Big data: architectures and data analytics

Spark Mllib - Introduction to Machine Learning

Credits to:

Elena Baralis, DAUIN, Politecnico di Torino
Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Dr. Christoph F. Eick - Director UH Data Analysis and Intelligent Systems Lab, University of Houston
Ethem Alpaydin - Department of Computer Engineering, Bogaziçi University

What is Machine Learning?

- Machine learning is **programming** computers to **optimize a performance** criterion using **example data or past experience**
- There is no need to “learn” to calculate payroll
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (routing on a computer network)
  - Solution needs to be adapted to particular cases (user biometrics)

What is Machine Learning?

- Data is cheap and abundant (data warehouses,...); knowledge is expensive and scarce.
- Build a model that is *a good and useful approximation* to the data.

What is Machine Learning?

- ML algorithms:
  - Improve their performance
  - at some task
  - with experience
- Role of Statistics: inference from a sample
- Role of Computer science: efficient algorithms to
  - Solve the optimization problem
  - Representing and evaluating the model for inference

Growth of Machine Learning

Machine learning has been out for more than 50 years, but in the last 10 years usage has exploded:

- Improved machine learning algorithms
- Improved data capture, networking, faster computers
- New sensors / IO devices
- It turns out to be difficult to extract knowledge from human experts→*failure of expert systems in the 1980's.*

ML – classification

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Clustering
- Reinforcement Learning
- Itemset and Association Rule Analysis
- ...

