Analysis of AI Systems

CS 640

Margrit Betke

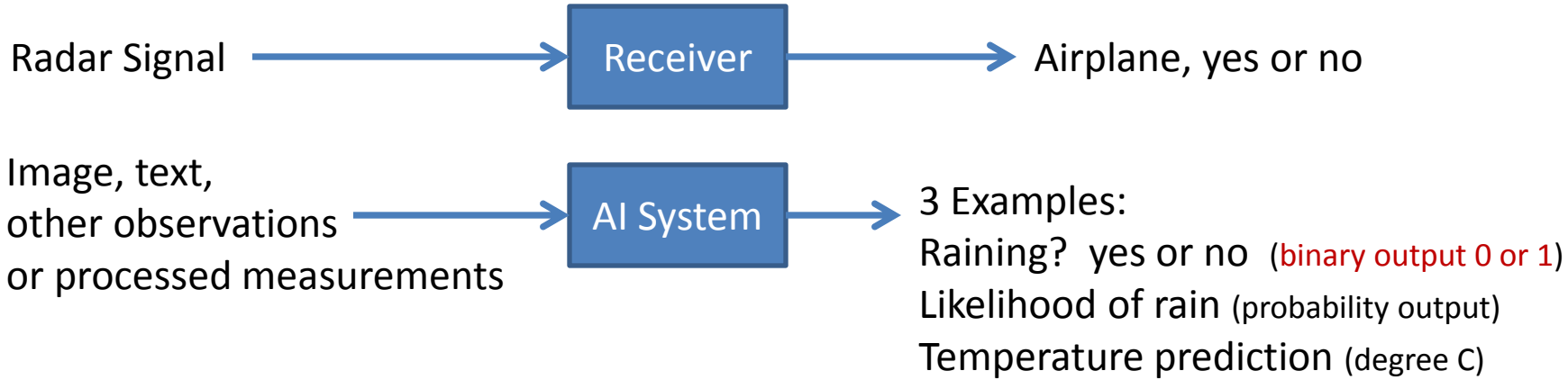
Lecture 2

September 8, 2020

ROC Analysis

ROC = receiver operating characteristics (historic name from radar signal processing)

ROC Analysis = Method to organize, visualize, and evaluate results of an AI system



Confusion Matrix for Binary Output Case

"Truth" =
Ground truth =
Gold standard =
Actual class

AI System output =
Hypothesis =
Predicted class

	1	0
1	**True Positive (TP)**	**False Positive (FP)**
0	**False Negative (FN)**	**True Negative (TN)**

Confusion Matrix for Binary Output Case

Example with 20 samples

"Truth" =
Ground truth =
Gold standard =
Actual class

AI System output =
Hypothesis =
Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

Confusion Matrix for Binary Output Case

"Truth" =
Ground truth =
Gold standard =
Actual class

AI System output =
Hypothesis =
Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

1st step of analyzing the confusion matrix:
Check that sum of matrix entries = number of samples used to test AI system

Confusion Matrix for Binary Output Case

"Truth" =
Ground truth =
Gold standard =
Actual class

AI System output =
Hypothesis =
Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

Good System?

Confusion Matrix for Binary Output Case

"Truth" =
Ground truth =
Gold standard =
Actual class

AI System output =
Hypothesis =
Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

Good System? We want high values in diagonal of matrix.

Confusion Matrix for Binary Output Case

"Truth" =
Ground truth =
Gold standard =
Actual class

AI System output =
Hypothesis =
Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

2nd step of analyzing the confusion matrix:
Compute sum of diagonal entries and compare that with total number of samples

Confusion Matrix for Binary Output Case

TP+TN=6+8=14
Total = 20

14/20 = 0.7

"Truth" =
Ground truth =
Gold standard =
Actual class

AI System output =
Hypothesis =
Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

2nd step of analyzing the confusion matrix:
Compute sum of diagonal entries and compare that with total number of samples

Confusion Matrix for Binary Output Case

TP+TN=6+8=14

Total = 20

Accuracy of AI System:

14/20 = 0.7

"Truth" =
Ground truth =
Gold standard =
Actual class

AI System output =
Hypothesis =
Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

