technische universität
dortmund

Statistical Methods of Data Analysis

Data Mining Part 1

Prof. Dr. Dr. Wolfgang Rhode Dr. Maximilian Linhoff

2023

Overview

Data Mining

Typical Exercises in Data Mining

Motivation

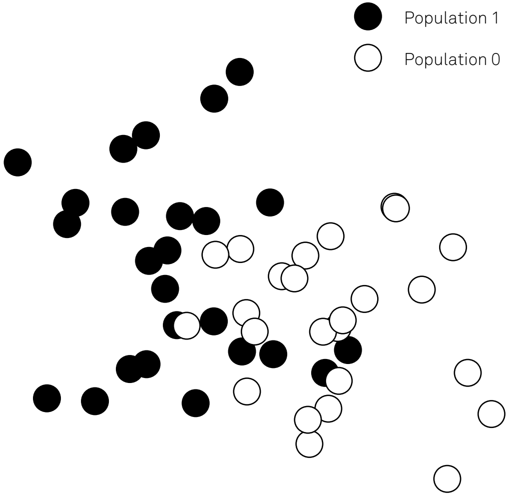
- The goal is to divide the points into two populations

Motivation

- The goal is to divide the points into two populations
- Known element affiliation in Monte Carlo

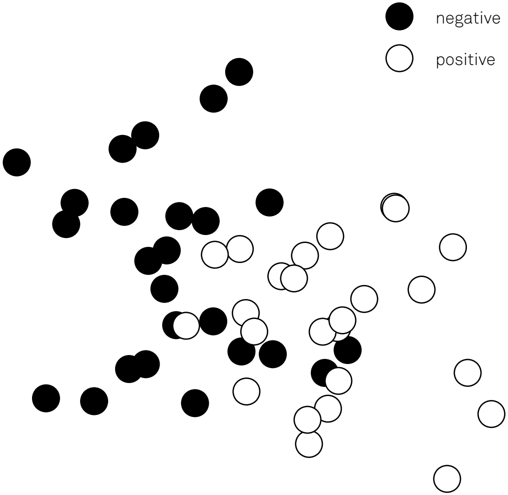
Motivation

- The goal is to divide the points into two populations
- Known element affiliation in Monte Carlo



Motivation

- The goal is to divide the points into two populations
- Known element affiliation in Monte Carlo

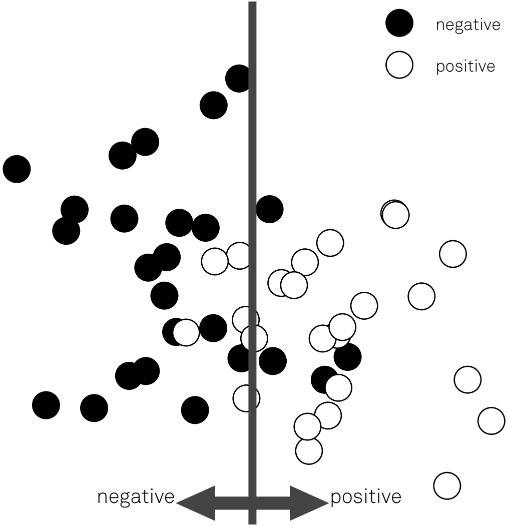


Motivation

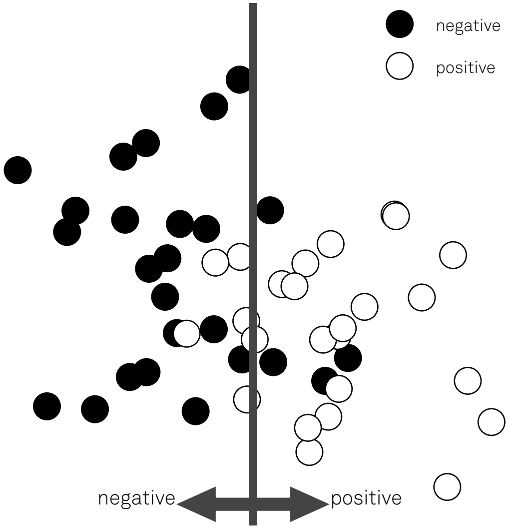
- The goal is to divide the points into two populations
- Known element affiliation in Monte Carlo
- Idea: search for "best" one-dimensional cut in Monte Carlo

