Feature selection and extraction

Petr Pošík

Czech Technical University in Prague
Faculty of Electrical Engineering
Dept. of Cybernetics

Feature selection **2**

Univariate methods of feature selection **6**

Multivariate methods of feature selection **11**

Feature extraction **15**

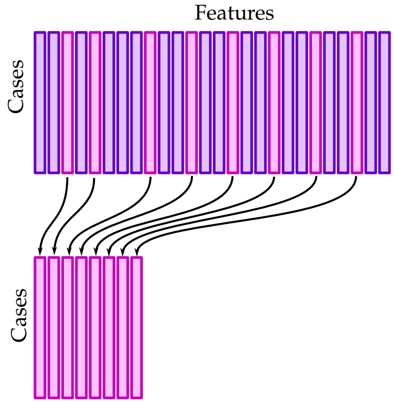
Conclusions **17**

Motivation

Why? To reduce overfitting which arises

- when we have a regular data set ($|T| > D$), but too flexible model, and/or
- when we have a high-dimensional data set with not enough data ($|T| < D$).

Data sets with thousands or millions of variables (features) are quite usual these days: we want to choose only those, which are needed to *construct simpler, faster, and more accurate models*.



Example: Fisher's Iris data

Complete enumeration of all combinations of features:
LOO Xval Error: Leave-one-out crossvalidation error

	Input features				LOO Xval Error	
	SL	SW	PL	PW	Dec. tree	3-NN
# inputs					100.0 %	100.0 %
1 input	x				26.7 %	28.7 %
		x			41.3 %	47.3 %
			x		6.0 %	8.0 %
				x	5.3 %	4.0 %
2 inputs	x	x			23.3 %	24.0 %
	x		x		6.7 %	5.3 %
	x			x	5.3 %	4.0 %
		x	x		6.0 %	6.0 %
		x		x	5.3 %	4.7 %
			x	x	4.7 %	5.3 %
3 inputs	x	x	x		6.7 %	7.3 %
	x	x		x	5.3 %	5.3 %
	x		x	x	4.7 %	3.3 %
		x	x	x	4.7 %	4.7 %
All inputs	x	x	x	x	4.7 %	4.7 %

- **Decision tree** reaches its lowest error (4.7 %) whenever PL and PW are among the inputs; it is able to choose them for decision making, more features do not harm.
- **3-NN** itself does not contain any feature selection method, it uses all features available. *The lowest error is usually not achieved when using all inputs!*

Classification of feature selection methods

Classification based on the number of variable considered together:

- **Univariate methods, variable ranking:** consider the input variables (features, attributes) one by one.
- **Multivariate methods, variable subset selection:** consider whole groups of variables together.

Classification based on the use of the ML model in the feature selection process:

- **Filter:** selects a subset of variables independently of the model that shall subsequently use them.
- **Wrapper:** selects a subset of variables taking into accoiunt the model that shall use them.
- **Embedded method:** the feature selection method is built in the ML model (or rather its training algorithm) itself (e.g. decision trees).

Univariate methods of feature selection

Variable ranking

- The main or auxiliary technique in many more complex methods.
- Simple and scalable, often works well in practice.

Incomplete list of methods usable for various combinations of input and output variable.

	Output variable Y	
Input variable X	Nominal	Continuous
Nominal	Confusion matrix analysis $p(Y)$ vs. $p(Y\|X)$ χ^2-test of independence Inf. gain (see decision trees)	T-test, ANOVA ROC (AUC) discretize Y (see the left column)
Continuous	T-test, ANOVA ROC (AUC) logistic regression discretize X (see the top row)	correlation regression discretize Y (see the left column) discretize X (see the top row)

- All the methods provide a score which can be used to rank the input variables according to the “size of relationship” with the output variable.
- Statistical tests provide the so-called p-values (attained level of significance); these may serve to judge the absolute “importance” of an attribute.