2nd step of analyzing the confusion matrix:
Compute sum of diagonal entries and compare that with total number of samples

Confusion Matrix for Binary Output Case

"Truth" =
Ground truth =
Gold standard =
Actual class

AI System output =
Hypothesis =
Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

Positive samples = TP+FN = 8

Negative samples = FP+FN = 12

How sensitive is the classifier in finding the positives?

true positive rate **= tp =**
TP/(TP+FN) = 6/8 = ¾
= *recall* = *sensitivity*

"Truth" =
Ground truth =
Gold standard =
Actual class

AI System output =
Hypothesis =
Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

Positive samples =
P= TP+FN =
8

"Truth" =
Ground truth =
Gold standard =
Actual class

false positive rate **= fp =**
FP/(FP+TN) = 4/12 = 1/3

AI System output =
Hypothesis =
Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

Negative samples =
FP+TN =
12

true positive rate **= tp =**
TP/(TP+FN) = 6/8 = ¾
= *recall* = *sensitivity*

false positive rate **= fp =**
FP/(FP+TN) = 4/12 = 1/3

"Truth" =
Ground truth =
Gold standard =
Actual class

AI System output =
Hypothesis =
Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

Positive samples = TP+FN = 8

Negative samples = FP+TN = 12

How specific is the classifier in finding the negatives?

Instead of *fp*, we sometimes focus on

1 – *fp* = *specificity*
TN/(FP+TN) = 8/12 = 2/3

"Truth" =
Ground truth =
Gold standard =
Actual class

AI System output =
Hypothesis =
Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

Negative samples =
FP+TN =
12

How precise is the classifier in finding the positives?

true positive rate **= tp = TP/(TP+FN) = 6/8 = ¾ = *recall* = *sensitivity***

precision **= TP/(TP+FP) = 6/10 = 3/5**

"Truth" = Ground truth = Gold standard = Actual class

AI System output = Hypothesis = Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

Positive hypotheses = **TP+FP** = 10

Positive samples = TP+FN = 8

F1 Score

true positive rate **= tp = TP/(TP+FN) = 6/8 = ¾ = *recall* = *sensitivity***

precision **= TP/(TP+FP) = 6/10 = 3/5**

F1 score = 2 recall x precision/(recall + precision)
= 2 x 3/4 x 3/5 / (3/4+3/5) = 9/10 / (27/20) = 2/3

AI System output = Hypothesis = Predicted class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

Positive hypotheses = **TP+FP** = 10

Positive samples = TP+FN = 8

Terms to remember:

ROC

Classifier

Predictor

Likelihood

Ground truth, gold standard

Hypothesis

Accuracy

False positive rate & False negative rate

Recall & Precision

Sensitivity & Specificity

F1 score

Building an ROC curve for an AI System: One classifier at time

TP+TN=6+8=14

Total = 20

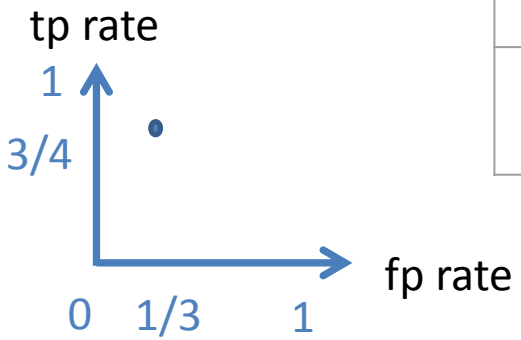
Accuracy of AI System:

