Algorithmics for Data Mining

Master in Innovation and Research in Informatics
FIB, UPC

Department of Computer Science

Spring 2020

0. Course Presentation

Personnel

José Luis Balcázar

- jose.luis.balcazar@upc.edu
- Omega 255 (2nd floor), 93 413 7847

Additionally, we plan for Prof. Josep Carmona to cover Business Process Mining.

Logistics

Schedule in the Racó with the initial plans.

- Quite low registration this year (about half the usual).
- The two half groups seem overkill, but the timing is not compatible.
- Could we reach an agreement on a more sensible decision?

Logistics

Schedule in the Racó with the initial plans.

- Quite low registration this year (about half the usual).
- The two half groups seem overkill, but the timing is not compatible.
- Could we reach an agreement on a more sensible decision?

Additional personal conversations as needed:

- Usually available after each of our sessions;
- recommended (but not enforced) to warn me in advance by email;
- many alternative slots for appointments, again by email.

Written Support

Link to the evolving slides:
`www.cs.upc.edu/~balqui/slidesADM2020.pdf`

Link will be made available also from the Racó.

Several books available in the Main Library BRGF
(please take initiative, look for them, browse through them...)
and also freely online (like `this one`, or also `that one`...).

Mainly, individually agreed research papers for state-or-the-art advances on each topic.

Evaluation, I

Papers to turn in, one optional oral presentation

I want four intermediate grades from each student in order to decide the final grade.

- Four written assignments ("your papers"), or

Evaluation, I

Papers to turn in, one optional oral presentation

I want four intermediate grades from each student in order to decide the final grade.

- Four written assignments ("your papers"), or
- Three written assignments ("your papers") and one oral presentation of one of them.

Evaluation, I

Papers to turn in, one optional oral presentation

I want four intermediate grades from each student in order to decide the final grade.

- Four written assignments ("your papers"), or
- Three written assignments ("your papers") and one oral presentation of one of them.
  - Which option?

Evaluation, I

Papers to turn in, one optional oral presentation

I want four intermediate grades from each student in order to decide the final grade.

- Four written assignments ("your papers"), or
- Three written assignments ("your papers") and one oral presentation of one of them.
  - Which option?
  - Which topics?

Evaluation, I

Papers to turn in, one optional oral presentation

I want four intermediate grades from each student in order to decide the final grade.

- Four written assignments ("your papers"), or
- Three written assignments ("your papers") and one oral presentation of one of them.
  - Which option?
  - Which topics?
  - You negotiate all this with me!

Evaluation, I

Papers to turn in, one optional oral presentation

I want four intermediate grades from each student in order to decide the final grade.

- Four written assignments ("your papers"), or
- Three written assignments ("your papers") and one oral presentation of one of them.
  - Which option?
  - Which topics?
  - You negotiate all this with me!
- There will be no formal exam.

Evaluation, I

Papers to turn in, one optional oral presentation

I want four intermediate grades from each student in order to decide the final grade.

- Four written assignments ("your papers"), or
- Three written assignments ("your papers") and one oral presentation of one of them.
  - Which option?
  - Which topics?
  - You negotiate all this with me!
- There will be no formal exam.
- Papers to be uploaded through the Racó.

Evaluation, I

Papers to turn in, one optional oral presentation

I want four intermediate grades from each student in order to decide the final grade.

- Four written assignments ("your papers"), or
- Three written assignments ("your papers") and one oral presentation of one of them.
  - Which option?
  - Which topics?
  - You negotiate all this with me!
- There will be no formal exam.
- Papers to be uploaded through the Racó.
  - The first one, around one month from now

Evaluation, I

Papers to turn in, one optional oral presentation

I want four intermediate grades from each student in order to decide the final grade.

- Four written assignments ("your papers"), or
- Three written assignments ("your papers") and one oral presentation of one of them.
  - Which option?
  - Which topics?
  - You negotiate all this with me!
- There will be no formal exam.
- Papers to be uploaded through the Racó.
  - The first one, around one month from now (probably, march 11);

Evaluation, I

Papers to turn in, one optional oral presentation

I want four intermediate grades from each student in order to decide the final grade.

