

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-of-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.

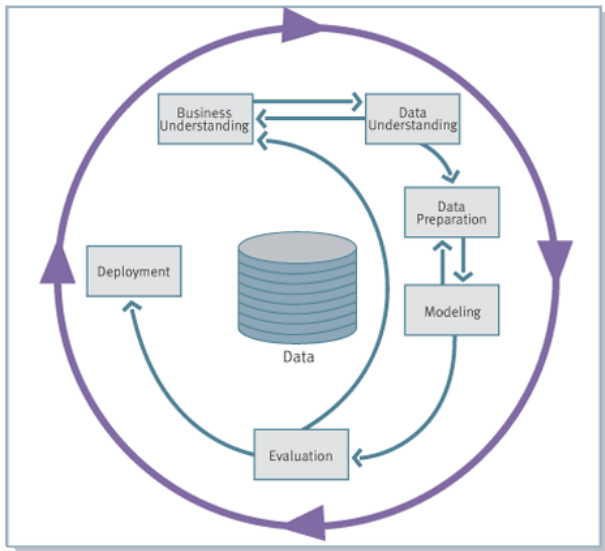
Evaluation, II

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 adjustments; **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 appropriate 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

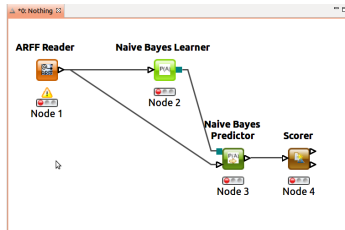
- ▶ **Programming** or CLI's: mostly “verbal”, visualization basically reduced to graphics of the results of analysis;
- ▶ Relatively static, traditional **GUIs** (declining): buttons to load data and run algorithms, configuration tabs. . .

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.



Practical Data Analysis, II

Tools: Specific proposals

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 cflowflows, 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](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/](http://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,
 - ▶ Coverttype,
 - ▶ (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**, $\text{supp}(A)$: number of observations where A holds;

Probabilistic Tools

Recap

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

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

Probabilistic Tools

Recap

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

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

Probabilistic Tools

Recap

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

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

4. Independence:

Probabilistic Tools

Recap

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

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

4. Independence:

$$\text{Pr}(A \wedge B) = \text{Pr}(A) * \text{Pr}(B),$$

Probabilistic Tools

Recap

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

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

4. Independence:

$$\begin{aligned}\text{Pr}(A \wedge B) &= \text{Pr}(A) * \text{Pr}(B), \\ \text{Pr}(A \mid B) &= \text{Pr}(A),\end{aligned}$$

Probabilistic Tools

Recap

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

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

4. Independence:

$$\text{Pr}(A \wedge B) = \text{Pr}(A) * \text{Pr}(B),$$

$$\text{Pr}(A \mid B) = \text{Pr}(A),$$

$$\text{Pr}(B \mid A) = \text{Pr}(B);$$

Probabilistic Tools

Recap

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

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

4. Independence:

$$\text{Pr}(A \wedge B) = \text{Pr}(A) * \text{Pr}(B),$$

$$\text{Pr}(A \mid B) = \text{Pr}(A),$$

$$\text{Pr}(B \mid A) = \text{Pr}(B);$$

5. Bayes Rule: $\text{Pr}(A|B) = \text{Pr}(B|A) * \text{Pr}(A)/\text{Pr}(B)$.

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

→

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

→

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

Probability-based predictive models

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

Probability-based predictive models

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”).

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

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

$$\begin{aligned} Pr(C|A_1 \dots A_n) &= \\ Pr(A_1 \dots A_n|C) * Pr(C) / Pr(A_1 \dots A_n) \end{aligned}$$

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:**

$$\begin{aligned} Pr(A_1 \dots A_n|C) * Pr(C) &= \\ Pr(A_1|C) * \dots * Pr(A_n|C) * Pr(C) \end{aligned}$$