14/20 = 0.7

false positive rate = 1/3

true positive rate = 3/4

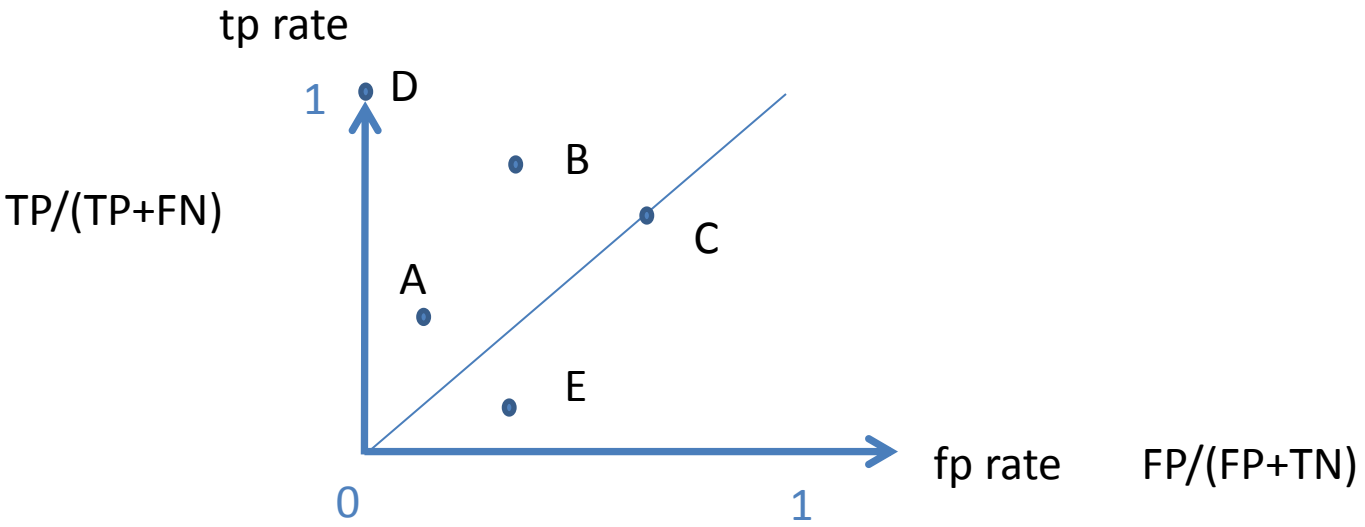
ROC Curve has 1 point:



"Truth" =
Ground truth =
Gold standard =
Actual class

	1	0
1	**True Positive (TP): 6**	**False Positive (FP): 4**
0	**False Negative (FN): 2**	**True Negative (TN): 8**

Comparing Classifiers



Classifier A:

Classifier B:

Classifier C:

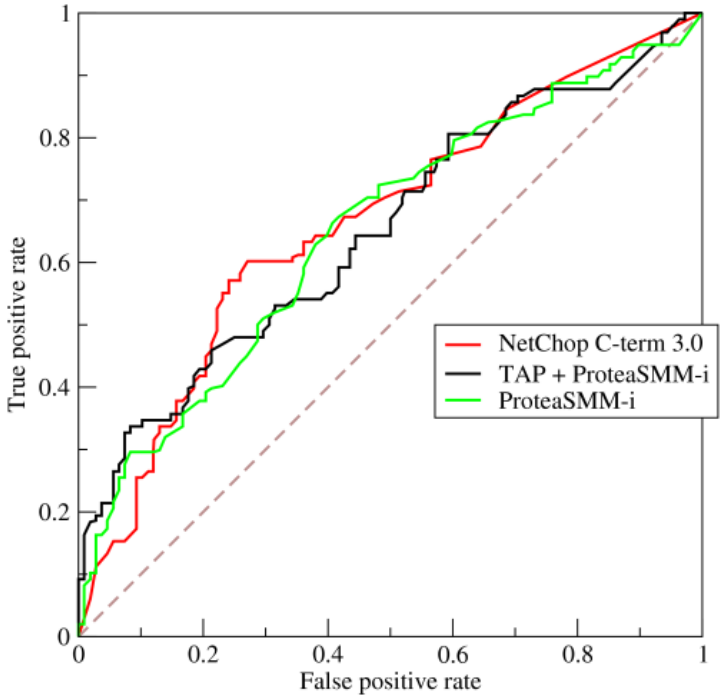
Classifier D:

Classifier E:

