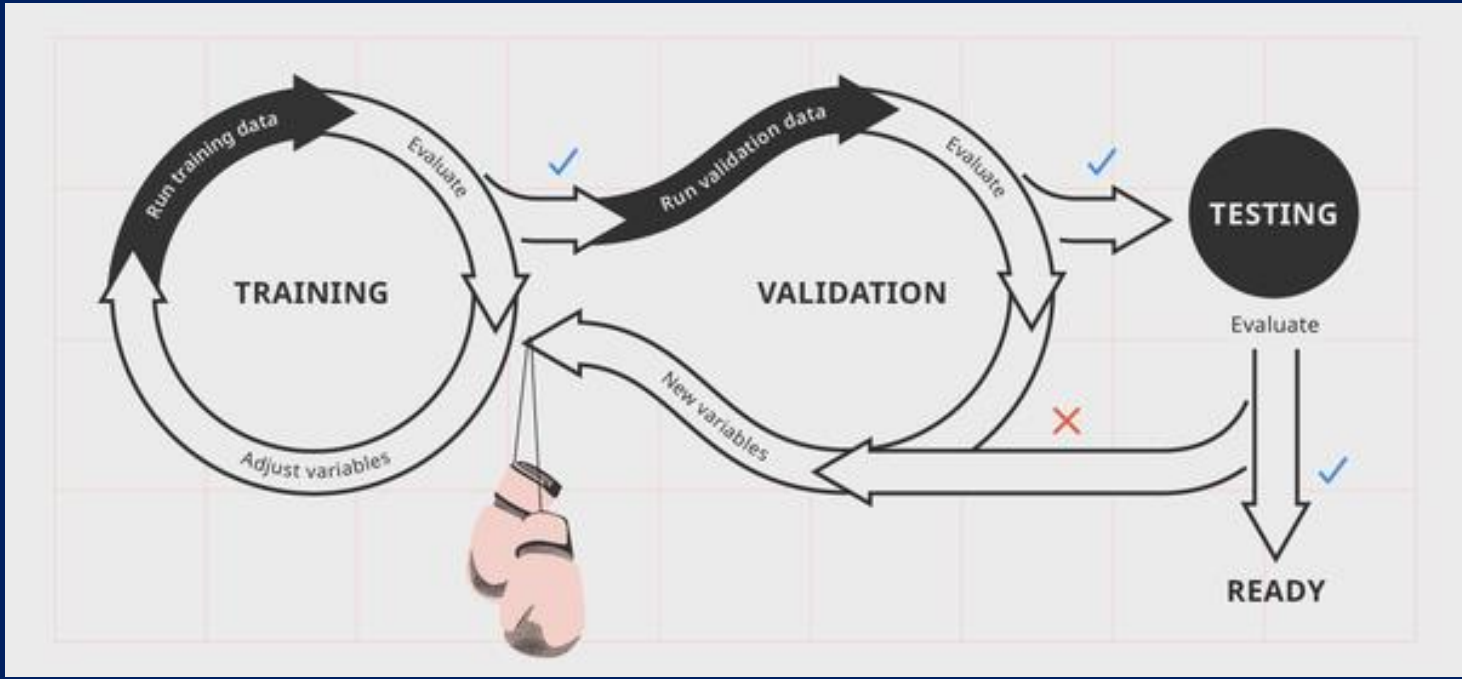
MCSE0007: Machine Learning



Training and Testing, Evaluation

Class Presentations on Machine Learning by Prof. Anand Singh Jalal

An Example Application

- An emergency room in a hospital measures 17 variables (e.g., blood pressure, age, etc) of newly admitted patients.
- A decision is needed: whether to put a new patient in an Intensive-Care Unit (ICU).
- Due to the high cost of ICU, those patients who may survive less than a month are given higher priority.
- Problem: to predict high-risk patients and discriminate them from low-risk patients.

Another Application

- A credit card company receives thousands of applications for new cards. Each application contains information about an applicant,
  - ✓ age
  - ✓ Marital status
  - ✓ annual salary
  - ✓ outstanding debts
  - ✓ credit rating
  - ✓ etc.
- Problem: To decide whether an application should approved, or to classify applications into two categories, approved and not approved.

