CS396: Selected CS2

(Deep Learning for visual recognition)

Spring 2022

Dr. Wessam EL-Behaidy

Associate Professor, Computer Science Department,
Faculty of Computers and Artificial Intelligence,
Helwan University.

Lectures (Course slides) are based on Stanford course :
Convolutional Neural Networks for Visual Recognition (CS231n):
http://cs231n.stanford.edu/index.html

Lecture 10:
Classifier Performance Metrics

Accuracy in a Classification Model

Accuracy is measured as the percentage of predicted results that match the expected results.

Ex: if there are 1000 results and 850 predicted results match the expected results, then the accuracy is 85%

Problem with accuracy metric (measure):

Skewed classes basically refer to a dataset, wherein the number of training examples belonging to one class outnumbers heavily the number of training examples belonging to the other.

Consider a binary classification (cancer is labeled 1 and not cancer labeled 0), where a cancerous patient is to be detected based on some features.
- only 1 % of the data provided has cancer positive.
If a system naively gives the prediction as all 0’s, still the prediction accuracy will be 99%.

Problem with accuracy metric (measure):

Irrelevant features

The data may be fitted against a feature that is not relevant.

Ex:
In image classification, if all images of one class have small/similar background, the model may match based on the background, not the object in the image.

Commonly used Metrics

Accuracy is only one metric.

Other metrics commonly used are:

- Precision
- Recall (Sensitivity)
- Specificity
- F1-score
- ROC AUC

Confusion Matrix

- The confusion matrix is a performance measurement technique that visualizes the accuracy of a classifier by comparing the actual and predicted classes.
- It is called a confusion matrix because it shows how confused the model is between the classes.
- The class of interest is commonly called **the positive class**, and the rest **negative class**

	Predicted (False)	Predicted (True)
Actual (False)	True Negative (TN)	False Positive (FP)
Actual (True)	False Negative (FN)	True Positive (TP)

Binary Confusion Matrix

		-	+
		Predicted (False)	Predicted (True)
-	Actual (False)	True Negative (TN)	False Positive (FP)
+	Actual (True)	False Negative (FN)	True Positive (TP)

Interested class is Class True

- True Positive (TP) : The outcome is correctly classified as positive.
- False Negative (FN) : The outcome is incorrectly classified as negative when it is positive.
- False positive (FP): The outcome is incorrectly classified as positive when it is negative.
- True Negative (TN): The outcome is correctly classified as negative.

Example of Confusion Matrix

Predicted Label

True Label		Daisy	Tulip
	Daisy	9	1
	Tulip	2	8

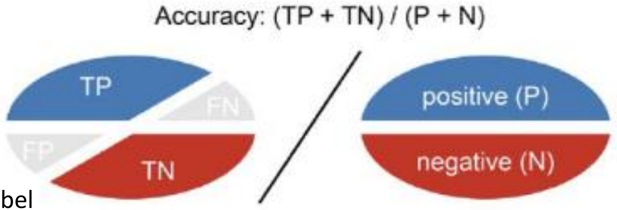
If class "Daisy" is the positive class, so:

TP=9 FN=1

FP=2 TN=8

Accuracy

- Accuracy is calculated as the number of all correct predictions divided by the total number of the dataset
- The best ACC is 1.0, whereas the worst is 0.0



Interested class is
Class 0

Predicted Label (columns), True Label (rows)

	0	1
0	9	1
1	2	8

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

- Accuracy = (9+8) / (9+2+1+8) = 0.85

Precision

- Precision is calculated as the number of correct positive predictions divided by the total number of positive predictions
- The best precision is 1.0, whereas the worst is 0.0

Precision: TP / (TP + FP)

Predicted Label

True Label		0	1
	0	9	1
	1	2	8

$$\text{Precision} = \frac{tp}{tp + fp}$$

- Precision = 9 / (9+2) = 0.81

Recall

- Sensitivity = Recall = True Positive Rate
- Recall is calculated as the number of correct positive predictions divided by the total number of positives
- The best recall is 1.0, whereas the worst is 0.0

