Lecture 06

Model Evaluation

Supervised vs. Unsupervised Learning

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Prediction Problems: Classification vs. Numeric Prediction

- Classification
  - predicts categorical class labels
  - Discrete (0 or 1)
  - Nominal (Male/ Female, Embarked in Titanic dataset)

- Numeric Prediction
  - Predicts unknown or missing values (Continuous values)
- Typical applications
  - Credit/loan approval:
  - Medical diagnosis: if a tumor is cancerous or benign
  - Fraud detection: if a transaction is fraudulent
  - Estimate of a price of a house

Classification—A Two-Step Process

- Model construction: describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of tuples used for model construction is training set
  - The model is represented as classification rules, decision trees, or mathematical formulae
- Model usage: for classifying future or unknown objects
  - Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set (otherwise overfitting)
  - If the accuracy is acceptable, use the model to classify new data
- Note: If *the test set* is used to select models, it is called validation (test) set

Data set

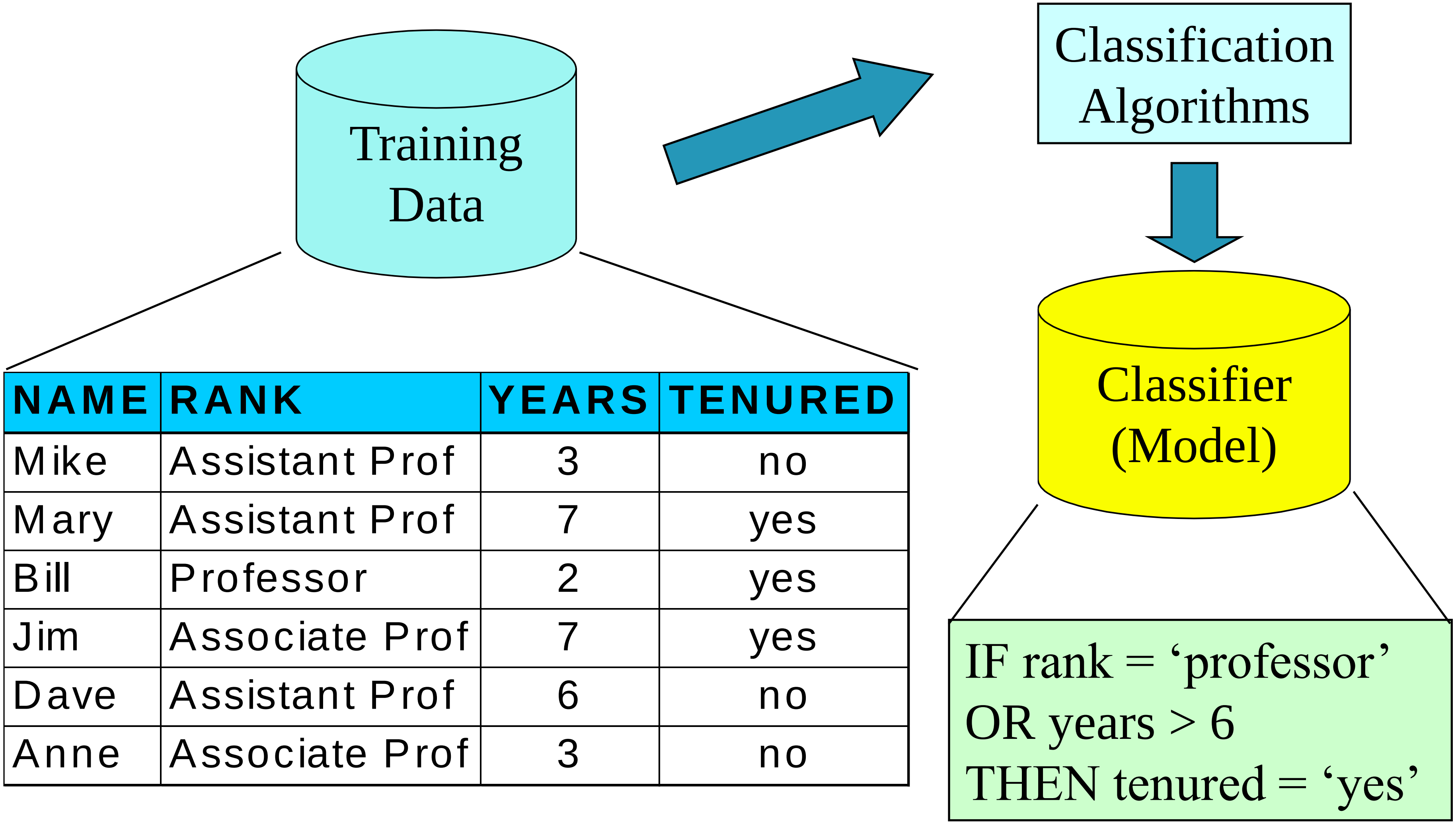
Training Data

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Test Data

NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Process (1): Model Construction



