L'art de l'utilisation des données

Partie I: Ground truth, gold standard, baseline, et autre objets « divins »*

Sylvie Ratté, Ph.D.
École de technologie supérieure

*This Spanish and English version of this seminar was given at UTPL in March 2014 as part of a grant from the Prometeo program of the Senescyt (Ecuadorian Government)

Contentu

Definitions

- Validation: refers to the internal consistency (i.e., a logical problem)
- Verification: deals with justification of knowledge claims

Barlas and Carpenter (1990).

Philosophical backgrounds

- Logical-empiricist: validation is a strictly formal, algorithmic, reductionist, and 'confrontational' process, where the model is either true of false. The validation becomes a matter of formal accuracy rather than practical use. This approach is appropriate for closed problems that have right or wrong answers associated with them, like mathematical expressions or algorithms.

- (Functional-Holistic) Relativist: validation is a semiformal and communicative process, where validation is seen as a gradual process of building confidence in the usefulness of the new knowledge (with respect to a purpose).

Pedersen et al. (2000)

It's all about convincing

- Valid model: A valid model is assumed to be only one of many possible ways of describing a real situation. No particular representation is superior to all others in any absolute sense, although one could prove to be more effective. No model can claim absolute objectivity, for every model carries in it the modeler's world view. Models are not true or false but lie on a continuum of usefulness.

- Confidence and usefulness: Model validation is a gradual process of building confidence in the usefulness of a model; validity cannot reveal itself mechanically as a result of some formal algorithms

- Validation is a Conversation: Validation is a matter of social conversation, because establishing model usefulness is a conversational matter.

Barlas and Carpenter (1990).

Data mining position

- We define scientific knowledge within the field of [DM] as socially justifiable belief according to the Relativistic School of Epistemology.

- We do so due to the open nature of [DM], where new knowledge is associated with heuristics and non-precise representations, thus knowledge validation becomes a process of building confidence in its usefulness with respect to a purpose.

Pedersen et al. (2000)

Conclusion

- Your experiments must be convincing

  Explore model behaviour, compare to other models, use statistical tests, compare using graphical displays

- Your datasets must be convincing

  It is usually difficult, time consuming, and costly to obtain appropriate, accurate, and sufficient data, and is often the reason that attempts to valid a model fail.

Sargent (2005)

- Barlas, Y., Carpenter, S., & Carpenter, S. (1990). Philosophical roots of model validation : two paradigms. System Dynamics Review, 6(2), 148–166.
- Pedersen, K., Bailey, R., Allen, J. K., & Mistree, F. (2000). Validating Design Methods & Research: The Validation Square. In DETC 2000 ASME Design Engineering Technical Conferences (pp. 1–12). Baltimore, Maryland.
- Sargent, Robert G. 2005. Verification and validation of simulation models. In *Proceedings of the 37th conference on Winter simulation* (WSC '05). Winter Simulation Conference, 130-143.
- John Hopkins University - Data Science Specialization

2. Dataset is the name of the DM game

A. Facts

B. Results and Replicability

C. Datasets' existence

D. Conclusion

Facts

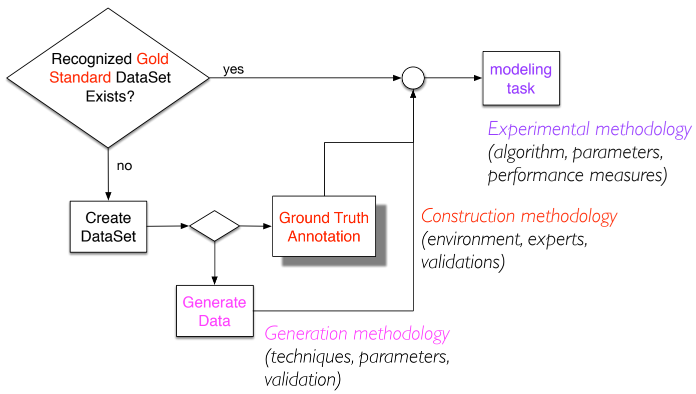
1. there is a lot of data out there
2. labeled data is rare
3. quality data is scarce;
4. training data frequently relies on human expertise

Even UCI repository contains bad datasets.

Results and Replicability

1. lack of reproducibility has been warned against repeatedly (Keogh & Kasetty 2003; Sonnenburg et al. 2007; Pedersen 2008)
2. has been highlighted as one of the most important challenges (Hirsh 2008)
3. some major conferences have started to require that all submitted research be fully reproducible (Manilescu et al. 2008)
4. a huge database for reproducible results has been put in place (Vanschoren et al. 2012)

Datasets' existence



*The meaning and definition of « Gold Standard » and « Ground Truth » vary from domains (e.g. medicine vs machine learning)

Conclusion

- Your experiments must be convincing

  In part II of this seminar, we will say more about evaluation models

- Your datasets must be convincing

  In the next section, we will go deeper into the subject of dataset construction

- Gold Standard and Ground Truth

  Any Ground Truth can become a Gold Standard. It is just a matter of how many people will accept it and use it. The more convincing you are, the better it is.

- Hirsh, H. (2008). Data mining research: Current status and future opportunities. Statistical Analysis and Data Mining, 1(2), 104–107.
- Keogh, E., & Kasetty, S. (2003). On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Mining and Knowledge Discovery, 7(4), 349–371.
- Manolescu, I., Afanasiev, L., Arion, A., Dittrich, J., Manegold, S., Polyzotis, N., Schnaitter, K., Senellart, P., & Zoupanos, S. (2008). The repeatability experiment of SIGMOD 2008. ACM SIGMOD Record, 37(1), 39–45.
- Pedersen, T. (2008). Empiricism is not a matter of faith. Computational Linguistics, 34, 465–470.
- Sonnenburg, S., Braun, M., Ong, C., Bengio, S., Bottou, L., Holmes, G., LeCun, Y., Muller, K., Pereira, F., Rasmussen, C., Ratsch, G., Scholkopf, B., Smola, A., Vincent, P., Weston, J., & Williamson, R. (2007). The need for open source software in machine learning. Journal of Machine Learning Research, 8, 2443–2466.
- Vanschoren, J., Blockeel, H., Pfahringer, B., & Holmes, G. (2012). Experiment databases. Machine Learning, 87(2), 127–158.

