Analysis of AI Systems

CS 640
Margrit Betke

Lecture 2
September 7, 2023
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

3 Examples:
- Raining? yes or no (binary output 0 or 1)
- Likelihood of rain (output between 0 and 1)
- Temperature prediction (degree C)
ROC Analysis

Image, text, other observations or processed measurements

AI System

4 Examples:

1) Raining? yes or no (binary output 0 or 1)  
   Predictor Type: Binary classifier

2) Likelihood of rain (output between 0 and 1)  
   Threshold on likelihood is typically used to make binary decision

3) Rain? Sun? Snow?  
   “One-hot” Output: (1,0,0) => rain  
   or Likelihood score: (0.1, 0.7, 0.2) => sun  
   Predictor Type: Multiclass Classifier

4) Temperature prediction (degree C)  
   Predictor Type: Regressor
Evaluating a Regressor

"Truth" = $y_1, \ldots, y_n$
Ground truth = $y_1, \ldots, y_n$
Gold standard = $y_1, \ldots, y_n$
Actual value = $y_1, \ldots, y_n$

AI System output = $\hat{y}_1, \ldots, \hat{y}_n$
Hypothesis = $\hat{y}_1, \ldots, \hat{y}_n$
Predicted value = $\hat{y}_1, \ldots, \hat{y}_n$
Evaluating a Regressor, e.g., Temperature Predictor

"Truth" = $Y_1, \ldots, Y_n$
Ground truth = $Y_1, \ldots, Y_n$
Gold standard = $Y_1, \ldots, Y_n$
Actual value = $Y_1, \ldots, Y_n$

AI System output = $\hat{Y}_1, \ldots, \hat{Y}_n$
Hypothesis = $\hat{Y}_1, \ldots, \hat{Y}_n$
Predicted value = $\hat{Y}_1, \ldots, \hat{Y}_n$

Compare measured temperature with Predicted temperature:

**Error:** $y_i - \hat{y}_i$

e.g., $80\text{F} - 78\text{F} = 2\text{F error}$
Evaluating a Regressor, e.g., Temperature Predictor

"Truth" = Ground truth = Gold standard = Actual value = \( y_1, \ldots, y_n \)

AI System output = Hypothesis = Predicted value = \( \hat{y}_1, \ldots, \hat{y}_n \)

Compare measured temperature with Predicted temperature:

Error: \( y_i - \hat{y}_i \)

e.g., 80F – 85F = -5F error

Need error measure that handles positive and negative differences!
Evaluating a Regressor, e.g., Temperature Predictor

"Truth" = $Y_1, \ldots, Y_n$
Ground truth = $Y_1, \ldots, Y_n$
Gold standard = $Y_1, \ldots, Y_n$
Actual value = $Y_1, \ldots, Y_n$

AI System output = $\hat{Y}_1, \ldots, \hat{Y}_n$
Hypothesis = $\hat{Y}_1, \ldots, \hat{Y}_n$
Predicted value = $\hat{Y}_1, \ldots, \hat{Y}_n$

Compare measured temperature with predicted temperature:

Error: $| y_i - \hat{y}_i |$

Absolute Error?
Evaluating a Regressor, e.g., Temperature Predictor

"Truth" = \( Y_1, \ldots, Y_n \)
Ground truth = \( Y_1, \ldots, Y_n \)
Gold standard = \( \hat{Y}_1, \ldots, \hat{Y}_n \)
Actual value = \( Y_1, \ldots, Y_n \)
Al System output = \( \hat{Y}_1, \ldots, \hat{Y}_n \)
Hypothesis = \( \hat{Y}_1, \ldots, \hat{Y}_n \)
Predicted value = \( Y_1, \ldots, Y_n \)

Compare measured temperature with Predicted temperature:

Error: \( ( y_i - \hat{y}_i )^2 \)

Squared error is preferred. Why?
Evaluating a Regressor over full dataset:

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

AI System output = 
Hypothesis = 
Predicted value = 

\( \hat{y}_1, \ldots, \hat{y}_n \)

Mean Squared Error:

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]

Or

Root Mean Squared Error:

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]
Confusion Matrix for Binary Output Case

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td>0</td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>

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

AI System output = Hypothesis = Predicted class
Confusion Matrix for Binary Output Case

Example with 20 samples

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>0</td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>
### Confusion Matrix for Binary Output Case

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive ((TP): 6)</td>
<td>False Positive ((FP): 4)</td>
</tr>
<tr>
<td>0</td>
<td>False Negative ((FN): 2)</td>
<td>True Negative ((TN): 8)</td>
</tr>
</tbody>
</table>

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

AI System output = Hypothesis = Predicted class

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>0</td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>

