Advanced Data Analysis and Machine Learning: Lecture

Data Preprocessing, Feature and Model Selection

Lasse Lensu

2015-09-14

Outline

Data characteristics

- Different sources of information represent a target object or process. In many cases, the data arises from a direct or indirect measurement action.
- The data from the different sources has (naturally) different characteristics.
- From the viewpoint of further processing and usage of the data, the data can be generally characterised with the following attributes:
    - Encoding of each data element/item/set (integer/real/complex, bits per sample, sampling rate, . . .).
    - Sources of noise and signal-to-noise ratio (SNR).
    - Number of dimensions in the data item/set.
- If the data from multiple sources is to be combined, special care must be taken to characterise and preprocess/transform the data into a form enabling appropriate data fusion and information gain.

Normalisation and redistribution

LUT Lappeenranta University of Technology

- Problem: original raw data/features do not have desired characteristics for the data analysis task.
- Examples:
  - One or more data variables dominate the variation in the data set.
  - One or more data variables have skewed value distribution.
  - Two or more variables correlate significantly.
- If the data/features are preprocessed appropriately, the advantages include the following: unbiased, more representative data/features, faster model parameter estimation/learning and more representative models/better generalization.

Data normalisation

LUT Lappeenranta University of Technology

- Minmax-scaling (of features):

$$x_k^{min} = \min_i x_{ki}, \quad x_k^{max} = \max_i x_{ki}, \quad k = 1, 2, \ldots, l$$

$$\hat{x}_{ik} = \frac{x_{ik} - x_k^{min}}{x_k^{max} - x_k^{min}}$$

Data normalisation

- Minmax-scaling (of features):

$$x_k^{min} = \min_i x_{ki}, \quad x_k^{max} = \max_i x_{ki}, \quad k = 1, 2, \ldots, l$$

$$\hat{x}_{ik} = \frac{x_{ik} - x_k^{min}}{x_k^{max} - x_k^{min}}$$

- One-to-one mapping between the original and normalised values which does not cause distortion to the data distribution.
- The result can be sensitive to outliers.
- Future data can have a different value range: samples can be used despite this, can be clipped to the original range, or the samples can be ignored. Alternatives: reserve "space" for the out-of-range values or squashing.

Data normalisation

LUT Lappeenranta University of Technology

- Mean and variance normalization/standardization (of features):

$$\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik}, \quad k = 1, 2, \ldots, l$$

$$\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2$$

$$\hat{x}_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k}$$

Data normalisation

LUT Lappeenranta University of Technology

- Mean and variance normalization/standardization (of features):

$$\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik}, \quad k = 1, 2, \ldots, l$$

$$\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2$$

$$\hat{x}_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k}$$

- One-to-one mapping between the original and normalised values which does not cause distortion to the data distribution.
- Dependent on the number of samples available to estimate the mean and STD $\Rightarrow$ can still have problems related to future data with out-of-range values.

Data normalisation

- Softmax-scaling

$$
\begin{aligned}
y_{ik} &= \frac{x_{ik} - \bar{x}_k}{\sigma_k} \\
\hat{x}_{ik} &= \frac{1}{1 + e^{-y_{ik}}}
\end{aligned}
$$

Data normalisation

LUT Lappeenranta University of Technology

- Softmax-scaling

$$
\begin{aligned}
y_{ik} &= \frac{x_{ik} - \bar{x}_k}{\sigma_k} \\
\hat{x}_{ik} &= \frac{1}{1 + e^{-y_{ik}}}
\end{aligned}
$$

- Possible that very large numbers do not have unique normalised value.
- Can be controlled by λ specifying the linear portion, for example, $y_{ik} = \frac{x_{ik} - \bar{x}_k}{(\lambda/2\pi)\sigma_k}$.

Data redistribution

LUT
Lappeenranta
University of Technology

- Problems:
  - Methods generally expect the data distributions to be uniform or normal.
  - Significantly varying densities can cause problems.

Data redistribution

LUT Lappeenranta University of Technology

- Problems:
  - Methods generally expect the data distributions to be uniform or normal.
  - Significantly varying densities can cause problems.
- Possibilities for data preprocessing:
  - Requantisation: samples close to each other are merged.
  - Redistribution: sample distribution is transformed to be either equalised (target distribution is uniform), or specified (target distribution any type).

