training set too well
- It is too complex or uses too many parameters
- Generally performs poorly with testing set
- Ends up modeling noise rather than data characteristics

- **Under-fitting**

- The opposite problem
- Algorithm or model does not classify the training set well at all
- It is too simple or uses too less parameters
- Generally performs poorly with testing set
- Ends up modeling overall data characteristics instead of per class

- **Bias-variance tradeoff**

- **Bias** measures errors in the model learnt (under-fitting)
- **Variance** measures errors when training set is perturbed (over-fitting)
- Low bias generally implies higher variance and vice versa

# Errors

- **Positives** ( $P$ ): Objects that are “true” answers (for a class)
- **Negatives** ( $N$ ): Objects that are not answers:  $N = D - P$

# Errors

- **Positives** ( $P$ ): Objects that are “true” answers (for a class)
- **Negatives** ( $N$ ): Objects that are not answers:  $N = D - P$
- For a particular classification algorithm  $\mathcal{A}$ ,
  - $P'$ : Objects returned as positive by  $\mathcal{A}$
  - $N'$ : Objects not returned as positive by  $\mathcal{A}$

# Errors

- **Positives** ( $P$ ): Objects that are “true” answers (for a class)
- **Negatives** ( $N$ ): Objects that are not answers:  $N = D - P$
- For a particular classification algorithm  $\mathcal{A}$ ,
  - $P'$ : Objects returned as positive by  $\mathcal{A}$
  - $N'$ : Objects not returned as positive by  $\mathcal{A}$
- **True Positives** ( $TP$ ):

# Errors

- **Positives** ( $P$ ): Objects that are “true” answers (for a class)
- **Negatives** ( $N$ ): Objects that are not answers:  $N = D - P$
- For a particular classification algorithm  $\mathcal{A}$ ,
  - $P'$ : Objects returned as positive by  $\mathcal{A}$
  - $N'$ : Objects not returned as positive by  $\mathcal{A}$
- **True Positives** ( $TP$ ): Answers returned by  $\mathcal{A}$ :  $P \cap P'$

# Errors

- **Positives** ( $P$ ): Objects that are “true” answers (for a class)
- **Negatives** ( $N$ ): Objects that are not answers:  $N = D - P$
- For a particular classification algorithm  $\mathcal{A}$ ,
  - $P'$ : Objects returned as positive by  $\mathcal{A}$
  - $N'$ : Objects not returned as positive by  $\mathcal{A}$
- **True Positives** ( $TP$ ): Answers returned by  $\mathcal{A}$ :  $P \cap P'$
- **True Negatives** ( $TN$ ):



# Errors

- **Positives** ( $P$ ): Objects that are “true” answers (for a class)
- **Negatives** ( $N$ ): Objects that are not answers:  $N = D - P$
- For a particular classification algorithm  $\mathcal{A}$ ,
  - $P'$ : Objects returned as positive by  $\mathcal{A}$
  - $N'$ : Objects not returned as positive by  $\mathcal{A}$
- **True Positives** ( $TP$ ): Answers returned by  $\mathcal{A}$ :  $P \cap P'$
- **True Negatives** ( $TN$ ): Non-answers not returned by  $\mathcal{A}$ :  $N \cap N'$

# Errors

- **Positives** ( $P$ ): Objects that are “true” answers (for a class)
- **Negatives** ( $N$ ): Objects that are not answers:  $N = D - P$
- For a particular classification algorithm  $\mathcal{A}$ ,
  - $P'$ : Objects returned as positive by  $\mathcal{A}$
  - $N'$ : Objects not returned as positive by  $\mathcal{A}$
- **True Positives** ( $TP$ ): Answers returned by  $\mathcal{A}$ :  $P \cap P'$
- **True Negatives** ( $TN$ ): Non-answers not returned by  $\mathcal{A}$ :  $N \cap N'$
- **False Positives** ( $FP$ ):