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

AI System output = Hypothesis = Predicted class

Good System?
### Confusion Matrix for Binary Output Case

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>TP</strong>: 6</td>
<td><strong>FP</strong>: 4</td>
</tr>
<tr>
<td>0</td>
<td><strong>FN</strong>: 2</td>
<td><strong>TN</strong>: 8</td>
</tr>
</tbody>
</table>

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

**AI System output = Hypothesis = Predicted class**

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

$TP + TN = 6 + 8 = 14$
### Confusion Matrix for Binary Output Case

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td><strong>0</strong></td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>

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

**AI System output = Hypothesis = Predicted class**

2\textsuperscript{nd} 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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>0</td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>

TP + TN = 6 + 8 = 14
Total number of samples = 20

14 versus 20: Is this a good system?

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

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

AI System output = Hypothesis = Predicted class
Confusion Matrix for Binary Output Case

TP+TN=6+8=14
Total number of samples = 20
Accuracy of AI System:
14/20 = 0.7

AI System output = Hypothesis = Predicted class

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

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td><strong>0</strong></td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>

**Truth** =

Ground truth =

Gold standard =

Actual class =

AI System output =

Hypothesis =

Predicted class =

Positive samples =

\[ TP + FN = 8 \]

Negative samples =

\[ FP + FN = 12 \]
How sensitive is the classifier in finding the positives?

True positive rate = \( tp = \frac{TP}{TP+FN} = \frac{6}{8} = \frac{3}{4} \)

= recall = sensitivity

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>0</td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>

Positive samples = \( P = TP+FN = 8 \)
**True Positive (TP):** 6
**False Positive (FP):** 4
**False Negative (FN):** 2
**True Negative (TN):** 8

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

**false positive rate** = \(fp = \frac{FP}{FP+TN} = \frac{4}{12} = \frac{1}{3}\)

**AI System output = Hypothesis = Predicted class**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td><strong>0</strong></td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>

Negative samples = \(FP+TN = 12\)
true positive rate = tp = TP/(TP+FN) = 6/8 = \(\frac{3}{4}\)
= recall = sensitivity

false positive rate = fp = FP/(FP+TN) = 4/12 = \(\frac{1}{3}\)

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>0</td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>

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

$$1 - fp = specificity$$

$$\frac{TN}{FP+TN} = \frac{8}{12} = \frac{2}{3}$$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>

Negative samples = $FP+TN = 12$
How precise is the classifier in finding the positives?

**true positive rate** = \( tp = \frac{TP}{TP+FN} = \frac{6}{8} = \frac{3}{4} \)

= **recall** = **sensitivity**

**precision** = \( \frac{TP}{TP+FP} = \frac{6}{10} = \frac{3}{5} \)

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>0</td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>

Positive samples = \( TP+FN = 8 \)

Positive hypotheses = \( TP+FP = 10 \)
F1 Score

**true positive rate** = tp = \( \frac{TP}{TP+FN} \) = 6/8 = \( \frac{3}{4} \)

= **recall** = **sensitivity**

**precision** = \( \frac{TP}{TP+FP} \) = 6/10 = \( \frac{3}{5} \)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>0</td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>

F1 score = \( 2 \times \text{recall} \times \text{precision} / (\text{recall} + \text{precision}) \) = \( 2 \times \frac{3}{4} \times \frac{3}{5} / (\frac{3}{4} + \frac{3}{5}) \) = \( \frac{2}{3} \)

Positive hypotheses = \( TP+FP = 10 \)

Positive samples = \( TP+FN = 8 \)
**F1 Score**

*true positive rate* = \( tp = \frac{TP}{(TP+FN)} = \frac{6}{8} = \frac{3}{4} = 0.75 \)

= *recall* = *sensitivity*

*precision* = \( \frac{TP}{(TP+FP)} = \frac{6}{10} = \frac{3}{5} = 0.6 \)

F1 score = \( 2 \times \frac{3}{4} \times \frac{3}{5} / \left( \frac{3}{4} + \frac{3}{5} \right) = \frac{2}{3} = 0.667 \)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>0</td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>
**Balanced Accuracy**

**true positive rate** = \( tp = \frac{TP}{(TP+FN)} = \frac{6}{8} = \frac{3}{4} = 0.75 \)  
= **recall** = sensitivity

\( 1 - fp = specificity \)

\( TN/(FP+TN) = \frac{8}{12} = 0.67 \)

AI System output = Hypothesis = Predicted class

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>0</td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>

Balanced Accuracy = \( \frac{(\text{sensitivity + specificity})}{2} = \frac{(3/4 + 2/3)}{2} = \frac{0.75 + 0.67}{2} = \frac{2}{3} = 0.708 \)
Accuracy vs. F1 Score vs. Balanced Accuracy