3. Tasks when the dataset exists

A. Common situations

B. Baseline performance

C. Experimental Methodology

D. Conclusion

Common Situations

- New models (algorithms, techniques)
- Have to compare to others

Your task

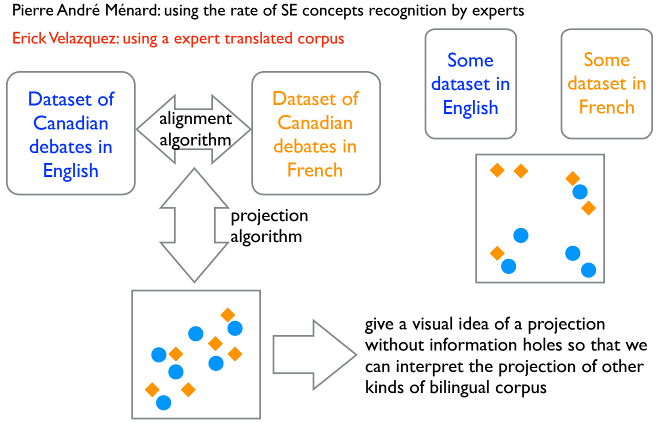
- Convince by your results that your model is better
- Use the Gold Standard or a Ground Truth for comparison
- Compare with a Baseline performance

Baseline performances

- A baseline is a previous result in controlled conditions you can compare to.
- It is usually expressed in terms of numbers and visualized in a graph form
- What to do with it:
  - try to beat it
  - try to be as close as possible because the baseline represents (almost) « perfection »
- Potential baselines (among many many more):
  - The results when a human expert is doing the task on the Gold Standard or the Ground Truth
  - The results when a basic algorithm is applied on the GS or GT (ex. using Naive Bayes as a baseline)
  - The results of your algorithm on a « near » perfect dataset as a way to explore how your algorithm will behave
  - The results of a recognized algorithm (e.g. Shih & Liu 2006)

Baseline performances examples

- PhD Students and their baselines:
  - Pierre André Ménard: using the rate of SE concepts recognition by experts
  - Erick Velazquez: using a expert translated corpus



Experimental Methodology

- Are you following a protocol defined by others?
- How are you setting the parameters for your method? (more on this in part II)
- How are you setting the parameters of the compared methods? (to be sure you are not favouring yours) - I will come back on this aspect in part II.
- Did any methods have to be modified to be applied to the Dataset? How? Convince the audience that this modification is objective.
- What are your performance measures? more on this in part II.

Conclusion

- The existence of a recognized dataset facilitates the experimental process
- The contribution is a new model, technique or algorithm
- Example of PhD Student in that situation:
  - Jose R. Pasillas Díaz (LiNCS, ÉTS)
    He is proposing two new Anomaly Detection Algorithms
    He is using 10 recognized datasets
    He compares his methods with the actual best 4

- Grosse, R.; Johnson, M.K.; Adelson, E.H.; Freeman, W.T., "Ground truth dataset and baseline evaluations for intrinsic image algorithms," Computer Vision, 2009 IEEE 12th International Conference on , vol., no., pp.2335,2342, Sept. 29 2009-Oct. 2 2009.
- Manohar, V., Soundararajan, P., & Raju, H. (2006). Performance Evaluation of Object Detection and Tracking in Video. ACCV, LNCS 2852, 151–161.
- Manohar, V., Soundararajan, P., Korzhova, V., Boonstra, M., Goldgof, D., Kasturi, R., Garofolo, J. (2007). A Baseline Algorithm for Face Detection and Tracking in Video. In Proceedings of the SPIE (pp. 1–11).
- Ménard, Pierre André, Concept exploration and discovery from business document for software engineering projects using dual mode filtering, PhD Thesis, ÉTS, 2014.
- Pasillas, J., Ratté, S. A novel approach for combining heterogeneous unsupervised anomaly detection techniques based on similarity measures, to be submitted to IEEE Transactions on Data Mining.
- Shih, Peichung, and Chengjun Liu. "Improving the face recognition grand challenge baseline performance using color configurations across color spaces." Image Processing, 2006 IEEE International Conference on. IEEE, 2006.

Common Situations

- New domains with maybe a mix of new and old techniques
- Have to compare to similar others (but they used a different dataset)

Your task

- Convince that your dataset (Ground Truth or else) is good
- Convince that your results are good

Construction Methodology (labeling new data)

- Never be your own expert!
- How do you choose experts (or other turks) to label your data?
- How is the annotation setup? (software, tools, task description)
- How to you validate the labels (the annotations)?
- What are you doing if the experts don't agree?

see Pustejovsky & Stubbs (2013)

Construction Methodology (creating data)

- Statistical Generation:
  What statistical laws are you using? On what ground?
- Generating Datasets Using known data:
  How to you insert known data inside a dataset? Sampling, repetition. Very useful in information retrieval.

Examples from PhD Students

- Pierre André Ménard (LiNCS, ÉTS)
  Concepts extraction in Software Engineering
  - Uses 7 experts to annotate SE documents for concepts
- Faten Mhiri (LiNCS, ÉTS)
  Identification of cardiac vessels in images of new borns
  - Uses medical experts to annotate images
- Paul Laurier & Frédéric Monchamps (LiNCS, ÉTS)
  Detection of potential mass murderers on the Internet
  - Uses a multiple sampling of Web pages inside which they are inserting pages known to be suspicious

Conclusion

- Dataset construction is hard
- The contribution is a new domain, problem, technique or algorithm, and a ground truth
- The realization of a good Ground Truth can be the theme of a paper (Grosse 2009, Manohar 2006, Turetsky & Ellis 2003)

- Grosse, R.; Johnson, M.K.; Adelson, E.H.; Freeman, W.T., 2009. Ground truth dataset and baseline evaluations for intrinsic image algorithms. Computer Vision, 2009 IEEE 12th International Conference on, 2335-2342.
- Manohar, V., Soundararajan, P., & Raju, H. 2006. Performance Evaluation of Object Detection and Tracking in Video. ACCV, LNCS 2852, 151–161.
- Ménard, Pierre André, Concept exploration and discovery from business document for software engineering projects using dual mode filtering, PhD Thesis, ÉTS, 2014.
- Pustejovsky, J., Stubbs, A. 2013. Natural Language Annotation for Machine Learning. O'Reilly
- Turetsky, R. J., & Ellis, D. P.W. 2003. Ground-Truth Transcriptions of Real Music from Force-Aligned MIDI Syntheses. In International Symposium on Music Information Retrieval.