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

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

we predict

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

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

Autonomous Learning Topics, I

Proposals to explore on yourself

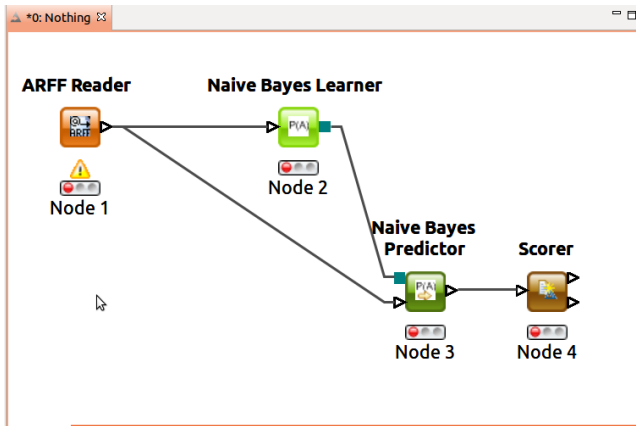
Some of these, if worked out in sufficient depth, may provide us with one of your four papers for the evaluation of the course.

1. Learn all details of the usage of Naïve Bayes predictors in various systems like R, KNIME, scikit-learn... (including the notion and usage of the Laplace correction).
2. Write your own implementation of MAP and Naïve Bayes in your favorite programming language (or, even better, in a programming language you don't master yet but want to practice further with).

How to Test a Predictor, I

On the original data?

Resubstitution error



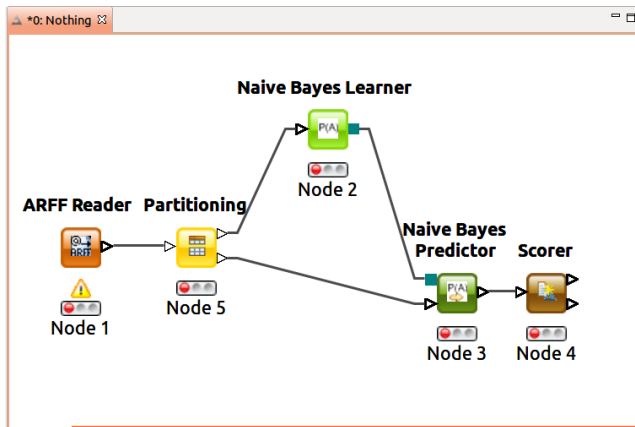
Far too **optimistic!**

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.

How to Test a Predictor, IV

The idea of x-val

Cross-validation

The key intuitions to get the maximum information from scoring the predictor:

- ▶ To run a scorer on a maximum of data instances, we wish **exactly one prediction** per training instance.
- ▶ Let's make sure that each instance is used exactly once for testing.
- ▶ Let's run **several prediction rounds**: each instance will be used for testing in exactly **one** round.
- ▶ The instances used for testing in one round are used for training in **all** the other rounds.

Cross-validation, I

Basic description

Partition the available data

into N disjoint subsets called **folds**. (Often, $N = 10$.)

- ▶ Each instance goes exactly into **one fold**.
- ▶ Run the learner and predictor N times.
- ▶ For each fold i ($1 \leq i \leq N$), the learned is **trained on the union** of all folds **except** fold i , and is then used to obtain predictions on **all** the instances of fold i as **test**.

Cross-validation, II

Options

Further precisions:

- ▶ How many folds?
 - ▶ **Leave-One-Out** X-validation:
 - ▶ one fold per instance;
 - ▶ most often unacceptably slow;
 - ▶ exhibits often too large variance to be reliable.
 - ▶ Very standard approach: **10 folds**.
- ▶ Construct folds as instances come in? **Randomize** instead?
 - ▶ Keep present **reproducibility**!
- ▶ Potential problems if some values of the class attribute are **infrequent**. How to solve this?

Cross-validation, III

Stratification

Stratified X-validation

means that the folds are constructed in such a way that **all the values** of the class attribute are as **evenly split** as possible.

- ▶ Ensures even presence of **all** labels in **all** folds.
- ▶ Turns out to reduce the variance of the computed approximate accuracy.