Accuracy \(\text{= } \frac{\text{TP+TN}}{\text{everything}} \text{ = } 0.700\)

F1 Score \(\text{= } \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \text{ = } 0.667\)

Balanced Accuracy \(\text{= } \frac{\text{sensitivity} + \text{specificity}}{2} \text{ = } 0.708\)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>0</td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>
Accuracy vs. F1 Score vs. Balanced Accuracy

Accuracy = \frac{TP+TN}{everything} = 0.999
F1 Score = \frac{2 \text{ recall} \times \text{ precision}}{\text{recall} + \text{precision}} = 0.667
Balanced Accuracy = \frac{\text{sensitivity} + \text{specificity}}{2} = 0.833

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>0</td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8000</td>
</tr>
</tbody>
</table>
Accuracy vs. F1 Score vs. Balanced Accuracy

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive (TP):</td>
<td>6000</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>False Negative (FN):</td>
<td>2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>

Accuracy = \(\frac{(TP+TN)}{\text{everything}}\) = 0.999

F1 Score = \(\frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}\) = 0.999

Balanced Accuracy = \(\frac{\text{sensitivity} + \text{specificity}}{2}\) = 0.833
Terms to remember:

- ROC
  - Ground truth, gold standard
  - Hypothesis

- Classifier
  - Accuracy, Balanced Accuracy, F1 score

- Predictor
  - False positive rate & False negative rate

- Likelihood
  - Recall & Precision
  - Sensitivity & Specificity
Building an ROC curve for an AI System: One classifier at time

TP+TN=6+8=14
Total number of samples = 20

**Accuracy of AI System:**
14/20 = 0.7

false positive rate = 1/3
true positive rate = 3/4

ROC curve has 1 point:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP): 6</td>
<td>False Positive (FP): 4</td>
</tr>
<tr>
<td>0</td>
<td>False Negative (FN): 2</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>

"Truth" = Ground truth = Gold standard = Actual class
Good Classifier?

false positive rate = $1/3$
true positive rate = $1/3$

ROC curve has 1 point:

<table>
<thead>
<tr>
<th>Truth</th>
<th>1 True Positive (TP): 4</th>
<th>0 False Positive (FP): 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>False Negative (FN): 8</td>
<td>True Negative (TN): 8</td>
</tr>
</tbody>
</table>
Comparing Classifiers

Classifier A:  Classifier D:
Classifier B:  Classifier E:
Classifier C:

See paper by Fawcett
Each colored line shows the behavior of a binary classifier when a parameter is changed.

Example for the rain prediction classifier:
The parameter could be the threshold $T$ on the likelihood of its prediction for rain:
- Likelihood $> T$ Predict “rain”
- $\leq T$ Predict “no rain”
ROC Curve: Classifier

Real 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
On the Quest to Interpret Web Image Content: Salient Object Subitizing

Jianming Zhang, Shugao Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, et al.,

CVPR 2015
IJCV 2017

Salient Object Subitizing

Task:
Predict the existence and number of salient objects in a scene

Solution:
GoogLeNet CNN called SOS
Zhang et al.

~ 69% accurate

0 or 1 object:
>90% accurate
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

<table>
<thead>
<tr>
<th></th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>100%</td>
<td>15%</td>
<td>10%</td>
<td>7%</td>
</tr>
<tr>
<td>Class 2</td>
<td>0%</td>
<td>80%</td>
<td>10%</td>
<td>3%</td>
</tr>
<tr>
<td>Class 3</td>
<td>0%</td>
<td>3%</td>
<td>80%</td>
<td>70%</td>
</tr>
<tr>
<td>Class 4</td>
<td>0%</td>
<td>2%</td>
<td>0%</td>
<td>20%</td>
</tr>
</tbody>
</table>

AI System Output = Hypothesis = Predicted class
### Confusion Matrix for Multiple Classes

<table>
<thead>
<tr>
<th></th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>100%</td>
<td>15%</td>
<td>10%</td>
<td>7%</td>
</tr>
<tr>
<td>Class 2</td>
<td>0%</td>
<td>80%</td>
<td>10%</td>
<td>3%</td>
</tr>
<tr>
<td>Class 3</td>
<td>0%</td>
<td>3%</td>
<td>80%</td>
<td>70%</td>
</tr>
<tr>
<td>Class 4</td>
<td>0%</td>
<td>2%</td>
<td>0%</td>
<td>20%</td>
</tr>
</tbody>
</table>

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

AI System Output = Hypothesis = Predicted class
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.