5. Dataset construction and validation

Facts

a) You cannot be the expert (maybe at the beginning but not for publishing)

b) One expert is not enough

c) Experts rarely agree

  i) They don't agree on the label

  ii) They agree on the label (maybe you gave them only one!), but they don't agree on the thing to label

(ii) is very frequent in text mining and corpus linguistics in general

Examples

dirección del Departamento de Recursos Humanos

Annotators don't agree on the coverage of labels

Ménard & Ratté (submitted)

This movie was all right. The special effects were good, but the plot didn't make a lot of sense. The actors were funny, which helped, but the music was really distracting.

Annotators don't agree on the label

Pustejovsky & Stubbs (2012)

Inter-annotator agreement (IAA)

- Also called Inter-rater reliability/agreement
  - Kappa metrics (Cohen's kappa, Fleiss Kappa, etc.)
  - Pyramid evaluation (inspired by summarization evaluation)
- For publication:
  - explain the setup (conditions, tools, etc.)
  - explain how the experts were chosen
  - what is the IAA and if its new, justify it thoroughly = convince

Small example of an IAA

- Cohen's kappa
  - 2 annotators: A and B
  - 3 labels: positive, neutral, negative
  - 250 movie reviews

		B	B	B
		positive	neutral	negative
A	positive	54	28	3
A	neutral	31	18	23
A	negative	0	21	72

κ	Agreement level
< 0	poor
0.01–0.20	slight
0.21–0.40	fair
0.41–0.60	moderate
0.61–0.80	substantial
0.81–1.00	perfect

relative observed agreement

expected agreement

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

Pr(a) = (54+18+72)/250 = 0.576

Apos = (54+28+3)/250 = 0.34

Bpos = (54+31+0)/250 = 0.34

pos = Apos * Bpos = 0.1156

neu = 0.077184

neg = 0.145824

Pr(e) = pos + neu + neg = 0.338608

k = 0.36

Pustejovsky & Stubbs (2012)

Conclusion

- One expert is not enough
- Think about what to do if the experts don't agree
- Understand and explain carefully your IAA
- Look at what the people are using in your domain; if nothing exist, rely at first on Kappa Metrics

- Grosse, R.; Johnson, M.K.; Adelson, E.H.; Freeman, W.T., 2009. Ground truth dataset and baseline evaluations for intrinsic image algorithms. Computer Vision, 2009 IEEE 12th International Conference on, 2335-2342.
- Landis, J.R., and G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33(1): 159– 174.
- Manohar, V., Soundararajan, P., & Raju, H. 2006. Performance Evaluation of Object Detection and Tracking in Video. ACCV, LNCS 2852, 151–161.
- Ménard, P.A. 2014 Concept exploration and discovery from business document for software engineering projects using dual mode filtering, PhD Thesis, ÉTS.
- Ménard, P.A., Ratté, S. (submitted) Hybrid extraction method for French complex nominal multiword expressions. Lectures Notes in Software Engineering
- Ménard, P.A., Ratté, S. (submitted) Concept extraction from business documents for software engineering projects. Automated Software Engineering.
- Ménard, P.A., Ratté, S. 2011. Classifier-based Acronym Extraction for business documents, Knowledge and Information Systems.
- Nenkova, A., Passonneau,, R. 2004. Evaluating content selection in summarization: The pyramid method. Proceedings of HLT-NAACL.
- Pustejovsky, J., Stubbs, A. 2013. Natural Language Annotation for Machine Learning. O'Reilly

6. General Conclusion

A. Relativistic point of view

B. Experiments and Datasets must be convincing

C. If you are proposing a new technique for an old problem, use recognized dataset GS o GT

D. If you are contributing a new problem for which there is no dataset, try to give a Ground Truth and a baseline

E. If you are using experts, take great care to IAA evaluation

F. Finally, for the sake of science, make your entire process and dataset public

Tareas

A. Give an example of a logicist-empiricist model of a problem and an example of a relativist model of a problem

B. Daniel wants to invent a new algorithm to solve the detection of breast cancer. He is using an UCI dataset. John has realized that some students are making illegal access to the database center of the university. He wants to use a well-known anomaly detection algorithm to catch them. Explain the difference between the contexts from the point of view of dataset and experimentation.

C. John's dataset is a Gold Standard? Explain your answer.

D. What kind of baseline could be appropriate for John? What kind of baseline could be appropriate for Daniel?

E. Calculate the Cohen's Kappa coefficient of the confusion matrix presented on the next slide. Discuss the result according to the table proposed in Landis & Koch (1977)

- 2 annotators, A & B
- 3 labels, positive, neutral, negative
- 250 chat messages from students to one professor in a course about philosophy!

		B	B	B
		positive	neutral	negative
A	positive	54	8	23
A	neutral	31	38	3
A	negative	0	11	82

The Art of Dataset Usage

Part II: The quest for perfection: dataset, results and unattainable heavens*

Sylvie Ratté, Ph.D.
École de technologie supérieure

*This seminar was given at UTPL in March 2014 as part of a grant from the Prometeo program of the Senescyt (Ecuadorian Government)

Content

What is data quality?