Predictor Evaluation, I

Simplest case first: binary accuracy

Confusion matrix

(also known as **Contingency matrix**):

- ▶ True positives (positive prediction, hit)
- ▶ False positives (positive prediction, fail: false alarm)
- ▶ True negatives (negative prediction, hit)
- ▶ False negatives (negative prediction, fail)

Accuracy, hit ratio:

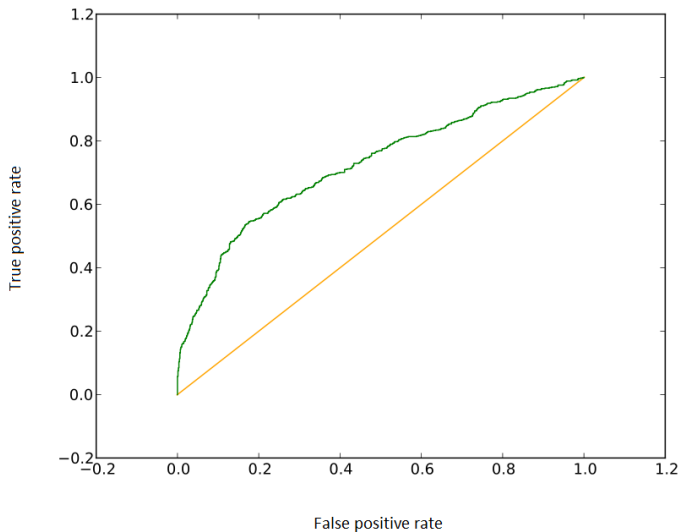
Number of hits divided by total number of predictions.

Warnings:

- ▶ Note that our reference to the true label is only indirect.
- ▶ Simple generalization to an $n \times n$ confusion matrix if the problem at hand consists of n class values.
- ▶ Some problems may suggest (or require) to weight differently the false positives and the false negatives.

ROC space and ROC curves

The curve is formalized subsequently



ROC Space

Predictors lead to points in ROC space

Consider the unit square:

Top left will mean performing quite well.

- ▶ The x coordinate is the **false positive rate**: the ratio of false positives to negative labels.
- ▶ The y coordinate is the **true positive rate**: ratio of true positives to positive labels.

The various regions of ROC space

Each has an intuitive meaning:

- ▶ Half-square below the main diagonal,
- ▶ around the center,
- ▶ near the corners...

The ROC Curve, I

Some predictors provide further information

Ranked predictions:

Predictors that may “bet” on pairs of observations, effectively sorting them.

- ▶ For instance, MAP and Naïve Bayes have several options:
 - ▶ Higher probability for the “positive” class value;
 - ▶ Larger difference of probabilities with respect to other class values. . .
- ▶ Regression-based predictors inherit the real line ordering;
- ▶ *Information Retrieval* algorithms are often able to order observations according to the expected relevance.

The ROC Curve, II

For predictors that are able to rank their observations

Tweak the predictor (usually by **thresholding** or by **sorting** all the n observations), so as to classify as negative exactly k points.

ROC curve:

(Receiver/Relative Operating Characteristics).

for each k from 0 to n ,

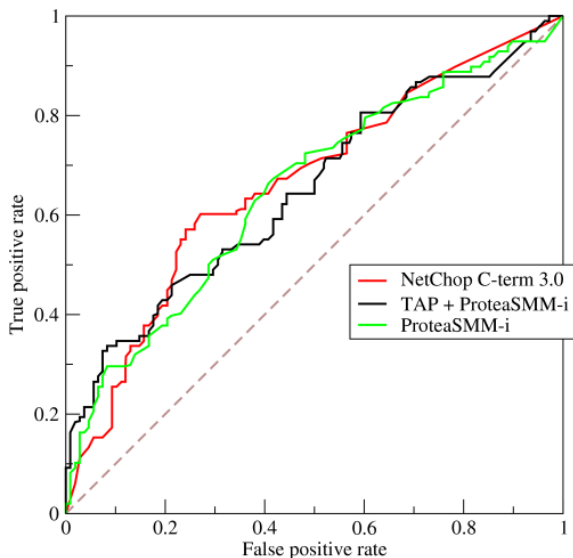
plot the ROC space point corresponding to predicting negatively to k cases (e.g. the k lowest-ranked observations).

We get a curve from $(0,0)$, where we reject everything and there are no false positives, all the way to $(1,1)$ where we accept everything and there are no false negatives.

