Introduction to Supervised Learning

Introduction

Marcelo S. Lauretto

Escola de Artes, Ciências e Humanidades,
Universidade de São Paulo
marcelolauretto@usp.br

Lima - Peru

Personal Information

- Marcelo de Souza Lauretto
  - Web page: www.each.usp.br
  - email: marcelolauretto@usp.br
  - CV: http://lattes.cnpq.br/2488734578237992
- Formation:
  - Undergraduation: Computer Science, Federal University of Mato Grosso do Sul (1992)
  - Master degree: Computer Science, University of Sao Paulo (1996)
  - Doctoral degree: Bioinformatics, University of Sao Paulo (2007)
- Areas of interest:
  - Machine Learning (Classification Trees)
  - Biostatistics
  - Quantitative Microbial Risk Assessment
  - Forecasting
  - Bayesian Tests

About this Mini Course

- Machine Learning: basic concepts and algorithms
  - Classification Trees
  - Naïve Bayes
- Classification performance evaluation
- Basics on R
- Common issues:
  - Imbalanced datasets
  - Feature selection
  - Missing data
  - Muticlass decomposition
- Exercices in R
- Public datasets
- Case studies: data analysis

Data vs. Information

- Society produces huge amounts of data
  - Sources: business, science, medicine, economics, environment, sports, ...
- Potentially valuable resource;
- Raw data is useless $\rightarrow$ information extraction needed
  - Data: recorded facts about objects
  - Information: patterns underlying the data
- Useful Patterns allow us to make nontrivial predictions on new data

Importance of Information

- Example 1: *in vitro* fertilization
  - Given: embryos described by 60 features
  - Problem: selection of embryos that will survive
  - Data: historical records of embryos and outcome
- Example 2: cow culling
  - Given: cows described by 700 features
  - Problem: selection of cows that should be culled
  - Data: historical records and farmers' decisions
- Example 3: credit scoring
  - Given: customer loan applications described by 30 features
  - Problem: rating the creditworthiness of each customer
  - Data: historical records of loan customers and respective outcome (payment/default)

Machine Learning Techniques

- How are patterns expressed?

  Two extremes:

  - *Black-box* representation: structure incomprehensible by a human being (or by people which do not know anything about the generating algorithm)
  - *White-box* representation: its construction reveals the structure of the pattern

- Our focus in this course: algorithms for acquiring structural descriptions from examples

- Structural descriptions: represent patterns explicitly

  - Can be used to predict outcome in new situation
  - Can be used to understand and explain how prediction is derived (may be even more important)

- Methods originate from artificial intelligence, statistics, and research on databases

Structural descriptions

- Example: if-then rules

If tear production rate = reduced
  then recommendation = none
Otherwise, if age = young and astigmatic = no
  then recommendation = soft

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
...	...	...	...	...

Machine Learning: Goal

- Goal of Machine learning:
  - Given the input representation, to provide a *concept description*
- Input:
  - Concept: what we expect to be learned
    - Ex: learning how to discriminate between good and bad loan customers
  - Instances: the individual, independent examples of a concept
  - Attributes: measuring aspects of an instance
- Output:
  - Concept description
    - Ex: a decision tree for deciding if a new loan applicant shall be a good or bad customer
  - Predictions for new instances not seen before

Machine Learning: Classification

- A **classifier** is a set of rules, commands or functions built with the goal of predicting the class of an object, on the basis of their observed *attributes* or *features*.
- The classifier construction (also called *induction*) may be performed via *supervised learning*, *unsupervised learning* or *semi-supervised learning*.
- In *Supervised learning*, the classifier is constructed from a set of examples which classes are already known.
- In *Unsupervised learning*, class labels are not provided. The goal is to partition the set of examples in *clusters* (or classes) with:
  - high internal homogeneity (examples in the same cluster must be similar each to other);
  - high external heterogeneity (examples in distinct clusters must be different each to other).

- In *Semisupervised learning*, the input contains both unlabeled and labeled data.

  The basic approach consists in the following steps:

  - Construct a classifier using the labeled examples;
  - Use this classifier to compute the class probabilities for the unlabeled data;
  - Construct a new classifier using the complete dataset (using the predicted classes as labels for the unlabeled data);
  - Continue until the process converges.

  In other words, this approach may be seen as an iterative clustering, where starting points and cluster labels are obtained from the labeled data.

Attributes

- In our context of machine learning, each object is represented by a set of *attributes* (also called *fields*, *variables* or *features*).
- An **attribute** is a quantity describing an instance.
- Attributes are usually grouped into the following types:
  - *Categorical* attributes: only assume a finite number of discrete values.
    They may be divided into:
    - Nominal
    - Ordinal
  - *Quantitative* (or *numerical*) attributes: are usually a subset of real numbers, where there is a measurable difference between the possible values.
    They may be divided into:
    - Continuous
    - Discrete

Categorical Attributes

- *Categorical attributes*: only assume a finite number of discrete values. They may be divided into:
  - *Nominal:* there is no ordering between the attibute values.
    Ex: color, blood type, marital status, religion, etc.
  - *Ordinal:* there is an ordering between attribute values, but their differences are not measurable.
    Ex: level of education, social class, degree of agreement with a statement, satisfaction level with a product, disease severity, etc.

