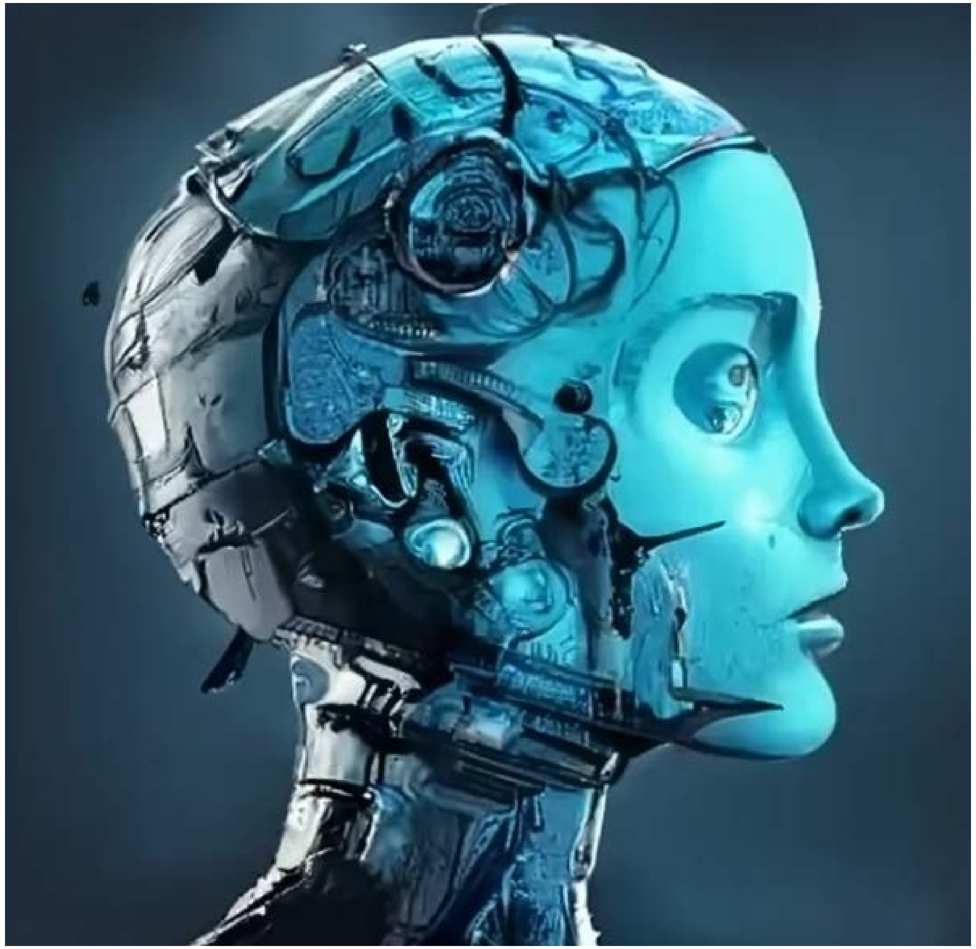
Machine Intelligence

Classification Metrics

Dr. Cory Merkel



How Should We Measure Classification Performance?

- Imagine that we have a binary classification problem:

$$y, \hat{y} \in \{0,1\}$$

Not Cancer   Cancer

  - Suppose we have 100 samples, $\mathbf{u}^{(p)} = \left[tumor\_size^{(p)} \;\; tumor\_density^{(p)}\right]$
  - Split data into train and test sets with 80% for training and 20% for testing
- After training our model, we find that our training and test accuracies (# of correctly classified/size of set) are both 95%
  - Let's say the next best reported accuracy is only 85%!
    - ❑Can we assume we have created a superior model?

How Should We Measure Classification Performance?

- Can we assume we have created a superior model?
  - No! We need more information.

- Suppose that only 5% of our data correspond to cancer
  - We could get 95% accuracy just by saying that everything is benign
    - Just looking at accuracy may be misleading
  - In this case, we would say that we have a *class imbalance*

- Can we break down our results into more than just one number?

Confusion Matrix

- A standard method is to use a *confusion matrix.*
  - For binary classification, a confusion matrix is a 2-by-2 table:

		Hypothesis ($\hat{y}$)	
		1	0
Ground Truth (y)	1	TP: True Positive	FN: False Negative
	0	FP: False Positive	TN: True Negative

- Fill in each cell with the number of occurrences from your classifier
  - Want large numbers on diagonal and small numbers off of diagonal

Confusion Matrix

- True Positive: Number of positive (expected value of 1) examples correctly classified
- True Negative: Number of negative (expected value of 0) examples correctly classified
- False Negative: Number of positive examples incorrectly classified
- False Positive: Number of negative examples incorrectly classified

		Hypothesis ($\hat{y}$)	
		1	0
Ground Truth (y)	1	TP: True Positive	FN: False Negative
	0	FP: False Positive	TN: True Negative