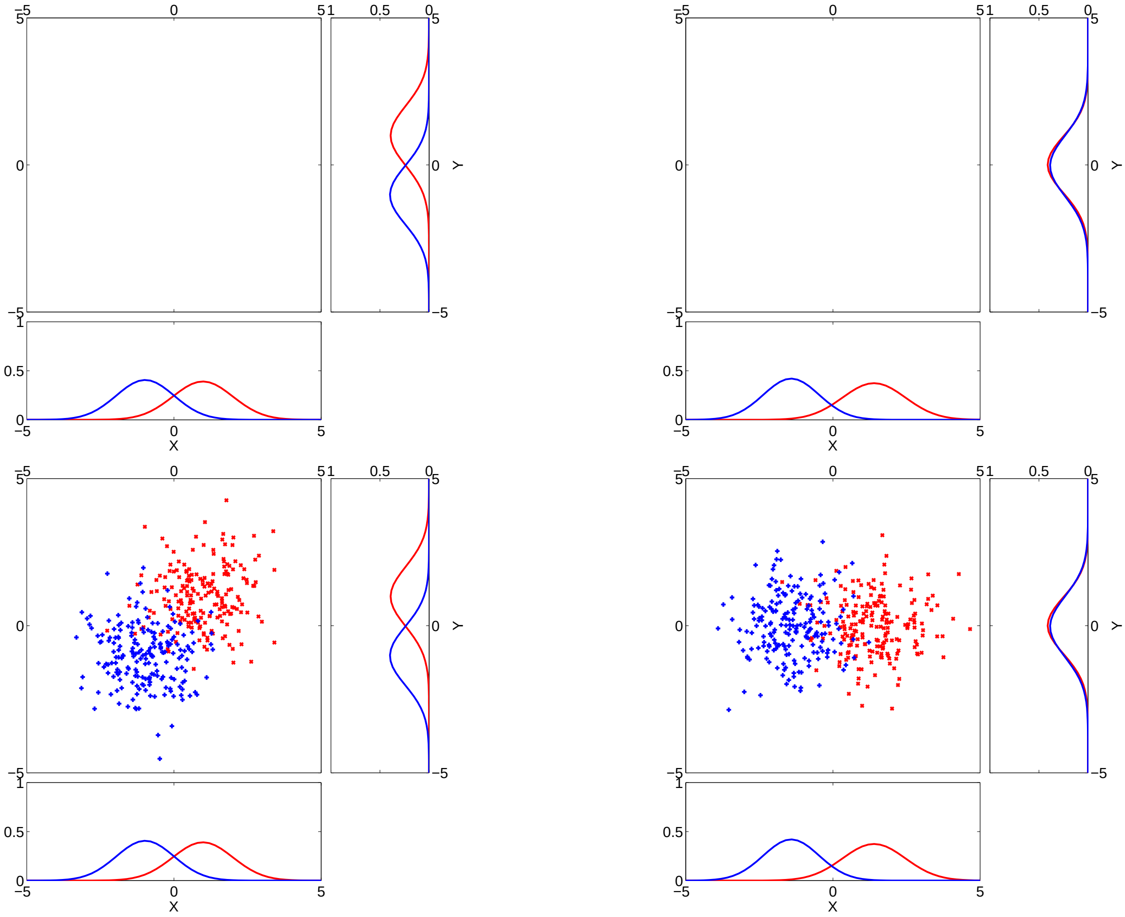
However, we can make many mistakes when relying on univariate methods!

Redundant variables?

Redundant variable

- does not bring any new information about the dependent variable.

Are we able to judge the redundancy of a variable looking just on 1D projections?



