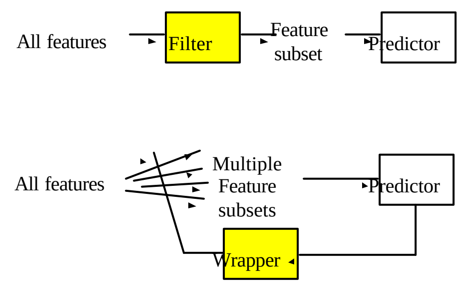
Lecture 8: Wrappers

Isabelle Guyon
guyoni@inf.ethz.ch

Chapter 2: Assessment methods
Chapter 4: Search strategies

Filters and Wrappers



Filters

Methods:

- Criterion: Measure feature/feature subset "relevance"
- Search: Usually order features (individual feature ranking or nested subsets of features)
- Assessment: Use statistical tests

Results:

- Are (relatively) robust against overfitting
- May fail to select the most "useful" features

Wrappers

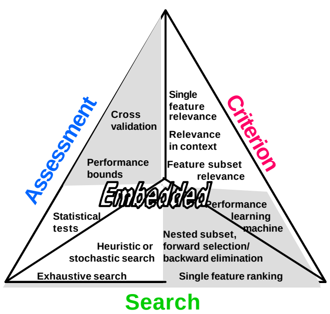
Methods:

- Criterion: Measure feature subset "usefulness"
- Search: Search the space of all feature subsets
- Assessment: Use cross-validation

Results:

- Can in principle find the most "useful" features, but
- Are prone to overfitting

Three “Ingredients”



Assessment Methods

How good are the feature subsets we have selected?

- Classical statistics:
  – Perform statistical tests.
- Machine learning:
  – Use a training set and a validation set.

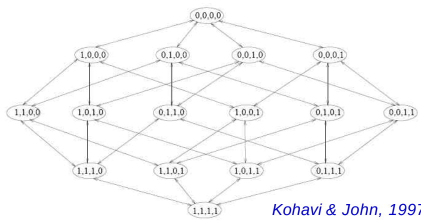
Part I: Search methods

Mostly for wrappers but also for filters

Wrapper Setting

- For simplicity, in this part of the lecture we will consider the wrapper setting in which
  – data are split into one training and one validation set
  – a feature subset is assessed by the validation performance of a classifier training on the training set using that feature subset
- Other setting are possible
  – not using a classifier (filter combined with search)
  – using a classifier with cross-validation or performance bounds for assessment

Exhaustive Search

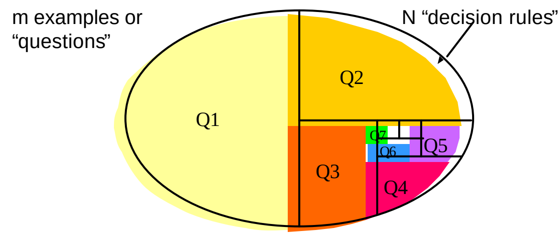


Kohavi & John, 1997

For n features, 2^n possible feature subsets!

2^n trainings

Statistical Complexity



Game of 20 questions: if you ask the questions correctly, you rule out ½ of the remaining possibilities at each question => the solution is found in $m_{opt}=\log_2 N$ questions.

Vapnik’s Bounds

1) m examples, N decision rules, learning without training error, generalization error rate bound:

$$E_{gene} \leq \frac{\ln N - \ln \alpha}{m}, \quad \text{with proba } (1-\alpha)$$

1) same but E_{tr} is the training error:

$$E_{gene} \leq E_{tr} + \mathrm{sqrt}\left(\frac{\ln N - \ln(\alpha/2)}{2\,m}\right)$$

Wrapper Complexity

- For simplicity, we will call C=log N the “complexity” of learning from a finite number of decision rules.
- The generalization error is governed by C/m.
- In our setting, **N is the number of feature subsets** to select from, **m is the number of *validation set* examples**.
- For the exhaustive search $N=2^n$ hence the generalization error is governed by n/m. We can only afford searching a number of features of the order on m.