- Four written assignments ("your papers"), or
- Three written assignments ("your papers") and one oral presentation of one of them.
    - Which option?
    - Which topics?
    - You negotiate all this with me!
- There will be no formal exam.
- Papers to be uploaded through the Racó.
    - The first one, around one month from now (probably, march 11);
    - the second one, just before the Easter break;

Evaluation, I

Papers to turn in, one optional oral presentation

I want four intermediate grades from each student in order to decide the final grade.

- Four written assignments ("your papers"), or
- Three written assignments ("your papers") and one oral presentation of one of them.
  - Which option?
  - Which topics?
  - You negotiate all this with me!
- There will be no formal exam.
- Papers to be uploaded through the Racó.
  - The first one, around one month from now (probably, march 11);
  - the second one, just before the Easter break;
  - the deadlines for third and fourth will depend on whether you give a presentation.

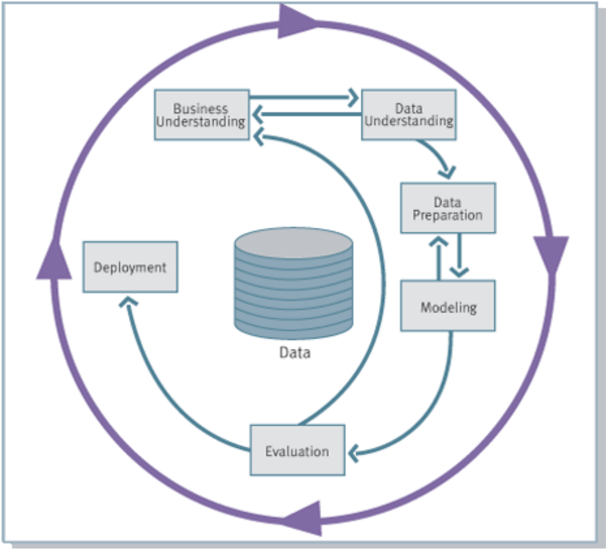
Evaluation, II

Expected characteristics

- "Your papers" must have a substantial content related to the topic of the course.
- Teamwork allowed, but:
  - Not for a paper that acts as basis of an oral presentation, and
  - your sets of coworkers on different papers must be disjoint.
- At least one of them (recommended: the first one) is to be on usage of a Data Mining tool for some Data Mining task.
- Under the previous conditions, the more your papers resemble original research papers, the better.
- Ask me if in need of clarification or if you want to propose some justified variant (I am likely to accept it).

CRISP-DM

Industry-designed diagram (1996)



Course Contents

Difficulty

Some of you may be attending, or have already listened to, courses similar to this one.

- We all must accept that there will be duplicities.
- Want most of these to still turn out to be useful!
  - By refreshing known but forgotten content,
  - By expanding the understanding,
  - By deepening the understanding.

Approximate topic guidance

- Book: The "Top Ten" Algorithms in Data Mining, `http://crcpress.com/product/isbn/9781420089646`,
- Preceding survey paper with same title, `http://link.springer.com/article/10.1007/s10115-007-0114-2`,
- plus a few variations and deeper considerations.

Taxonomy of Modeling Tools in Data Mining

Careful: not universal

- Predictive Models (always "supervised"):
  - Classification (Discrimination): non-numeric, unstructured prediction space
  - Categorization and Multiclassification: non-numeric, structured prediction space
  - Ranking: non-numeric prediction on a total ordering
  - Regression (Interpolation): numeric prediction space
    - Linear,
    - Polynomial,
    - ...
- Descriptive Models (possibly "unsupervised"):
  - Humanly interpretable predictors,
  - Clustering,
  - Pattern mining:
    - Frequent sets, frequent closures,
    - Association rule mining,
    - Pattern set mining...

Relational Data

Most common for starters

Relational data:

- Structured in tuples of attribute/value pairs.
- Akin to a SQL table.
- Often reformulated as a cloud of points in R^n.
- To predict: the value of one chosen "class" attribute.

Toy Relational Data

A simple and somewhat famous example that probably you have seen before

outlook	temperature	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

(Source today: Witten & Frank "Data Mining".)

Transactional Data, I

Alternative context, usual for pattern mining

Each observation is seen as a data structure on itself.

On the basis of a set of atomic items:

- Simplest (and most common) case: each observation is a set. (Analogy: documents as sets of terms.)
- Slight sophistication: multiplicity is relevant (but is likely to need adjustements; analogy: tfidf-like weights...).
- Further sophistications!