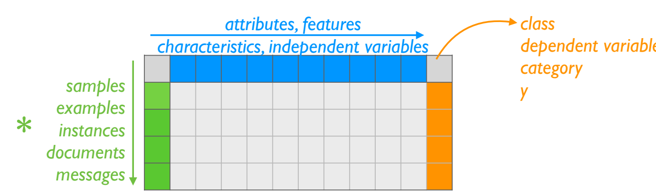
a) *Accuracy: attribute values (human errors, willingly putting incorrect values (e.g. date of birth, missing values turning like 9999))
b) *Completeness vs incompleteness
c) *Consistency vs inconsistency
d) *Believability: data trusted by users
e) Interpretability: understandability

*Problems more related to « pure » data mining (that has to deal with data integration from multiple databases.

Do I have enough data?*

a) There is no magic number.. but cover your classes
b) Rule of thumbs: examples >= 10 x attributes†
c) Depends of the number of classes
d) Depends on the model
e) Explore with what you have, add examples later

†Rarely happens



$f(x_1, x_2, \ldots, x_n) = y$

Exploring the dataset

a) Take time to calculate basic statistics

b) Take time to visualize graphs

c) Indication of heterogenous variability (difference in variability across samples)

d) Draw a learning curve to track down bias and variance (coming back on this in 4)

   i) Let's say you have two datasets, one for learning (A), the other to test (B)

   ii) Results are very good with A, horrible with B: high variance - sign of overfitting (too few examples, too many features)

   iii) Results are not good with A and also not good on B: bias (too few features, wrong features)

Exploring the attributes and their values

1. Qualitative attributes: Nominal, Binary, Ordinal
2. Quantitative attributes: Discrete, Continuous
3. What to look for before modeling:
   a) Detection of Outliers: points that appear to be isolated (might be errors or genuine values)

   b) Asymmetry in the distribution, skewness (long tails)

   c) Clusters

   d) Non-linear bivariate relationships

Conclusion

- Be sure you have enough data
- Take the time to explore your dataset
- Explore the features you are using
- Take time to draw some basic graphs

- Beleites, C. and Neugebauer, U. and Bocklitz, T. and Krafft, C. and Popp, J.: Sample size planning for classification models. Analytica Chimica Acta, 2013, 760, 25-33.
- Han, J., Kamber, M., Pei, J. 2012. Data Mining: Concepts and Techniques, Morgan Kaufmann (Elsevier).
- Maindonald, J., Braun, J. Data Analysis and Graphics Using R - An Example-Based Approach, 2nd edition, Cambridge University Press.
- Nisbet, R., Elder, J., Miner, G. 2009. Hanbook of Statistical Analysis and Data Mining Applications, Academic Press.
- Osborne, J.W. 2013. Best Practices in Data Cleaning. Sage.
- Pustejovsky, J., Stubbs, A. 2013. Natural Language Annotation for Machine Learning. O'Reilly.
- Witten, I.H., Frank, E., Hall, M.A. 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition, Morgan Kaufmann (Elsevier).

2. Some useful statistics

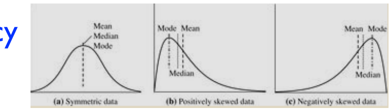
A. Where most of the values fall?

B. How are the data spread out?

C. Show me the problems

D. Conclusion

Where most of the values fall?



- Measures of central tendency
  - ✦ Mean = 696/12 = 58
  - ✦ Median (the middle value of the ordered list): 52 or 56. Take the average: 54
  - ✦ Mode (the most frequent value): 52 and 70 (bimodal)
  - ✦ Mid-range (the average between the min and the max): (30 + 110)/2 = 70.

    30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.
- Variance and standard deviation are well know and scalable to large dataset

How are the data spread out?

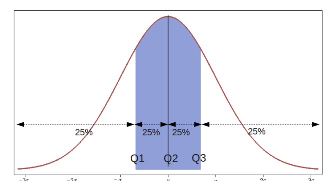
- Measures of dispersion
  - ✦ Range (the difference between the max and the min)
  - ✦ Quartiles (points taken at regular intervals to cover 25%, 50%, 75% of the data): gives an indication of the center, spread and shape.
    Q1= 47, Q2 ≃ median, Q3 = 63.
  - ✦ Interquartile range (IQR): the distance between the first and third quartile.
    IQR = 63 - 47 = 16.

    30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

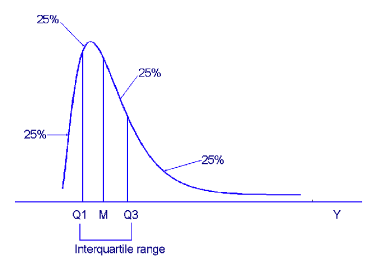
    Q_1 Q_2 Q_3

How are the data spread out?

- Quartiles and Normal distribution



- Quartiles and Skewed distribution



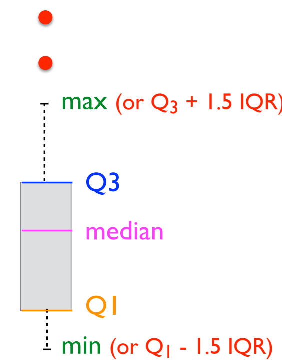
Cai, E. The Chemical Statistician. http://chemicalstatistician.wordpress.com/2013/03/31/checking-for-normality-with-quantile-ranges-and-the-standard-deviation/

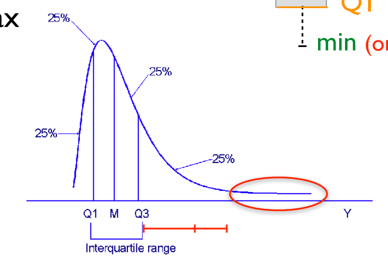
Show me the problems

- Some useful graphs for dataset exploration
  - Boxplots: outliers, changes in variability
  - Quantile plots: specific distribution, outliers, skewness
  - Scatterplots: relationship between pair of features, clusters, outliers, non linearity
  - Histograms: skewness

Boxplots

- Measures of dispersion
  - Rule of thumbs to identify outliers: check values at least 1.5 IQR above Q3 or below Q1.
  - Five-number summary: Min Q1, Median, Q3, Max



