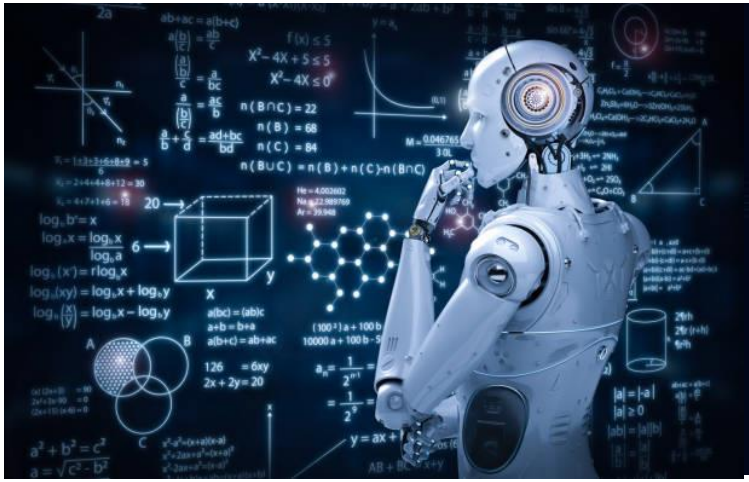
Welcome…

Classification: Performance Metrics-Confusion Matrix, ROC, AUC, Macro and Micro Averages

CS 797Q
Fall 2024

09/30/2024

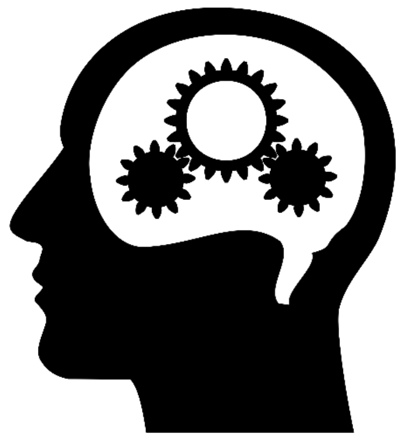


Categories of Data Analysis Techniques

Category	Techniques Covered	Problem to be solved
Association Rules	Apriori	Relationships between items
Clustering	K-Means Clustering DB Scan	Grouping of similar items Identification of structures
Classification	K-nearest Neighbor Decision Trees Random Forests Logistic Regression Naive Bayes Support Vector Machines Neural Networks	Assignment of labels to objects
Regression	Linear Regression Ridge Lasso	Relationship between outcome and inputs
Time Series Analysis	ARMA	Identification of temporal structures Forecasting of temporal processes
Text Mining	Bag-of-Words Stemming/Lemmatization TF-IDF	Analysis of textual data