Training Data

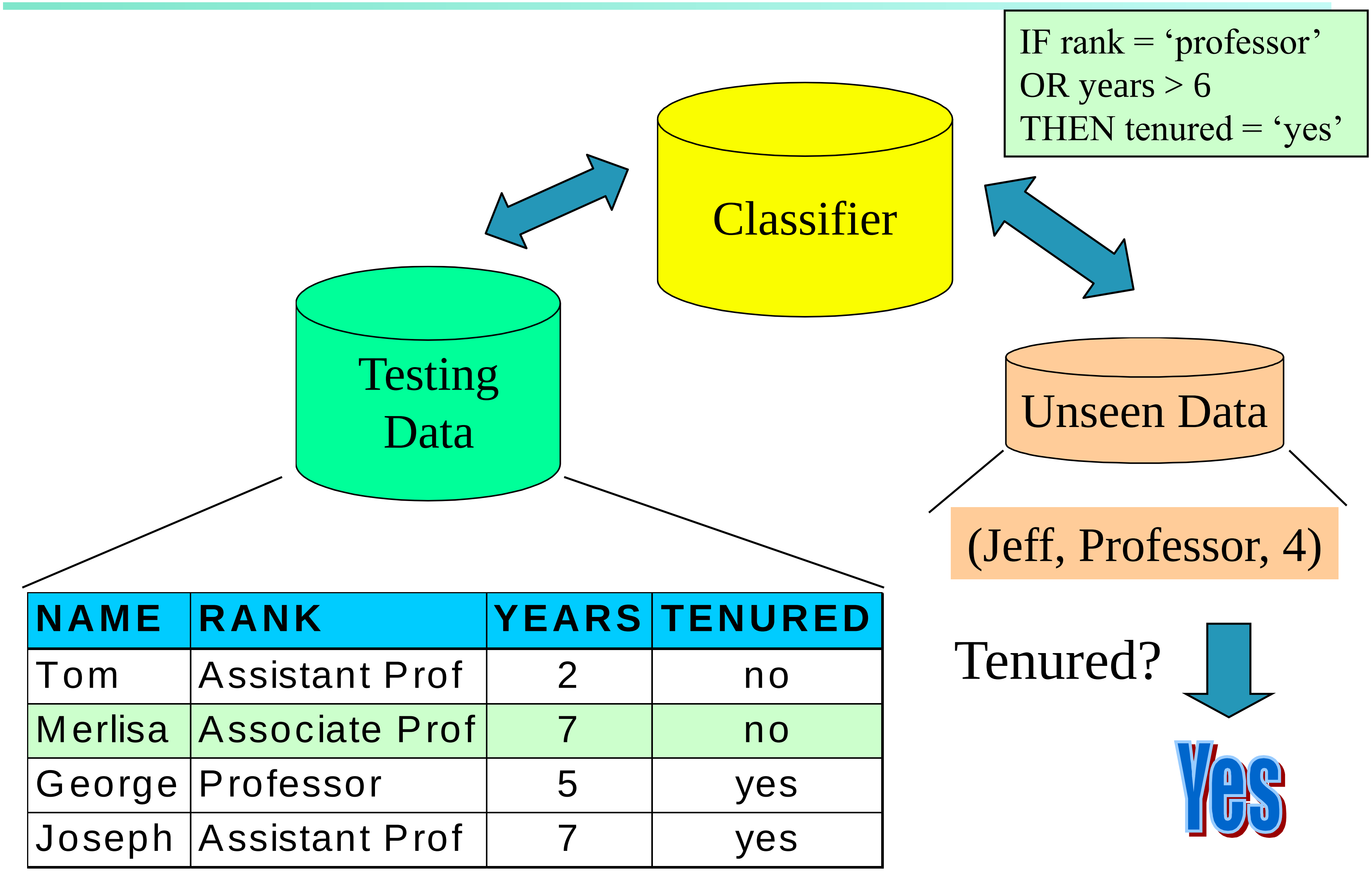
Classification Algorithms

Classifier (Model)

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

Process (2): Using the Model in Prediction



IF rank = 'professor'
OR years > 6
THEN tenured = 'yes'

(Jeff, Professor, 4)

NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Model Evaluation Techniques

- Methods for evaluating a model's performance are divided into two categories
  - Holdout
  - Cross-Validation

- Both methods use unseen test set because if test records are already used in training, the model will simply learn the whole training set and will therefore predict the correct label for all the training set. This is **overfitting.**

Holdout

- Data is divided into three subsets
  - Training set
  - Validation set: Used to assess the performance of the model built in training phase. It provides a test platform for fine-tuning a model’s parameters and selecting the best performing model.
  - Test set: Used to assess the likely future performance of a model.

Cross-Validation

- K-fold cross-validation
  - K equal size subsamples
  - One of the k subsets is used as a test/ validation set and other k-1 subsets are used as a training set. (This is repeated k times).
  - Every data point gets to be in a test set exactly once and gets to be in a training set k-1 times.

Classification Metrics

- When performing the classification predictions, there are four types of outcomes that could occur.
  - **True Positives** are when you predict an observation belongs to a class and it actually belongs to that class