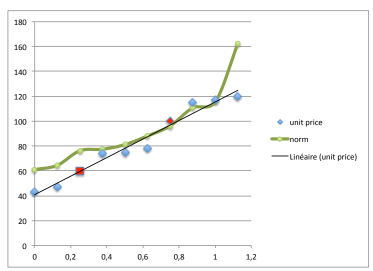


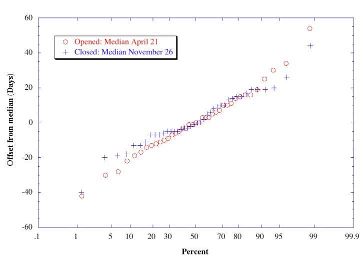
Quartiles plots

Univariate (1 feature at a time)

Associate % of data to a value in the data by increment of 1/n % (n being the size of the dataset)

In green, some points I generate with R (rnorm function) to get a normal distribution with the same mean and standard deviation as my data.





Possible usage. Superpose two distributions of the same feature to look for outliers. (here in the upper corner).

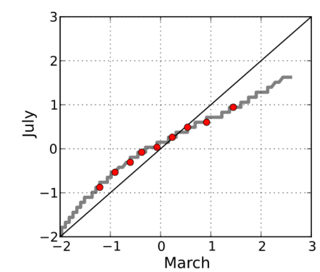
http://en.wikipedia.org/wiki/File:State_Route_20.png

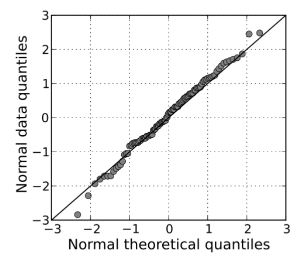
Q-Q plots

univariate, 2 distributions

use the same feature and plot each quartile against the corresponding quartile of the other distribution.

http://en.wikipedia.org/wiki/File:Ohio_temps_qq.svg





Common usage: plot the dataset against a normal distribution

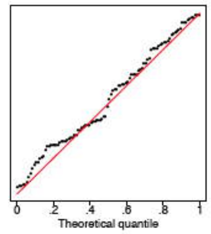
http://en.wikipedia.org/wiki/File:Normal_normal_qq.svg

q-q plots test whether a sample is from a specified distribution. If the points do not fall close to a straight line there is evidence that the sample is not from the specified distribution.

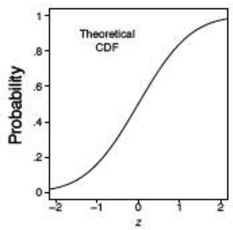
http://onlinestatbook.com/2/advanced_graphs/q-q_plots.html

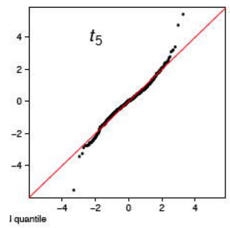
Just remember

this is the « look » of a uniform distribution



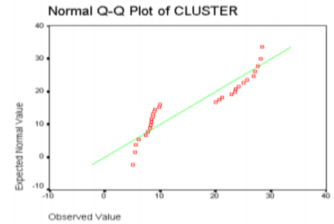
this is the « look » of a normal distribution



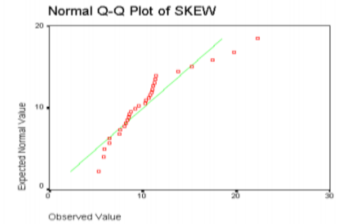


Some examples

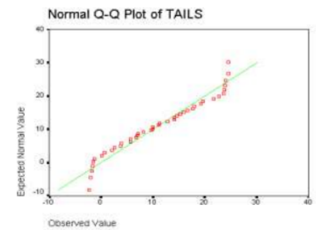
http://math.bu.edu/people/nkatenka/MA115_FALL2010/QQPlot.pdf



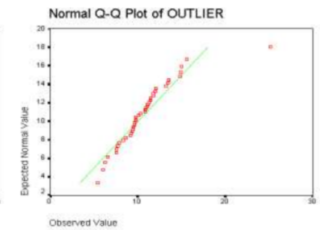
seems to have 2 clusters



seems to be skewed



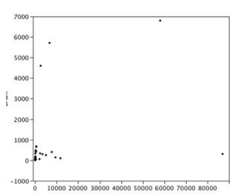
seems to have heavy tails

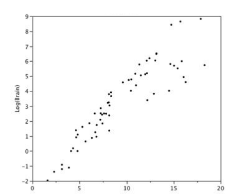


normal (except for the outlier)

Final note

The log transformation can be used to make highly skewed distributions less skewed.





http://onlinestatbook.com/2/transformations/log.html

Conclusion

- Take the time to calculate at least basic statistics and attempt an interpretation
- Take the time to study the disperson of your features
- Take the time to explore your features with simple graphs (these could potentially be used in a paper)

- Han, J., Kamber, M., Pei, J. 2012. Data Mining: Concepts and Techniques, Morgan Kaufmann (Elsevier).
- Maindonald, J., Braun, J. Data Analysis and Graphics Using R - An Example-Based Approach, 2nd edition, Cambridge University Press.
- Nisbet, R., Elder, J., Miner, G. 2009. Hanbook of Statistical Analysis and Data Mining Applications, Academic Press.
- Osborne, J.W. 2013. Best Practices in Data Cleaning. Sage.
- Witten, I.H., Frank, E., Hall, M.A. 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition, Morgan Kaufmann (Elsevier).