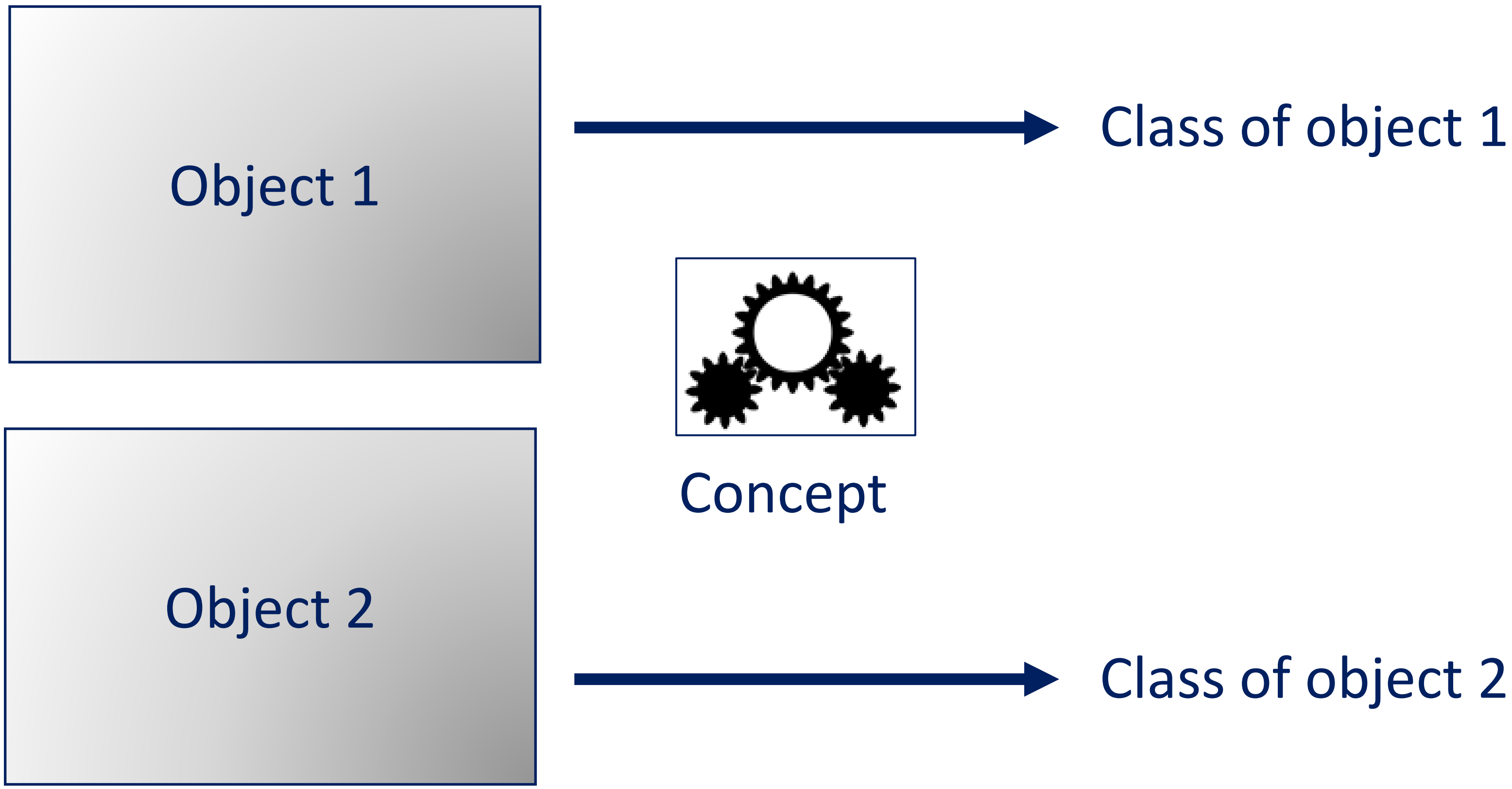
Example of Classification

This is a whale



This is a bear

The General Problem



The Formal Problem

- Object space
  - $O = \{object_1, object_2, \dots\}$
  - Often infinite
- Representations of the objects in a feature space
  - $\mathcal{F} = \{\phi(o), o \in O\}$
- Set of classes
  - $C = \{class_1, \dots, class_n\}$
- A *target concept* that maps objects to classes
  - $h^*: O \to C$
- Classification
  - Finding an approximation of the target concept

The „Whale“ Hypothesis

- Why do we know this is a whale?

