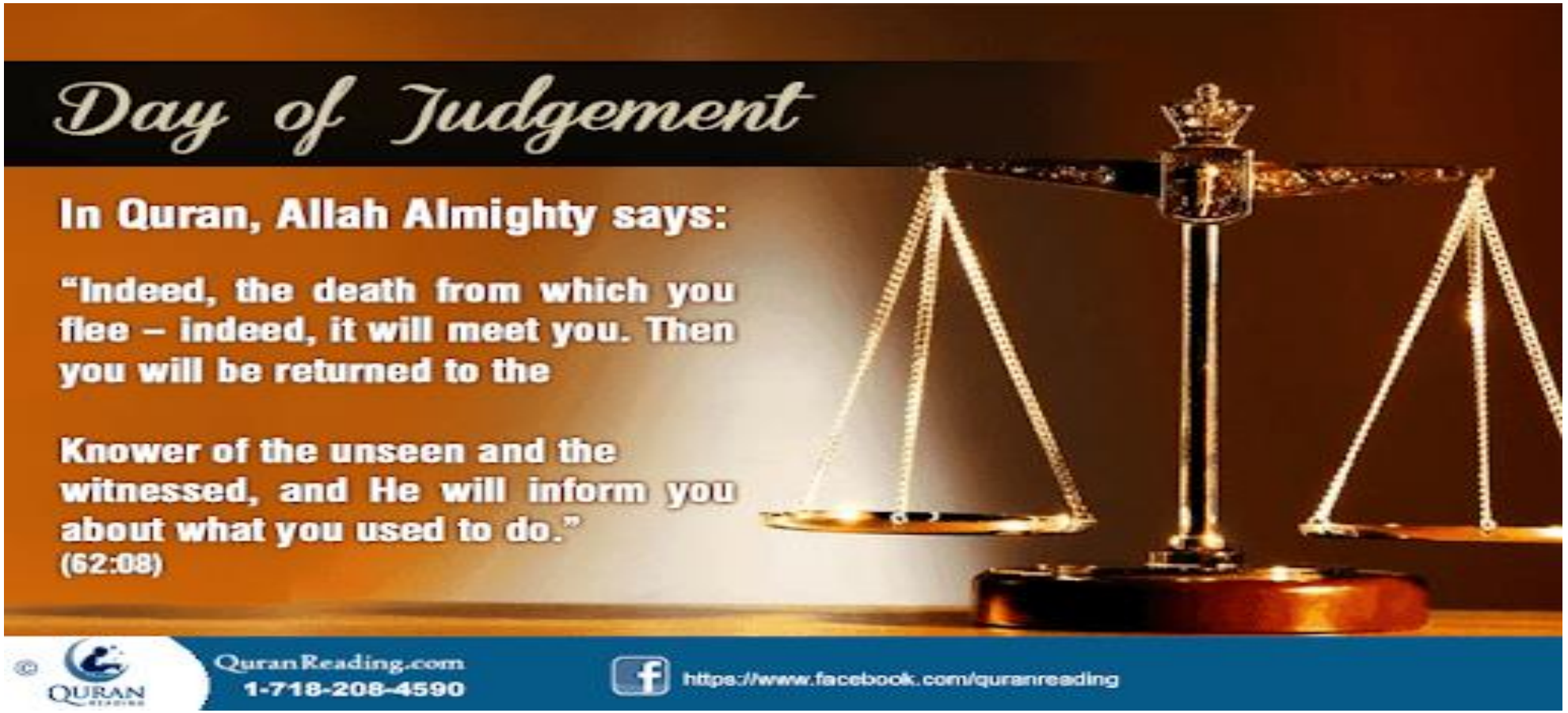
Data Mining

SUPERVISED LEARNING PERFORMANCE EVALUATION MEASURES

Prof. Dr. Hikmat Ullah Khan
Department of Information Technology

UNIVERSITY OF SARGODHA, SARGODHA

Lesson from Holy Quran



Naïve Bayes revisited

- Have you computed any other example of Naïve Bayes?
  - Online Source
  - Synthetic Example
- Have you checked how to implement the Naïve Bayes?
- Have you checked how to run Naïve Bayes in
  - WEKA
  - C#
  - R
  - Python
  - RapidMiner/any other toll you are working on. ?

Try it Again?