The ROC Curve, III

Source: Wikipedia, 2009

Further Examples of ROC curves:



The Area Under the ROC Curve, AUC

Fashionable but dangerous

Motivation:

ROC Curves often do not lead to a clear winner among several choices of a classifier.

- ▶ AUC reduces each classifier's performance on a dataset to a single number.
- ▶ Thus allowing us to compare classifiers.
- ▶ **However**, it corresponds to weighting differently the false positive errors than the false negative errors,

The Area Under the ROC Curve, AUC

Fashionable but dangerous

Motivation:

ROC Curves often do not lead to a clear winner among several choices of a classifier.

- ▶ AUC reduces each classifier's performance on a dataset to a single number.
- ▶ Thus allowing us to compare classifiers.
- ▶ **However**, it corresponds to weighting differently the false positive errors than the false negative errors,
- ▶ and the weights **depend on the classifier**.

The Area Under the ROC Curve, AUC

Fashionable but dangerous

Motivation:

ROC Curves often do not lead to a clear winner among several choices of a classifier.

- ▶ AUC reduces each classifier's performance on a dataset to a single number.
- ▶ Thus allowing us to compare classifiers.
- ▶ **However**, it corresponds to weighting differently the false positive errors than the false negative errors,
- ▶ and the weights **depend on the classifier**.
- ▶ Thus, we should **avoid** that usage.
- ▶ See Hand (Machine Learning Journal, 2009) for further explanations and alternatives.

Predictor Evaluation, II

Sometimes accuracy is insufficient

Alternative quantities:

- ▶ Confidence of “positive label” \implies “positive prediction”:
Sensitivity (**recall** in *IR*): ratio of true positives to all positively labeled cases;
- ▶ Confidence of “positive prediction” \implies “positive label”:
Precision: ratio of true positives to all positively predicted cases;
- ▶ Confidence of “negative label” \implies “negative prediction”:
Specificity: ratio of true negatives to all negatively labeled cases.

For you to think about:

- ▶ How do these notions connect with the axes in ROC space?
- ▶ Express accuracy as a linear combination of sensitivity and specificity, and interpret the weights.

Predictor Evaluation, III

Train set, test set, and validation set

Even if we evaluate just accuracy, we may have two frequent reasons for the evaluation.

- ▶ We may just want to know (an approximate assessment of) the accuracy of our predictor.
 - ▶ A hold-out train/test approach will do,
 - ▶ but cross-validation will do better.
- ▶ Or, we may want to have (approximate assessments of) the accuracies of several predictors, in order to choose the one with best accuracy.
- ▶ **But:** what if we want **both**?
 - ▶ Then the cross-validation accuracy is unreliable!
 - ▶ We chose the best among several possibilities, hence it is “biased towards optimism”.
 - ▶ We should combine both: hold out a subset for final validation, and run cross-validation on the rest.

Autonomous Learning Topics, II

Proposals to explore on yourself

4. Find out why people tend to use 10-fold X-validation schemes (link to a research paper of 1995; also the presentation slides are available and quite interesting).
5. Write your own implementation of a ROC curve visualizer.
6. There are several possibilities to handle ties in the construction of a ROC curve. Find bibliography on ROC curves and investigate whether these references mention this issue.
7. Ask the instructor for the link to the paper by David Hand which criticizes the AUC measure and offers an alternative, and try to construct your own summary and intuitive explanation of the research findings reported there.

Prediction on Transactional Data

Still on simple predictors

Our discussion of predictors has assumed **relational data**.

(Often, real-valued vectors.)

What if we are to predict on **transactional data**?

A very common application: **classification on texts**.

Like:

- ▶ Spam detection,
- ▶ sentiment analysis
(movie reviews, tweets...),
- ▶ news classification...

From Texts to Transactional Data

A wide set of issues

Preprocessing:

Very important! However, not covered in this course.

- ▶ Stemming: mapping each word to its stem.
- ▶ Feature selection: only keep **things that matter**
(remove **stop words**, maybe punctuation...).
- ▶ ...

Terms as attributes:

With values *true* or *false*: **binary representation**.

Terms as items:

Texts as **transactions**: but,

- ▶ do **repetitions** matter?
- ▶ does **order** matter? (**Sequence** mining!, not covered either.)