Has a fin

Blue background

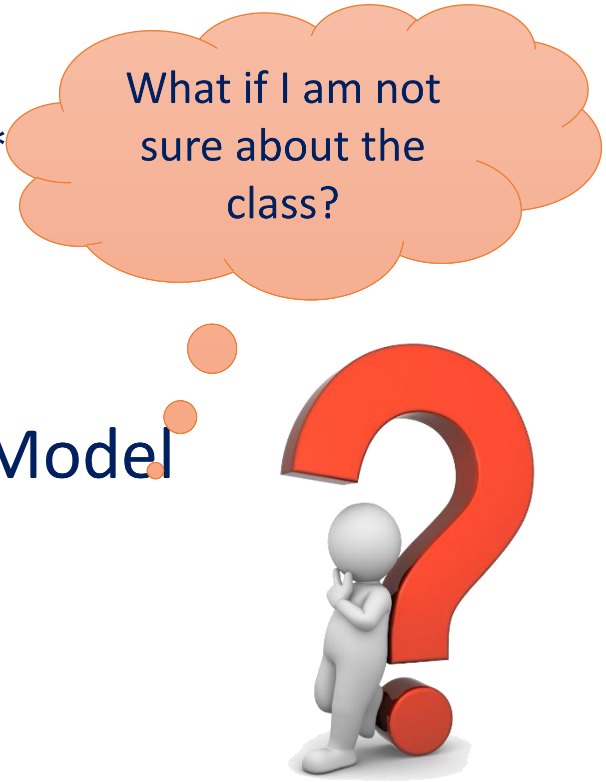
Oval body

Black top, white bottom

Hypothesis: Objects with fins, an oval general shape that are black on top and white on the bottom in front of a blue background are whales.

The Hypothesis

- A hypothesis maps features to classes
  - $h: \mathcal{F} \rightarrow C$
  - $h: \phi(o) \rightarrow C$
- Approximation of the target concept h^*
  - $h^*(o) \approx h(\phi(o))$
- Hypothesis = Classifier = Classification Model



Classification using Scores

- A numeric score for each class $c \in C$
- Often a probability distribution
  - $h': \phi(o) \to [0,1]^{|C|}$
  - $\|h'(\phi(o))\|_1 = 1$
- Example
  - Three classes: „whale“, „bear“, „other“
  - $h'\big(\phi(\text{"whalepicture"})\big) = (0.7, 0.1, 0.2)$

score whale

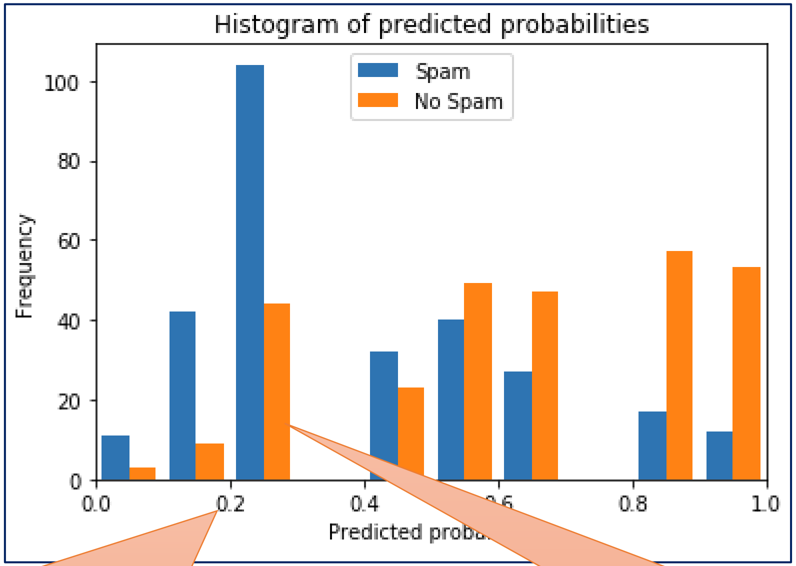
score bear

score other

- Standard approach:
  - Classification is class with highest score

Thresholds for Scores

- Different thresholds also possible



Threshold of 0.2 would miss “Spam” but better identify “No Spam”

Many “No Spam” incorrectly detected as spam if “highest” score is used

Quality of Hypothesis

- Goal: Approximation of the targ
  - $h^*(o) \approx h(\phi(o))$

How do you evaluate
$h^*(o) \approx h(\phi(o))$

→ Use Test Data
  - Structure is the same as training data
  - Apply hypothesis

$\boldsymbol{\phi(o)}$					$\boldsymbol{h^*(o)}$	$\boldsymbol{h(\phi(o))}$
hasFin	shape	colorTop	colorBottom	background	class	prediction
true	oval	black	black	blue	**whale**	**whale**
false	rectangle	brown	brown	green	**bear**	**whale**
…	…	…	…	…	…	