Motivation

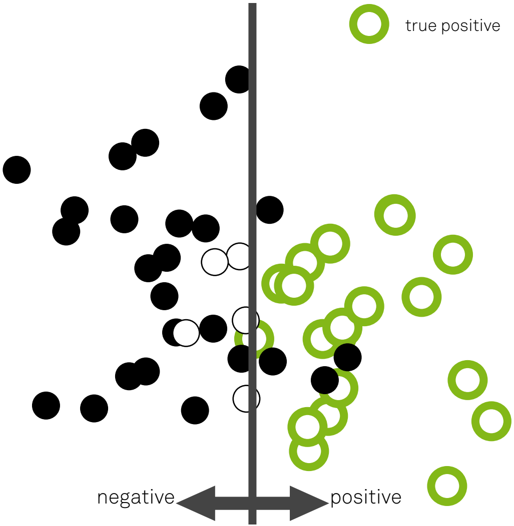
- The goal is to divide the points into two populations
- Known element affiliation in Monte Carlo
- Idea: search for “best” one-dimensional cut in Monte Carlo



Motivation: What is the best cut?



Motivation: What is the best cut?



true positive (tp)

- “positive” elements in “positive” range

Motivation: What is the best cut?

true positive (tp)

- “positive” elements in “positive” range

false negative (fn)

- “positive” elements in “negative” range

Motivation: What is the best cut?

true positive (tp)

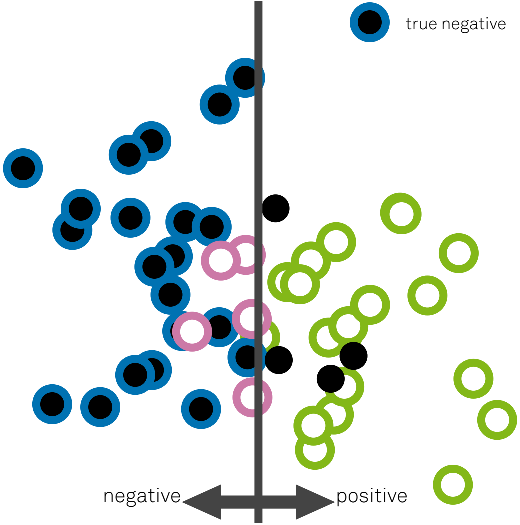
- “positive” elements in “positive” range

false negative (fn)

- “positive” elements in “negative” range

true negative (tn)

- “negative” elements in “negative” range



Motivation: What is the best cut?

true positive (tp)

- “positive” elements in “positive” range

false negative (fn)

- “positive” elements in “negative” range

true negative (tn)

- “negative” elements in “negative” range

false positive (fp)

- “negative” elements in “positive” range

Motivation

- What is the best cut?
  - Quality measure for two populations:

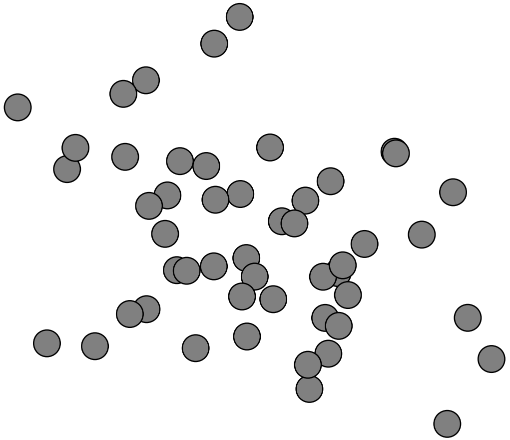
$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fn + fp}$$