  - **True Negatives** are when you predict an observation doesn’t belong to a class and it actually doesn’t belong to that class
  - **False Positives** occur when you predict an observation belongs to a class when it really does not.
  - **False Negatives** occur when you predict an observation does not belong to a class when in fact it does.

Classifier Evaluation Metrics: Confusion Matrix

Confusion Matrix:

Actual class\Predicted class	1	0
1	**True Positives (TP)**	**False Negatives (FN)**
0	**False Positives (FP)**	**True Negatives (TN)**

Example of Confusion Matrix:

Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	**6954**	**46**	7000
buy_computer = no	**412**	**2588**	3000
Total	7366	2634	10000

Errors (Type-1 and Type-2 Error)

- Type-1 Error: FP
- Type-2 Error: FN

- Example: Biometric
  - Type-1: Possibility of acceptance even with a wrong / unauthorized match
  - Type-2: Possibility of rejection even with an authorized match
  - Which one is sensitive:
    - Type-1

Errors (Type-1 and Type-2 Error)

- Type-1 Error: FP
- Type-2 Error: FN

- Example: Construction model of a bridge
  - Type-1: Predicting that a model is correct when it is not
  - Type-2: Predicting that a model is not correct when it is correct
  - Which one is sensitive:
    - Type-1

Errors (Type-1 and Type-2 Error)

- Type-1 Error: FP
- Type-2 Error: FN

- Example: Medical trials for a drug which is a cure of cancer
  - Type-1: Predicting that a cure is found when it is not the case
  - Type-2: Predicting that a cure is not found when in fact it is the case
  - Which one is sensitive:
    - Type-2 as it could be discarded as no cure and a cure can save millions of lives.