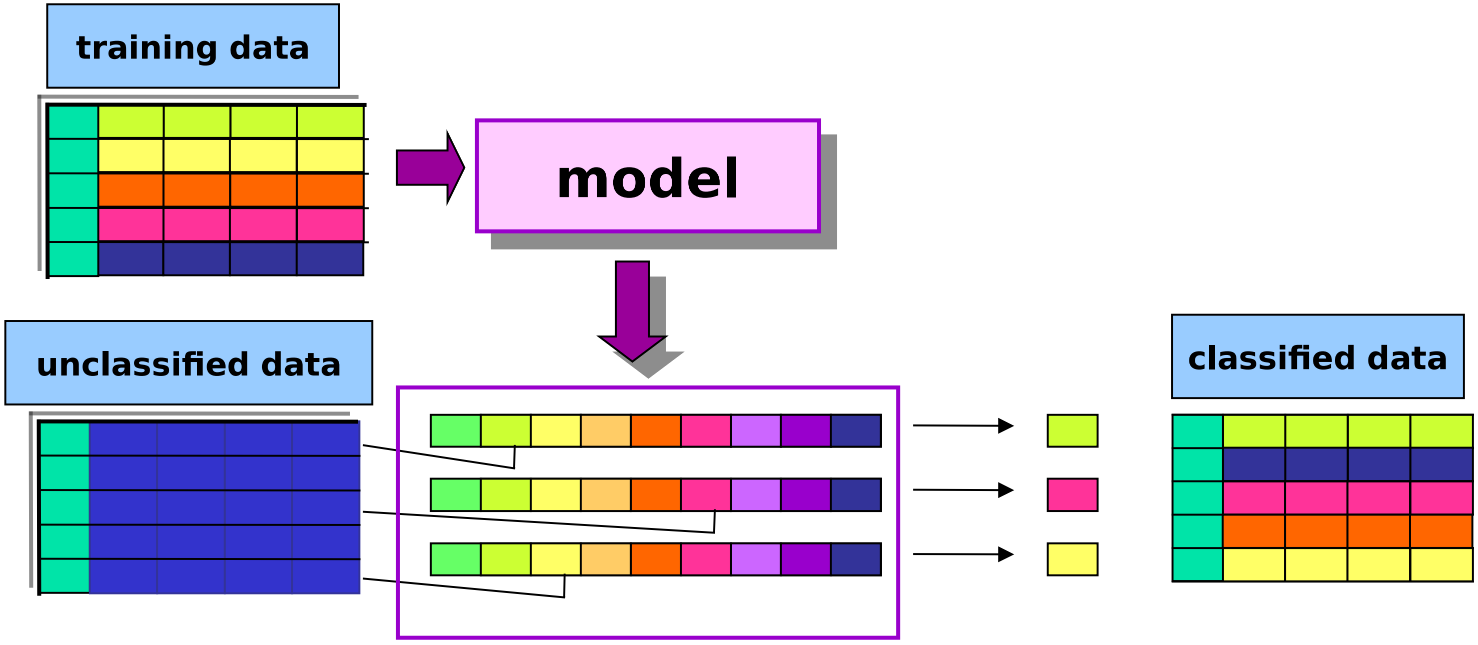
ML – classification

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Clustering
- Reinforcement Learning
- Itemset and Association Rule Analysis
- ...

What we will see with Spark MLlib

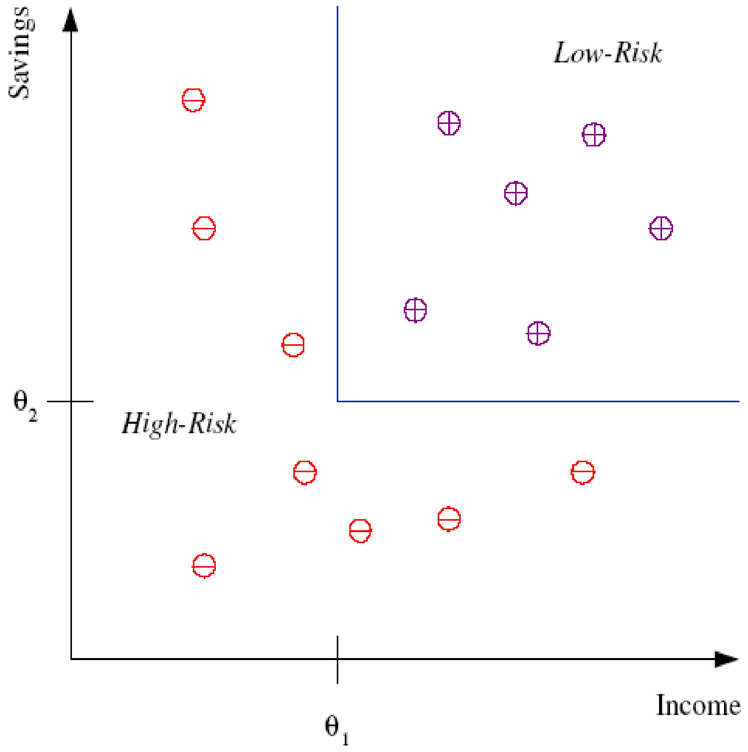
Classification

Objective: prediction of a class label



Classification: example

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their *income* and *savings*



Discriminant: IF *income* > θ_1 AND *savings* > θ_2

THEN low-risk ELSE high-risk

Model

Classification: Face Recognition

Training examples of a person

Test images

AT&T Laboratories, Cambridge UK
http://www.uk.research.att.com/facedatabase.html

Classification: Applications

Aka Pattern recognition

- Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
- Character recognition: Different handwriting styles.
- Speech recognition: Temporal dependency.
  - Use of a dictionary or the syntax of the language.
  - Sensor fusion: Combine multiple modalities; eg, visual (lip image) and acoustic for speech
- Medical diagnosis: From symptoms to illnesses
- Web Advertising: Predict if a user clicks on an ad on the Internet.

