

Machine Learning for Cities

Prof. Luis Gustavo Nonato

NYU / CUSP - GX 5006

January 24, 2017

Course Structure

Content

Course Structure

Course Structure

Content

- Principal Component Analysis

Course Structure

Content

- Principal Component Analysis
- Regression: Part I and II

Course Structure

Content

- Principal Component Analysis
- Regression: Part I and II
- Clustering and Classification

Course Structure

Content

- Principal Component Analysis
- Regression: Part I and II
- Clustering and Classification
- SVM for Classification

Course Structure

Content

- Principal Component Analysis
- Regression: Part I and II
- Clustering and Classification
- SVM for Classification
- Tree-based Regr. and Classif.

Course Structure

Content

- Principal Component Analysis
- Regression: Part I and II
- Clustering and Classification
- SVM for Classification
- Tree-based Regr. and Classif.
- Graph-based Methods

Course Structure

Content

- Principal Component Analysis
- Regression: Part I and II
- Clustering and Classification
- SVM for Classification
- Tree-based Regr. and Classif.
- Graph-based Methods
- Bayesian Networks

Course Structure

Content

- Principal Component Analysis
- Regression: Part I and II
- Clustering and Classification
- SVM for Classification
- Tree-based Regr. and Classif.
- Graph-based Methods
- Bayesian Networks



ANALYTICS | BIG DATA | HADOOP | DATA PLUMBING | DATAVIZ | JOBS

The Algorithms Every Data
Scientist Should Know

- Principal Component Analysis
- Regression: Part I and II
- Clustering and Classification
- SVM for Classification
- Tree-based Repr. and Classif.
- Graph-based Methods
- Bayesian Networks



Data Science Central

ANALYTICS **BIG DATA** **HADOOP** **DATA PLUMBING** **DATAVIZ** **JOBS**

The Algorithms Every Data Scientist Should Know

```

graph LR
    ML[Machine Learning] --> R[regression]
    ML --> B[bayesian]
    ML --> REG[regularization]
    ML --> IB[instance based]
    ML --> DT[decision tree]

    R --> R1[Ordinary Least Square Regression (OLS)]
    R --> R2[Linear Regression]
    R --> R3[Logistic Regression]
    R --> R4[unregularized regression]
    R --> R5[Bayesian Adaptive Regression Spline (BARS)]
    R --> R6[Locally Estimated Scatterplot Smoothing (LOESS)]
    R --> R7[Jackknife Regression]

    B --> B1[Native Bayes]
    B --> B2[Naïve Bayes]
    B --> B3[Naïve Bayes with Multinomial Naïve Bayes]
    B --> B4[Averaged One-Dependence Estimators (AOE)]
    B --> B5[Bayesian Network (BN)]
    B --> B6[Bayesian Belief Network (BBN)]
    B --> B7[Conditional Random Fields (CRF)]

    REG --> REG1[Ridge Regression]
    REG --> REG2[Least Absolute Shrinkage and Selection Operator (LASSO)]
    REG --> REG3[Class's non-]
    REG --> REG4[Least-Angle Regression (LARS)]

    IB --> IB1[k-Nearest Neighbour (KNN)]
    IB --> IB2[Learning Nearest Neighbour (LNN)]
    IB --> IB3[Self-Organizing Map (SOM)]
    IB --> IB4[Locally Weighted Learning (LWL)]

    DT --> DT1[Classification and Regression Tree (CART)]
    DT --> DT2[Random Decision Forests (RDF)]
    DT --> DT3[C4.5 and C4.5 Improved versions of Quinlan's ID3 algorithm]
    DT --> DT4[Chi-squared Automatic Interaction Detector (CHAID)]
    DT --> DT5[Decision Stump]
    DT --> DT6[Random Forests]
    DT --> DT7[Conditional Decision Forest]
  