The Confusion Matrix

- Table of actual values versus prediction

Actual class

Predicted Class

	whale	bear	other
whale	29	1	3
bear	2	22	13
other	4	11	51

Two whales were incorrectly predicted as bears

Binary Classification

- Many problems are binary
    - Will I get my money back?
    - Is this credit card fraud?
    - Will my paper be accepted?
    - …
- Can all be formulated as either being in a class or not

→Labels *true* and *false*

The Binary Confusion Matrix

Actual class

Predicted Class

	true	false
true	True Positives (TP)	False Positives (FP)
false	False Negatives (FN)	True Negatives (TN)

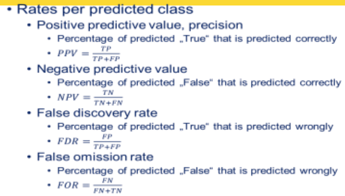
- False positives are also called Type I error
- False negatives are also called Type II error

Binary Performance Metrics (1)

		Actual class	
		true	false
Predicted Class	true	True Positives (TP)	False Positives (FP)
	false	False Negatives (FN)	True Negatives (TN)

- Rates per actual class
  - True positive rate, recall, sensitivity
    - Percentage of actually „True“ that is predicted correctly
    - $TPR = \frac{TP}{TP+FN}$
  - True negative rate, specificity
    - Percentage of actually „False“ that is predicted correctly
    - $TNR = \frac{TN}{TN+FP}$
  - False negative rate
    - Percentage of actually „True“ that is predicted wrongly
    - $FNR = \frac{FN}{FN+TP}$
  - False positive rate
    - Percentage of actually „False“ that is predicted wrongly
    - $FPR = \frac{FP}{FP+TN}$

Binary Performance Metrics (2)

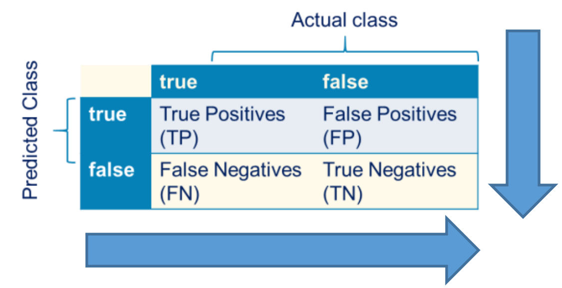


- Rates per predicted class
  - Positive predictive value, precision
    - Percentage of predicted „True“ that is predicted correctly
    - $PPV = \frac{TP}{TP+FP}$
  - Negative predictive value
    - Percentage of predicted „False“ that is predicted correctly
    - $NPV = \frac{TN}{TN+FN}$
  - False discovery rate
    - Percentage of predicted „True“ that is predicted wrongly
    - $FDR = \frac{FP}{TP+FP}$
  - False omission rate
    - Percentage of predicted „False“ that is predicted wrongly
    - $FOR = \frac{FN}{FN+TN}$

- Rates per predicted class
  - Positive predictive value, precision
    - Percentage of predicted „True“ that is predicted correctly
    - $PPV = \frac{TP}{TP+FP}$
  - Negative predictive value
    - Percentage of predicted „False“ that is predicted correctly
    - $NPV = \frac{TN}{TN+FN}$
  - False discovery rate
    - Percentage of predicted „True“ that is predicted wrongly
    - $FDR = \frac{FP}{TP+FP}$
  - False omission rate
    - Percentage of predicted „False“ that is predicted wrongly
    - $FOR = \frac{FN}{FN+TN}$