Sensitivity: TP / P

Predicted Label

True Label

	0	1
0	9	1
1	2	8

$$\text{Recall} = \frac{tp}{tp + fn}$$

- Recall = 9 / (9+1) = 0.9

Specificity

Specificity= True Negative Rate

Specificity is calculated as the number of correct negative predictions divided by the total number of negatives

Specificity: TN / N

Predicted Label

True Label	0	1
0	9	1
1	2	8

$$\text{Specificity} = \frac{tn}{tn+fp}$$

- Specificity= 8/(8+2)= 0.8

F1-Score

F1-score also known as **F-score** and **F-measure**

F1 Score considers both precision and recall.

It is the harmonic mean (average) of the precision and recall.

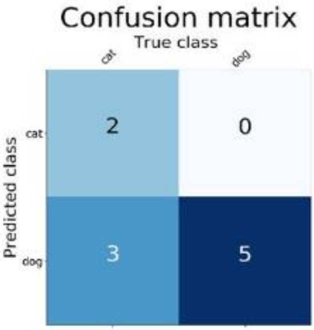
F1 Score is best if there is some sort of balance between precision (p) & recall (r) in the system.

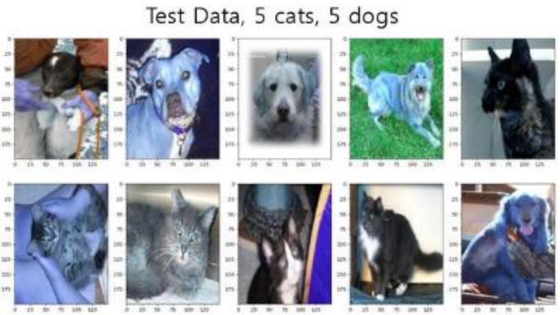
Oppositely F1 Score isn't so high if one measure is improved at the expense of the other. For example, *if Precision is 1 & Recall is 0, F1 score is 0.*

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Example 1

- **Example: The example to classify whether images contain either a dog or a cat**
  - The training data contains 25000 images of dogs and cats;
  - The training data 75% of 25000 images; (25000*0.75 = 18750)
  - Validation data 25% of training data; (25000*0.25 = 6250)





- **Precision** = 2/(2 + 0) * 100% = 100%
- **Recall** = 2/(2 + 3) * 100% = 40%
- **Accuracy** = (2 + 5)/(2 + 0 + 3 + 5) * 100% = 70%

Confusion Matrix with Three Classes

True Positives

		Predicted class		
		Apple	Orange	Pear
Actual class	Apple	50	5	50
	Orange	10	50	20
	Pear	5	5	0

The model correctly classified 50 apples and 50 oranges.

True Negatives for Class Apple

		Predicted class		
		Apple	Orange	Pear
	Apple	50	5	50
Actual class	Orange	10	50	20
	Pear	5	5	0

The model correctly classified 75 cases as not belonging to class apple.

True Negatives for Class Orange

		Predicted class		
		Apple	Orange	Pear
Actual class	Apple	50	5	50
	Orange	10	50	20
	Pear	5	5	0

The model correctly classified 105 cases as not belonging to class orange.

True Negatives for Class Pear

		Predicted class		
		Apple	Orange	Pear
Actual class	Apple	50	5	50
	Orange	10	50	20
	Pear	5	5	0

The model correctly classified 115 cases as not belonging to class pear.

False Positives of Class Apple

		Predicted class	
	Apple	Orange	Pear
Actual class: Apple	50	5	50
Actual class: Orange	10	50	20
Actual class: Pear	5	5	0

The model incorrectly classified 15 cases as apples.

False Positives of Class Orange

		Predicted class		
		Apple	Orange	Pear
	Apple	50	5	50
Actual class	Orange	10	50	20
	Pear	5	5	0

The model incorrectly classified 10 cases as oranges.

False Positives of Class Pear

		Predicted class		
		Apple	Orange	Pear
Actual class	Apple	50	5	50
	Orange	10	50	20
	Pear	5	5	0

The model incorrectly classified 70 cases as pears.

