# Machine Learning for Cities

Prof. Luis Gustavo Nonato

NYU / CUSP - GX 5006

January 24, 2017

# Course Structure

# Course Structure

## Content

# 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

Data Science Central

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

The Algorithms Every Data Scientist Should Know

# Course Structure

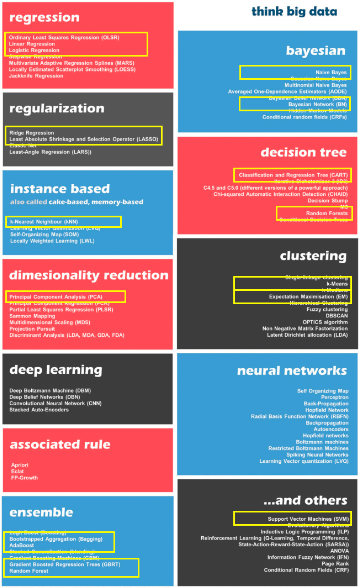
## Content

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

Data Science Central

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

The Algorithms Every Data Scientist Should Know



# 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**

# Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1. **Predictive or Supervised**
    - Regression: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$

# 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: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

# 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: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

   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: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

   Given $\mathbf{x}$ predict $y$
2. **Descriptive or Unsupervised**

# 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: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

   Given $\mathbf{x}$ predict $y$
2. **Descriptive or Unsupervised**
   - Clustering

# 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: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

   Given $\mathbf{x}$ predict $y$
2. **Descriptive or Unsupervised**
   - Clustering
   - Manifold Learning

# 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: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

   Given $\mathbf{x}$ predict $y$
2. **Descriptive or Unsupervised**
   - Clustering
   - Manifold Learning
   - Relationship between instances

# 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: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

   Given $\mathbf{x}$ predict $y$
2. **Descriptive or Unsupervised**
   - Clustering
   - Manifold Learning
   - Relationship between instances

   Is there "any structure" in the data?

## 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: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

   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 coved topics (Neural Networks for instance)

# Recap

Summary of the Lecture:

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