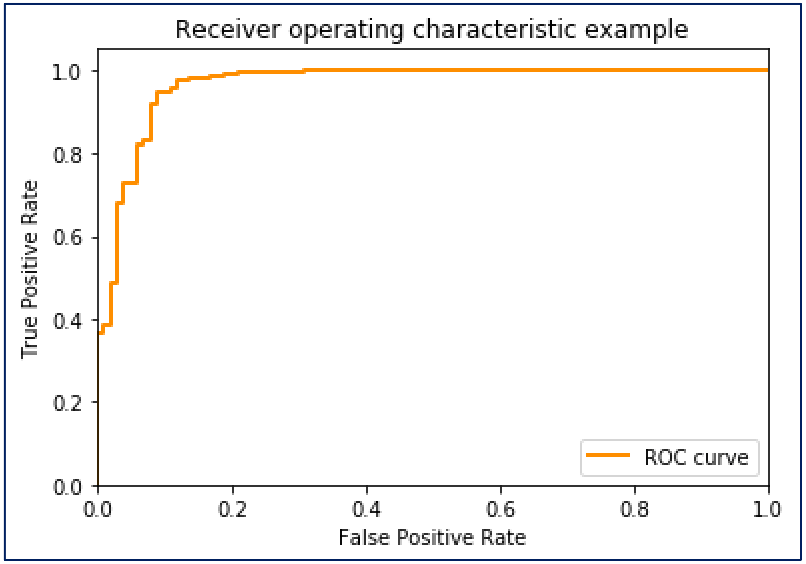
Binary Performance Metrics (3)

- Metrics that take „everything“ into account
  - Accuracy
    - Percentage of data that is predicted correctly
    - $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
  - F1 measure
    - Harmonic mean of precision and recall
    - $F_1 = 2\frac{precision \times recall}{precision+recall}$
  - Matthews correlation coefficient (MCC)
    - Chi-squared correlation between prediction and actual values
    - $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$



Receiver Operator Characteristics (ROC)

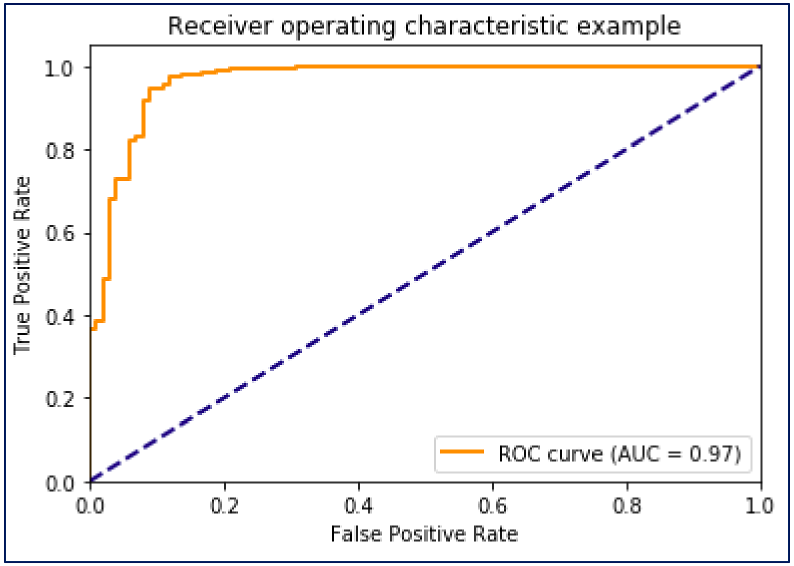
- Plot of true positive rate (TPR) versus false positive rate (FPR)
- Different TPR/FPR values possible due to thresholds for scores



Area Under the Curve (AUC)

- Large Area = Good Performance

- Accounts for tradeoffs between TPR and FPR



Micro and Macro Averaging

- Metrics not directly applicable for more than two classes
  - Accuracy is the exception
- Micro Averaging
  - Expand formulas to use individual positive, negative examples for each class
- Macro Averaging
  - Assume one class as true, combine all other as false
  - Compute metrics for all such combinations
  - Take average
- Example for the true positive rate:
  - $TPR_{micro} = \frac{\sum_{c \in C} TP_c}{\sum_{c \in C} TP_c + \sum_{c \in C} FN_c}$
  - $TPR_{macro} = \frac{\sum_{c \in C} \frac{TP_c}{TP_c + FN_c}}{|C|}$