- problem: identify people as good or bad from their appearance

	sex	mask	cape	tie	ears	smokes	class
			training data				
batman	male	yes	yes	no	yes	no	Good
robin	male	yes	yes	no	no	no	Good
alfred	male	no	no	yes	no	no	Good
penguin	male	no	no	yes	no	yes	Bad
catwoman	female	yes	no	no	yes	no	Bad
joker	male	no	no	no	no	no	Bad
			test data				
batgirl	female	yes	yes	no	yes	no	??
riddler	male	yes	no	no	no	no	??

Evaluation

- HOW
  - Comparison b/w Human and Method
  - Method is simple
  - Method is intuitive

Confusion Matrix

- Confusion matrix (confuses or combines actual vs predictions)
- All Performance Evaluation Measures are based on this simple Matric

Prediction based on Algorithms

Predicted Values

Actual Values (Ground Truth)	C_1	$\neg C_1$
C_1	**True Positives (TP)**	**False Negatives (FN)**
$\neg C_1$	**False Positives (FP)**	**True Negatives (TN)**

Learning how Confusion Matrix is Prepared

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0
Result	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN

Confusion Matrix

Confusion Matrix:

Actual Values (Ground Truth)	Predicted Values: C_1	Predicted Values: $\neg C_1$
C_1	**True Positives (TP)**	**False Negatives (FN)**
$\neg C_1$	**False Positives (FP)**	**True Negatives (TN)**

Example of Confusion Matrix:

Actual Values (Ground Truth)	Predicted Values: buy_computer = yes	Predicted Values: buy_computer = no	Total
buy_computer = yes	**TP=6954**	**FN=46**	7000
buy_computer = no	**FP=412**	**TN=2588**	3000
Total	7366	2634	10000

Confusion Matrix

Actual Values (Ground Truth)	Predicted Values: buy_computer = yes	Predicted Values: buy_computer = no	Total
buy_computer = yes	TP=6954	FN=46	7000
buy_computer = no	FP=412	TN=2588	3000
Total	7366	2634	10000

Can you answer these questions:

a) How many computers were bought?

b) What is the value of prediction about sale of computers?

c) What is the numbers of computers which were actually bought and the algorithm also predicted it correctly?

d) What is the total number of instances in the data?

Classifier Evaluation Metrics: Accuracy,

- **Classifier Accuracy,** or recognition rate:
- percentage of test set tuples that are correctly classified

Accuracy = (TP + TN)/TP+TN+FP+FN

OR **Accuracy = (TP + TN)/All**

- **Error rate:** *1 – accuracy*, or

Error rate = (FP + FN)/All

Compute Accuracy and Error

Actual Values (Ground Truth)	Predicted Values: C_1	Predicted Values: $\neg C_1$
C_1	**True Positives (TP)**	**False Negatives (FN)**
$\neg C_1$	**False Positives (FP)**	**True Negatives (TN)**

Example of Confusion Matrix:

Actual Values (Ground Truth)	Predicted Values: buy_computer = yes	Predicted Values: buy_computer = no	Total
buy_computer = yes	**TP=6954**	**FN=46**	7000
buy_computer = no	**FP=412**	**TN=2588**	3000
Total	7366	2634	10000

Compute Accuracy and Error

Predicted Values

Actual Values (Ground Truth)	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	TP=6954	FN=46	7000
buy_computer = no	FP=412	TN=2588	3000
Total	7366	2634	10000

Accuracy = 6954+2588 / 10000 = 0.95

The values will always be between 0 and 1.

Usually shown in terms of Percentage (95%)

Error = 1 – 0.95 = 0.05

Classifier Evaluation Metrics: Precision and Recall

- **Precision**: The ratio of correctly classified (patients with the disease) (*TP*) to the total patients predicted to have the disease (*TP+FP*).

$$precision = \frac{TP}{TP + FP}$$

- *Precision = 6954 / 6954 +412 = 6954/7366*

Predicted Values

Actual Values (Ground Truth)	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	TP=6954	FN=46	7000
buy_computer = no	FP=412	TN=2588	3000
Total	7366	2634	10000

Classifier Evaluation Metrics: Precision and Recall

- **Recall:** the ratio of correctly classified diseased patients (*TP*) divided by total number of patients who have actually the disease.
- Perfect score is 1.0
- Inverse relationship between precision & recall

$$recall = \frac{TP}{TP + FN}$$

- *Recall = 6954/ (6954 +46)= 6954 /7000*

Predicted Values

Actual Values (Ground Truth)	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	TP=6954	FN=46	7000
buy_computer = no	FP=412	TN=2588	3000
Total	7366	2634	10000

Classifier Evaluation Metrics: F-measures

- ***F* measure (*F_1* or *F*-score)**:
- harmonic mean of precision and recall,

$$F_1 = \frac{2pr}{p + r}$$

F_1-score is the harmonic mean of precision and recall.

$$F_1 = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

Precision vs Recall vs F-Measure (Example)

- Compute
- Precision, Recall , and F-Measure
- and
- then Compare the results

	Recurrence	No Recurrence
Recurrence	10	13
No Recurrence	75	188
	85	201

Do we need another Performance Evaluation Measures?