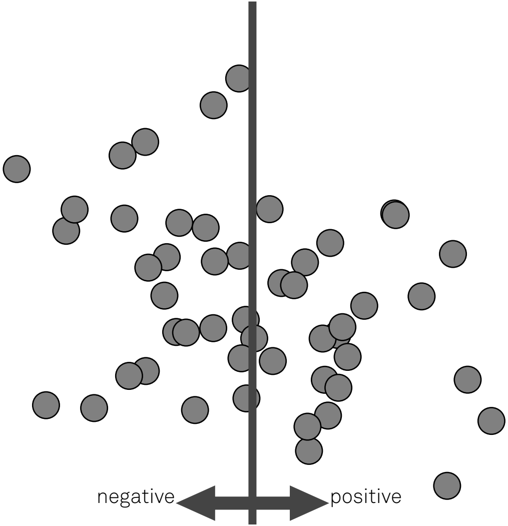
Motivation

- The goal is to divide the points into two populations
- Known element affiliation in Monte Carlo
- Idea: search for “best” one-dimensional cut in Monte Carlo



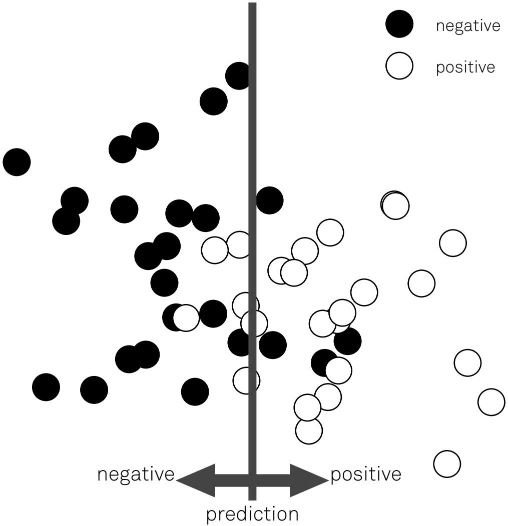
Motivation

- The goal is to divide the points into two populations
- Known element affiliation in Monte Carlo
- Idea: search for “best” $(n-1)$-dimensional cut in Monte Carlo



Motivation

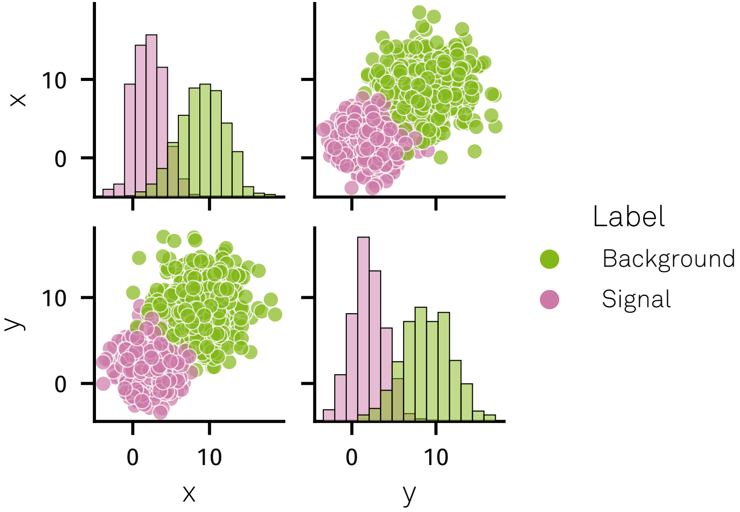
- The goal is to divide the points into two populations
- Known element affiliation in Monte Carlo
- Idea: search for “best” $(n-1)$-dimensional cut in Monte Carlo



Example

- Exercise: Separation of two populations
  - "Signal"
  - "Background"
- Elements of both populations are described via value pairs (x, y)
  - Background: Gaussian distribution with mean $(8, 8)$ and standard deviation $(2.5, 2.5)$
  - Signal: Gaussian distribution with mean $(2, 2)$ and standard deviation $(1.5, 1.5)$
- Search for "best" one-dimensional cut (separating hyperplane)