We will return to transactional data every now and then; but, for the time being, we work mostly with relational data.

Missing Topics

(Some of) The most important notions we are not discussing

- Time Series (very important in practice);
- Visual Analytics;
- OLAP;
- Data Streams;
- Neural Models (hint at connection at the approprate time);
- ...

Practical Data Analysis, I

Tools: Programming, GUIs, and workflows

Approaches

- Programming or CLI's: mostly "verbal", visualization basically reduced to graphics of the results of analysis;

Practical Data Analysis, I

Tools: Programming, GUIs, and workflows

Approaches

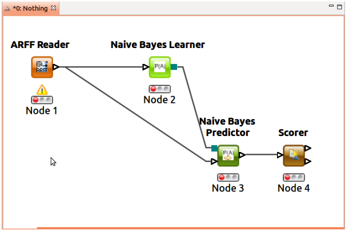
- Programming or CLI's: mostly "verbal", visualization basically reduced to graphics of the results of analysis;
- Relatively static, traditional GUIs (declining): buttons to load data and run algorithms, configuration tabs. . .

Practical Data Analysis, I

Tools: Programming, GUIs, and workflows

Approaches

- Programming or CLI's: mostly "verbal", visualization basically reduced to graphics of the results of analysis;
- Relatively static, traditional GUIs (declining): buttons to load data and run algorithms, configuration tabs. . .
- Workflows: very much visual; everything (or almost) is handled graphically: movable icons, contextual menus to configure. . . — may be successful with managers.



Who's who

`Recent poll` from `http://www.kdnuggets.com` (or navigate `http://www.kdnuggets.com` → `Polls`...)

- Tools with a different originary purpose:
  - Python, R, EXCEL, SQL...
- More or less traditional GUI:
  - Weka Explorer, FRIDA...
- Workflow-based:
  - KNIME, RapidMiner, Weka Knowledge Flows, Orange...
  - Cloud-supported `clowdflows`, not very mature yet but you are welcome to give it a try.
- Omitted from this course: Visual Analytics tools (Tableau, Spotfire, Qlik...)

Practical Data Analysis, III

About most datasets

To keep in mind:

Blindly feeding the data into your data analysis tools is unlikely to work well!

A substantial amount of reading and thinking must be spent in preprocessing and transformation.

`https://www.kdnuggets.com/2015/05/data-science-inconvenient-truth.html`

Practical Data Analysis, IV

Where to explore for datasets

Main dataset sources:

- `mldata.org`,
- `https://www.kaggle.com/competitions`,
- the classical `archive.ics.uci.edu/ml/`:
  - Car evaluation (synthetic),
  - Mushroom (semi-synthetic),
  - Adult (a.k.a. "census income"),
  - Congressional Voting Records,
  - Contraceptive Method Choice,
  - Covertype,
  - (Statlog) German Credit Scoring,
  - (Statlog) Shuttle...

Additional data sources for the politically motivated:
`http://databank.worldbank.org`

(and plenty of others out there!)

1. Intro to KNIME

Lab Session 1, I

http://KNIME.org

Get KNIME working on your machine!

- On Linux, only installation necessary is uncompressing the tarball.
- Self-installer on Windows: run it, keep going…
- Folder for your workflows: maybe on cloud?

Lab Session 1, IV

KNIME Nodes

Learn to:

- read in data;
- transform data matrices:
  - handle sorting criteria for visualizing tables,
  - identify and change the types of columns,
  - perform other data manipulation operations:
    column/row filters, group-by, join, sampling...
  - handle collection columns;
- get a glimpse of the basic statistics of your data;
- visualize and plot data;
  - create interactive tables, hilite instances, and propagate the highlighter marks,
  - create and manipulate scatter plots,
  - handle colors, sizes, and shapes,
  - create histograms, line plots, box plots...

Count on a bit of help from the instructor when necessary.