3. I finally have a dataset

Data preparation

- Results of Data Integration:
  - a lot, but really, a lot of Cleaning (See Osborne 2013)
  - Transformation: changing the way values are expressed
  - Imputation: what to do with missing values?
  - Reduction: (human-like) if the features are too many in the DB
  - Derivation: sometimes
  - Sampling: sometimes
- Balancing: all classes should (or not) be balanced?
- Filtering: what to do with outliers and other unwanted data?
- Reduction: reduction of the number of features (curse of dimensionality)
- Transformation: changing the way values are expressed
- Weighting: it's more a question of optimization

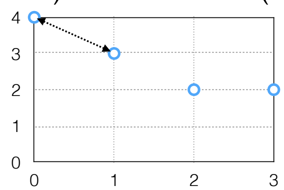
Representative and balanced dataset

- Representative
  - A dataset is always a sample of a reality: ensure that it is representative of the « full range of variability in the population » (Biber 1993)
- Balanced
  - A well-balanced dataset contains sufficient examples representative of every class
  - Of course, if your task is to detect outliers or anomalies, we don't want a so-well balanced dataset!

Do I have too many features?

- Preliminaries
  - an instance can be seen as a vector or, if you prefer, as a list of coordinates for a point in space
  - easy to calculate the distance between two points (two instances): Euclidian distance (among others) can be used:

	f	f
e	1	3
e	2	2
e	0	4
e	3	2



$$\sqrt{(f_1^1 - f_3^1)^2 + (f_1^2 - f_3^2)^2}$$

  - the generalization to k dimensions is straightforward:

$$\sqrt{(f_1^1 - f_3^1)^2 + (f_1^2 - f_3^2)^2 + \ldots + (f_1^k - f_3^k)^2}$$

Curse of dimensionality

- *k* is the number of dimensions (the number of features):
- the expected distance to the nearest neighbor goes up dramatically with *k* unless the size of the training data set increases exponentially with *k*.
- data distributed uniformly in a hypercube of dimension *k*, the probability that a point is within a distance of 0.5 units from the center is :

$$\frac{\pi^{k/2}}{2^{k-1} \cdot \Gamma\left(k/2\right)}$$

distance=0.5/1M of instances		
k	prob	points
2	0,785398163	785398,1634
3	0,523598776	523598,7756
4	0,308425138	308425,1375
5	0,164493407	164493,4067
10	0,002490395	2490,39457
20	2,46114E-08	0,02461137
30	2,04103E-14	2,04103E-08
40	3,27848E-21	3,27848E-15

When is this truly dramatic?

- When you are trying to find similarities
- When you are using clustering
- When you are using a non-parametric method like Knn

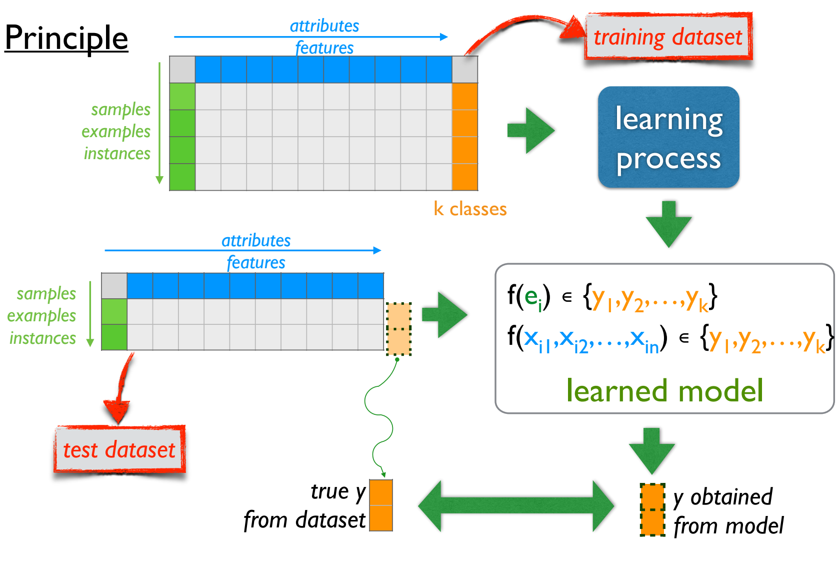
Conclusion

- Do the necessary cleaning of the data
- Leave the « special » cleaning for the optimization phase
- Be sure to have a representative and well-balanced dataset
- Try to limit, if possible, the number of features. Otherwise, think about using a dimension reduction technique.
- NEVER think that you can interpret the result of a model if you are cursed by dimensionality!

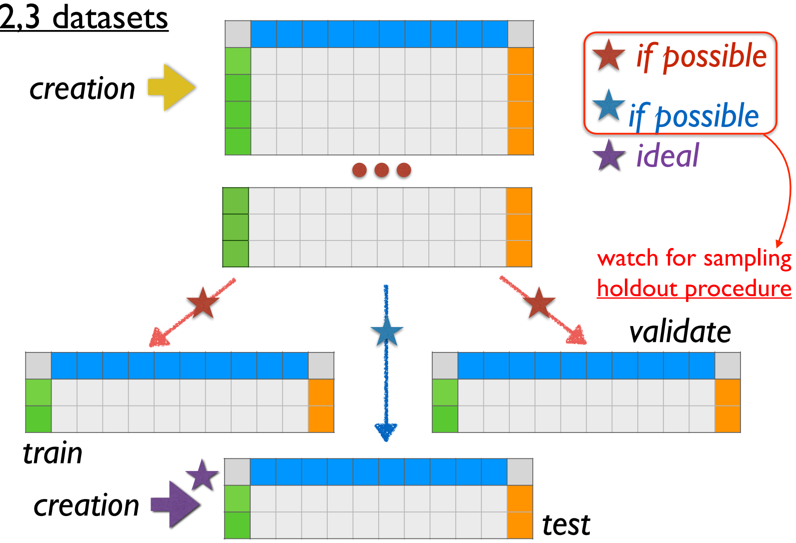
- Han, J., Kamber, M., Pei, J. 2012. Data Mining: Concepts and Techniques, Morgan Kaufmann (Elsevier).
- Maindonald, J., Braun, J. Data Analysis and Graphics Using R - An Example-Based Approach, 2nd edition, Cambridge University Press.
- Nisbet, R., Elder, J., Miner, G. 2009. Hanbook of Statistical Analysis and Data Mining Applications, Academic Press.
- Osborne, J.W. 2013. Best Practices in Data Cleaning. Sage.
- Patel, Nitin, Data Mining, Lecture 1 : k-Nearest Neighbor Algorithms for Classification and Prediction, Spring 2003. (Massachusetts Institute of Technology: MIT OpenCouseWare).
- Witten, I.H., Frank, E., Hall, M.A. 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition, Morgan Kaufmann (Elsevier).