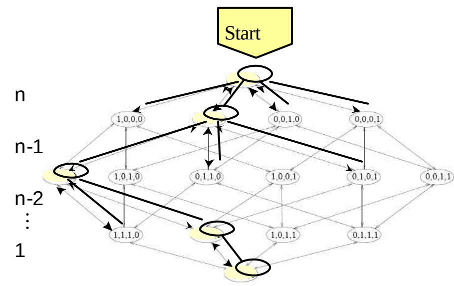
Nested Subset Methods

Nested subset methods perform a greedy search:

At each step add or remove a single feature to best improve (or least degrade) the cost function.

- **Backward elimination:**
  Start with all features, progressively remove (never add).
- **Forward selection:**
  Start with an empty set, progressively add (never remove).

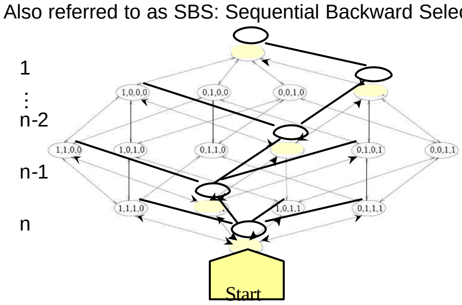
Forward Selection



Also referred to as SFS: Sequential Forward Selection

Backward Elimination

Also referred to as SBS: Sequential Backward Selection



Computational Complexity

Imagining one split training/validation, n features:

- Step 1: Train n classifiers
- Step 2: Train (n-1) classifiers
- …
- Step n: Train 1 classifier

➔n+(n-1)+…+1 = n(n+1)/2 trainings

But: forward selection starts with small feature subsets, so cheaper if stopped early.

Statistical Complexity

- N= n(n+1)/2 feature subsets considered.
- C=log N ~ log n^2 = 2 log n
- Generalization error governed by C/m ~ (log n) / m
- Much better than exhaustive search: we can afford to search a number of features n **exponential** in the number of validation examples m.

Comparison

1) In **Feature Ranking**:
- There is no search.
- A total order of features is formed
- This also defined nested subsets.
- To determine the optimum number of features, one can used the performances of a classifier, but the only n trainings are performed i.e. C = n .

2) Some **Embedded Methods** are also nested subset methods (performing forward selection or backward elimination). But at each step, they "consider" only the addition or removal of ONE feature so, n trainings are performed i.e. C = n .

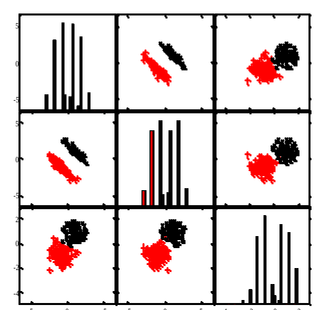
Complexity Comparison

Generalization_error ≤ Validation_error + ε(C / m)

Method	Number of subsets tried	Complexity C
Exhaustive search wrapper	2^n	n
Nested subsets greedy wrapper	n(n+1)/2	log n
Feature ranking or embedded methods	n	log n

m: number of validation examples, n: number of features.

Forward or Backward?



A Few Variants and Extensions

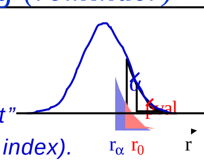
- **Beam search:** keep k best path at each step.
- **GSFS:** generalized sequential forward selection – when (n-k) features are left try all subsets of g features i.e. $\binom{n-k}{g}$ trainings. More trainings at each step, but fewer steps.
- **PTA(l,r):** plus l , take away r – at each step, run SFS l times then SBS r times.
- **Floating search** (SFFS and SBFS): One step of SFS (resp. SBS), then SBS (resp. SFS) as long as we find better subsets than those of the same size obtained so far. Any time, if a better subset of the same size was already found, switch abruptly.

Stochastic Search

- Simulated Annealing:
  - Make a step in feature space, compute ΔE
  - If $\Delta E<0$, accept the change
  - Otherwise, accept the change with probability $\exp(-\Delta E/T)$
  - Progressively “cool down”.
- Genetic Algorithms:
  - Keep a “population” of candidates (not just one)
  - A bit vector defining a feature subset is a “chromosome”
  - A “mutation” is a bit flip
  - A “cross-over” is obtained by cutting two chromosomes and swapping their tails.