2. Brief Probability Review

Probabilistic Tools

Recap

1. Probability space, events, random variables;

Probabilistic Tools

Recap

1. Probability space, events, random variables;
2. Empirical frequency as approximate probability,

Probabilistic Tools

Recap

1. Probability space, events, random variables;
2. Empirical frequency as approximate probability,
   - support, $supp(A)$: number of observations where A holds;

Probabilistic Tools

Recap

1. Probability space, events, random variables;
2. Empirical frequency as approximate probability,
   - support, $supp(A)$: number of observations where A holds;
   - probability $Pr(A)$, normalized support in $[0, 1]$: divide by total number of observations;

Probabilistic Tools

Recap

1. Probability space, events, random variables;
2. Empirical frequency as approximate probability,
   - support, $supp(A)$: number of observations where A holds;
   - probability $Pr(A)$, normalized support in $[0, 1]$: divide by total number of observations;
   - sometimes we normalize into $[0, 100]$ and indicate %.

Probabilistic Tools

Recap

1. Probability space, events, random variables;
2. Empirical frequency as approximate probability,
   - support, $supp(A)$: number of observations where A holds;
   - probability $Pr(A)$, normalized support in $[0, 1]$: divide by total number of observations;
   - sometimes we normalize into $[0, 100]$ and indicate %.
3. Conditional probability,

$$conf(A \to B) = Pr(B \mid A) = \frac{supp(AB)}{supp(A)};$$

Probabilistic Tools

Recap

1. Probability space, events, random variables;
2. Empirical frequency as approximate probability,
   - support, $supp(A)$: number of observations where A holds;
   - probability $Pr(A)$, normalized support in $[0, 1]$: divide by total number of observations;
   - sometimes we normalize into $[0, 100]$ and indicate %.
3. Conditional probability,
   - confidence: empirical approximation to the conditional probability in "implicational" form,

$$conf(A \to B) = Pr(B \mid A) = \frac{supp(AB)}{supp(A)};$$

Probabilistic Tools

Recap

1. Probability space, events, random variables;
2. Empirical frequency as approximate probability,
   - support, $supp(A)$: number of observations where A holds;
   - probability $Pr(A)$, normalized support in $[0, 1]$: divide by total number of observations;
   - sometimes we normalize into $[0, 100]$ and indicate %.
3. Conditional probability,
   - confidence: empirical approximation to the conditional probability in "implicational" form,
   $$conf(A \to B) = Pr(B \mid A) = \frac{supp(AB)}{supp(A)};$$
4. Independence:

Probabilistic Tools

Recap

1. Probability space, events, random variables;
2. Empirical frequency as approximate probability,
   - support, $supp(A)$: number of observations where A holds;
   - probability $Pr(A)$, normalized support in $[0, 1]$: divide by total number of observations;
   - sometimes we normalize into $[0, 100]$ and indicate %.
3. Conditional probability,
   - confidence: empirical approximation to the conditional probability in "implicational" form,

   $$conf(A \to B) = Pr(B \mid A) = \frac{supp(AB)}{supp(A)};$$

4. Independence:

   $$Pr(A \wedge B) = Pr(A) * Pr(B),$$

Probabilistic Tools

Recap

1. Probability space, events, random variables;
2. Empirical frequency as approximate probability,
   - support, $supp(A)$: number of observations where A holds;
   - probability $Pr(A)$, normalized support in $[0, 1]$: divide by total number of observations;
   - sometimes we normalize into $[0, 100]$ and indicate %.
3. Conditional probability,
   - confidence: empirical approximation to the conditional probability in "implicational" form,
   $$conf(A \to B) = Pr(B \mid A) = \frac{supp(AB)}{supp(A)};$$
4. Independence:
   $Pr(A \land B) = Pr(A) * Pr(B)$,
   $Pr(A \mid B) = Pr(A)$,

Probabilistic Tools

Recap

1. Probability space, events, random variables;
2. Empirical frequency as approximate probability,
   - support, $supp(A)$: number of observations where A holds;
   - probability $Pr(A)$, normalized support in $[0,1]$: divide by total number of observations;
   - sometimes we normalize into $[0,100]$ and indicate %.
3. Conditional probability,
   - confidence: empirical approximation to the conditional probability in "implicational" form,

$$conf(A \to B) = Pr(B \mid A) = \frac{supp(AB)}{supp(A)};$$

4. Independence:

$$Pr(A \wedge B) = Pr(A) * Pr(B),$$
$$Pr(A \mid B) = Pr(A),$$
$$Pr(B \mid A) = Pr(B);$$