4. Evaluate the results

Principle



1,2,3 datasets



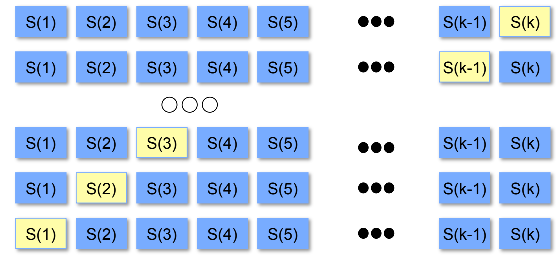
1,2,3 datasets (what to watch for)

- Very common split: 1/3 for testing, 2/3 for training
- If holdout is chosen (either a split in 2 or in 3): Watch for sampling
- Assure stratification: that examples of each class appear in both in the training set and the test/validation set(s)
- resubstitution error: the error rate on the training set
- test error: the error rate on the test set

Cross-validation

train test

- Testing on the training set is NEVER a good indicator of error rate

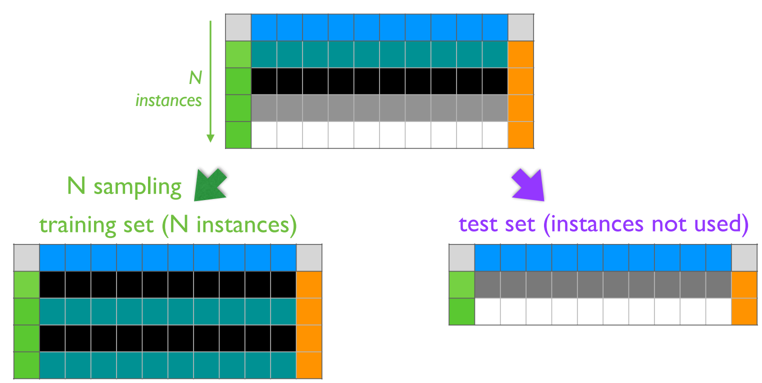


- Final results: average of the k results
- Method of choice when the dataset cannot be split efficiently between a train set and a test set.
- Usual in the literature to use 10 folds cross-validation (k=10)
- Repeat m times (m*k results)

Leave-one out and bootstrap

- Leave-one out is a cross-validation where k=1
  - ✦ Advantage 1: maximize the number of instances used for training
  - ✦ Advantage 2: the method is deterministic, no sampling, repeating it always gives the same results
  - ✦ Disadvantage 1: Computational cost of N training (N being the number of instances)
  - ✦ Disadvantage 2: Does not assure stratification (in fact, it assures it will not be stratified)

- Bootstrap is sampling a dataset with replacement (the most common is called 0.632 bootstrap)



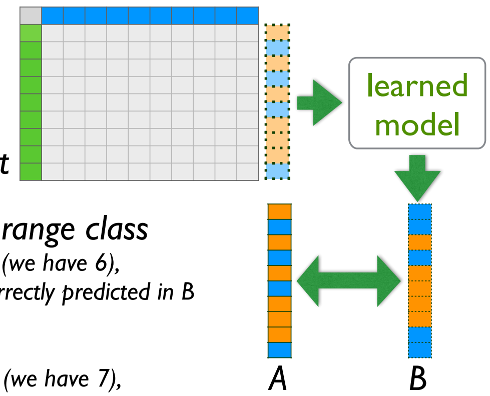
- **0.632 Bootstrap**
  - probability of being picked: $\frac{1}{n}$
  - probability of NOT being picked: $1 - \frac{1}{n}$

$$\underbrace{(1-\tfrac{1}{n})\cdot(1-\tfrac{1}{n})\cdot\ldots\cdot(1-\tfrac{1}{n})} = (1-\tfrac{1}{n})^n \approx e^{-1} = 0.368$$

  - **test set**: about 36,8 % (0.368) **training set**: about 63,2 % (0.632)
  - The error rate = 0.632 x testError + 0.368 x resubstitutionError
  - The bootstrap is repeated m times and the results are averaged
  - The best way to evaluate when the dataset is small

Basic measures

test



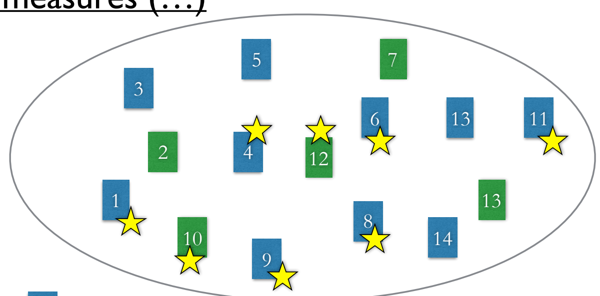
Let's consider, the orange class

4/6 • *of all the* ■ *in A (we have 6), how many are correctly predicted in B*
recall

4/5 • *of all the* ■ *in B (we have 7), how many are actually orange in A*
precision

Basic measures (...)

A

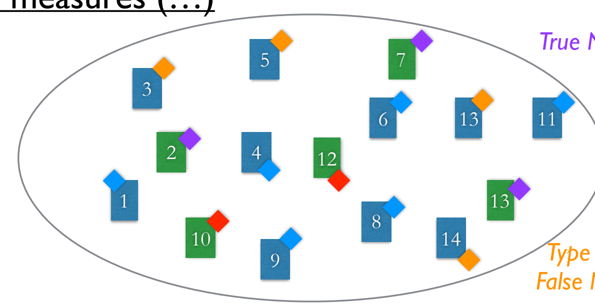


of all the ■ *that exists, how many did the algorithm actually retrieved?*
recall 6/10

The retrieval algorithm gets : 1 4 6 8 9 11 10 12 B

how many are good here?
precision 6/8

Basic measures (...)