# Errors

- **Positives** ( $P$ ): Objects that are “true” answers (for a class)
- **Negatives** ( $N$ ): Objects that are not answers:  $N = D - P$
- For a particular classification algorithm  $\mathcal{A}$ ,
  - $P'$ : Objects returned as positive by  $\mathcal{A}$
  - $N'$ : Objects not returned as positive by  $\mathcal{A}$
- **True Positives** ( $TP$ ): Answers returned by  $\mathcal{A}$ :  $P \cap P'$
- **True Negatives** ( $TN$ ): Non-answers not returned by  $\mathcal{A}$ :  $N \cap N'$
- **False Positives** ( $FP$ ): Non-answers returned by  $\mathcal{A}$ :  $N \cap P'$

# Errors

- **Positives ( $P$ )**: Objects that are “true” answers (for a class)
- **Negatives ( $N$ )**: Objects that are not answers:  $N = D - P$
- For a particular classification algorithm  $\mathcal{A}$ ,
  - $P'$ : Objects returned as positive by  $\mathcal{A}$
  - $N'$ : Objects not returned as positive by  $\mathcal{A}$
- **True Positives ( $TP$ )**: Answers returned by  $\mathcal{A}$ :  $P \cap P'$
- **True Negatives ( $TN$ )**: Non-answers not returned by  $\mathcal{A}$ :  $N \cap N'$
- **False Positives ( $FP$ )**: Non-answers returned by  $\mathcal{A}$ :  $N \cap P'$
- **False Negatives ( $FN$ )**:

# Errors

- **Positives ( $P$ )**: Objects that are “true” answers (for a class)
- **Negatives ( $N$ )**: Objects that are not answers:  $N = D - P$
- For a particular classification algorithm  $\mathcal{A}$ ,
  - $P'$ : Objects returned as positive by  $\mathcal{A}$
  - $N'$ : Objects not returned as positive by  $\mathcal{A}$
- **True Positives ( $TP$ )**: Answers returned by  $\mathcal{A}$ :  $P \cap P'$
- **True Negatives ( $TN$ )**: Non-answers not returned by  $\mathcal{A}$ :  $N \cap N'$
- **False Positives ( $FP$ )**: Non-answers returned by  $\mathcal{A}$ :  $N \cap P'$
- **False Negatives ( $FN$ )**: Answers not returned by  $\mathcal{A}$ :  $P \cap N'$

# Errors

- **Positives ( $P$ )**: Objects that are “true” answers (for a class)
- **Negatives ( $N$ )**: Objects that are not answers:  $N = D - P$
- For a particular classification algorithm  $\mathcal{A}$ ,
  - $P'$ : Objects returned as positive by  $\mathcal{A}$
  - $N'$ : Objects not returned as positive by  $\mathcal{A}$
- **True Positives ( $TP$ )**: Answers returned by  $\mathcal{A}$ :  $P \cap P'$
- **True Negatives ( $TN$ )**: Non-answers not returned by  $\mathcal{A}$ :  $N \cap N'$
- **False Positives ( $FP$ )**: Non-answers returned by  $\mathcal{A}$ :  $N \cap P'$
- **False Negatives ( $FN$ )**: Answers not returned by  $\mathcal{A}$ :  $P \cap N'$

$$P = TP \cup FN \quad N = TN \cup FP \quad P' = TP \cup FP \quad N' = TN \cup FN$$

# Errors

- **Positives ( $P$ )**: Objects that are “true” answers (for a class)
- **Negatives ( $N$ )**: Objects that are not answers:  $N = D - P$
- For a particular classification algorithm  $\mathcal{A}$ ,
  - $P'$ : Objects returned as positive by  $\mathcal{A}$
  - $N'$ : Objects not returned as positive by  $\mathcal{A}$
- **True Positives ( $TP$ )**: Answers returned by  $\mathcal{A}$ :  $P \cap P'$
- **True Negatives ( $TN$ )**: Non-answers not returned by  $\mathcal{A}$ :  $N \cap N'$
- **False Positives ( $FP$ )**: Non-answers returned by  $\mathcal{A}$ :  $N \cap P'$
- **False Negatives ( $FN$ )**: Answers not returned by  $\mathcal{A}$ :  $P \cap N'$

$$P = TP \cup FN \quad N = TN \cup FP \quad P' = TP \cup FP \quad N' = TN \cup FN$$