Probabilistic Tools

Recap

1. Probability space, events, random variables;
2. Empirical frequency as approximate probability,
   - support, $supp(A)$: number of observations where A holds;
   - probability $Pr(A)$, normalized support in $[0, 1]$: divide by total number of observations;
   - sometimes we normalize into $[0, 100]$ and indicate %.
3. Conditional probability,
   - confidence: empirical approximation to the conditional probability in "implicational" form,

$$conf(A \to B) = Pr(B \mid A) = \frac{supp(AB)}{supp(A)};$$

4. Independence:

$$Pr(A \wedge B) = Pr(A) * Pr(B),$$
$$Pr(A \mid B) = Pr(A),$$
$$Pr(B \mid A) = Pr(B);$$

5. Bayes Rule: $Pr(A|B) = Pr(B|A) * Pr(A)/Pr(B)$.

Numerical Spaces and Expectation

Main property: Linearity

If random outcomes allow for the operations of addition and of multiplication by a real number (for instance, real vectors), we can use probabilities to compute expectations, that is, weighted averages:

$$E[A] = \sum_{x}(x * Pr[A = x])$$

Numerical Spaces and Expectation

Main property: Linearity

If random outcomes allow for the operations of addition and of multiplication by a real number (for instance, real vectors), we can use probabilities to compute expectations, that is, weighted averages:

$$E[A] = \sum_x (x * Pr[A = x])$$

Properties:

- Linearity of expectation: $E[\sum_i \alpha_i * A_i] = \sum_i (\alpha_i * E[A_i])$.
- For independent events, commuting with product: $E[A * B] = E[A] * E[B]$ provided $Pr(A \wedge B) = Pr(A) * Pr(B)$.

Counterintuitive Facts About Probability, I

“Rosencrantz and Guildenstern are dead” (Link)

Some context: http://en.wikisource.org/wiki/The_Tragedy_of_Hamlet,_Prince_of_Denmark/Act_5

Counterintuitive Facts About Probability, I

"Rosencrantz and Guildenstern are dead" (Link)

Some context: http://en.wikisource.org/wiki/The_Tragedy_of_Hamlet,_Prince_of_Denmark/Act_5

Starting scene of the movie: ''Heads''.

Counterintuitive Facts About Probability, I

"Rosencrantz and Guildenstern are dead" (Link)

Some context: `http://en.wikisource.org/wiki/The_Tragedy_of_Hamlet,_Prince_of_Denmark/Act_5`

```
Starting scene of the movie: ''Heads''.
```

Recap:

- 79 times (92 in the theater play), a fair coin has been tossed along the way.
- All of them came up heads.
- Surely the probability of the next cointoss is higher for tails!

Counterintuitive Facts About Probability, I

"Rosencrantz and Guildenstern are dead" (Link)

Some context: http://en.wikisource.org/wiki/The_Tragedy_of_Hamlet,_Prince_of_Denmark/Act_5

Starting scene of the movie: ''Heads''.

Recap:

- 79 times (92 in the theater play), a fair coin has been tossed along the way.
- All of them came up heads.
- Surely the probability of the next cointoss is higher for tails!

  Actually, no.

  They are independent events!
- Related:
  - http://en.wikipedia.org/wiki/Ludic_fallacy.
  - "Bayesian" point of view: infer that the coins are not fair.

Counterintuitive Facts About Probability, II

Three doors in TV

Monty Hall paradox:

There are three doors. All participants know the rules:

- Behind one door there is a prize ("the car"). Behind the others, less desirable items ("big pumpkins", "goats").
- You choose one door.
- Monty Hall opens one door, different from the one you have chosen: the prize is not there.
- Then he asks you: do you want to switch?

Counterintuitive Facts About Probability, II

Three doors in TV

Monty Hall paradox:

There are three doors. All participants know the rules:

- Behind one door there is a prize ("the car"). Behind the others, less desirable items ("big pumpkins", "goats").
- You choose one door.
- Monty Hall opens one door, different from the one you have chosen: the prize is not there.
- Then he asks you: do you want to switch?

Is it better to switch? Is it better to stick?

Counterintuitive Facts About Probability, II

Three doors in TV

Monty Hall paradox:

There are three doors. All participants know the rules:

- Behind one door there is a prize ("the car"). Behind the others, less desirable items ("big pumpkins", "goats").
- You choose one door.
- Monty Hall opens one door, different from the one you have chosen: the prize is not there.
- Then he asks you: do you want to switch?

Is it better to switch? Is it better to stick?