```

Machine Learning

- regression**
 - Ordinary Least Square Regression (OLS)
 - Linear Regression
 - Logistic Regression
 - unregularized regression
 - Bayesian Adaptive Regression Spline (BARS)
 - Locally Estimated Scatterplot Smoothing (LOESS)
 - Jackknife Regression
- bayesian**
 - Native Bayes
 - Naïve Bayes
 - Naïve Bayes with Multinomial Naïve Bayes
 - Averaged One-Dependence Estimators (AOE)
 - Bayesian Network (BN)
 - Bayesian Belief Network (BBN)
 - Conditional Random Fields (CRF)
- regularization**
 - Ridge Regression
 - Least Absolute Shrinkage and Selection Operator (LASSO)
 - Class's non-
 - Least-Angle Regression (LARS)
- instance based**
also called case-based, memory-based
 - k-Nearest Neighbour (KNN)
 - Learning Nearest Neighbour (LNN)
 - Self-Organizing Map (SOM)
 - Locally Weighted Learning (LWL)
- decision tree**
 - Classification and Regression Tree (CART)
 - Random Decision Forests (RDF)
 - C4.5 and C4.5 Improved versions of Quinlan's ID3 algorithm
 - Chi-squared Automatic Interaction Detector (CHAID)
 - Decision Stump
 - Random Forests
 - Conditional Decision Forest

Course Structure

This course has two sections running in parallel:

Course Structure

This course has two sections running in parallel:

- 70% of overlap in terms of content

Course Structure

This course has two sections running in parallel:

- 70% of overlap in terms of content
- The same number of assignments, exams and projects

Course Structure

This course has two sections running in parallel:

- 70% of overlap in terms of content
- The same number of assignments, exams and projects
- The same level in terms of requirements.

Course Structure

Assignments, exam, and final projects (gradings)

Course Structure

Assignments, exam, and final projects (gradings)

- Four assignments (40% of the grade)

Course Structure

Assignments, exam, and final projects (gradings)

- Four assignments (40% of the grade)
- Mid term exam (20% of the grade)

Course Structure

Assignments, exam, and final projects (gradings)

- Four assignments (40% of the grade)
- Mid term exam (20% of the grade)
- Final project (40% of the grade)

Course Structure

Assignments, exam, and final projects (gradings)

- Four assignments (40% of the grade)
- Mid term exam (20% of the grade)
- Final project (40% of the grade)

All material will be in NYU Classes !!

Course Structure

Assignments, exam, and final projects (gradings)

- Four assignments (40% of the grade)
- Mid term exam (20% of the grade)
- Final project (40% of the grade)

All material will be in NYU Classes !!

[CUSP-GX-5006-classesmaterial.pdf](#) will be updated every week !!

Course Structure

Working in groups.

Course Structure

Working in groups.

- Assignments and Final Project can be done in group
(at most 3 people)

Course Structure

Working in groups.

- Assignments and Final Project can be done in group
(at most 3 people)
- Mid term exam is individual

Course Structure

Working in groups.

- Assignments and Final Project can be done in group (at most 3 people)
- Mid term exam is individual

The deliverable for each assignment will be a short report describing methods and achievements/results.

Course Structure

Working in groups.

- Assignments and Final Project can be done in group (at most 3 people)
- Mid term exam is individual

The deliverable for each assignment will be a short report describing methods and achievements/results.

A model for the report will be made available.

Course Structure

Working in groups.

- Assignments and Final Project can be done in group (at most 3 people)
- Mid term exam is individual

The deliverable for each assignment will be a short report describing methods and achievements/results.

A model for the report will be made available.

Each team can propose a theme/problem for the Final Project.

Course Structure

Working in groups.

- Assignments and Final Project can be done in group (at most 3 people)
- Mid term exam is individual

The deliverable for each assignment will be a short report describing methods and achievements/results.

A model for the report will be made available.

Each team can propose a theme/problem for the Final Project.

The team name, members, and the proposed problem should be submitted for approval no later than March 21st.

Course Structure

Working in groups.

- Assignments and Final Project can be done in group (at most 3 people)
- Mid term exam is individual

The deliverable for each assignment will be a short report describing methods and achievements/results.

A model for the report will be made available.

Each team can propose a theme/problem for the Final Project.

The team name, members, and the proposed problem should be submitted for approval no later than March 21st.