Prediction: Regression

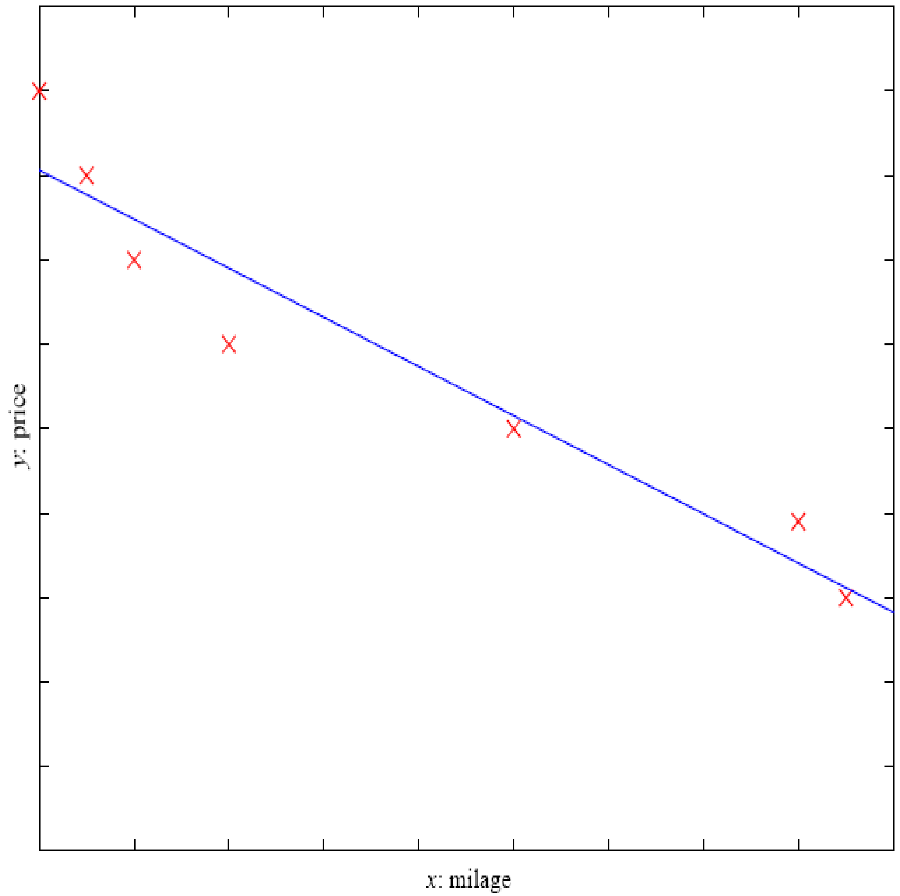
- Example: Price of a used car
- x : car attributes

  y : price

  $y = g(x \mid \theta)$

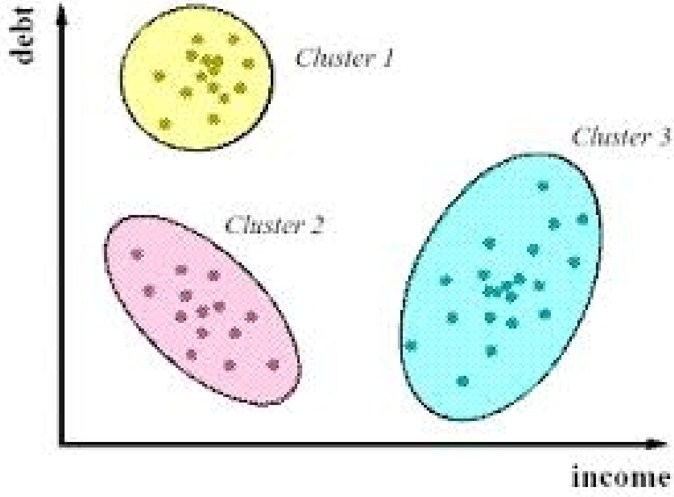
  $g(\ )$ model,

  θ parameters



Unsupervised Learning

- Learning “what normally happens”
- No output
- Clustering: Grouping similar instances
- Other applications: Summarization, Association Analysis



Reinforcement Learning

- Topics:
  - Policies: what actions should an agent take in a particular situation
  - Utility estimation: how good is a state (→used by policy)
- No supervised output but delayed reward
- Credit assignment problem (what was responsible for the outcome)
- Applications:
  - Game playing
  - Robot in a maze

Learning Associations

- Basket analysis:

  $P(Y \mid X)$ probability that somebody who buys X also buys Y where X and Y are products/services.