See paper by Fawcett

ROC Curve: Classifier

Example: Three predictors of peptide cleaving

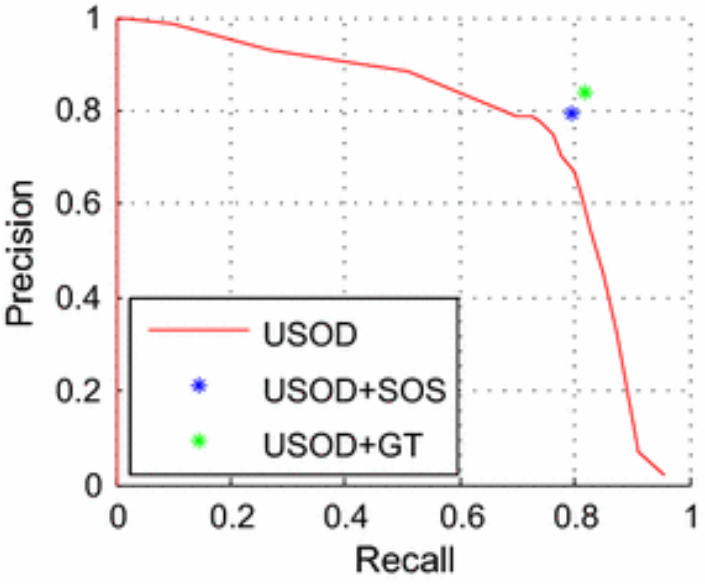


By BOR at the English language Wikipedia, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=10714489

Comparing Classifiers

Some researchers prefer to draw precision/recall curves instead of tp/fp curves:

Precision
=TP/(TP+FP)



Recall = tp = TP/(TP+FN)

Plot from one of my research papers (Zhang et al, IJCV 2017):

Classifier: Salient object subitizing (SOS)
= predicts the number (1, 2, 3, and 4+) of salient objects in an image
USOD stands for “unconstrained object detection method”
(parameter: number of detection windows)
GT = ground truth

Confusion Matrix for Multiple Classes

"Truth" = Ground truth = Gold standard = Actual class

AI System Output = Hypothesis = Predicted class

	Class 1	Class 2	Class 3	Class 4
Class 1	100%	15%	10%	7%
Class 2	0%	80%	10%	3%
Class 3	0%	3%	80%	70%
Class 4	0%	2%	0%	20%

Confusion Matrix for Multiple Classes

"Truth" = Ground truth = Gold standard = Actual class

AI System Output = Hypothesis = Predicted class

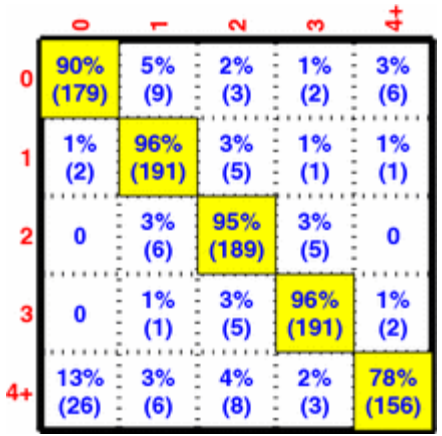
	Class 1	Class 2	Class 3	Class 4
Class 1	100%	15%	10%	7%
Class 2	0%	80%	10%	3%
Class 3	0%	3%	80%	70%
Class 4	0%	2%	0%	20%

Confusion Matrix for Multiple Classes

Note: Rows and columns of a confusion matrix may be reversed
Reporting only percentages and not actual number is usually NOT a good practice.

Example of a multi-class confusion matrix in one of my papers (Zhang et al, IJCV 2017):

Each row corresponds to a ground-truth category label. The percentage reported is the average proportion of images of the category A (row number) labeled as category B (column number). For over 90% images, predicted labels are consistent with the ground-truth labels.



	0	1	2	3	4+
0	90% (179)	5% (9)	2% (3)	1% (2)	3% (6)
1	1% (2)	96% (191)	3% (5)	1% (1)	1% (1)
2	0	3% (6)	95% (189)	3% (5)	0
3	0	1% (1)	3% (5)	96% (191)	1% (2)
4+	13% (26)	3% (6)	4% (8)	2% (3)	78% (156)