The first correct answer right away is actually another question:

Counterintuitive Facts About Probability, II

Three doors in TV

Monty Hall paradox:

There are three doors. All participants know the rules:

- Behind one door there is a prize ("the car"). Behind the others, less desirable items ("big pumpkins", "goats").
- You choose one door.
- Monty Hall opens one door, different from the one you have chosen: the prize is not there.
- Then he asks you: do you want to switch?

Is it better to switch? Is it better to stick?

The first correct answer right away is actually another question: What do we mean by "better"?

But, for a sensible notion of "better", it is better to switch.

Counterintuitive Facts About Probability (III)

Expectation of linearity

Simpson's Paradox:

Somebody has performed a survey.

- All along North Alderonia, vegetarians are more common among blue-eyed people than among non-blue-eyed people.

Counterintuitive Facts About Probability (III)

Expectation of linearity

Simpson's Paradox:

Somebody has performed a survey.

- All along North Alderonia, vegetarians are more common among blue-eyed people than among non-blue-eyed people.
- Also all along South Alderonia, vegetarians are more common among blue-eyed people than among non-blue-eyed people, as well.

Counterintuitive Facts About Probability (III)

Expectation of linearity

Simpson's Paradox:

Somebody has performed a survey.

- All along North Alderonia, vegetarians are more common among blue-eyed people than among non-blue-eyed people.
- Also all along South Alderonia, vegetarians are more common among blue-eyed people than among non-blue-eyed people, as well.

We can infer that, all along both Alderonias, vegetarianism occurs more often among blue-eyed people than among the rest.

Counterintuitive Facts About Probability (III)

Expectation of linearity

Simpson's Paradox:

Somebody has performed a survey.

- All along North Alderonia, vegetarians are more common among blue-eyed people than among non-blue-eyed people.
- Also all along South Alderonia, vegetarians are more common among blue-eyed people than among non-blue-eyed people, as well.

We can infer that, all along both Alderonias, vegetarianism occurs more often among blue-eyed people than among the rest.

No! We cannot make that inference. It is possible that the comparison of the ratios gets reversed upon considering the whole population.

Counterintuitive Facts About Probability (IV)

Don't place too much confidence on confidence

Dataset CMC (Contraceptive Method Choice)

A "partial implication" of over 10% support and 90% confidence:

`near-low-wife-education` `no-contraception-method`

$\longrightarrow$

`good-media-exposure`

Seems like a reliable "partial implication".

Counterintuitive Facts About Probability (IV)

Don't place too much confidence on confidence

Dataset CMC (Contraceptive Method Choice)

A "partial implication" of over 10% support and 90% confidence:

`near-low-wife-education` `no-contraception-method`

$\longrightarrow$

`good-media-exposure`

Seems like a reliable "partial implication".

But the support of "`good-media-exposure`" is over 92%.

The "correlation" is actually negative!

3. Predictors and their Evaluation

Probabilistic Prediction

Probability-based predictive models

Probabilistic prediction

In a merely frequentist sense: counting;

- when is the prediction to be issued?
    - before seeing anything?

Probabilistic prediction

In a merely frequentist sense: counting;

- when is the prediction to be issued?
  - before seeing anything?
    "a priori" predictor: the most common value for the class (*ZeroR* predictor);

Probabilistic prediction

In a merely frequentist sense: counting;

- when is the prediction to be issued?
  - before seeing anything?
    "a priori" predictor: the most common value for the class
    (*ZeroR* predictor);
  - after seeing all values for all non-class attributes?
    "a posteriori" predictor: the most common value for the class, conditioned to the values seen
    (*MAP* predictor, for "maximum a posteriori").

$$\arg\max_C\{Pr(C|A_1 \ldots A_n)\}$$

MAP Prediction

Unfortunately infeasible

A small case:

Task of binary classification:

- Assume ten attributes with four values each;
- Then we need to store 2^{20} conditional probabilities;
- and we need to estimate 2^{20} conditional probabilities.

Rule of thumb:

Ten or more observations per parameter to estimate might be still far from sufficient, but are necessary anyway; with less, don't even dream.

Conditional Independence Assumption

One way out

Bayes rule

Applied to $\arg\max_C \{Pr(C|A_1 \ldots A_n)\}$:

$$Pr(C|A_1 \ldots A_n) =$$
$$Pr(A_1 \ldots A_n|C) * Pr(C)/Pr(A_1 \ldots A_n)$$

We can forget about the divisor, as it is the same for all values of C and does not modify the max.