Quantitative (or Numeric) Attributes

- *Quantitative* (or *numeric*) attributes: are usually a subset of real numbers, where there is a measurable difference between the possible values.

  Quantitative attribute may be:

  - *Continuous*: resultant of measurement processes, assuming therefore values in a certain interval of the set of real numbers. Ex: time, distance, temperature, glucose concentration in blood, etc.
  - *Discrete*: usually resulting of counting processes (integers). Ex: number of children, frequency of events in a fixed time interval, etc.

- In practical problems, integers are usually treated as continuous.

Notation and Basic Definitions

- We denote by $\mathcal{U}$ the *universe set*, that is, the set of observable objects in the current problem (domain) of interest.
- We consider that each element of $\mathcal{U}$ is described by a set of M attributes (or features) $a_1, \ldots, a_M$.
- The vector $\boldsymbol{x} = (x_1, x_2 \ldots x_M)$ represents the values of attributes $a_1, \ldots, a_M$, for a given element of $\mathcal{U}$.

  This vector is usually called the *attribute vector* (or *feature vector*) of the element.
- We denote by $\mathcal{X}_j$ the domain (or set of possible values) of a_j.
- The cartesian product $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_M$ is called *attribute space* (called also *feature space*) and corresponds to the set of all possible attribute vectors.

Notation and Basic Definitions

- In the context of supervised learning, we assume the existence of a partition of universe set $\mathcal{U}$ in K disjoint and non-empty sets $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_K$. Each of these subsets corresponds to one *class*. Here, we denote classes by their respective indexes $k = 1, 2, \ldots, K$.
- A *training set*, denoted by $\mathcal{L}$, is a set of N observed examples*,

$$\mathcal{L} = \{(\boldsymbol{x}_{i,\bullet}, y_i), \quad i = 1, 2, \ldots, N\} , \tag{1}$$

  where:

  - $\boldsymbol{x}_{i,\bullet} = (x_{i,1}, x_{i,2}, \ldots, x_{i,M}) \in \mathcal{X}$ and $y_i \in \{1, 2, \ldots, K\}$ denote, respectively, the attribute vector and the class of example of index i;
  - $x_{i,j}$ denotes the value of attribute a_j for example i.

- Assumption: the N observed examples are *independent*

Notation and Basic Definitions

- A hypothetical training set $\mathcal{L}$: The mail reading problem

Autor	Assunto	Tamanho	Ler em casa?
conhecido	novo	curto	sim
desconhecido	novo	longo	sim
desconhecido	antigo	curto	não
conhecido	antigo	curto	sim
conhecido	novo	longo	sim
conhecido	antigo	longo	sim
desconhecido	antigo	longo	não
desconhecido	novo	longo	sim
conhecido	antigo	curto	sim
conhecido	novo	curto	sim
desconhecido	antigo	longo	não
conhecido	novo	curto	sim
conhecido	antigo	longo	sim
conhecido	novo	longo	sim

Exemplo

atributos

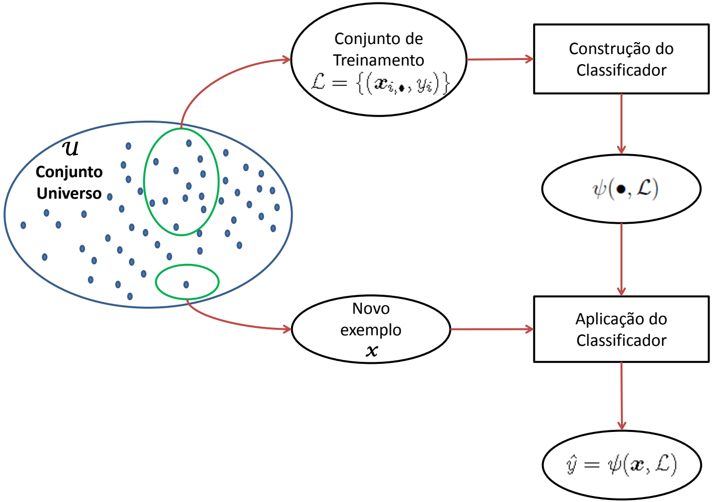
classe

Notation and Basic Definitions

- An classifier induced from the training set $\mathcal{L}$, denoted by $\psi(\bullet, \mathcal{L})$, is a function which assigns, for every attribute vector $\boldsymbol{x} \in \mathcal{X}$, a class of $\{1 \ldots K\}$:
$$\psi(\bullet, \mathcal{L}) : \mathcal{X} \rightarrow \{1 \ldots K\}. \tag{2}$$
- The application of the classifier ψ on a new object (represented by its attribute vector $\boldsymbol{x}$) provides its predicted class (denoted by $\hat{y}$):
$$\hat{y} = \psi(\boldsymbol{x}, \mathcal{L}) \tag{3}$$

Notation and Basic Definitions

- The machine learning general scheme:



Some Problem Examples

- The weather problem (fictious)
  - To discover the conditions that are suitable for playing some unspecified game.
- The contact lenses problem (fictious)
  - Problem: To recommend the type of lenses (soft/hard/none) on the basis of patient features.
- Irises: a classic numeric dataset
  - Contains 50 examples of each of three types of plant: *Iris setosa*, *Iris versicolor*, and *Iris virginica*.
  - Attributes: Sepal length, sepal width, petal length, petal width.

