Title:	Feature Selection
Course:	Data Mining
Instructor:	Claudio Sartori
Date:	13/04/2023
Master:	Data Science and Business Analytics
Academic Year:	2022/2023

BOLOGNA BUSINESS SCHOOL

Alma Mater Studiorum Università di Bologna

Synonims

At least in this document

- feature
- attribute
- column

Problems with attributes I

[Witten et al.(2011)Witten, Frank, and Hall] Ch 7 and 11 and also this

the significance of attributes for the purposes of data mining can vary highly

irrelevant alteration they can alter the results of some mining algorithm, in particular in case of insufficient control of overfitting

redundant some attributes can be strongly related to other useful attributes

Problems with attributes II

[Witten et al.(2011)Witten, Frank, and Hall] Ch 7 and 11 and also this

alteration some mining algorithms (e.g. Naive Bayes) are strongly influenced by strong correlations between attributes

Problems with attributes III

[Witten et al.(2011)Witten, Frank, and Hall] Ch 7 and 11 and also this

confounding some attributes can be misleading

- **hidden effect** on the outcome variable
  - **example** in a study on weight gain, physical exercise, age and sex, the sex can be confounding if in the available data the ages of males and females have very different ranges

mixed effect one attribute could be strongly related to the class in 65% of the cases and random in the other cases

Why feature selection/creation

Sometimes less is better (by Rohan Rao)

Sometimes:

- It enables the machine learning algorithm to train faster
- It reduces the complexity of a model and makes it easier to interpret
- It improves the accuracy of a model if the right subset is chosen
- It reduces overfitting.

It may be the case that a specific selection action obtain only one of the above effects

Supervised or not?

unsupervised a lot of methods available e.g. for clustering see this: Feature Selection for Clustering: A Review

- feature transformation techniques, such as PCA, can have the effect of reducing the number of features

supervised consider the relationship between each attribute and the *class*

- Filter methods (i.e. Scheme–Independent Selection)
- Scheme–Dependent Selection
  - Wrapper methods
  - Embedded methods have their own built-in feature selection methods (e.g. *Lasso* and *Ridge* regression)

Filter methods (Scheme–Independent Selection)

- Assessment based on general characteristics of data
- Select the subset of attributes independently from the mining model that will be used
  - e.g. build a decision tree and consider the attributes near the root of the tree, then use the selected attributes for building a classifier with another method
  - e.g. select a subset of attributes that individually correlate to the class, but but have a little intercorrelation[1]

1 See Symmetric Uncertainty in the Data module for correlation between nominal attributes

Some filter methods

Pearson''s Correlation A measure for quantifying linear dependence between two continuous variables X and Y; value from -1 to +1

LDA Linear Discriminant Analysis is used to find a linear combination of features that characterizes or separates two or more classes

ANOVA Analysis Of VAriance is similar to LDA except for the fact that it is operated using one or more categorical independent features and one continuous dependent feature

Chi-Square Is a statistical test applied to the groups of categorical features to evaluate the likelihood of correlation or association between them using their frequency distribution

Filter Methods – Synopsis

Feature Response	*Continuous*	*Categorical*
Continuous	Pearson's Correlation	LDA
Categorical	ANOVA	Chi-Square

Set of all features $\rightarrow$ Selecting the best subset $\rightarrow$ Learning algorithm $\rightarrow$ Performance

Wrapper methods

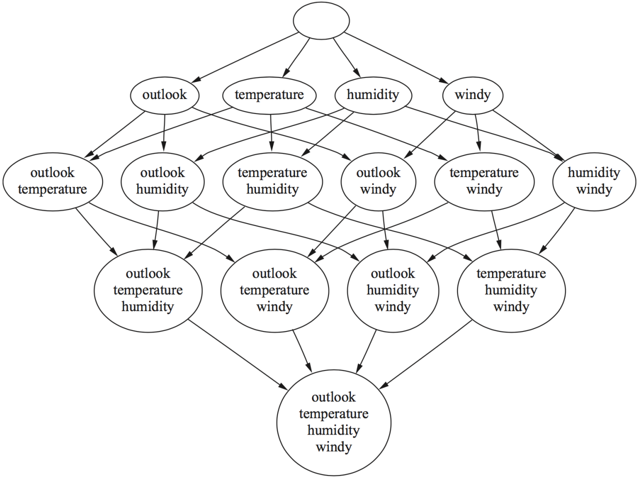
- Try to use a subset of features and train a model using them
- Based on the inferences that we draw from the previous model, we decide to add or remove features from your subset
- The problem is essentially reduced to a search problem

		Selecting the best subset		
Set of all features $\rightarrow$	Generate a subset	$\rightarrow$ / $\leftarrow$	Learning algorithm	$\rightarrow$ Performance

One wrapper method

Search the Attribute Space

- e.g. Weather dataset
- Search greedily the space
- For each subset test the performance of the chosen classification model
- Computation intensive