  Example: P (chips | beer) = 0.7

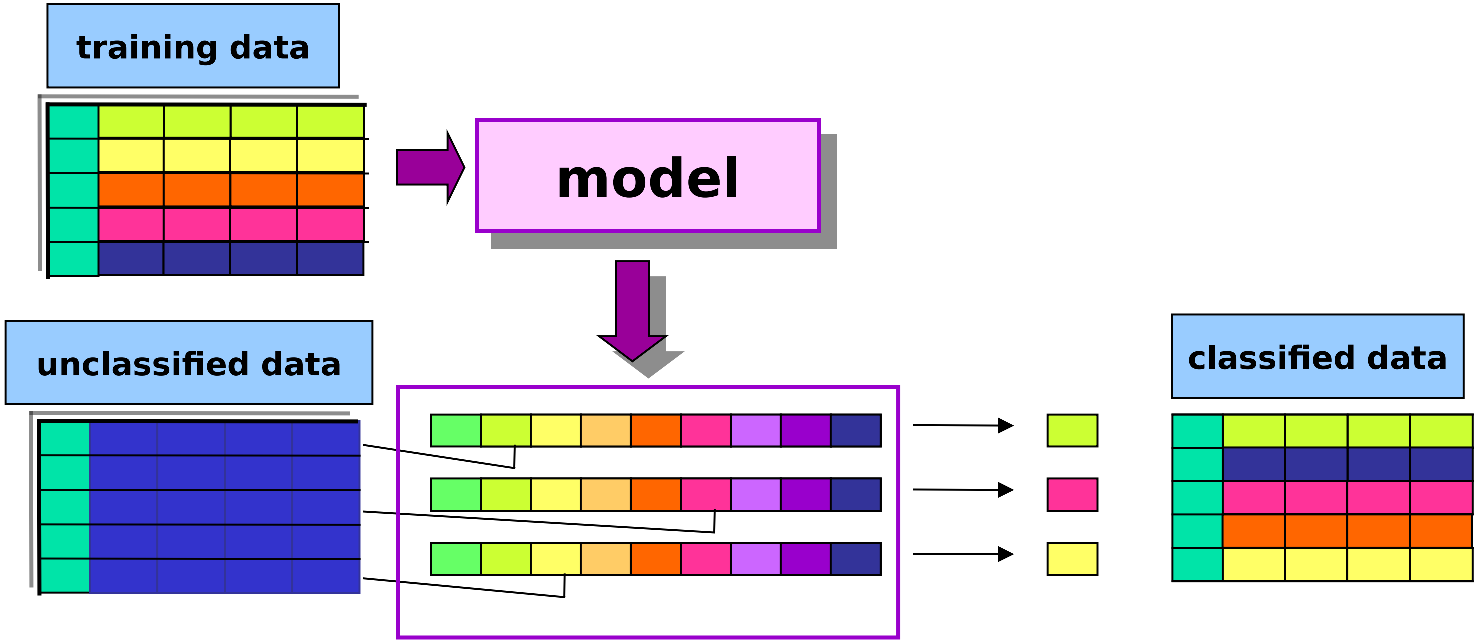
Market-Basket transactions

TID	*Items*
1	**Bread, Milk**
2	**Bread, Diaper, Beer, Eggs**
3	**Milk, Diaper, Beer, Coke**
4	**Bread, Milk, Diaper, Beer**
5	**Bread, Milk, Diaper, Coke**

Classification - Model performance evaluation

Classification

Objective: prediction of a class label



Classification: definition

- Given
  - a collection of class labels
  - a collection of data objects labelled with a class label
- Find a descriptive profile of each class, which will allow the assignment of unlabeled objects to the appropriate class