Overview

- Like human learning from past experiences.
- But, A computer does not have “experiences”.
- A computer system learns from data, which represent some “past experiences” of an application domain.
- Our focus: learn a target function that can be used to predict the values of a discrete class attribute, e.g., approve or not-approved, and high-risk or low risk.
- The task is commonly called: Supervised learning, classification, or inductive learning.

Overview …

- Data: A set of data records (also called examples, instances or cases) described by
  - *k* attributes: A_1, A_2, … A_k.
  - a class: Each example is labelled with a pre-defined class.
- Goal: To learn a classification model from the data that can be used to predict the classes of new (future, or test) cases/instances.

An example: data (loan application)

Approved or not

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**

An example: the learning task

- Learn a classification model from the data
- Use the model to classify future loan applications into
  - ✓ Yes (approved) and
  - ✓ No (not approved)
- What is the class for following case/instance?

Age	Has_Job	Own_house	Credit-Rating	Class
young	false	false	good	?

Overview …
Supervised vs. Unsupervised Learning

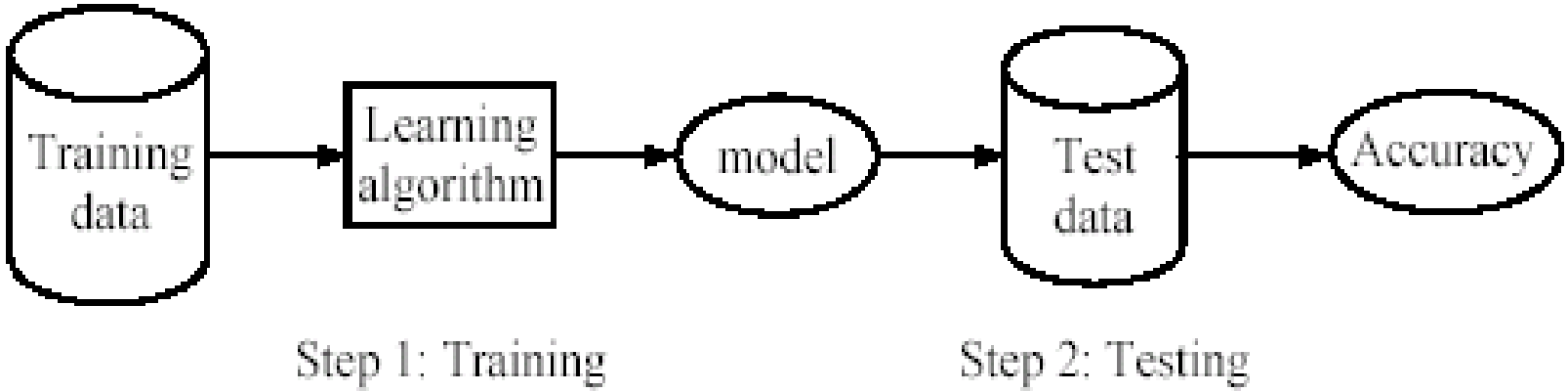
- Supervised learning: classification is seen as supervised learning from examples.
  - Supervision: The data (observations, measurements, etc.) are labeled with pre-defined classes.
  - Test data are classified into these classes too.
- Unsupervised learning (clustering)
  - Class labels of the data are unknown
  - Given a set of data, the task is to establish the existence of classes or clusters in the data

Overview …

Supervised learning process: two steps

- Learning (training): Learn a model using the training data
- Testing: Test the model using unseen test data to assess the model accuracy

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}},$$



Evaluation Methods

- **Holdout set**: The available data set *D* is divided into two disjoint subsets,
  - ✓ the *training set* D_{train} (for learning a model)
  - ✓ the *test set* D_{test} (for testing the model)
- **Important:** training set should not be used in testing and the test set should not be used in learning.
  - **Unseen test set provides a unbiased estimate of accuracy.**
- The test set is also called the holdout set. (the examples in the original data set *D* are all labeled with classes.)
- This method is mainly used when the data set *D* is large.