Confusion Matrix

		Hypothesis ($\hat{y}$)	
		1	0
Ground Truth (y)	1	TP: True Positive	FN: False Negative
	0	FP: False Positive	TN: True Negative

- Accuracy can be derived from the CM:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

- Other useful metrics:

$$Sensitivity = \frac{TP}{TP+FN}$$ (A.K.A. Recall, True Positive Rate, TPR)

$$Specificity = \frac{TN}{TN+FP}$$ (A.K.A. True Negative Rate, TNR)

$$Precision = \frac{TP}{TP+FP}$$ (A.K.A. Positive Predictive Value, PPV)

$$Negative\ Predictive\ Value = \frac{TN}{TN+FN}$$ (A.K.A. NPV)

- Notice that rows correspond to TPR and TNR, while columns correspond to PPV and NPV)

Confusion Matrix

		Hypothesis ($\hat{y}$)	
		1	0
Ground Truth (y)	1	TP: True Positive	FN: False Negative
	0	FP: False Positive	TN: True Negative

- We can also look at a harmonic mean of TPR and PPV(Recall and Precision)
  - We call this the F1 score

$$F1\ Score = \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$$

Example

- For our example of cancer vs. no cancer classification, if our classifier always chooses no cancer:

$Accuracy = 95\%$

$Recall = 0\%$

$Specificity = 100\%$

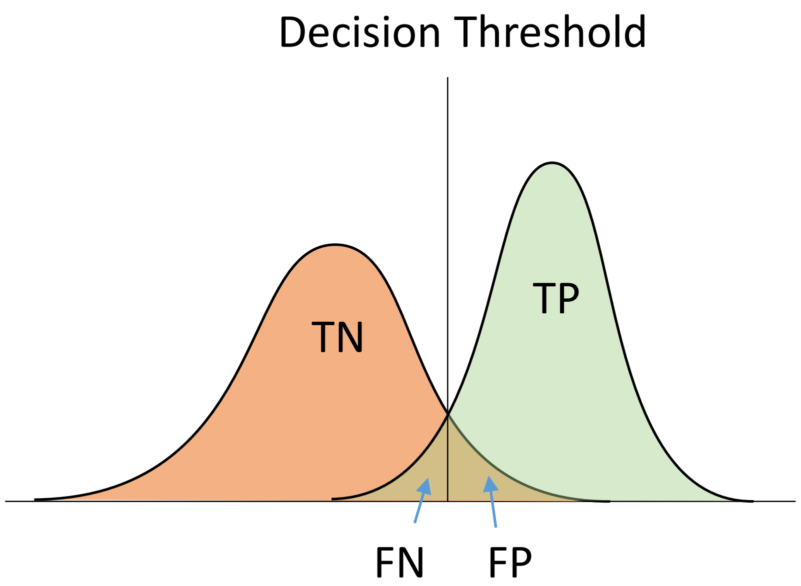
$Precision = 0/0$ (treat as undefined or 0)

$NPV = 95\%$

$F1\ Measure = 0$

<table>
<tr><td colspan="2"></td><td colspan="2">Hypothesis ($\hat{y}$)</td></tr>
<tr><td colspan="2"></td><td>1</td><td>0</td></tr>
<tr><td rowspan="2">Ground Truth (y)</td><td>1</td><td>0</td><td>5</td></tr>
<tr><td>0</td><td>0</td><td>95</td></tr>
</table>