Now we assume independence conditioned to the class value:

$$Pr(A_1 \ldots A_n|C) * Pr(C) =$$
$$Pr(A_1|C) * \ldots * Pr(A_n|C) * Pr(C)$$

Naïve Bayes

Rather good for such a simple approach

Precompute $Pr(A_i|C)$ for each value of each attribute conditioned to the class value; do it through the empirical frequency.

Instead of predicting

$$\arg\max_C \{Pr(C|A_1 \ldots A_n)\},$$

we predict

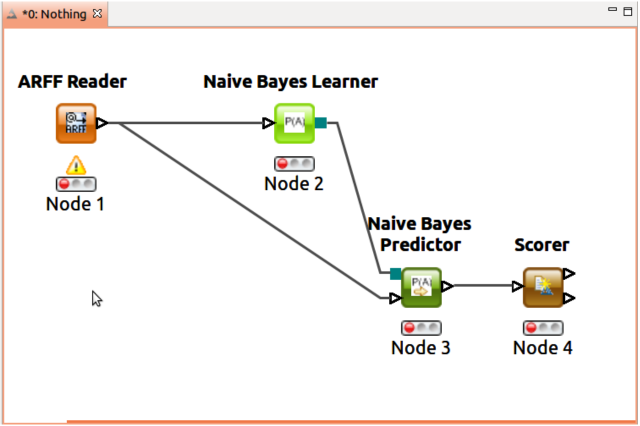
$$\arg\max_C \{Pr(A_1|C) * \ldots * Pr(A_n|C) * Pr(C)\}$$

Variant: the "Laplace correction" makes up for cases that might be potentially missing; some tools (like Weka) apply it (without warning).

How to Test a Predictor, I

On the original data?

Resubstitution error



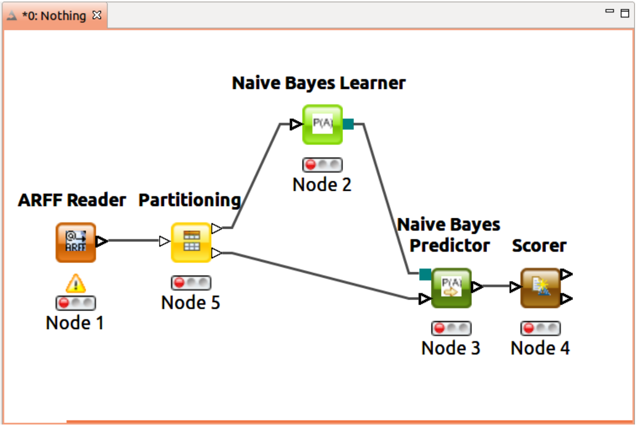
Far too optimistic!

How to Test a Predictor, II

On holdout data?

Test error

after training on a different subset.



How to Test a Predictor, III

Advantages and disadvantages

Resubstitution error

- Employs data to the maximum.

How to Test a Predictor, III

Advantages and disadvantages

Resubstitution error

- Employs data to the maximum.
- However, it cannot detect overfitting:
  - A predictor overfits when it adjusts very closely to peculiarities of the specific instances used for training.
  - Overfitting may hinder predictions on unseen instances.

How to Test a Predictor, III

Advantages and disadvantages

Resubstitution error

- Employs data to the maximum.
- However, it cannot detect overfitting:
  - A predictor overfits when it adjusts very closely to peculiarities of the specific instances used for training.
  - Overfitting may hinder predictions on unseen instances.

Holdout data

- Requires us to balance scarce instances into two tasks: training and test.
- Usual: train with $2/3$ of the instances — but, which ones?
- It does not sound fully right that some available data instances are never seen for training.

How to Test a Predictor, III

Advantages and disadvantages

Resubstitution error

- Employs data to the maximum.
- However, it cannot detect overfitting:
  - A predictor overfits when it adjusts very closely to peculiarities of the specific instances used for training.
  - Overfitting may hinder predictions on unseen instances.

Holdout data

- Requires us to balance scarce instances into two tasks: training and test.
- Usual: train with $2/3$ of the instances — but, which ones?
- It does not sound fully right that some available data instances are never seen for training.
- It sounds even worse that some are never used for testing.