- In statistics,
  - **Type I error**:  $FP$
  - **Type II error**:  $FN$

# Confusion Matrix

- **Confusion matrix** visually represents the information
- Rows indicate “true” answers:  $P$  and  $N$
- Columns indicate those returned by  $\mathcal{A}$ :  $P'$  and  $N'$
- Shows which error is more

Sets		Returned by $\mathcal{A}$	
		Positives $P'$	Negatives $N'$
True answers	Positives $P$	$TP$	$FN$
	Negatives $N$	$FP$	$TN$



# Confusion Matrix for Multiple Classes

- Is more useful when extended for multiple classes
- Shows which classes are confused more against which other classes

Sets		Predicted by $\mathcal{A}$		
		Class $C'_1$	Class $C'_2$	Class $C'_3$
True answers	Class $C_1$	5	3	0
	Class $C_2$	2	3	1
	Class $C_3$	0	2	9

# Error Parameters or Performance Metrics

<i>Parameter</i>	<i>Interpretation</i>	<i>Formula</i>
Precision	Proportion of positives in those returned by $\mathcal{A}$	

# Error Parameters or Performance Metrics

<i>Parameter</i>	<i>Interpretation</i>	<i>Formula</i>
Precision	Proportion of positives in those returned by $\mathcal{A}$	$\frac{ TP }{ TP \cup FP } = \frac{ TP }{ P' }$
Recall or Sensitivity or True positive rate or Hit rate	Proportion of positives returned by $\mathcal{A}$	

# Error Parameters or Performance Metrics

<i>Parameter</i>	<i>Interpretation</i>	<i>Formula</i>
Precision	Proportion of positives in those returned by $\mathcal{A}$	$\frac{ TP }{ TP \cup FP } = \frac{ TP }{ P' }$
Recall or Sensitivity or True positive rate or Hit rate	Proportion of positives returned by $\mathcal{A}$	$\frac{ TP }{ TP \cup FN } = \frac{ TP }{ P }$
Specificity or True negative rate	Proportion of negatives not returned by $\mathcal{A}$	

# Error Parameters or Performance Metrics

<i>Parameter</i>	<i>Interpretation</i>	<i>Formula</i>
Precision	Proportion of positives in those returned by $\mathcal{A}$	$\frac{ TP }{ TP \cup FP } = \frac{ TP }{ P' }$
Recall or Sensitivity or True positive rate or Hit rate	Proportion of positives returned by $\mathcal{A}$	$\frac{ TP }{ TP \cup FN } = \frac{ TP }{ P }$
Specificity or True negative rate	Proportion of negatives not returned by $\mathcal{A}$	$\frac{ TN }{ TN \cup FP } = \frac{ TN }{ N }$
False positive rate	Proportion of negatives returned by $\mathcal{A}$	

# Error Parameters or Performance Metrics

<i>Parameter</i>	<i>Interpretation</i>	<i>Formula</i>
Precision	Proportion of positives in those returned by $\mathcal{A}$	$\frac{ TP }{ TP \cup FP } = \frac{ TP }{ P' }$
Recall or Sensitivity or True positive rate or Hit rate	Proportion of positives returned by $\mathcal{A}$	$\frac{ TP }{ TP \cup FN } = \frac{ TP }{ P }$
Specificity or True negative rate	Proportion of negatives not returned by $\mathcal{A}$	$\frac{ TN }{ TN \cup FP } = \frac{ TN }{ N }$
False positive rate	Proportion of negatives returned by $\mathcal{A}$	$\frac{ FP }{ TN \cup FP } = \frac{ FP }{ N }$
False negative rate	Proportion of positives not returned by $\mathcal{A}$	