Another View



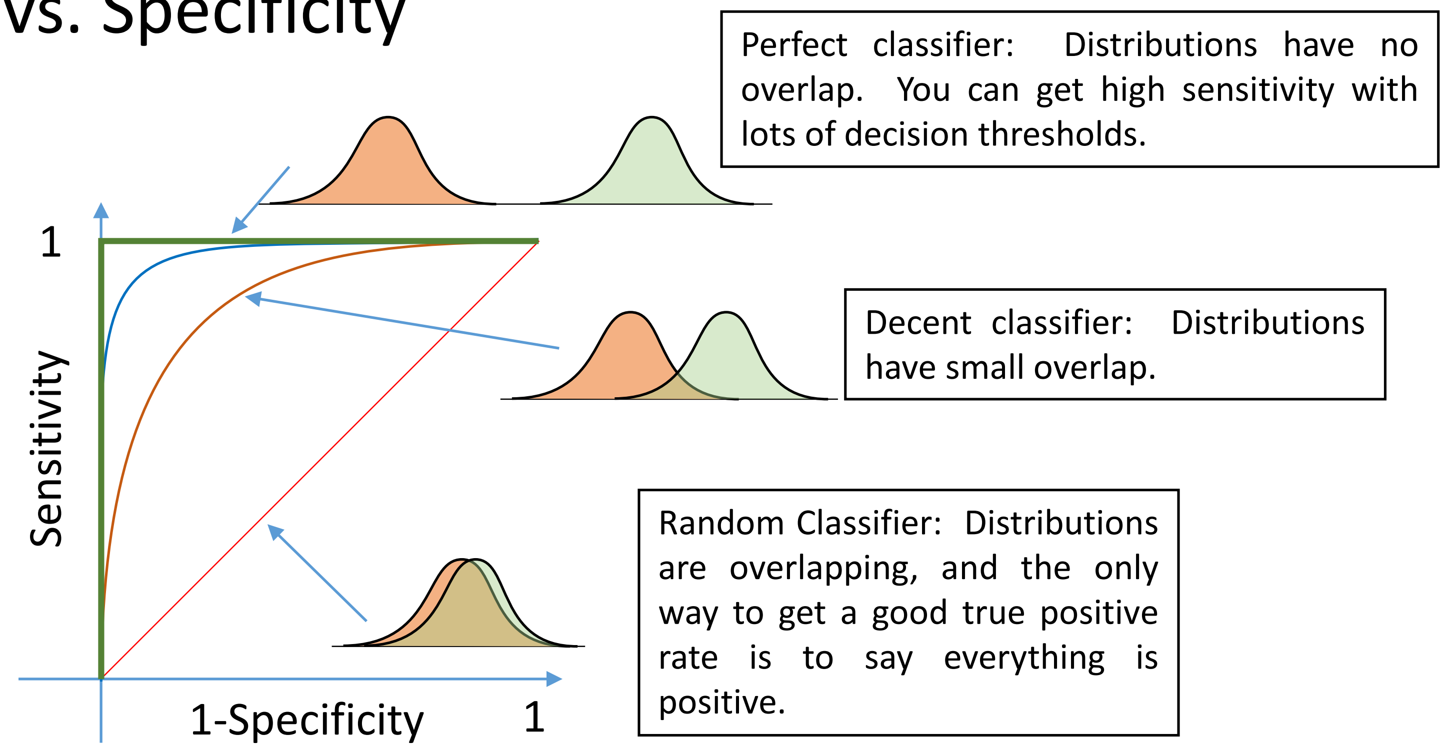
$$Sensitivity = \frac{TP}{TP+FN}$$

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

- Very low decision threshold will give high sensitivity, but poor specificity
- Very high decision threshold will give poor sensitivity, but high specificity
- Note that as the distributions become further apart, we have more decision thresholds that relate to both high sensitivity **and** high specificity

Receiver Operating Characteristics

- Plot the Sensitivity vs. Specificity



- Note that better classifiers have higher areas under the curve (AUC). Therefore, we can use AUC as a measure of the model performance under different decision thresholds.

Multi-Class Confusion Matrices

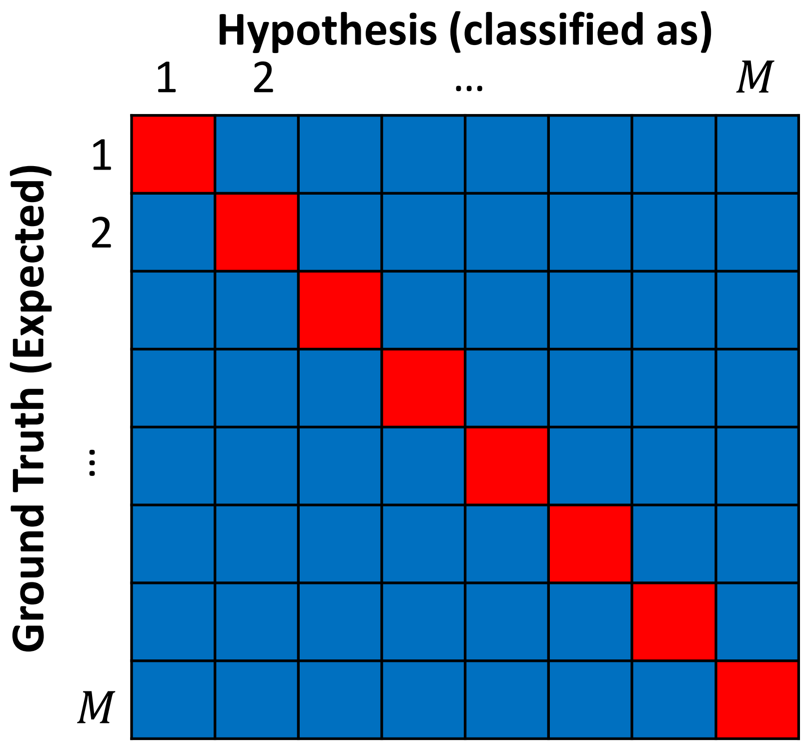
- When we have more than two classes, we can have an $M \times M$ confusion matrix (M is the number of classes). We can use color to represent the relative size of the entries.

Hypothesis (classified as)

Ground Truth (Expected)	1	2		…				M
1								
2								
⋮								
M								

Multi-Class Confusion Matrices

- Still want high numbers (hot colors) on diagonal and low numbers (cool colors) off diagonal:



- Accuracy is the trace of the matrix divided by the sum of all elements