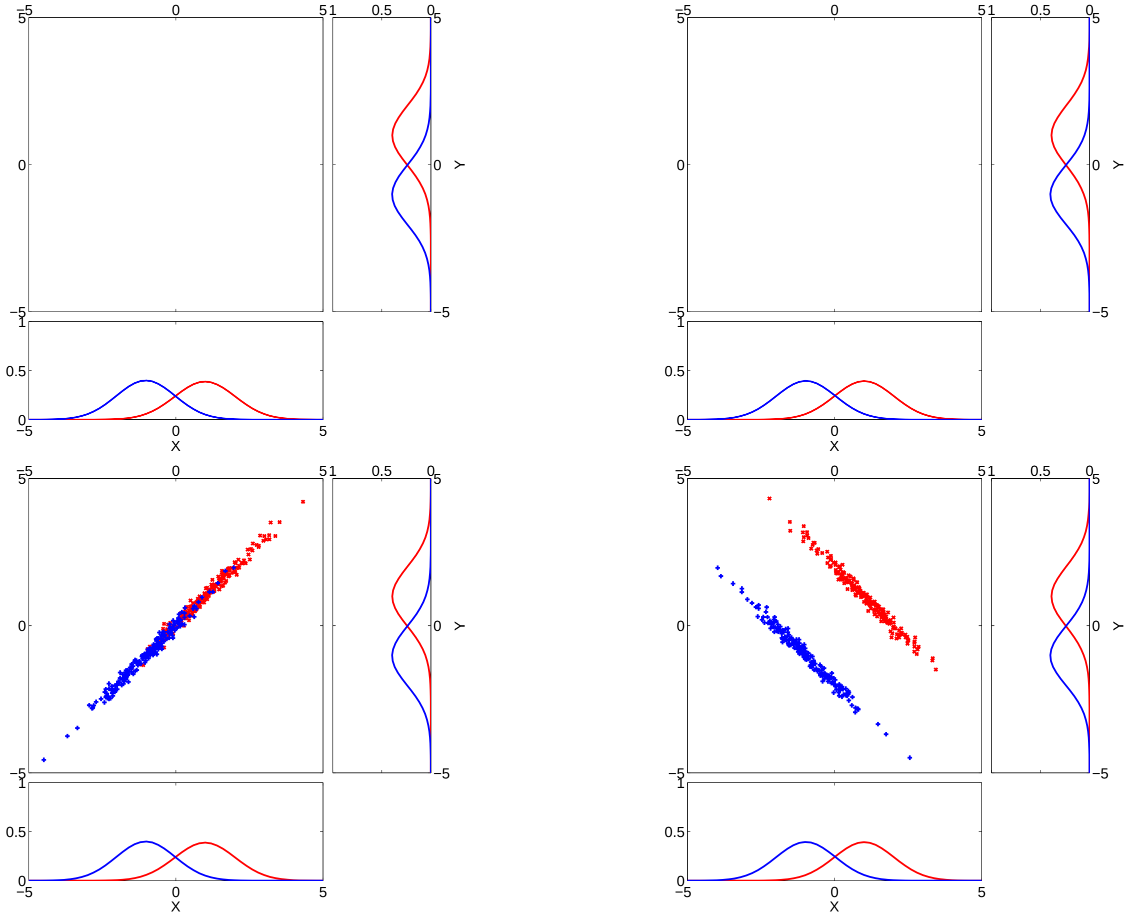
- Based on the 1D projections, it seems that both variables on the left have similar relationship with the class. (So one of them is redundant, right?) On the right, one variable seems to be useless (Y), the other (X) seems to carry more information about the class than each of the variables on the left (the “peaks” are better separated).
- The situation on the right is the same as the situation on the left, only rotated. If we decided to throw away one of the variables on the left, we wouldn’t be able to create the situation on the right.

Correlation influence on redundancy?

In the last slide:

- for a given class, the variables were not correlated, but
- the variables were correlated due to the positions of the Gaussian clouds.

How does correlation inside the classes affect redundancy?