If a team is not able to find a theme/problem I can suggest one.

Class Dynamics

Our lessons will be divided in two parts:

Class Dynamics

Our lessons will be divided in two parts:

- 1 Theoretical content

Class Dynamics

Our lessons will be divided in two parts:

- 1 Theoretical content
- 2 Lab: from theory to practice

Class Dynamics

Our lessons will be divided in two parts:

- 1 Theoretical content
- 2 Lab: from theory to practice

Put ML algorithms to working in practice, using real data sets !!

Class Dynamics

Our lessons will be divided in two parts:

- 1 Theoretical content
- 2 Lab: from theory to practice

Put ML algorithms to working in practice, using real data sets !!

No deep knowledge about python will be required.

Class Dynamics

Our lessons will be divided in two parts:

- 1 Theoretical content
- 2 Lab: from theory to practice

Put ML algorithms to working in practice, using real data sets !!

No deep knowledge about python will be required.

The code will be as simple as possible and easily understandable, even for students not so experienced in computer programming.

Machine Learning: What and Why?

Machine Learning: What and Why?

“Given an email, is it a spam?”

Machine Learning: What and Why?

“Given an email, is it a spam?”

We can hardly come up with an algorithm to answer this question without some knowledge about what characterize a spam.

Machine Learning: What and Why?

“Given an email, is it a spam?”

We can hardly come up with an algorithm to answer this question without some knowledge about what characterize a spam.

Data can make up for the lack of knowledge !!

Machine Learning: What and Why?

“Given an email, is it a spam?”

We can hardly come up with an algorithm to answer this question without some knowledge about what characterize a spam.

Data can make up for the lack of knowledge !!

It is assumed that there is a “hidden” model/process capable of analyzing the content of an email and then decides whether it is a spam or not.

Machine Learning: What and Why?

“Given an email, is it a spam?”

We can hardly come up with an algorithm to answer this question without some knowledge about what characterize a spam.

Data can make up for the lack of knowledge !!

It is assumed that there is a “hidden” model/process capable of analyzing the content of an email and then decides whether it is a spam or not.

The goal is to *learn* the hidden model/process from data !!

Taxonomy

There are a variety of learning methods.

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1 Predictive or Supervised

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1 Predictive or Supervised

- Regression: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1 Predictive or Supervised

- Regression: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1 Predictive or Supervised

- Regression: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Given \mathbf{x} predict y

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1 Predictive or Supervised

- Regression: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Given \mathbf{x} predict y

2 Descriptive or Unsupervised

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1 Predictive or Supervised

- Regression: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Given \mathbf{x} predict y

2 Descriptive or Unsupervised

- Clustering

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1 Predictive or Supervised

- Regression: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Given \mathbf{x} predict y

2 Descriptive or Unsupervised

- Clustering
- Manifold Learning

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1 Predictive or Supervised

- Regression: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Given \mathbf{x} predict y

2 Descriptive or Unsupervised

- Clustering
- Manifold Learning
- Relationship between instances

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1 Predictive or Supervised

- Regression: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Given \mathbf{x} predict y

2 Descriptive or Unsupervised

- Clustering
- Manifold Learning
- Relationship between instances

Is there “any structure” in the data?

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1 Predictive or Supervised

- Regression: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification: $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Given \mathbf{x} predict y

2 Descriptive or Unsupervised

- Clustering
- Manifold Learning
- Relationship between instances

Is there “any structure” in the data?

The taxonomy above is simplistic, there are methods that do not properly fit in any of those two categories, as for example semi-supervised and reinforcement learning methods.

Recap

Summary of the Lecture:

Recap

Summary of the Lecture:

- This course covers important concepts and machine learning techniques

Recap

Summary of the Lecture:

- This course covers important concepts and machine learning techniques
- There are, though, relevant not covered topics (Neural Networks for instance)

Recap

Summary of the Lecture:

- This course covers important concepts and machine learning techniques
- There are, though, relevant not covered topics (Neural Networks for instance)
- We will adopt a very practical approach, with real data and applications.