# Error Parameters or Performance Metrics

<i>Parameter</i>	<i>Interpretation</i>	<i>Formula</i>
Precision	Proportion of positives in those returned by $\mathcal{A}$	$\frac{ TP }{ TP \cup FP } = \frac{ TP }{ P' }$
Recall or Sensitivity or True positive rate or Hit rate	Proportion of positives returned by $\mathcal{A}$	$\frac{ TP }{ TP \cup FN } = \frac{ TP }{ P }$
Specificity or True negative rate	Proportion of negatives not returned by $\mathcal{A}$	$\frac{ TN }{ TN \cup FP } = \frac{ TN }{ N }$
False positive rate	Proportion of negatives returned by $\mathcal{A}$	$\frac{ FP }{ TN \cup FP } = \frac{ FP }{ N }$
False negative rate	Proportion of positives not returned by $\mathcal{A}$	$\frac{ FN }{ TP \cup FN } = \frac{ FN }{ P }$
Accuracy	Proportion of positives returned and negatives not returned by $\mathcal{A}$	

# Error Parameters or Performance Metrics

<i>Parameter</i>	<i>Interpretation</i>	<i>Formula</i>
Precision	Proportion of positives in those returned by $\mathcal{A}$	$\frac{ TP }{ TP \cup FP } = \frac{ TP }{ P' }$
Recall or Sensitivity or True positive rate or Hit rate	Proportion of positives returned by $\mathcal{A}$	$\frac{ TP }{ TP \cup FN } = \frac{ TP }{ P }$
Specificity or True negative rate	Proportion of negatives not returned by $\mathcal{A}$	$\frac{ TN }{ TN \cup FP } = \frac{ TN }{ N }$
False positive rate	Proportion of negatives returned by $\mathcal{A}$	$\frac{ FP }{ TN \cup FP } = \frac{ FP }{ N }$
False negative rate	Proportion of positives not returned by $\mathcal{A}$	$\frac{ FN }{ TP \cup FN } = \frac{ FN }{ P }$
Accuracy	Proportion of positives returned and negatives not returned by $\mathcal{A}$	$\frac{ TP \cup TN }{ D }$
Error rate	Proportion of positives not returned and negatives returned by $\mathcal{A}$	



# Error Parameters or Performance Metrics

<i>Parameter</i>	<i>Interpretation</i>	<i>Formula</i>
Precision	Proportion of positives in those returned by $\mathcal{A}$	$\frac{ TP }{ TP \cup FP } = \frac{ TP }{ P' }$
Recall or Sensitivity or True positive rate or Hit rate	Proportion of positives returned by $\mathcal{A}$	$\frac{ TP }{ TP \cup FN } = \frac{ TP }{ P }$
Specificity or True negative rate	Proportion of negatives not returned by $\mathcal{A}$	$\frac{ TN }{ TN \cup FP } = \frac{ TN }{ N }$
False positive rate	Proportion of negatives returned by $\mathcal{A}$	$\frac{ FP }{ TN \cup FP } = \frac{ FP }{ N }$
False negative rate	Proportion of positives not returned by $\mathcal{A}$	$\frac{ FN }{ TP \cup FN } = \frac{ FN }{ P }$
Accuracy	Proportion of positives returned and negatives not returned by $\mathcal{A}$	$\frac{ TP \cup TN }{ D }$
Error rate	Proportion of positives not returned and negatives returned by $\mathcal{A}$	$\frac{ FP \cup FN }{ D }$