Bernoulli Naïve Bayes

The direct transformation for binary representation

Terms as **attributes** with boolean values:

- ▶ N = total number of terms,
- ▶ x_i is **boolean** (presence or absence of term),
- ▶ $P(x_i|y)$ is the probability of finding term i in a training observation labeled y , **if** $x_i = \text{True}$,
- ▶ $P(x_i|y)$ is the probability of **not** finding term i in a training observation labeled y , **if** $x_i = \text{False}$.

Bernoulli Naïve Bayes

The direct transformation for binary representation

Terms as **attributes** with boolean values:

- ▶ N = total number of terms,
- ▶ x_i is **boolean** (presence or absence of term),
- ▶ $P(x_i|y)$ is the probability of finding term i in a training observation labeled y , **if** $x_i = \text{True}$,
- ▶ $P(x_i|y)$ is the probability of **not** finding term i in a training observation labeled y , **if** $x_i = \text{False}$.

That is:

if the ratio of observations labeled y that have term i , among the total of observations labeled y , is p , then $P(x_i|y)$ is either

- ▶ p , for $x_i = \text{True}$, or
- ▶ $1 - p$, for $x_i = \text{False}$.

Multinomial Naïve Bayes

Appropriate for bag-of-words representation

Count only occurrences:

- ▶ Ignore **absence** of the items/terms,
- ▶ ignore also their order, but
- ▶ take into account **repetitions**:
 - ▶ N = size of the observation (transaction) on which we predict,
 - ▶ x_i are the actual items in that observation,
 - ▶ some of them may be the same,
 - ▶ $P(x_i|y)$ is the **proportion** of item x_i among **all** items in **all** transactions of class y .

$$P(x_i|y) = \frac{\text{frequency of item } x_i \text{ in transactions of class } y}{\text{total of items in transactions of class } y}$$

Additional Considerations, I

Must be taken into account!

Upon implementation

A couple of relevant ideas:

- ▶ Multiplying together probabilities: very small numbers, high risk of **underflow**!
Work instead with their logarithms (often negative!):
addition instead of multiplication.
- ▶ **Always** apply a **Laplace correction**, unless you are pretty sure that all counts are positive
(it is like adding implicitly to the training observations new, artificial ones that make sure that no counts remain at zero).

Additional Considerations, II

Numerical attributes?

Naive Bayes Learner View - 0:4 - Naive Bayes Learner

File

Class counts for Score

Class:	bad	good
Count:	300	700

Total count: 1000

Threshold to used for zero probabilities: 1.0E-4

P(Foreign worker | class=?)

Class/Foreign worker	No	Yes
bad	4	296
good	33	667
Rate:	4%	96%

P(Personal status and sex | class=?)

Class/Personal status and sex	female (divorced/separated/married)	male (divorced/separated)	male (married/widowed)	male (single)
bad	109	20	25	146
good	201	30	67	402
Rate:	31%	5%	9%	55%

P(Purpose | class=?)

Class/Purpose	business	car (new)	car (used)	domestic appliances	education	furniture/equipment	others	radio/television	repairs	retraining
bad	34	89	17	4	22	58	5	62	8	1
good	63	145	86	8	28	123	7	218	14	8
Rate:	10%	23%	10%	1%	5%	18%	1%	28%	2%	1%

Additional Considerations, II

Numerical attributes, standard approach: **parametric** view assuming Gaussians

Naive Bayes Learner View - 0.4 - Naive Bayes Learner

File

Gaussian distribution for Credit amount per class value

	bad	good
Count:	300	700
Mean:	3938.12667	2885.45714
Std. Deviation:	3535.81898	2401.47228
Rate:	30%	70%

Gaussian distribution for Duration in months per class value

	bad	good
Count:	300	700
Mean:	24.86	19.20714
Std. Deviation:	13.28264	11.07956
Rate:	30%	70%

P(Foreign worker | class=?)

Class/Foreign worker	No	Yes
bad	4	298
good	33	667
Rate:	4%	96%

P(Personal status and sex | class=?)

Class/Personal status and sex	female (divorced/separated/married)	male (divorced/separated)	male (married/widowed)	male (single)
bad	109	20	25	148
good	201	30	67	402
Rate:	31%	5%	9%	55%

P(Purpose | class=?)