Part III:
Assessment Methods:
Machine Learning Viewpoint

Hypothesis Testing *(reminder)*

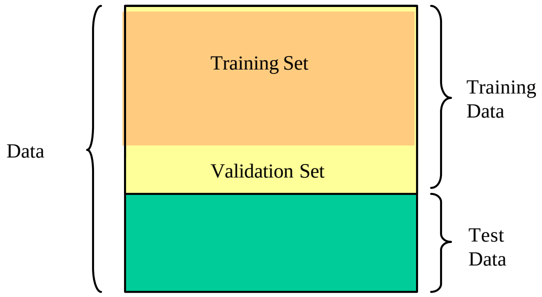


Ingredients:

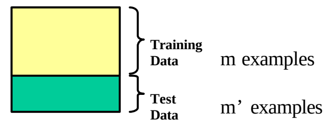
- A “null hypothesis” H_0.
  “H_0: The feature is irrelevant”
- A test statistic R *(relevance index).*
- A distribution of R if H_0 is true *(null distribution)* Proba(R > ε).
- A risk value α and its corresponding threshold r_α, such that α=Proba(R > r_α).
- A realization r_0 of R from the training samples.

If $r_0 > r_\alpha$, reject H_0, with risk α of being wrong.

ML viewpoint

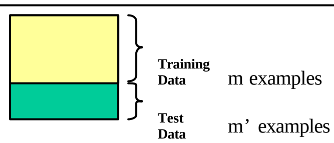


Variance of test error rate



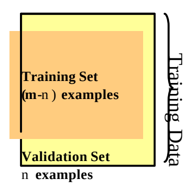
- i.i.d. errors.
- 2-class classification case: probability of error E, m' independent Bernouilli trials.
- The number of errors is distributed according to the Bionomial law of expected value m'E and variance m'E(1-E).
- The error rate (average number of errors) has variance **E(1-E)/m'**. [because var(aX)=a^2var(X)]

What size test set?

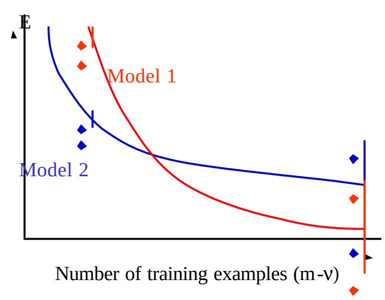


- Variance of test error rate σ^2= **E(1-E)/m'**.
  If E<<1, $\sigma^2 \cong E/m'$. (1)
- Choose a given coefficient of variance $\sigma/E=0.1$, that is $\sigma^2/E^2 = 0.01$. (2)
- Combining (1) and (2):
  1/m'E=0.01
  m'=100/E

What size validation set?



- Single split.
- Variance of E: E(1-E)/ν
- Tradeoff bias/variance.



Test of significance

- What difference in error rate between 2 classifiers is statistically significant?
- McNemar paired test:
  - assume classifier 1 is better
  - ν_i=number of errors classifier i makes that the other classifier does not make.
  - if $E_2 - E_1 \geq (z_\alpha/\nu)\mathrm{sqrt}(\nu_1 + \nu_2)$ reject H_0 of equality of error rates with risk α.
  - one sided risk α=0.01, z_α=2.33.

Single Split

- Advantage: i.i.d errors. We can easily compute error bars and perform statistical tests.
- Disadvantage:
  – Small number of validation examples: large error bar.
  – Large number of validation examples, small number of training examples: large bias.

Cross-Validation

- Average over multiple splits
- Multiple splits with replacement (bootstrap)
- K-fold cross-validation
- Leave-one-out

Virtual LOO

- For some algorithms, it is possible to compute exactly (or approximately) the effect of removing one example on the loss function value of that example.
- Need to train only once!
- Examples:
  – Least square regression: exact formula.
  – Neural nets: approximate formula.
  – SVC: approximate formula.

Avoid biased CV!

- Wrong:
  - Rank the features with all the training set.
  - Use CV (e.g. virtual LOO) to select among subsets of variable size.
  - Cost: one training for each subset size.
- Correct:
  - Remove one example.
  - Rank the features.
  - Train on remaining examples and test on left out example for variable subset sizes.
  - Average the results for each subset size.
  - Cost: m training for each subset size.

Nested CV loops

- One should select both features and hyperparameters. Which should come first?
  - HP before feature selection
  - feature selection before HP
  - Both simultaneously
- Difficulty: both simultaneously is computationally expensive and requires a lot of data.