Outliers

- Outlier, *n. Statistics.* An observation whose value lies outside the set of values considered likely according to some hypothesis (usually one based on other observations); an isolated point. (OED)
  $\Rightarrow$ Unexpected, non-typical or out-of-range samples.
- Resulting from noise, human error or malfunctioning measurement/data processing equipment.
- The problem of defining outliers in a generally acceptable manner is nontrivial.

Outliers

- Outlier, *n. Statistics.* An observation whose value lies outside the set of values considered likely according to some hypothesis (usually one based on other observations); an isolated point. (OED)
  $\Rightarrow$ Unexpected, non-typical or out-of-range samples.
- Resulting from noise, human error or malfunctioning measurement/data processing equipment.
- The problem of defining outliers in a generally acceptable manner is nontrivial.
- Note: according to information theory [1], the most improbable events carry the most information!

Outlier detection

LUT Lappeenranta University of Technology

- For low-dimensional data, outliers are easy to identify visually.
- Possible task formulation: Given a set of n data points and k, the expected number of outliers, find the top k samples that are considerably dissimilar, exceptional or inconsistent with the rest of the data.
- Data models can be used to represent the "valid" data, but the representativeness/validity of the model can significantly affect the outlier detection result.

Outlier detection

- Possible solutions for outlier detection:
  - Statistical distribution-based methods:
    Assumed distribution and its parameters; working/alternative hypothesis; statistical significance.
  - Distance-based methods:
    At least a fraction of data points lie at a distance greater than a threshold.
  - Density-based methods:
    Non-uniform distributions; local outliers and reachability; degree of being an outlier (not binary).
  - Deviation-based methods:
    Subsets of data points; greatest reduction of dissimilarity metric or the statistical method within cubes.
  - Regression methods:
    Robust regression; residual error.

Variable/feature selection

- Dimensionality:
  - Variable/feature space volume is exponential w.r.t. the number of variables/features.
  - Large number of variables/features requires much data to be representative, and processing can be slow.
  - Adding more variables/features does not necessarily add more information (because of noise, correlation between variables).
- Goals:
  - Simpler models easier to interprete
  - Faster model building/analysis/training
  - Representativeness, and improved generalisation without overfitting

Variable/feature selection

LUT Lappeenranta University of Technology

- Selection is dependent on the data analysis task.
- Favorable properties for variables/features: invariance to occurring transformations.
- Two approaches:
    - Individual variable/feature selection
    - Variable/feature subset selection

Variable/feature selection

- General variable/feature selection methods:
  - Wrappers:
    Scoring of variables/features with a predictive model and error rate (of the hold-out subset).
  - Filters:
    Usefulness instead of error rate; proxy measure (mutual information[1], Pearson correlation coefficient, . . .).
  - Embedded methods:
    Selection through penalisation/exclusion of variables/features as part of the model building.
  - Heuristics:
    Combination of different approaches.

[1] joint dist. *vs.* factored marginal dist. product

Variable/feature subset selection

- Number of subsets large: for selecting l features from a total of m

$$\binom{m}{l}.$$

- Difficult problem to exhaustively select the best combination.

Variable/feature subset selection

LUT Lappeenranta University of Technology

- Number of subsets large: for selecting l features from a total of m

$$\binom{m}{l}.$$

- Difficult problem to exhaustively select the best combination.
- Possible solutions:
  - Greedy search (does not guarantee global optimum)
  - Nondeterministic (random sampling) methods

Method-independent selection

LUT
Lappeenranta
University of Technology

- Example context: Binary classification.
- Why do we need many different approaches?
- Generalization *vs.* overfitting?
- Improving classification performance by combining classifiers
- Improving classification performance by modifying training data

No classifier is superior – Occam's razor

LUT Lappeenranta University of Technology

- No classifier is superior over all problems.
- Occam's razor: "When you have two competing theories which make exactly the same predictions, the one that is simpler is the better."
  - In other words: "One should not make more assumptions than the minimum needed."
  - For a given set of observations or data, there is always an infinite number of possible models explaining those same data.
  - The simplest model should be selected because it minimizes the number of your incorrect assumptions.