Class/Purpose	business	car (new)	car (used)	domestic appliances	education	furniture/equipment	others	radio/television	repairs	retraining
bad	34	89	17	4	22	58	5	62	8	1
good	63	145	86	8	28	123	7	218	14	8
Rate:	10%	23%	10%	1%	5%	18%	1%	28%	2%	1%

Additional Considerations, III

Alternative approaches related to Naïve Bayes

Good candidates for your papers!

- ▶ A proposal to **discretize** numerical attributes into “bins” instead of fitting a Gaussian (and how to do that **smart**);
- ▶ trying to find out explicitly dependencies among the attributes: **Bayesian networks**;
- ▶ trying to account for dependencies among the attributes “implicitly”: **Hidden Naïve Bayes**.

4. Regression, Bias, and Variance

Approximating a Real Value

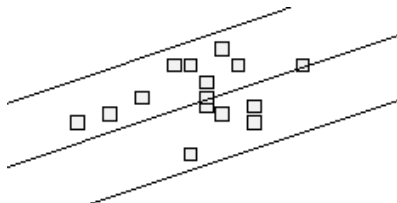
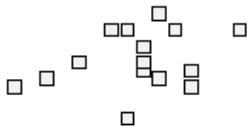
Analyzing sources of error

A rather common need:

- ▶ **Regression tasks:** predicting $f(x)$ given x and a sample of points $(x_i, f(x_i))$;
- ▶ **Estimating the expected accuracy** of a given classifier on unseen data, given how well it performs on sample data.
- ▶ Estimating other quantities related to properties of predictors.
- ▶ **Common difficulty:** the value to estimate depends on a number of explicit or implicit variables.

Regression: Prediction of Real Values

Rather: floats



Linear regression using minimum square error

The most classical and venerable predictor

A **linear predictor** is a line in 2D space
or a hyperplane in higher dimensions.

Absolute error is the difference between the value given by the hyperplane and the actual value.

By differentiating the expression that sums all the **squares** of the absolute errors and equating to zero, we can solve for the “best” hyperplane.

(Other options exist: **minimum margin** for one.)

The Intuition Of Variance

One source of prediction error

Why may the result be incorrect?

Variance:

Risk arising from the data.

- ▶ Data is seen as a sample;
- ▶ one cannot rule out the risk that the sample is a particularly bad one, just due to sheer bad luck;
- ▶ different samples may lead to different predictions — how different? Can the answer vary very much?
- ▶ This question is modeled by **variance** in the good old statistics sense: expected squared difference between the obtained values and their own mean.
- ▶ If the outcome is very concentrated around the mean outcome, there is little risk of being misled due to hitting a bad sample.

The Intuition Of Bias

Another source of prediction error

Why may the result be incorrect?

Bias:

Risk arising from your family of hypotheses.

- ▶ In a poor family of hypothesis, even the **best** one might not be very good;
- ▶ besides, the reference hypothesis is not the best one: it is the **expected** one with respect to the sample data; how good is it?

The Intuition Of Bias

Another source of prediction error

Why may the result be incorrect?

Bias:

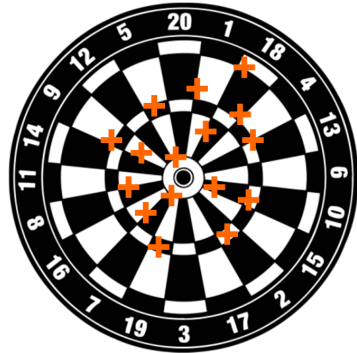
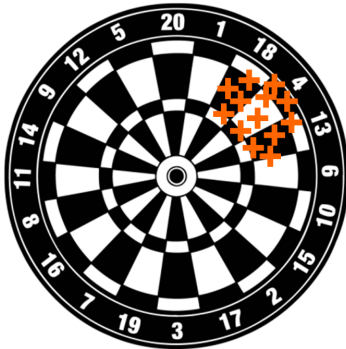
Risk arising from your family of hypotheses.

- ▶ In a poor family of hypothesis, even the **best** one might not be very good;
- ▶ besides, the reference hypothesis is not the best one: it is the **expected** one with respect to the sample data; how good is it?