Variance of CV

- We average over multiple splits, but now we do not know the error bar exactly anymore (non i.i.d. errors).
- LOO has a lot of variance. Often 10-fold CV is a good choice.
- Stdev(CV-results): overestimate error bar; Stderr(CV-results): underestimate error.

Multiple Testing

- When we compare N classifiers, we perform multiple tests. Our risk of being wrong increases. Remember Bonferroni's correction $\alpha \leftarrow \alpha/N$.
- This is the same story as the Vapnik bound: $E_{gene} \le \frac{\ln N - \ln \alpha}{m}$, with proba $(1-\alpha)$
- One should compare as few classifiers are possible:
  - Pre-rank the classifiers before your experiments
  - Of two classifiers performing similarly (within the error bar), prefer the classifier of lower rank.

Performance Prediction Challenge

Dataset	Size	Type	Features	Training Examples	Validation Examples	Test Examples
ADA	0.6 MB	Dense	48	4147	415	41471
GINA	19.4 MB	Dense	970	3153	315	31532
HIVA	7.6 MB	Dense	1617	3845	384	38449
NOVA	2.3 MB	Sparse binary	16969	1754	175	17537
SYLVA	15.6 MB	Dense	216	13086	1308	130858

http://www.modelselect.inf.ethz.ch/

Conclusion

- No training data split:
  - Use statistical tests or probe method to compute FPR=pval.
  - Set threshold of significance on FDR $\cong$ FPRn/n_{sc}
- Training data split(s):
  - One split: variance known E(1-E)/ν (but high), statistical tests can be performed.
  - Cross-validation: variance less high but not exactly known, statistical tests less rigorous.
  - Multiple comparisons: rank classifiers a priori.

Exercise Class

Homework 8: Dexter

- Baseline model: 5% BER
- Best challenge entries ~3% BER
- 1) Download the software for homework 7.
  2) Using the method you implemented for homework 7 or another method, try to outperform the baseline method on the Dexter dataset.
  3) Email a zip file your results to guyoni@inf.ethz.ch with subject "Homework8" no later than:
  Tuesday December 20th.

Tips for making a good slide presentation

Outline

- The contents
- The spirit
- The title slide
- The warm up
- The slides
- The flow
- The take home message
- The questions

Good Presentations:

1) **Good material**
   - know what you want to talk about
2) **Good slides**
   - informative but simple
3) **Good communication skills**
   - practice, practice, practice

Good Material

You need:

- Something **interesting** to communicate
- A **goal**
  – Get a job offer
- To **start**
  – Make an outline
  – Choose your title
  – Think of your opening joke

The art of being relevant

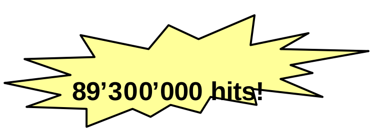
Isabelle Guyon
ETH Zürich
guyoni@inf.ethz.ch

A Good Title

- Short
- Informative
- A bit provocative

Better be Relevant

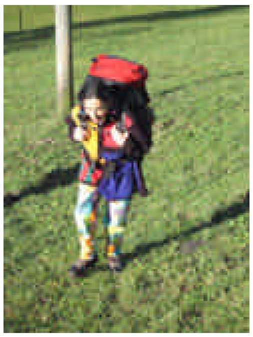
- Irrelevance: **1’610’000 hits in Google**
- Relevance:



A Good Start

- Come early
  - to make sure the projector works
  - to meet with your audience
- Thank your guest
  - and your collaborators
- Make a joke
  - if you can’t… show your outline

ME



- Twenty years of experience in hill climbing, always choosing the steepest ascent
- Three+one children (the third one is my husband)
- Judo black belt

A Good Spirit

- Provide a service
  - your audience is your customer
- Impress by your contents
  - no boasting
- Be nice
  - don't talk bad about others
  - acknowledge the work of others

Feature Irrelevance (variants)

Conditionally

- ~~Surely~~ irrelevant feature:

  $P(X_i, Y \,|\mathbf{X}^{-i}) = P(X_i \,|\mathbf{X}^{-i})P(Y \,|\mathbf{X}^{-i})$

  for all assignment of values to $\mathbf{X}^{-i}$

- Surely irrelevant feature:

  $P(X_i, Y \,|\mathbf{S}^{-i}) = P(X_i \,|\mathbf{S}^{-i})P(Y \,|\mathbf{S}^{-i})$

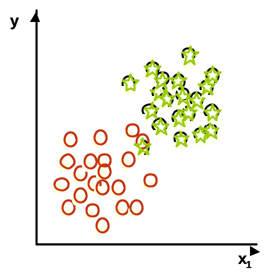
  for all $\mathbf{S}^{-i} \subseteq \mathbf{X}^{-i}$

  for all assignment of values to $\mathbf{S}^{-i}$

Informative Slides

- One topic per slide
  - have a slide title
- Go from the known to the unknown
  - start with a sentence, a picture, an idea people are familiar with
- Less is more
  - avoid busy slides, too many fonts
  - but, labels the axes!

Adding a variable...



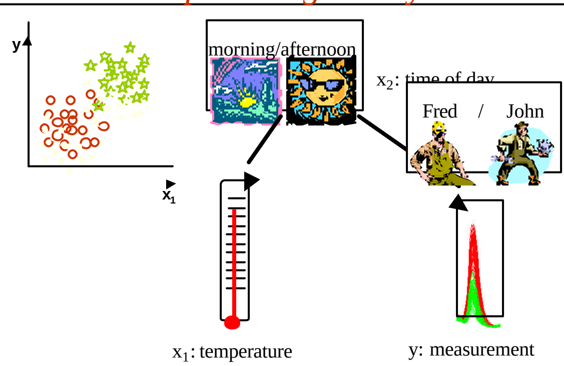
... can make another one irrelevant

Pictures

One picture is worth 10'000 words
"Un bon croquis vaut mieux qu'un long discours" (Napoléon)

- Use colors
  - ... but not too many
  - be consistent with color-coding
- Use animations
  - but only if necessary

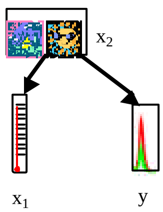
Explaining Away



Flow

- Progress smoothly
  - don't jump from one idea to the next
  - eventually repeat the last sentence/picture
- Progress logically
  - don't assume anything is self evident
  - go from the known to the unknown
- Progress slowly
  - one idea at a time
  - stop to breathe and get questions

Conditional Relevance



- We found that x_1 and y are correlated:

$P(X_1,Y) \neq P(X_1)P(Y)$

- But they are conditionally independent:

$P(X_1,Y|X_2=M) = P(X_1|X_2=M)P(Y|X_2=M)$

$P(X_1,Y|X_2=A) = P(X_1|X_2=A)P(Y|X_2=A)$

so … $P(X_i, Y|\mathbf{X}^{-i}) = P(X_i|\mathbf{X}^{-i})P(Y|\mathbf{X}^{-i})$

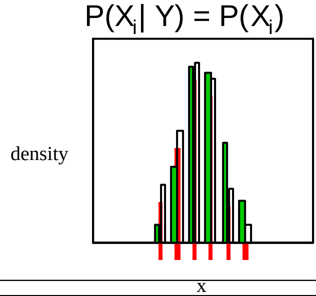
does not imply $P(X_i, Y|\mathbf{S}^{-i}) = P(X_i|\mathbf{S}^{-i})P(Y|\mathbf{S}^{-i})$ for $\mathbf{S}^{-i} \subset \mathbf{X}^{-i}$

Punchline

- Do not forget to SAY what should be concluded
- Nothing is self evident

Individual Irrelevance

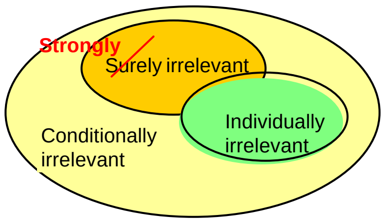
$$P(X_i, Y) = P(X_i)\,P(Y)$$
$$P(X_i|\,Y) = P(X_i)$$



Good Explanations

- Speak clearly
  - don't whisper
- Explain everything on the slide
  - what are the axes of the plots?
  - point at what you explain
- Get feed-back from the audience
  - make eye contacts
  - ask questions
- Rehearse your talk
  - preferably in front of friends
  - keep track of your time

What is Relevance?



Closing

- Don't forget the "take home message"
- Thank your audience
- Open up for questions
- Answer the questions with confidence (but don't lie)
- Verify you answered the question