False Negatives of Class Apple

		Predicted class		
		Apple	Orange	Pear
Actual class	Apple	50	5	50
	Orange	10	50	20
	Pear	5	5	0

The model incorrectly classified 55 cases as not belonging to class apple.

False Negatives of Class Orange

		Predicted class		
		Apple	Orange	Pear
	Apple	50	5	50
Actual class	Orange	10	50	20
	Pear	5	5	0

The model incorrectly classified 30 cases as not belonging to class orange.

False Negatives of Class Pear

		Predicted class	
	Apple	**Orange**	**Pear**
Actual class Apple	50	5	50
Orange	10	50	20
Pear	5	5	0

The model incorrectly classified 10 cases as not belonging to class pears.

Example 2

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{30+60+80}{300} = 170/300 = .556$$

$$Recall_{class=0} = \frac{TP_{class=0}}{TP_{class=0}+FN_{class=0}} = \frac{30}{30+20+10} = .5$$

$$Recall_{class=1} = \frac{TP_{class=1}}{TP_{class=1}+FN_{class=1}} = \frac{60}{60+50+10} = .5$$

$$Recall_{class=2} = \frac{TP_{class=2}}{TP_{class=2}+FN_{class=2}} = \frac{80}{80+20+20} = .667$$

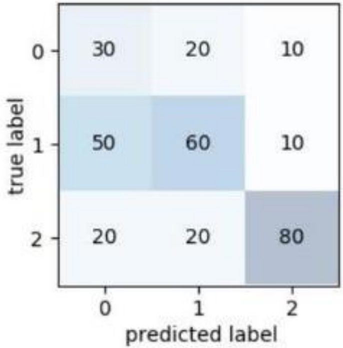
$$Recall = \frac{.5+.5+.667}{3} = 0.556$$

$$Precision_{class=0} = \frac{TP_{class=0}}{TP_{class=0}+FP_{class=0}} = \frac{30}{30+50+20} = .3$$

$$Precision_{class=1} = \frac{TP_{class=1}}{TP_{class=1}+FP_{class=1}} = \frac{60}{60+20+20} = .6$$

$$Precision_{class=2} = \frac{TP_{class=2}}{TP_{class=2}+FP_{class=2}} = \frac{80}{80+10+10} = 0.8$$

$$Precision = \frac{.3+.6+.8}{3} = 0.556$$



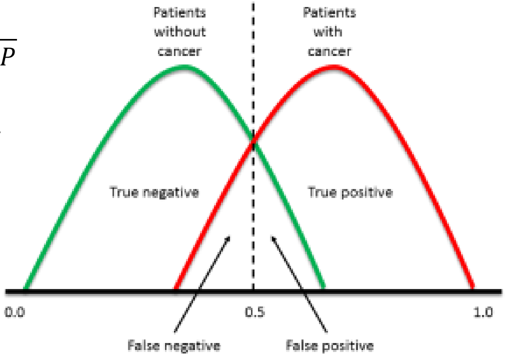
ROC and AUC

Threshold

Most classifiers produce a score, which is then thresholded to decide the classification. If a classifier produces a score between 0.0 (definitely negative) and 1.0 (definitely positive), it is common to consider anything over 0.5 as positive.

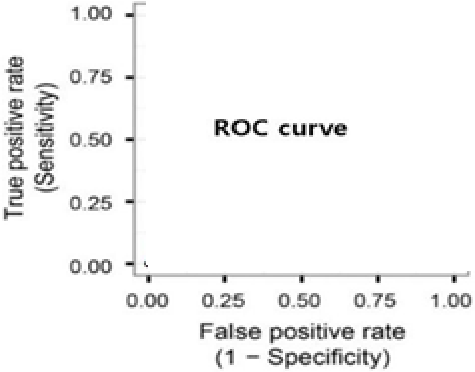
$$precision = \frac{TP}{TP + FP}$$

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



ROC Curve

- The Receiver Operating Characteristics(ROC) curve
- The ROC curve is a evaluation measure
  that is based on two basic evaluation measures
  - Specificity = True Negative Rate
  - Sensitivity = Recall = True Positive Rate



How to Plot ROC Curve?