Classification techniques

- Decision trees
- Random forests
- Neural Networks (Multilayer perceptron)
- Naïve Bayes
- Linear Support Vector Machines
- …

Classification techniques

- Decision trees
- Random forests
- Neural Networks (Multilayer perceptron)
- Naïve Bayes
- Linear Support Vector Machines
- …

All these are available in MLlib

Model evaluation

- Methods for performance evaluation
  - Partitioning techniques for training and validation sets

Definitions

- **Training set:** Collection of labeled data objects used to learn the classification model

- **Validation set:** Collection of labeled data objects used to validate the classification model, i.e., tune the parameters of a classifier

- **Test set:** A set of examples used only to assess the performance of a fully specified classifier

But ...

- Sometimes there is not proper test set
- In this case often “validation set” and “test set” are referred to the same set
- More on this topic later...

Methods of estimation

- Partitioning labeled data in
  - Training / validation /(test)
- Several partitioning techniques
  - Fixed split ratio
  - Cross validation
  - ...

Holdout

- Fixed partitioning
  - reserve 2/3 for training and 1/3 for validation
- Appropriate for large datasets
  - may be repeated several times
    - repeated holdout

Cross validation

- Cross validation
  - partition data into k disjoint subsets (i.e., folds)
  - k-fold: train on k-1 partitions, validate on the remaining one
    - repeat for all folds
  - reliable accuracy estimation, not appropriate for very large datasets
- Leave-one-out
  - cross validation for k=n
  - only appropriate for very small datasets

Evaluation of classification techniques

- Quality of the prediction
  - Confusion matrix
  - Accuracy
  - Precision, Recall, F-measure
- Efficiency
  - model building time
  - classification time
- Scalability
  - training set size
  - attribute number
- Robustness
  - noise, missing data
- Interpretability
  - model interpretability
  - model compactness

Methods for performance evaluation

- Objective
  - reliable estimate of performance
- Performance of a model may depend on other factors besides the learning algorithm
  - Class distribution
  - Cost of misclassification
  - Size of training and test sets

Metrics for model evaluation

- Evaluate the predictive accuracy of a model
- Confusion matrix
  - binary classifier

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Accuracy

- Most widely-used metric for model evaluation

$$\text{Accuracy} = \frac{\text{Number of correctly classified objects}}{\text{Number of classified objects}}$$

- Not always a reliable metric

Accuracy

- For a binary classifier

	PREDICTED CLASS	
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Limitations of accuracy

- Consider a binary problem
  - Cardinality of Class 0 = 9900
  - Cardinality of Class 1 = 100
- Model

$() \rightarrow$ *class 0*

  - Model predicts everything to be class 0
    - accuracy is 9900/10000 = 99.0 %
- Accuracy is misleading because the model does not detect any class 1 object

Limitations of accuracy

- Classes may have different importance
  - Misclassification of objects of a given class is more important
  - e.g., ill patients erroneously assigned to the healthy patients class
- Accuracy is not appropriate for
  - unbalanced class label distribution
  - different class relevance

Class specific measures

- Evaluate separately for each class C

$$\text{Recall (r)} = \frac{\text{Number of objects correctly assigned to C}}{\text{Number of objects belonging to C}}$$

$$\text{Precision (p)} = \frac{\text{Number of objects correctly assigned to C}}{\text{Number of objects assigned to C}}$$

- Maximize

$$\text{F - measure (F)} = \frac{2rp}{r + p}$$

Class specific measures

For a binary classification problem

on the confusion matrix, for the positive class

$$\text{Precision (p)} = \frac{a}{a + c}$$

$$\text{Recall (r)} = \frac{a}{a + b}$$

$$\text{F - measure (F)} = \frac{2rp}{r + p} = \frac{2a}{2a + b + c}$$