No classifier is superior – no free lunch

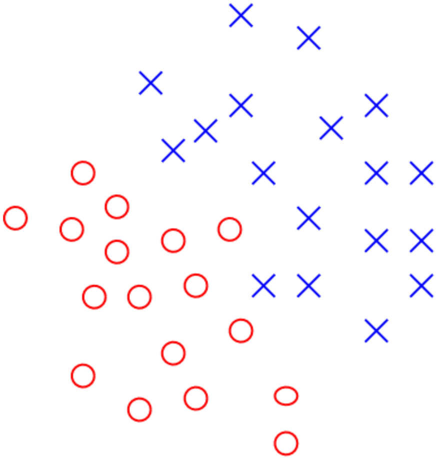
LUT Lappeenranta University of Technology

- No Free Lunch theorem: "No algorithm is superior to any other over all classification problems."
    - In the absence of prior information about a problem, there are no reasons to prefer one classifier over another.
    - Comparing generalization performance, there is no problem-independent best pattern recognition method.
    - Opposite viewpoint: The prior information (e.g. homogeneity of decision regions in the feature space) makes classifiers work.

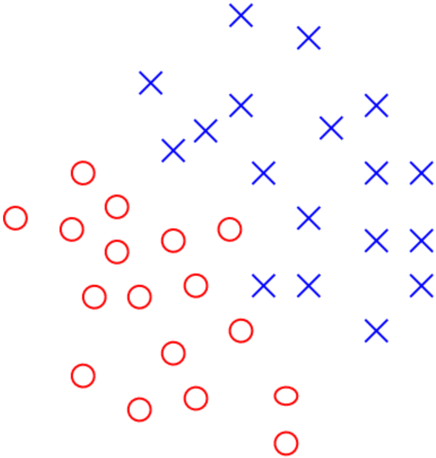
Bias and variance

- A training set is finite and it represents almost always a subset of all possible cases.
- Having few samples, it would be beneficial to use a classifier which is predictable with few samples (typically simple boundaries).
  - Predictability can differ even for classifiers with the same form of boundary.
- Having many samples, we can more confidently use classifiers with more complex boundaries.
- The total error rate of a classifier can be decomposed into two parts:
  - Error rate due to the average performance of the classifier
  - Error rate due to the variation of training set

Example data



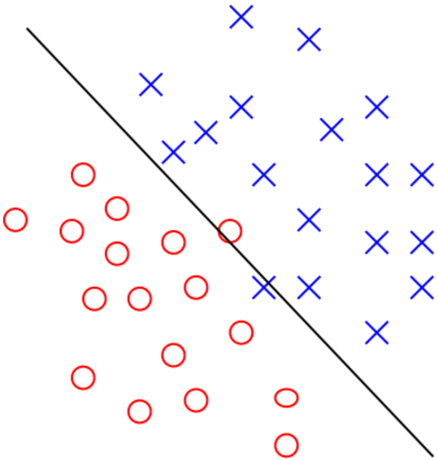
Example data



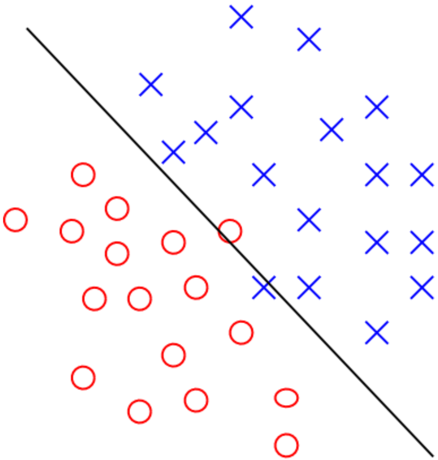
Next look into best possible classifier.

Minimum error classifier

LUT
Lappeenranta
University of Technology



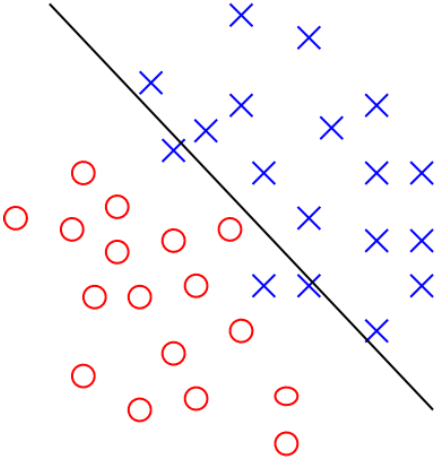
Minimum error classifier



This is the optimal classifier (minimum error).