- **Dynamic cut-off thresholds**

Instance	Yes	No	Actual
1	0.008	0.992	N
2	0.011	0.989	N
3	0.021	0.979	Y
4	0.009	0.991	N
5	0.014	0.986	N
6	0.015	0.985	N
7	0.012	0.988	N
8	0.015	0.985	Y

Cut-off = 0.020

Instance	Predict	Type
1	N	TN
2	N	TN
3	Y	TP
4	N	TN
5	N	TN
6	N	TN
7	N	TN
8	N	FN

TP=1	FN=1
FP=0	TN=6

Cut-off = 0.015

Instance	Predict	Type
1	N	TN
2	N	TN
3	Y	TP
4	N	TN
5	N	TN
6	Y	FP
7	N	TN
8	Y	TP

TP=2	FN=0
FP=1	TN=5

Cut-off = 0.010

Instance	Predict	Type
1	N	TN
2	Y	FP
3	Y	TP
4	N	TN
5	Y	FP
6	Y	FP
7	Y	FP
8	Y	TP

TP=2	FN=0
FP=4	TN=2

How to Plot ROC Curve?

- True positive rate (TPR) = $TP/(TP+FN)$ and False positive rate (FPR) = $FP/(FP+TN)$
- Use different cut-off thresholds (0.00, 0.01, 0.02,..., 1.00), calculate the TPR and FPR, and plot them into graph. That is receiver operating characteristic (ROC) curve.
- Example

TP=1	FN=1
FP=0	TN=6

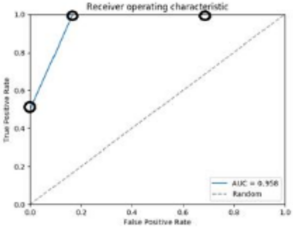
TPR = 0.5
FPR = 0

TP=2	FN=0
FP=1	TN=5

TPR = 1
FPR = 0.167

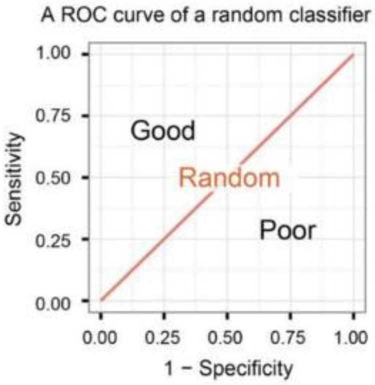
TP=2	FN=0
FP=4	TN=2

TPR = 1
FPR = 0.667



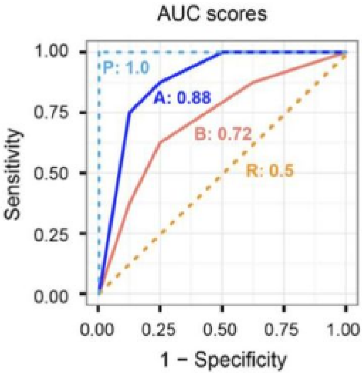
ROC Curve

- **A classifier with the random performance level always shows a straight line**
- **Two areas separated by this ROC curve**
  - ROC curves in the area with the top left corner indicate good performance levels
  - ROC curves in the other area with the bottom right corner indicate poor performance levels



AUC

- **AUC(Area under the ROC curve) score**
  - An advantage of using ROC curve is a single measure called AUC score
  - As the name indicates, it is an area under the curve calculated in the ROC space
  - Although the theoretical range of AUC score is between 0 and 1,
    the actual scores of meaningful classifiers are greater than 0.5,
    which is the AUC score of a random classifier
  - ROC curves clearly shows classifiers A
    outperforms classifier B



This lecture references

- https://www.slideshare.net/AndrewFerlitsch/machine-learning-accuracy-and-confusion-matrix
- https://www.slideshare.net/samuelbohman/confusion-matrix-explained
- https://www.guru99.com/confusion-matrix-machine-learning-example.html
- https://slideplayer.com/slide/15918720/