Example



Example

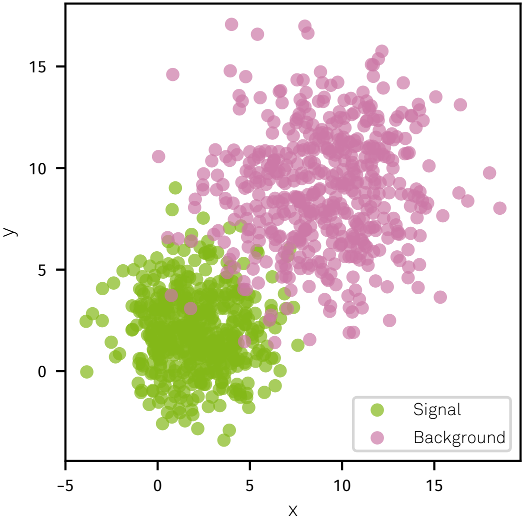
- Exercise: Separation of two populations
  - “Signal”
  - “Background”
- Elements of both populations are described via value pairs (x, y)
  - Background: Gaussian distribution with mean $(8, 8)$ and standard deviation $(2.5, 2.5)$
  - Signal: Gaussian distribution with mean $(2, 2)$ and standard deviation $(1.5, 1.5)$
- Search for “best” one-dimensional cut (separating hyperplane)
  - → Projection onto normal vector of the hyperplane must separate the classes “maximally”

Linear Fisher Discriminant Analysis

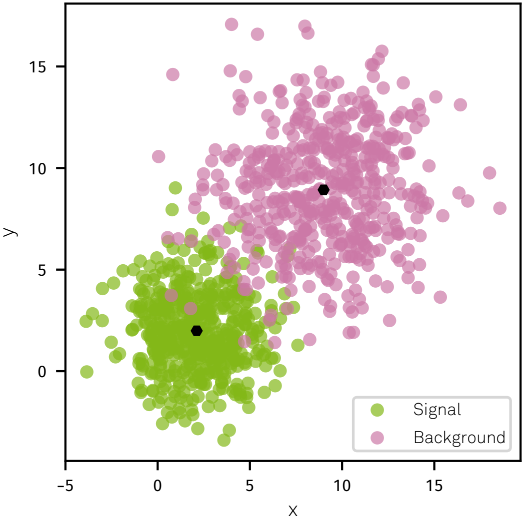
- To find a good projection $\vec{\lambda}(\vec{x}' = \vec{\lambda}^T\vec{x})$, a measure of separability must be defined
  - first (naive) idea:
    Distance of the mean values of the classes one the projection axis