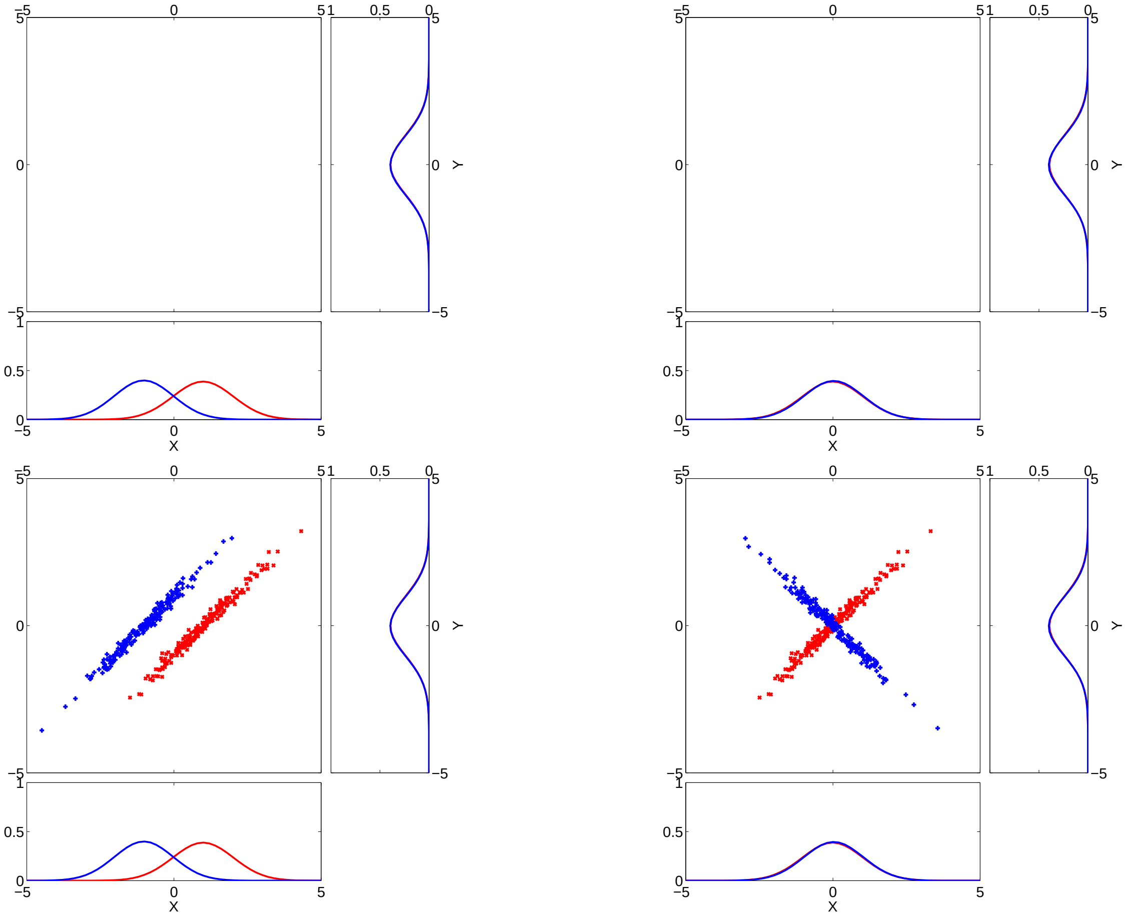
- The 1D projections to X a Y axes are the same on both figures.
- On the left, the variables are highly correlated, one is almost a linear function of the other, i.e. one of them is indeed redundant. On the right, the situation is completely different: both classes are nicely separated. If we decided to throw away one of them, we could not build a perfect classifier.

Useless variables?

Useless variable

- does not carry any information about the dependent variable; the output is independent of it.

Can we judge if a variable is useless just from 1D projections? Can a seemingly useless variable be usefull in combination with other variables?



- On the left, based on 1D projections, it seems that variable X carries some information about the class, while variable Y does not. On the right, seemingly, neither variable carries any information about the class.
- On the left: seemingly useless variable Y is useful in combination with X! On the right: although both variables were seemingly useless, together they allow us to build quite a good classifier!

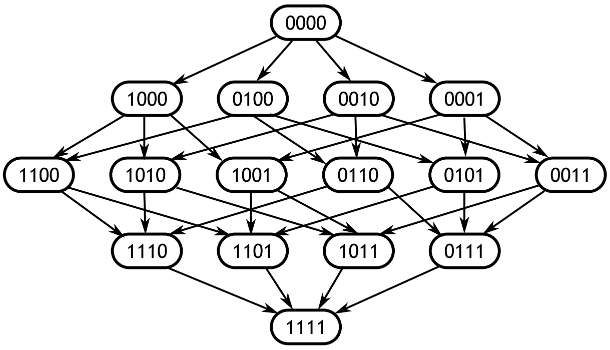
Multivariate methods of feature selection

Univariate methods may fail:

- They needn't recognize that a feature is important (in combination with other variable).
- They can select a group of variables which are dependent and carry similar (or the same) information about the output, i.e. it is sufficient to use only one (or a few) of these variables.