Evaluation Methods …

- **Validation set**: the available data is divided into three subsets,
  - ✓ a training set,
  - ✓ a validation set and
  - ✓ a test set.
- A validation set is used frequently for estimating parameters in learning algorithms.
- In such cases, the values that give the best accuracy on the validation set are used as the final parameter values.
- Cross-validation can be used for parameter estimating as well.

Evaluation Methods: Cross Validation

- **k-fold cross-validation**: The available data is partitioned into *k* equal-size disjoint subsets.
- Use each subset as the test set and combine the rest ***k*-1** subsets as the training set to learn a classifier.
- The procedure is run ***k*** times, which give ***k*** accuracies.
- The final estimated accuracy of learning is the average of the ***k*** accuracies.
- 10-fold and 5-fold cross-validations are commonly used.
- This method is used when the available data is not large.

Evaluation Methods: Cross Validation

Evaluation Methods: Cross Validation …

- cross-validation generates an approximate estimate of how well the learned model will do on "unseen" data
- by averaging over different partitions it is more robust than just a single train/validate partition of the data
- "k-fold" cross-validation is a generalization
  - ✓ partition data into disjoint validation subsets of size n/k
  - ✓ train, validate, and average over the v partitions
  - ✓ e.g., k=10 is commonly used
- k-fold cross-validation is approximately k times computationally more expensive than just fitting a model to all of the data

Classification Measures

- **Predictive accuracy**

$$\text{Accuracy} = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$

- **Efficiency**
  - ✓ time to construct the model
  - ✓ time to use the model
- **Robustness:** handling noise and missing values
- **Scalability:** efficiency in disk-resident databases
- **Interpretability:**
  - understandable and insight provided by the model
- **Compactness of the model:** size of the tree, or the number of rules.

Classification Measures

Confusion Matrix

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

TP: the number of correct classifications of the positive examples

FN: the number of incorrect classifications of the positive examples

FP: the number of incorrect classifications of the negative examples

TN: the number of correct classifications of the negative examples

Precision and Recall Measures

- Precision *p* is the number of correctly classified positive examples divided by the total number of examples that are classified as positive.
- Recall *r* is the number of correctly classified positive examples divided by the total number of actual positive examples in the test set.

$$p = \frac{TP}{TP + FP}. \qquad r = \frac{TP}{TP + FN}.$$

F1-value (also called F1-score)

It is hard to compare two classifiers using two measures.

F_1 score combines precision and recall into one measure

The F score is used to measure a test's accuracy

$$F_1 = \frac{2pr}{p+r} \quad , \quad F_1 = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

F1 Score is the weighted average of Precision and Recall.

The F score reaches the best value, meaning perfect precision and recall, at a value of 1. The worst F score, which means lowest precision and lowest recall, would be a value of 0.