$$D_{\text{naive}}(\vec{\lambda}) = |\vec{\mu}_1 - \vec{\mu}_2|$$

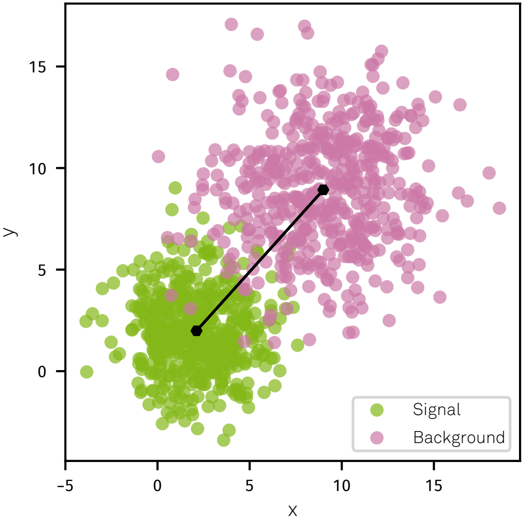
Example



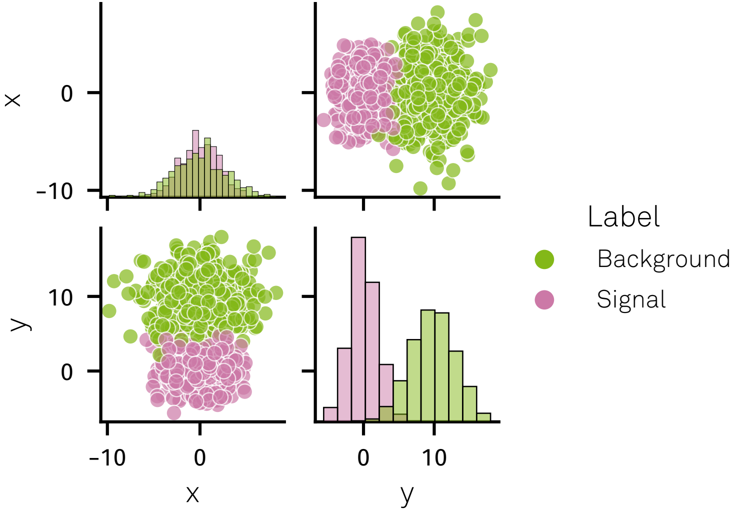
Example



Example



Example



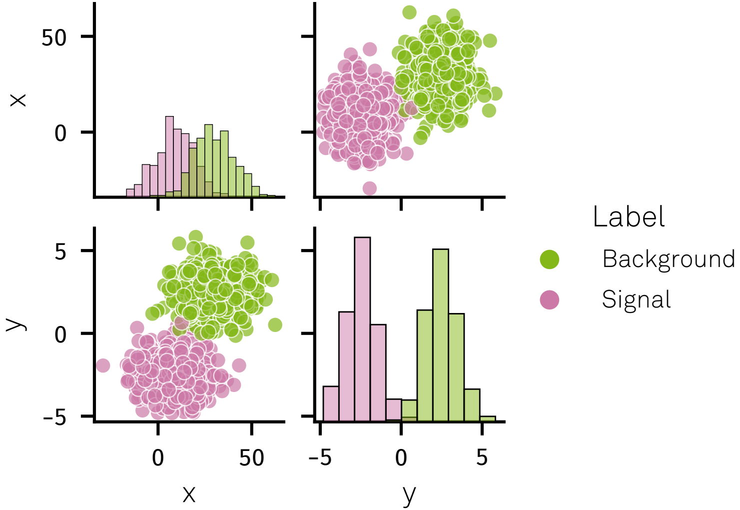
Linear Fisher Discriminant Analysis

- To find a good projection $\vec{\lambda}(\vec{x}' = \vec{\lambda}^T\vec{x})$, a measure of separability must be defined
  - first (naive) idea:
    Distance of the mean values of the classes one the projection axis

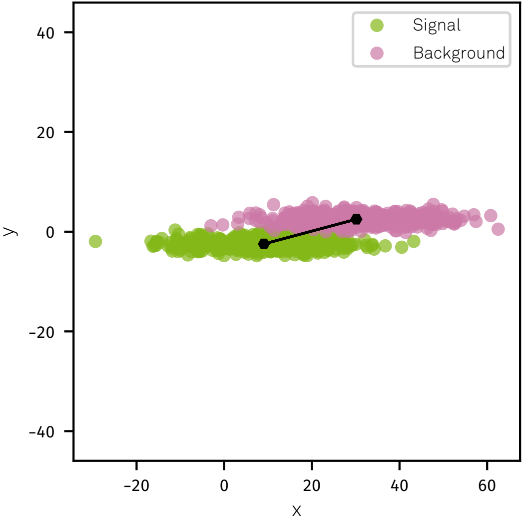
$$D_{\text{naive}}(\vec{\lambda}) = |\vec{\mu}_1 - \vec{\mu}_2|$$

Problem: Variance within the classes not taken into account!