But: avoiding bias error requires using rich families of hypotheses. . . which leads to high variance!

Visual Intuition

Of bias versus variance



Source: R. Gavaldà

A Very Simple Example

That may help understanding bias versus variance

A case with a single additional parameter.

We wish to estimate the average of a Gaussian
from a sample from it.

- ▶ It can be proved that the best estimation is the average of the sample:

We receive a couple dozen *floats*,
we are told they come from our Gaussian,
we estimate the true average by the empirical average. . .

A Very Simple Example

That may help understanding bias versus variance

A case with a single additional parameter.

We wish to estimate the average of a Gaussian from a sample from it.

- It can be proved that the best estimation is the average of the sample:

We receive a couple dozen *floats*,
we are told they come from our Gaussian,
we estimate the true average by the empirical average...
... using how many decimal places?

The data could be y values for very close x values if we are facing a standard regression problem with Gaussian noise.

Is it better to use very few decimal places? We use a Python (actually web-based Brython) program to try a few cases.

Bias and Variance: Formalization, I

Main ingredients at play

Context:

- ▶ A real value we want to predict, y ;
- ▶ A sample s which reveals some information about y ;
- ▶ An estimator $e(s)$ that tries to pinpoint y after seeing s .

As $e(s)$ depends on sample s , it is actually a random variable.

But note that y does not depend on s : from the point of view of the sample, y is a constant.

Bias and Variance: Formalization, II

Formalizing bias and variance

Variance:

Quadratic average error of $e(s)$ used as estimator of its own average $E[e(s)]$: $E[(e(s) - E[e(s)])^2]$.

Bias and Variance: Formalization, II

Formalizing bias and variance

Variance:

Quadratic average error of $e(s)$ used as estimator of its own average $E[e(s)]$: $E[(e(s) - E[e(s)])^2]$.

Bias:

Absolute expected error of e with respect to the true target: $|E[e(s)] - y|$. Note: $E[e(s)]$ and similar quantities, as well as the difference with y , are again independent of s .

Note the different “scale”:

we will square the bias to compensate for this.

Error Descomposition

Error is made of bias and variance

Let's **add up** variance and bias squared:

$$\begin{aligned} & E[(e(s) - E[e(s)])^2] + \\ & \quad (E[e(s)] - y)^2 = \\ & E[e(s)^2 - 2E[e(s)]e(s) + E[e(s)]^2] + \\ & \quad E[e(s)]^2 - 2yE[e(s)] + y^2 = \\ & E[e(s)^2] - 2E[e(s)]E[e(s)] + E[e(s)]^2 + \\ & \quad E[e(s)]^2 - 2yE[e(s)] + y^2 = \\ & E[e(s)^2] - 2yE[e(s)] + y^2 = \\ & E[e(s)^2] - E[2y e(s)] + E[y^2] = E[e(s)^2 - 2y e(s) + y^2] = \\ & E[(e(s) - y)^2] \end{aligned}$$

Error Descomposition

Error is made of bias and variance

Let's **add up** variance and bias squared:

$$\begin{aligned} & E[(e(s) - E[e(s)])^2] + \\ & \quad (E[e(s)] - y)^2 = \\ & E[e(s)^2 - 2E[e(s)]e(s) + E[e(s)]^2] + \\ & \quad E[e(s)]^2 - 2yE[e(s)] + y^2 = \\ & E[e(s)^2] - 2E[e(s)]E[e(s)] + E[e(s)]^2 + \\ & \quad E[e(s)]^2 - 2yE[e(s)] + y^2 = \\ & E[e(s)^2] - 2yE[e(s)] + y^2 = \\ & E[e(s)^2] - E[2y e(s)] + E[y^2] = E[e(s)^2 - 2y e(s) + y^2] = \\ & E[(e(s) - y)^2] \end{aligned}$$

They add up to the average quadratic error!

Consequences

Why prediction is difficult

Mean square error, bias, and variance:

- ▶ Rigid estimators, that is, with relatively limited possibilities for the result, risk converging to a result rather far away from the truth:

high error caused by **high bias**.