Difference between Filter and Wrapper methods

- Filter methods measure the relevance of features by their correlation with dependent variable while wrapper methods measure the usefulness of a subset of feature by actually training a model on it
- Filter methods are much faster compared to wrapper methods as they do not involve training the models. On the other hand, wrapper methods are computationally very expensive as well.
- Filter methods use statistical methods for evaluation of a subset of features while wrapper methods use cross validation
- Filter methods might fail to find the best subset of features in many occasions but wrapper methods can always provide the best subset of features
- Using the subset of features from the wrapper methods make the model more prone to overfitting as compared to using subset of features from the filter methods

Dimensionality reduction

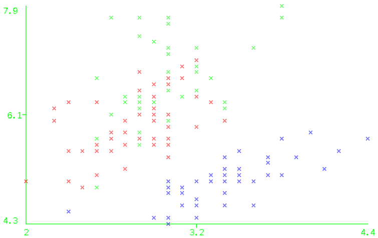
Instead of considering which subset of attributes is to be ignored it is possible to map the dataset into a new space with fewer attributes

PCA Principal Component Analysis

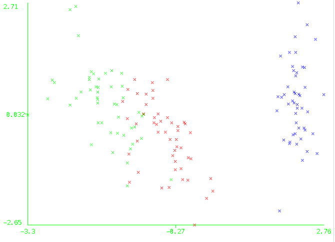
PCA

- Find a new (ordered) set of dimensions that better captures the variability of the data
  - the first one captures most of the variability
  - the second one is orthogonal to the first one and captures most of the remaining variability
  - ...
- The fraction of variance in data captured by each new variable is measured
- A small number of new variables can capture most of the variability

Iris dataset

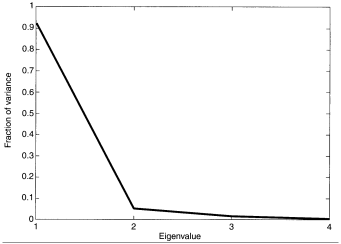


Sepalwidth/Sepallength Plot



PCA - first two components

95% variance



Fraction of variance for each principal component

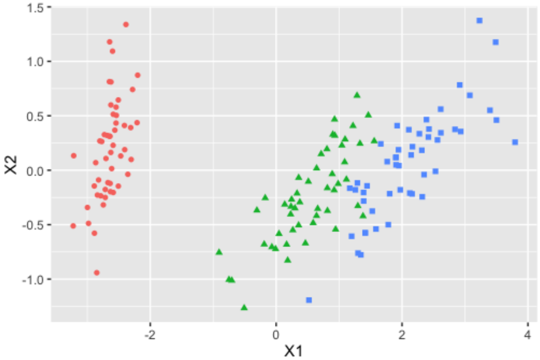
A few mathematical details

- Covariance matrix (positive semidefinite)
- Eigenvalue analysis
- Eigenvalues are positive and can be sorted in decreasing order
- Eigenvectors are sorted according to the eigenvalue order

MDS – Multi–Dimensional Scaling

A presentation technique

- Starting from the distances among the elements of the dataset
- Fits the projection of the elements into a *m* dimensional space in such a way that the distances among the elements are preserved
- Versions for non–metric and metric spaces



2D scaling for the Iris dataset

The *Scikit-learn* solution for feature selection

General structure

The main methods (there are more, somewhat different for the various estimators)

- `.fit`
  - Learn empirical variances from X
- `.fit_transform`
  - Fit to data, then transform it
- `.transform`
  - Reduce X to the selected features
- The main argument is X, the dataset

The baseline estimator

- `VarianceThreshold` = removing features with low variance
  - unsupervised
  - Example:
    - dataset with binary attributes
    - we decide to eliminate the features with a proportion 80-20 or more, $p = .8$ or more
    - a *bernoullian* experiment has variance $p * (1 - p)$
    - the threshold will be $.08 * (1 - .8) = .16$

Univariate feature selection

- Select the best set of features based on univariate statistical tests
- Consider the *original set of features* and the *target*
- For each feature, return a score and a pvalue
- Among the selection methods:
  - `SelectKBest`
    - removes all but the k highest scoring features
  - `SelectPercentile`
    - removes all but a user-specified highest scoring percentage of features

Score functions

Are used by the feature selector to evaluate how much a feature is useful to predict the target

- `mutual_info_classif` computes the Mutual Information, which is a generalisation of the *Information Gain*
- `f_classif`: Fisher test with ANOVA (analysis of variance)

Recursive Feature Elimination - RFE click to see the manual page

Feature ranking with recursive feature elimination

- Uses an external estimator to assign weights to features
- Considers smaller and smaller sets of features
- The estimator is trained on the initial set of features and the importance of each feature is obtained
- The least important features are pruned
- Stops when the desired number of features is reached

Bibliography I

- Ian H. Witten, Eibe Frank, and Mark Hall.
  Data Mining – Practical Machine Learning Tools and Techniques.
  Morgan Kaufman, 2011.
  ISBN 978-0-12-374856-0.