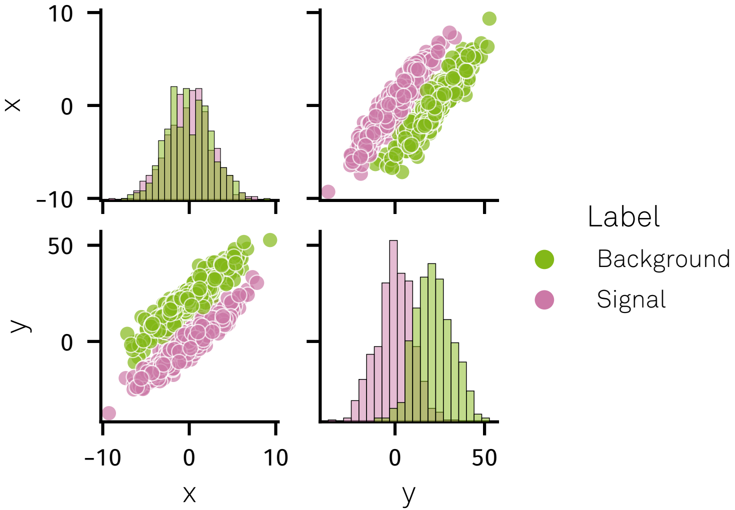
Example



Example



Example



Linear Fisher Discriminant Analysis

- To find a good projection $\vec{\lambda}(\vec{x}' = \vec{\lambda}^T\vec{x})$, a measure of separability must be defined
  - first (naive) idea:
    Distance of the mean values of the classes one the projection axis

$$D_{\text{naive}}(\vec{\lambda}) = |\vec{\mu}_1 - \vec{\mu}_2|$$

    Problem: Variance within the classes not taken into account!
  - Idea by Fisher:
    Square of the distance of the mean values of the classes on the projection axis normalized with the spread of the classes

$$D(\vec{\lambda}) = \frac{|\vec{\mu}'_1 - \vec{\mu}'_2|^2}{s'^2_1 + s'^2_2}$$

Linear Fisher Discriminant Analysis

- Optimal separation of two classes with n observables each by $(n-1)$-dimensional hyperplane
- The projection $\vec{\lambda}$ that maximizes $D(\vec{\lambda})$ is searched for
    1. Calculation of n-dimensional mean vectors

Linear Fisher Discriminant Analysis

1. Calculation of n-dimensional mean vectors
   - In general

$$\vec{\mu}_j = \begin{pmatrix} \bar{x}_{j,1} \\ \dots \\ \bar{x}_{j,\mathrm{n}} \end{pmatrix} = \frac{1}{N_j} \begin{pmatrix} \sum \bar{x}_{j,1,i} \\ \dots \\ \sum \bar{x}_{j,\mathrm{n},i} \end{pmatrix}$$

   - Example

$$\vec{\mu}_1 = \begin{pmatrix} \bar{x}_1 \\ \bar{y}_1 \end{pmatrix} = \frac{1}{N_1} \begin{pmatrix} \sum x_{1,i} \\ \sum y_{1,i} \end{pmatrix}$$

$$\vec{\mu}_2 = \begin{pmatrix} \bar{x}_2 \\ \bar{y}_2 \end{pmatrix} = \frac{1}{N_2} \begin{pmatrix} \sum x_{2,i} \\ \sum y_{2,i} \end{pmatrix}$$

Linear Fisher Discriminant Analysis

- Optimal separation of two classes with n observables each by $(n-1)$-dimensional hyperplane
- The projection $\vec{\lambda}$ that maximizes $D(\vec{\lambda})$ is searched for
    1. Calculation of n-dimensional mean vectors
    2. Calculation of scattering matrices

Linear Fisher Discriminant Analysis

2. Calculation of scattering matrices S_W and S_B
   - Scattering within classes (“within-class scatter matrix”)

     Total scattering: $S_W = \sum_j^{N_{\text{Klassen}}} S_j$

     Scattering of class j: $S_j = \sum_j^{n_j} (\vec{x}_i - \vec{\mu}_j)(\vec{x}_i - \vec{\mu}_j)^T$
   - Using this matrix $s_1'^2 + s_2'^2 = \vec{\lambda}^T S_W \vec{\lambda}$ since:

$$
\begin{aligned}
s_j'^2 &= \sum (\vec{x}' - \vec{\mu}')^2 = \sum (\vec{\lambda}^T \vec{x} - \vec{\lambda}^T \vec{\mu})^2 = \sum (\vec{\lambda}^T (\vec{x} - \vec{\mu}))^2 \\
&= \sum (\vec{\lambda}^T (\vec{x} - \vec{\mu}))(\vec{\lambda}^T (\vec{x} - \vec{\mu}))^T = \sum \vec{\lambda}^T (\vec{x} - \vec{\mu})(\vec{x} - \vec{\mu})^T \vec{\lambda} = \vec{\lambda}^T S_j \vec{\lambda}
\end{aligned}
$$