- ▶ Flexible estimators, with very many possibilities for the result, are likely to converge to the true value (low bias error) but even small sample perturbations will change the outcome:

high error caused by **high variance**.

- ▶ How to strike the best balance in terms of **rigidity** or **flexibility** of the estimators?
- ▶ Variance can be reduced if we have **large datasets**, but in many practical cases available datasets are very **far** from large enough.

Lab Session 2, I

First couple of predictors

MAP and Naïve Bayes

- ▶ Brief recap:
 - ▶ *MAP* predictor:
$$\arg \max_C \{Pr(C|A_1 \dots A_n)\}$$
 - ▶ *Naïve Bayes* predictor:
$$\arg \max_C \{Pr(A_1|C) * \dots * Pr(A_n|C) * Pr(C)\}$$
- ▶ **Today:** Watch them running!
 - ▶ Check confusion matrices,
 - ▶ view the internal parameters,
 - ▶ check ROC curves,
 - ▶ run comparisons...

Lab Session 2, I

First couple of predictors

MAP and Naïve Bayes

- ▶ Brief recap:
 - ▶ *MAP* predictor:
$$\arg \max_C \{Pr(C|A_1 \dots A_n)\}$$
 - ▶ *Naïve Bayes* predictor:
$$\arg \max_C \{Pr(A_1|C) * \dots * Pr(A_n|C) * Pr(C)\}$$
- ▶ **Today:** Watch them running!
 - ▶ Check confusion matrices,
 - ▶ view the internal parameters,
 - ▶ check ROC curves,
 - ▶ run comparisons. . .
- ▶ Run them on additional datasets.
- ▶ See the MAP predictor fail.

Lab Session 2, II

Using Predictors

File: <http://www.cs.upc.edu/~balqui/LabADM20190220.zip>

Parts of the Python code depend on `matplotlib` and are not runnable if you don't have it.

Explore first `LabADM20190220.py` which loads in the dataset `ds/weatherNominalTr.txt` and calls a predictor.

- ▶ Understand the code,
- ▶ uncomment `pr.show()`,
- ▶ uncomment the `print...` line inside the for loop,
- ▶ swap predictors to use `NaiveBayes` instead;
- ▶ now explore `compare.py` and understand the comparison.
- ▶ Explore on your own the code for both predictors and see how it fits what we discussed so far.

Lab Session 2, III

Comparing Predictors

Move on to other datasets.

- ▶ Change the main program so as to load in the `titanicTr` dataset instead;
- ▶ Redo the whole thing with both predictors: which one seems better?
- ▶ Add instructions to ask the predictors to predict on new tuples:

```
print pr.predict(('Class:1st','Sex:Female','Age:Child'))  
print pr.predict(('Class:Crew','Sex:Female','Age:Child'))
```

Explain what happens!

- ▶ Try predicting other attributes.
- ▶ Explore the other datasets:
 - ▶ which predictor seems better for each?
 - ▶ Try to make the MAP predictor fail again like before.

Lab Session 2, IV

Evaluating Predictors

Train/test split

- ▶ In the Python LabADM... file, replace the declaration `Data(filename)` with `Data(filename,75)` (or a different figure in $[0,100]$).
- ▶ Understand what happens; then move to `compare.py` and work likewise on it.
- ▶ Explore this train/test decomposition for other datasets on yourself, using `naivebayes.py`, `maxapost.py`, `compare.py`...

ROC Curves

If you have `matplotlib`, proceed to exploring ROC curves through the source `jlb_roc.py` using both predictors and varying the dataset.

Make sure to understand what happens!

Lab Session 2, V

Back to KNIME

On KNIME

redo parts of what you have done today (see Slides 35, 36):

- ▶ Manage to read in some dataset.
- ▶ Find the nodes for Naïve Bayes Learning and for Naïve Bayes Prediction (that uses the model learned).
- ▶ Find the Scorer node, that implements confusion matrices and accuracy evaluation, and compute resubstitution error.
- ▶ Add a Partitioning node to split the data into training set and test set, and compute the test set error.
- ▶ Try several partitioning strategies to check whether the test set error is stable.
- ▶ Find the ROC Curve node and show ROC curves of your predictors on your datasets.