Classifier Evaluation Metrics: Accuracy and Error Rate

A\P	1	0	
1	**TP**	**FN**	**P**
0	**FP**	**TN**	**N**
	P’	**N’**	**All**

- **Classifier Accuracy**: percentage of correct predictions for test data

  Accuracy = (TP + TN)/All

- **Error rate:** *1 – accuracy*, or

  Error rate = (FP + FN)/All

A\P	1	0	
1	**TP**	**FN**	**P**
0	**FP**	**TN**	**N**
	P’	**N’**	**All**

Class Imbalance Problem

- One class may be *rare*, e.g. fraud, or HIV-positive
- Significant *majority of the negative class* and minority of the positive class
- A model may be 99% accurate but 0% useful
- Recall ensures that we are not overlooking the people who have the disease. Its perfect score is 1

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

- Precision ensures that we are not misclassifying too many people as having the disease when they don’t.

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

Classifier Evaluation Metrics: Precision, Recall, and F-measures

A\P	1	0	
1	**TP**	**FN**	**P**
0	**FP**	**TN**	**N**
	P’	**N’**	**All**

- Inverse relationship between precision & recall

- ***F***$_{\beta}$: weighted measure of precision and recall
  - ß<1 focuses more on precision
  - ß>1 focuses more on recall

$$F_\beta = \frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$$

- F-Measure

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

Classifier Evaluation Metrics: Example

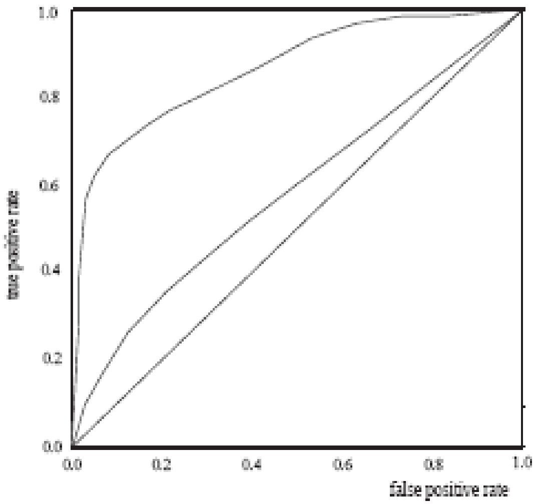
Actual Class\ Predicted class	Cancer =yes	Cancer=No
Cancer=yes	90	210
Cancer=no	140	9560

- *Precision* = 90/230 = 39.13%
- *Recall* = 90/300 = 30.00%

Model Selection: ROC Curves

A\P	1	0	
1	**TP**	**FN**	**P**
0	**FP**	**TN**	**N**
	P’	**N’**	**All**

- **ROC** (Receiver Operating Characteristics) curves: for visual comparison of classification models
- Originated from signal detection theory
- Shows the trade-off between the true positive rate and the false positive rate
- The area under the ROC curve is a measure of the accuracy of the model
- The closer to the diagonal line (i.e., the closer the area is to 0.5), the less accurate is the model



- ❑ Vertical axis represents the true positive rate
- ❑ Horizontal axis rep. the false positive rate
- ❑ The plot also shows a diagonal line
- ❑ A model with perfect accuracy will have an area of 1.0

TPR = TP/(TP+FN)
FPR = FP/(FP+TN)

Issues Affecting Model Selection

- **Accuracy**
  - classifier accuracy: predicting class label
- **Speed**
  - time to construct the model (training time)
  - time to use the model (classification/prediction time)
- **Robustness**: handling noise and missing values
- **Scalability**: efficiency in disk-resident databases
- **Interpretability**
  - understanding and insight provided by the model
- Other measures, e.g., goodness of rules, such as decision tree size or compactness of classification rules