# Single Measures

- Single measures capturing both precision and recall

# Single Measures

- Single measures capturing both precision and recall
- **F-score** or **F-measure** or **F1-score** is the *harmonic mean* of precision and recall

$$Fscore = \frac{2.Precision.Recall}{Precision + Recall}$$

# Single Measures

- Single measures capturing both precision and recall
- **F-score** or **F-measure** or **F1-score** is the *harmonic mean* of precision and recall

$$Fscore = \frac{2.Precision.Recall}{Precision + Recall}$$

- In terms of errors

$$Fscore = \frac{2.TP}{2.TP + FN + FP}$$

# Single Measures

- Single measures capturing both precision and recall
- **F-score** or **F-measure** or **F1-score** is the *harmonic mean* of precision and recall

$$Fscore = \frac{2.Precision.Recall}{Precision + Recall}$$

- In terms of errors

$$Fscore = \frac{2.TP}{2.TP + FN + FP}$$

- Similarly, **G-measure** is *geometric mean* of precision and recall

# Single Measures

- Single measures capturing both precision and recall
- **F-score** or **F-measure** or **F1-score** is the *harmonic mean* of precision and recall

$$Fscore = \frac{2.Precision.Recall}{Precision + Recall}$$

- In terms of errors

$$Fscore = \frac{2.TP}{2.TP + FN + FP}$$

- Similarly, **G-measure** is *geometric mean* of precision and recall
- **EER** or **Equal Error Rate** is when FP rate is equal to FN rate

# Weighting Precision versus Recall

- Suppose recall and precision are weighted at a ratio  $\alpha : (1 - \alpha)$
- F-score is the *weighted harmonic mean*

$$\frac{1}{F} = \alpha \cdot \frac{1}{P} + (1 - \alpha) \cdot \frac{1}{R}$$

# Weighting Precision versus Recall

- Suppose recall and precision are weighted at a ratio  $\alpha : (1 - \alpha)$
- F-score is the *weighted harmonic mean*

$$\frac{1}{F} = \alpha \cdot \frac{1}{P} + (1 - \alpha) \cdot \frac{1}{R}$$

- $\beta^2 = \frac{1-\alpha}{\alpha}$  measures the relative importance of precision over recall
  - $\alpha \in [0, 1]$  while  $\beta \in [0, \infty]$
  - $\beta > 1$  emphasizes precision, while  $\beta < 1$  emphasizes recall



# Weighting Precision versus Recall

- Suppose recall and precision are weighted at a ratio  $\alpha : (1 - \alpha)$
- F-score is the *weighted harmonic mean*

$$\frac{1}{F} = \alpha \cdot \frac{1}{P} + (1 - \alpha) \cdot \frac{1}{R}$$

- $\beta^2 = \frac{1-\alpha}{\alpha}$  measures the relative importance of precision over recall
  - $\alpha \in [0, 1]$  while  $\beta \in [0, \infty]$
  - $\beta > 1$  emphasizes precision, while  $\beta < 1$  emphasizes recall
- Using  $\beta^2$ , **weighted F-score** is

$$F = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}$$

- When  $\beta = 1$ , precision and recall are equally weighted ( $\alpha = 1/2$ )
- **F1-score** is the harmonic mean