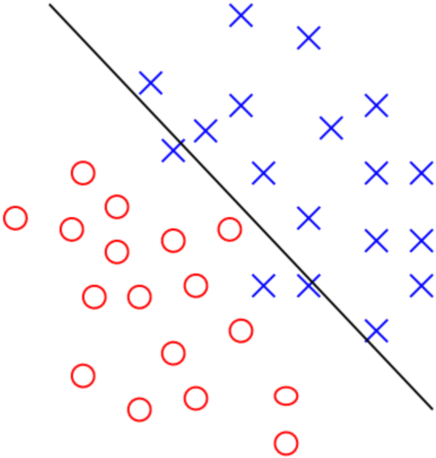
Worse classification performance

LUT
Lappeenranta
University of Technology



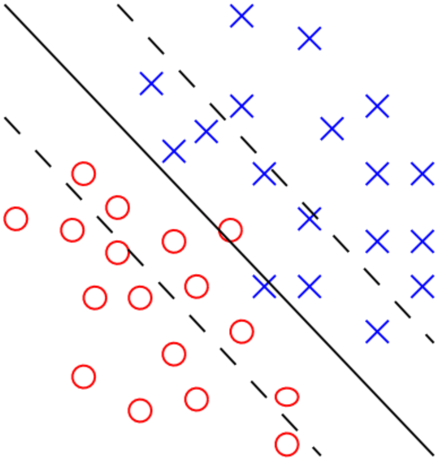
Worse classification performance

LUT
Lappeenranta
University of Technology



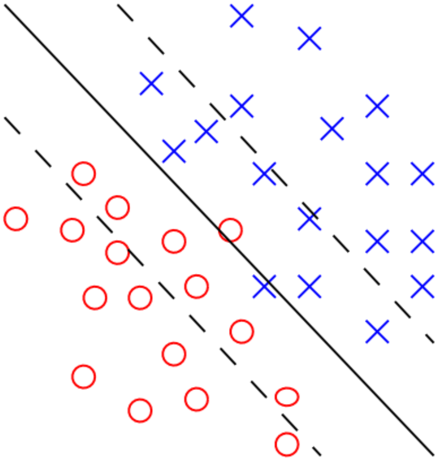
But what if these are the average performances?

Optimal on average, varies a lot



Optimal on average, varies a lot

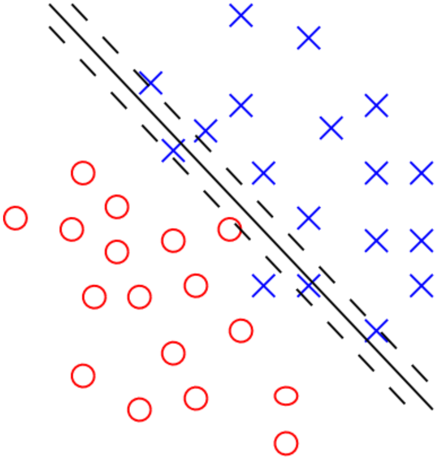
LUT
Lappeenranta
University of Technology



Optimal on average, but varies much for different training sets.

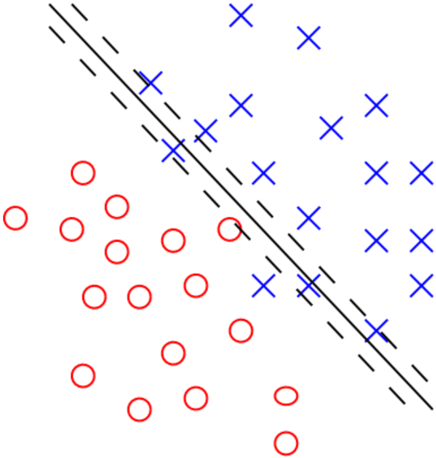
Worse average, but varies little

LUT
Lappeenranta
University of Technology



Worse average, but varies little

LUT Lappeenranta University of Technology



Worse on average, but smaller variance.

Bias and variance for classification

LUT
Lappeenranta
University of Technology

- Bias:
  - How far is the average performance from the optimal one?
- Variance:
  - How much does the performance vary over different training sets?
  - How much does the boundary vary over different training sets?
- Small variance often more important than small bias if the bias is relatively small.

Resampling

- Could we decrease the variance somehow?
- How to take an "average" classifier over different data sets?
- Resampling of the data produces several different data sets.

Summary

LUT Lappeenranta University of Technology

- Original raw data/features can have less than optimal characteristics for the data analysis task $\Rightarrow$ data preprocessing is needed.
- Data normalisation methods have different properties and consequences of using them should be understood.
- Outliers are non-typical samples in the data set, and a few approaches exist for their detection.
- Simpler models with fewer variables/features have benefits.

Claude E. Shannon.
A mathematical theory of communication.
Bell System Technical Journal, 27(3):379–423, 1948.