	Actual Positive	Acutal Negative
Predicted Positive	10	25
Predicted Negative	15	100

Compute Accuracy

	Actual Positive	Acutal Negative
Predicted Positive	0	25
Predicted Negative	0	125

Do we need another Performance Evaluation Measures?

	Actual Positive	Acutal Negative
Predicted Positive	10	25
Predicted Negative	15	100

Compute Accuracy

Classifier Evaluation Metrics: Sensitivity and Specificity

- **Class Imbalance Problem**:
  - One class may be *rare*,

    e.g. fraud, or HIV-positive
  - Significant *majority of the negative class* and minority of the positive class
  - **Sensitivity**: True Positive Rate (TPR or ?)
    - **Sensitivity = TP/(TP + FN)**
  - **Specificity**: True Negative Rate (TNR)
    - **Specificity = TN/(TN + FP)**

Receive operating characteristics curve

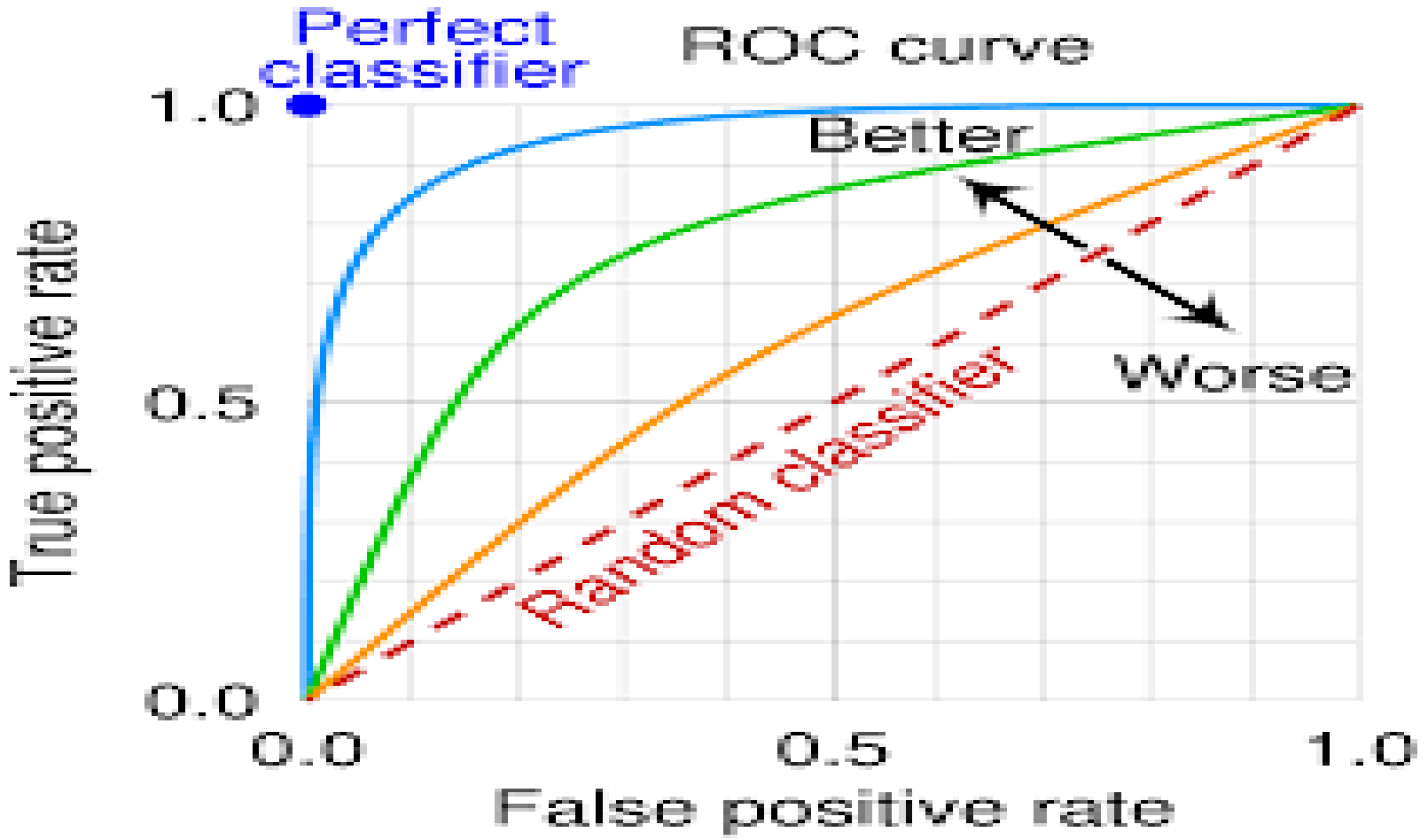
- It is commonly called the ROC curve.
- It is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR).
- True positive rate:
  - Ratio of What is True Positive w.r..t Positive