True Negative

Type 2 error
False Negative

Recall: $\frac{TP}{TP + FN}$

Precision: $\frac{TP}{TP + FP}$

Type 1 error
False Positive

1 4 6 8 9 11 10 12

True Positive

Basic measures (…)

predicted

	A	B	C	D	E			
A	20	2	2	3	3	30	→	20/30
B	1	20	4	0	0	25	→	20/25
C	0	0	10	0	0	10	→	10/10
D	5	10	5	30	10	60	→	30/60
E	2	3	0	0	20	25	→	20/25
	28	35	21	33	33			
	$\frac{20}{28}$	$\frac{20}{35}$	$\frac{10}{21}$	$\frac{30}{33}$	$\frac{20}{33}$			

Recall: $\frac{TP}{TP + FN}$

$\frac{100}{150}$ Accuracy

Precision: $\frac{TP}{TP + FP}$

Basic measures (…)

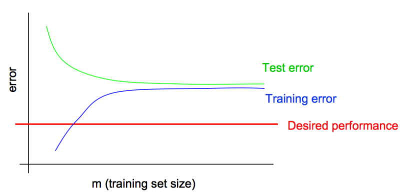
$$F_1 = 2 \cdot \frac{p \cdot r}{p + r}$$

$$F_\beta = \frac{(1+\beta^2) \cdot TP}{(1+\beta^2) \cdot TP + \beta^2 \cdot FN + FP}$$

F_2 the weights of the recall is higher

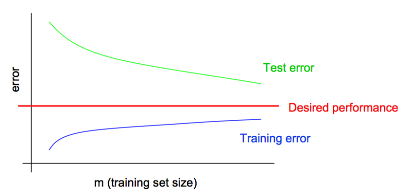
$F_{0.5}$ the weights of the precision is higher

Learning curves (bias and variance)



high bias
both errors are very high
also called underfitting

try more features

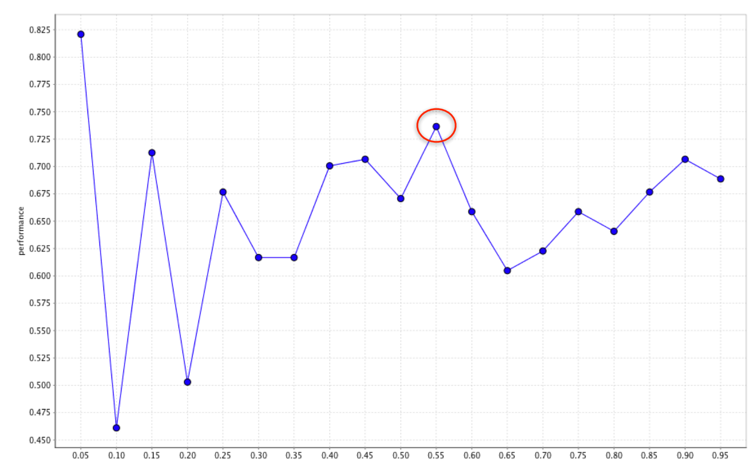


high variance
large gap between training and test
also called overfitting

try bigger dataset (more instances)
try reducing features

http://see.stanford.edu/materials/aimlcs229/ML-advice.pdf

Learning curves (rapidminer)



Conclusion

- Understand carefully what it means to learn from a dataset
- If your dataset is big enough to be split in three better, if you have one dataset constructed independently, better. But let's get realistic!
- At least, do a Cross-validation
- If using Leave-one out or bootstrap, justify it.
- Without hesitation, be prepare to explain recall vs precision. For your problem, which one is more important?
- Be aware of bias and variance effects on your results

- Geman, S., Bienenstock, E., Doursat, R. 1992. Neural networks and the bias/variance dilemma. Neural Computation 4, 1–58.
- Han, J., Kamber, M., Pei, J. 2012. Data Mining: Concepts and Techniques, Morgan Kaufmann (Elsevier).
- James, G. 2003. Variance and Bias for General Loss Functions, Machine Learning 51, 115-135.
- Ng, Andrew, Advice for applying Machine Learning, Stanford University, Course on Machine Learning. http://see.stanford.edu/materials/aimlcs229/ML-advice.pdf
- Pedregosa et al., 2011. Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830. http://www.astroml.org/sklearn_tutorial/practical.html
- Witten, I.H., Frank, E., Hall, M.A. 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition, Morgan Kaufmann (Elsevier).

5. General conclusion

A. Prepare with great care the dataset you will be using

B. Explore the both the examples and the attributes properties to understand them

C. If your dataset comes from a data integration process, consider cleaning it up. The transformation and the reduction can always come later. Never ever apply an algorithm blindly on whatever dataset in the hope that the problems will resolve by themselves

D. Be convincing when you present results. Be sure of the methods you used for testing. Understand where errors are coming from, even if your model is good (it will make a terrific discussion section!)

A. Consider the dataset « census-household »*. It contains the distribution of incomes of Americans since 1967. You will notice (in yellow) that is the data is already divided in quartiles. Using the average of all years (you must calculate that by yourself) 1) determines if there is any outliers 2) Are the values skewed? Can you give an interpretation of these data?

B. Use a boxplot to show if the « attribute » « income » contains outliers.

C. A student have painfully gathered 500 specialized NASA images that were separate in 4 classes. To describe perfectly those images, he is using 10000 attributes. What is your opinion about this dataset. Any suggestion?

D. For both situation, we will do the same exercise: Take the Training dataset and perform a cross-validation on it (using the classifier of your choice). Compare with the results when you are using the Training set only (no split, no cross-validation). Now take the Testing dataset as a test file. Discuss. (Note: you should have three sets of results here).

1. Training: segment-challenge.csv Testing: segment-test.csv
2. Training**: train1.csv Testing dataset: test1.csv

*Table H17 from: https://www.census.gov/hhes/www/income/data/historical/household/
**http://cseweb.ucsd.edu/~elkan/255/dm.pdf