Some Problem Examples

- CPU performance (numeric prediction)
  - To predict the relative performance of computer processing power.
  - Attributes: Cycle time, main memory (min and max),cache, channels (min and max).
- Labor negotiations
  - Canadian contract negotiations in 1987 and 1988: collective agreements reached in the business and personal services sector for organizations with at least 500 members (teachers, nurses, university staff, police, etc).
  - Classes: acceptable (agreements were accepted by both labor and management), unacceptable (offers that were not accepted by one party or agreements that had been significantly perturbed afterwards).
  - Attributes: Duration, wage increase in first, second and third year, cost of living adjustment, work hours per week, etc.
  - Many *missing values*

The Weather Problem

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
...	...	...	...	...

If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes

The Weather Data with Mixed Attributes

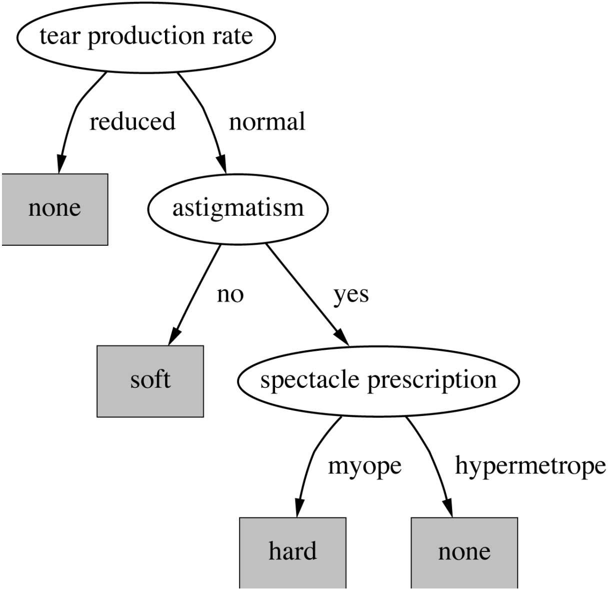
Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...	...	...	...	...

If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes

The Contact Lenses Data

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

A Decision Tree for the Contact Lenses Problem



Iris Flowers Classification

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					

If petal length < 2.45 then Iris setosa
If sepal width < 2.10 then Iris versicolor
...

Predicting CPU Performance

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

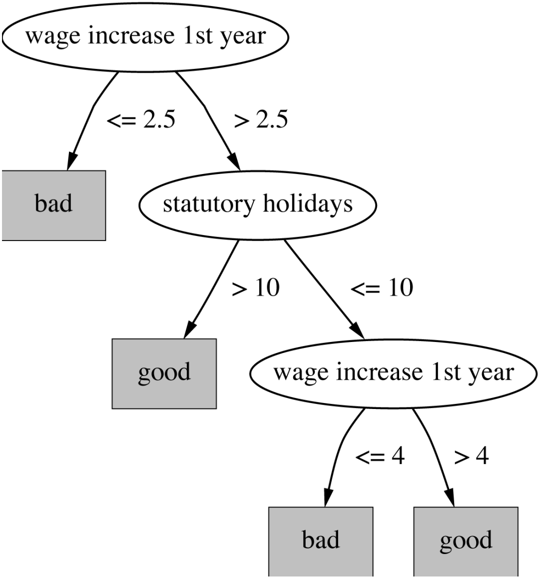
- Linear regression function:

$$PRP = -55.9 + 0.0489\ MYCT + 0.0153\ MMIN + 0.0056\ MMAX + 0.6410\ CACH - 0.2700\ CHMIN + 1.480\ CHMAX$$

Labor Negotiation Data

Attribute	Type	1	2	3	…	40
Duration	(Number of years)	1	2	3		2
Wage increase first year	Percentage	2%	4%	4.3%		4.5
Wage increase second year	Percentage	?	5%	4.4%		4.0
Wage increase third year	Percentage	?	?	?		?
Cost of living adjustment	{none,tcf,tc}	none	tcf	?		none
Working hours per week	(Number of hours)	28	35	38		40
Pension	{none,ret-allw, empl-cntr}	none	?	?		?
Standby pay	Percentage	?	13%	?		?
Shift-work supplement	Percentage	?	5%	4%		4
Education allowance	{yes,no}	yes	?	?		?
Statutory holidays	(Number of days)	11	15	12		12
Vacation	{below-avg,avg,gen}	avg	gen	gen		avg
Long-term disability assistance	{yes,no}	no	?	?		yes
Dental plan contribution	{none,half,full}	none	?	full		full
Bereavement assistance	{yes,no}	no	?	?		yes
Health plan contribution	{none,half,full}	none	?	full		half
Acceptability of contract	{good,bad}	bad	good	good		good

A Decision Tree for the Labor Data



Another Decision Tree for the Labor Data