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

Correct answer set =  $\{O_1, O_5, O_7\}$

Algorithm returns =  $\{O_1, O_3, O_5, O_6\}$

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P =$$

$$N =$$

$$TP =$$

$$TN =$$

$$FP =$$

$$FN =$$

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P = \{O_1, O_5, O_7\}$$

$$N =$$

$$TP =$$

$$TN =$$

$$FP =$$

$$FN =$$

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P = \{O_1, O_5, O_7\}$$

$$N = \{O_2, O_3, O_4, O_6, O_8\}$$

$$TP =$$

$$TN =$$

$$FP =$$

$$FN =$$

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P = \{O_1, O_5, O_7\}$$

$$N = \{O_2, O_3, O_4, O_6, O_8\}$$

$$TP = \{O_1, O_5\}$$

$$TN =$$

$$FP =$$

$$FN =$$

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P = \{O_1, O_5, O_7\}$$

$$N = \{O_2, O_3, O_4, O_6, O_8\}$$

$$TP = \{O_1, O_5\}$$

$$TN = \{O_2, O_4, O_8\}$$

$$FP =$$

$$FN =$$

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P = \{O_1, O_5, O_7\}$$

$$N = \{O_2, O_3, O_4, O_6, O_8\}$$

$$TP = \{O_1, O_5\}$$

$$TN = \{O_2, O_4, O_8\}$$

$$FP = \{O_3, O_6\}$$

$$FN =$$



# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P = \{O_1, O_5, O_7\}$$

$$N = \{O_2, O_3, O_4, O_6, O_8\}$$

$$TP = \{O_1, O_5\}$$

$$TN = \{O_2, O_4, O_8\}$$

$$FP = \{O_3, O_6\}$$

$$FN = \{O_7\}$$

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P = \{O_1, O_5, O_7\}$$

$$N = \{O_2, O_3, O_4, O_6, O_8\}$$

$$TP = \{O_1, O_5\}$$

$$TN = \{O_2, O_4, O_8\}$$

$$FP = \{O_3, O_6\}$$

$$FN = \{O_7\}$$

$$\therefore \text{Recall} = \text{Sensitivity} =$$

$$\text{Precision} =$$

$$\text{Specificity} =$$

$$\text{F-score} =$$

$$\text{Accuracy} =$$

$$\text{Error rate} =$$

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P = \{O_1, O_5, O_7\}$$

$$N = \{O_2, O_3, O_4, O_6, O_8\}$$

$$TP = \{O_1, O_5\}$$

$$TN = \{O_2, O_4, O_8\}$$

$$FP = \{O_3, O_6\}$$

$$FN = \{O_7\}$$

$$\therefore \text{Recall} = \text{Sensitivity} = 2/3 = 0.67$$

$$\text{Precision} =$$

$$\text{Specificity} =$$

$$\text{F-score} =$$

$$\text{Accuracy} =$$

$$\text{Error rate} =$$

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P = \{O_1, O_5, O_7\}$$

$$N = \{O_2, O_3, O_4, O_6, O_8\}$$

$$TP = \{O_1, O_5\}$$

$$TN = \{O_2, O_4, O_8\}$$

$$FP = \{O_3, O_6\}$$

$$FN = \{O_7\}$$

$$\therefore \text{Recall} = \text{Sensitivity} = 2/3 = 0.67$$

$$\text{Precision} = 2/4 = 0.5$$

$$\text{Specificity} =$$

$$\text{F-score} =$$

$$\text{Accuracy} =$$

$$\text{Error rate} =$$

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P = \{O_1, O_5, O_7\}$$

$$N = \{O_2, O_3, O_4, O_6, O_8\}$$

$$TP = \{O_1, O_5\}$$

$$TN = \{O_2, O_4, O_8\}$$

$$FP = \{O_3, O_6\}$$

$$FN = \{O_7\}$$

$$\therefore \text{Recall} = \text{Sensitivity} = 2/3 = 0.67$$

$$\text{Precision} = 2/4 = 0.5$$

$$\text{Specificity} = 3/5 = 0.6$$

$$\text{F-score} =$$

$$\text{Accuracy} =$$