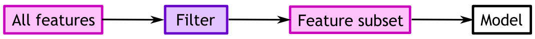
Multivariate feature selection is complex!

- D variables, 2^D variable subsets!



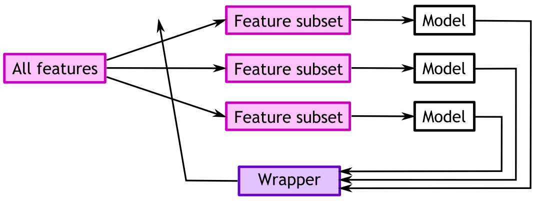
Filter vs. Wrapper

Filter: selects a subset of variables independently of the model that shall use them.



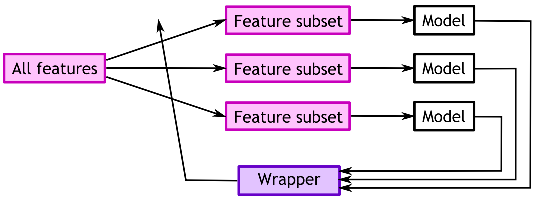
- It is a one-shot process (not iterative).
- It provides a set of "the most important" variables as the resulting subset *independently of the employed model*.

Wrapper: selects a subset of variables taking into account the model that will use them.



- It is an iterative process.
- In each iteration, several subsets of input variables is generated and *tested on the particular model type*.
- According to the success of the model for individual feature subsets, it is chosem which subsets will be tested in the next iteration.

Wrappers



Wrapper:

- A general method of feature selection.
- The model type and its learning algorithm treated as black box.

Before applying the wrapper feature selection we have to specify:

- What model type and which learning algorithm shall be used?
- How to evaluate the model accuracy?
  - Based on testing data, or using k-fold crossvalidation?
- How to search the space of possible feature subsets?
  - NP-hard problem.
  - Enumerative search is possible only for small number of features (e.g. the example with Iris data set).
  - Greedy search is often used (*forward selection* or *backward elimination*).
  - Branch and bound, simulated annealing, genetic algorithms, . . .

Feature extraction

Feature extraction

Better prediction models are often build using features derived from the original ones (aggregations, tranformations, etc.):

- When judging the overall health of a patient, is it better to know that she visited her physician in September 2008, October 2009, January 2010, February 2010 and in April 2010, or is it better to know that in 2008 and 2009, she visited the doctor only once each year, while in the first half on 2010, she already made 3 visits?
- When estimating the result of a chess play, is it better to know that the black king is at D1 while white queen at H4, or is it better to know that the white queen threatens the black king?

New variables are often derived (constructed) which are

- linear or non-linear functions of
- one, more, or all the input variables with the hope that they will have
- larger and cleaner relationship with the output variable.
- Domain knowledge is used very often.

Two different goals of feature extraction:

- data reconstruction (unsupervised methods)
- prediction improvement (supervised methods)

Methods:

- clustering (a group of similar variables is replaced with a single centroid)
- principal component analysis (PCA/SVD), projection pursuit, linear discriminant analysis (LDA), kernel PCA, . . .
- spectral transformations (Fourier, wavelet), . . .

Conclusions

Summary

- The selection of the optimal subset of input variables is NP-hard problem.
- Univariate methods are
  - simple and allow us to rank the input variables according to some measure of usefullnes for prediction,
  - work well in practice, but
  - can make fatal errors.
- Multivariate methods are
  - more resistent against the mistakes during selection, but
  - they are computationally much more demanding.
- We distinguish
  - filters,
  - wrappers, and
  - embedded methods.
- Feature selection merely selects a subset of the original variables, while feature extractio constructs new variables from the original ones.