Receiver Operating Characteristics Curve

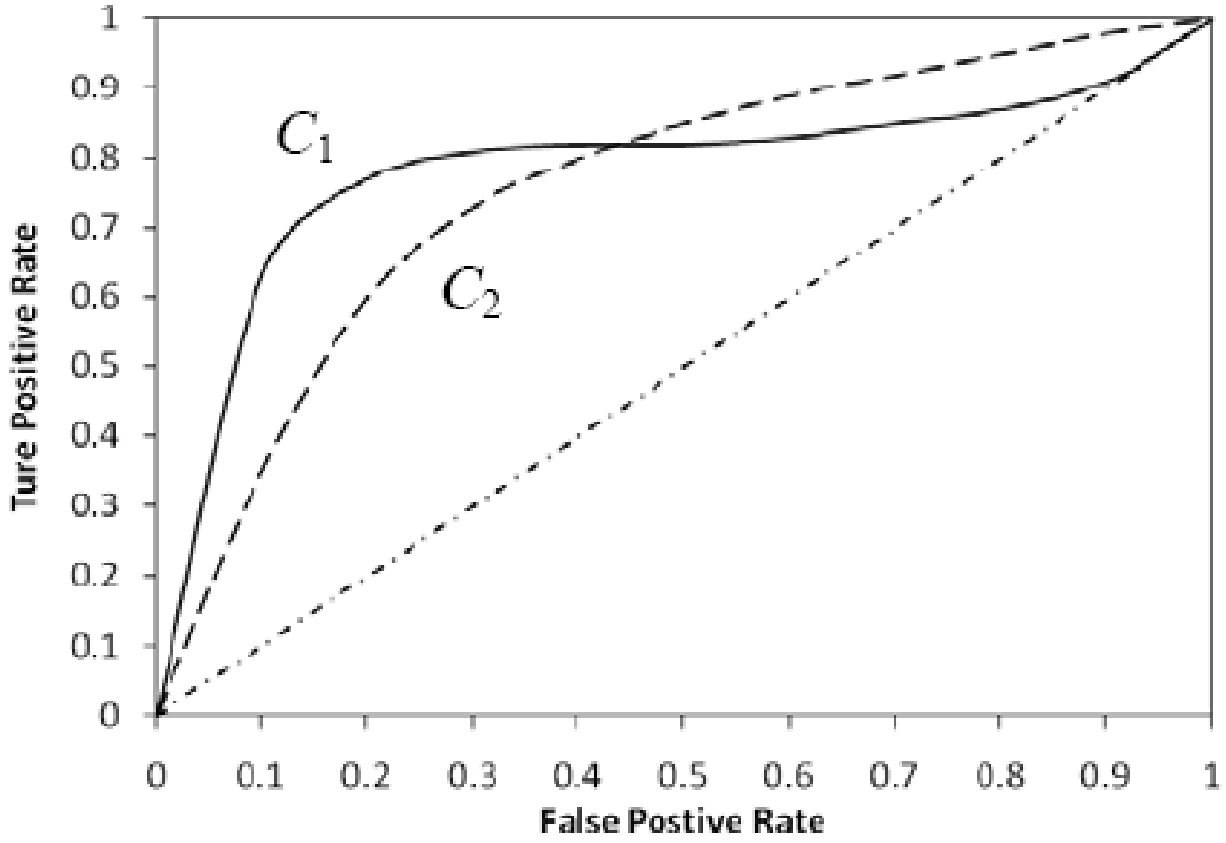
- It is commonly called the ROC curve.
- It is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR).
- True Positive Rate (TPR) :

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

- False Positive Rate (FPR):

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

Example ROC curves



ROC cuvers for two classifiers (C1 and C2) on the same data

Area Under the Curve (AUC)

- Which classifier is better, C_1 or C_2?
  - It depends on which region you talk about.
- Can we have one measure?
  - Yes, we compute the area under the curve (AUC)
- If AUC for C_i is greater than that of C_j, it is said that C_i is better than C_j.
  - If a classifier is perfect, its AUC value is 1
  - If a classifier makes all random guesses, its AUC value is 0.5.

Sensitivity and Specificity

- In statistics, there are two other evaluation measures:
  - Sensitivity: Same as TPR
  - Specificity: Also called True Negative Rate (TNR)

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

- Then we have

$$FPR = 1 - specificity$$

Summary

		Actual Label	
		Positive	Negative
Predicted Label	Positive	**True Positive (TP)**	**False Positive (FP)**
	Negative	**False Negative (FN)**	**True Negative (TN)**

Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions that are correct
Precision	TP / (TP + FP)	The percentage of positive predictions that are correct
Sensitivity (Recall)	TP / (TP + FN)	The percentage of positive cases that were predicted as positive
Specificity	TN / (TN + FP)	The percentage of negative cases that were predicted as negative

Question

Given the confusion matrix, find:

Classification Accuracy, Recall, Precision, F-measure

n = 165	Predicted: No	Predicted: Yes	
Actual: No	Tn =50	FP=10	60
Actual: Yes	Fn=5	Tp=100	105
	55	110	

Quiz

The number of correct classifications of the negative examples ……
The number of correct classifications of the positive examples ……
The number of incorrect classifications of the negative examples …….
The number of incorrect classifications of the positive examples ……

Any Questions ?