$$TPR = \frac{TP}{TP + FN}$$

- False positive rate (or Fallout):
  - Ratio of Fasle Positive w.r.t Negative

$$FPR = \frac{FP}{TN + FP}$$

Understanding ROC



Example ROC curves

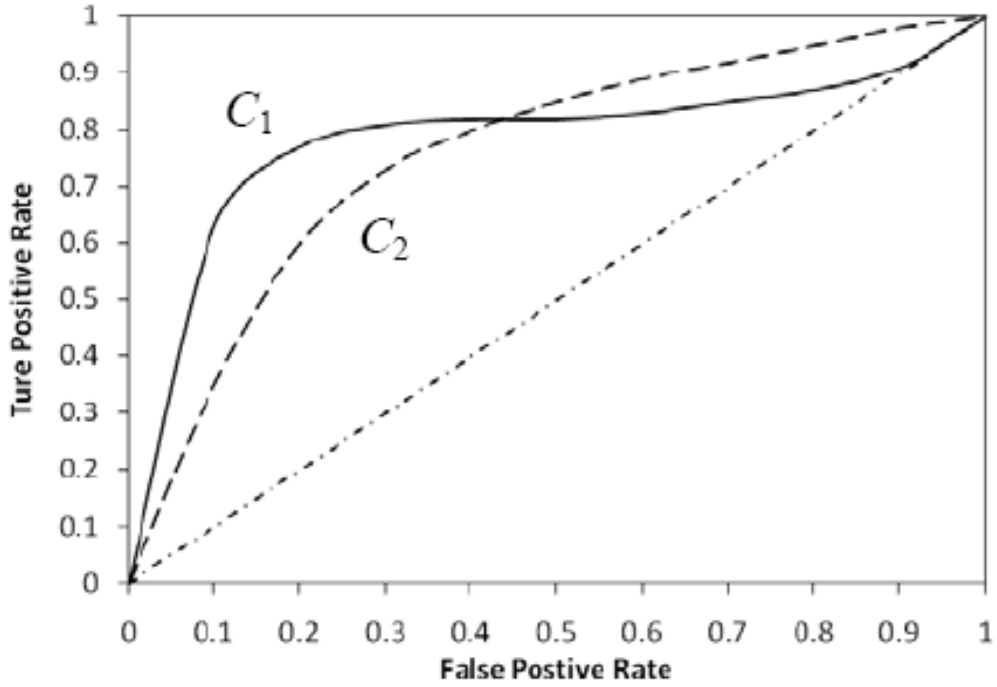


Fig. 3.8. ROC curves for two classifiers (C_1 and C_2) on the same data

Area under the curve (AUC)

- Which classifier is better, C_1 or C_2?
  - It depends on which region you talk about.
- Can we have one measure?
  - Yes, we compute the area under the curve (AUC)
- If AUC for C_i is greater than that of C_j, it is said that C_i is better than C_j.
  - If a classifier is perfect, its AUC value is 1
  - If a classifier makes all random guesses, its AUC value is 0.5.

Want to learn more

- Want to Compute Online

- There are many online Sources of Computation of these Measures
- http://www.marcovanetti.com/pages/cfmatrix/

- http://onlineconfusionmatrix.com/
- Learn Confusion Matric from Wikipedia
- https://en.wikipedia.org/wiki/Confusion_matrix


“The future belongs to those who prepare for it today”
Malcolm X