$$\text{Error rate} =$$

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P = \{O_1, O_5, O_7\}$$

$$N = \{O_2, O_3, O_4, O_6, O_8\}$$

$$TP = \{O_1, O_5\}$$

$$TN = \{O_2, O_4, O_8\}$$

$$FP = \{O_3, O_6\}$$

$$FN = \{O_7\}$$

$$\therefore \text{Recall} = \text{Sensitivity} = 2/3 = 0.67$$

$$\text{Precision} = 2/4 = 0.5$$

$$\text{Specificity} = 3/5 = 0.6$$

$$\text{F-score} = 4/7 = 0.571$$

$$\text{Accuracy} =$$

$$\text{Error rate} =$$

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P = \{O_1, O_5, O_7\}$$

$$N = \{O_2, O_3, O_4, O_6, O_8\}$$

$$TP = \{O_1, O_5\}$$

$$TN = \{O_2, O_4, O_8\}$$

$$FP = \{O_3, O_6\}$$

$$FN = \{O_7\}$$

$$\therefore \text{Recall} = \text{Sensitivity} = 2/3 = 0.67$$

$$\text{Precision} = 2/4 = 0.5$$

$$\text{Specificity} = 3/5 = 0.6$$

$$\text{F-score} = 4/7 = 0.571$$

$$\text{Accuracy} = 5/8 = 0.625$$

$$\text{Error rate} =$$

# Example

$$D = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

$$\text{Correct answer set} = \{O_1, O_5, O_7\}$$

$$\text{Algorithm returns} = \{O_1, O_3, O_5, O_6\}$$

$$\therefore P = \{O_1, O_5, O_7\}$$

$$N = \{O_2, O_3, O_4, O_6, O_8\}$$

$$TP = \{O_1, O_5\}$$

$$TN = \{O_2, O_4, O_8\}$$

$$FP = \{O_3, O_6\}$$

$$FN = \{O_7\}$$

$$\therefore \text{Recall} = \text{Sensitivity} = 2/3 = 0.67$$

$$\text{Precision} = 2/4 = 0.5$$

$$\text{Specificity} = 3/5 = 0.6$$

$$\text{F-score} = 4/7 = 0.571$$

$$\text{Accuracy} = 5/8 = 0.625$$

$$\text{Error rate} = 3/8 = 0.375$$

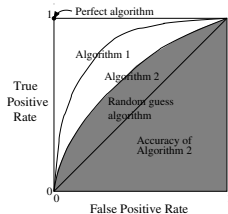
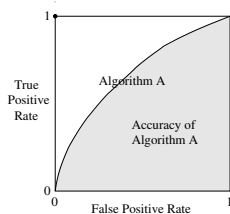


# ROC Curve

- Performance of an algorithm depends on parameters
- To assess over a range of parameters, ROC curve is used
  - 1 - Specificity (x-axis) versus Sensitivity (y-axis)
  - False positive rate (x-axis) versus True positive rate (y-axis)

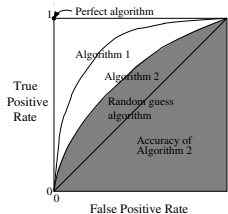
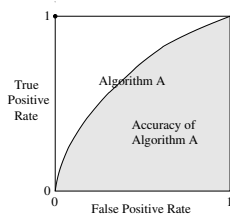
# ROC Curve

- Performance of an algorithm depends on parameters
- To assess over a range of parameters, **ROC** curve is used
  - 1 - Specificity (x-axis) versus Sensitivity (y-axis)
  - False positive rate (x-axis) versus True positive rate (y-axis)
- A random guess algorithm is a 45° line



# ROC Curve

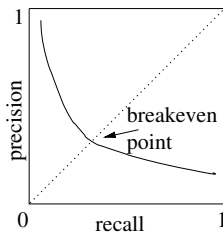
- Performance of an algorithm depends on parameters
- To assess over a range of parameters, **ROC** curve is used
  - 1 - Specificity (x-axis) versus Sensitivity (y-axis)
  - False positive rate (x-axis) versus True positive rate (y-axis)
- A random guess algorithm is a 45° line



- Area under the ROC curve (**AUC** or **AUROC**) measures **accuracy** (or **discrimination**)
- What AUC is good?
  - 0.9+: excellent; 0.8+: good; 0.7+: fair; 0.6+: poor; 0.6-: fail
- **EER** denotes the point in ROC where FP rate is equal to FN rate

# Precision-Recall Curve

- Precision versus recall



- Breakeven point** where precision is the same as recall