Linear Fisher Discriminant Analysis

2. Calculation of scattering matrices S_W and S_B
   - Scattering between classes (“between-class scatter matrix”)

$$S_B = (\vec{\mu}_1 - \vec{\mu}_2)(\vec{\mu}_1 - \vec{\mu}_2)^T$$

   - $|\vec{\mu}'_1 - \vec{\mu}'_2|^2 = \lambda^T S_B \lambda$ since:

$$\begin{aligned} |\vec{\mu}'_1 - \vec{\mu}'_2|^2 &= (\vec{\lambda}^T \vec{\mu}_1 - \vec{\lambda}^T \vec{\mu}_2)^2 \\ &= \vec{\lambda}^T (\vec{\mu}_1 - \vec{\mu}_2)(\vec{\mu}_1 - \vec{\mu}_2)^T \vec{\lambda} \\ &= \vec{\lambda}^T S_B \vec{\lambda} \end{aligned}$$

Linear Fisher Discriminant Analysis

2. Calculation of scattering matrices S_W and S_B
   - With the matrices S_W and S_B holds

$$D(\vec{\lambda}) = \frac{|\vec{\mu}_1' - \vec{\mu}_2'|^2}{s_1'^2 + s_2'^2} = \frac{\vec{\lambda}^T S_B \vec{\lambda}}{\vec{\lambda}^T S_W \vec{\lambda}}$$

   - This expression is to be maximized

$$\vec{\lambda}^* = \arg\max \left[\frac{\vec{\lambda}^T S_B \vec{\lambda}}{\vec{\lambda}^T S_W \vec{\lambda}} \right]$$

Linear Fisher Discriminant Analysis

- Optimal separation of two classes with n observables each by $(n - 1)$-dimensional hyperplane
- The projection $\vec{\lambda}$ that maximizes $D(\vec{\lambda})$ is searched for
  1. Calculation of n-dimensional mean vectors
  2. Calculation of scattering matrices
  3. Calculation of projection $\vec{\lambda}^*$

Linear Fisher Discriminant Analysis

3. Calculation of projection $\vec{\lambda}^*$ (part 1)
   - To show: $\vec{\lambda}^* = \arg\max \left[\frac{\vec{\lambda}^T S_B \vec{\lambda}}{\vec{\lambda}^T S_W \vec{\lambda}}\right] = S_W^{-1}(\mu_1 - \mu_2)$

     Differentiate $D(\vec{\lambda})$ and set equal to **0**:

$$\frac{\mathrm{d}}{\mathrm{d}\vec{\lambda}}\left[D(\vec{\lambda})\right] = \frac{\mathrm{d}}{\mathrm{d}\vec{\lambda}}\left[\frac{\vec{\lambda}^T S_B \vec{\lambda}}{\vec{\lambda}^T S_W \vec{\lambda}}\right] = 0$$

$$\Leftrightarrow \left[\vec{\lambda}^T S_W \vec{\lambda}\right]\frac{\mathrm{d}\vec{\lambda}^T S_B \vec{\lambda}}{\mathrm{d}\vec{\lambda}} - \left[\vec{\lambda}^T S_B \vec{\lambda}\right]\frac{\mathrm{d}\vec{\lambda}^T S_W \vec{\lambda}}{\mathrm{d}\vec{\lambda}} = 0$$

$$\Leftrightarrow \left[\vec{\lambda}^T S_W \vec{\lambda}\right] 2 S_B \vec{\lambda} - \left[\vec{\lambda}^T S_B \vec{\lambda}\right] 2 S_W \vec{\lambda} = 0$$

Linear Fisher Discriminant Analysis

3. Calculation of projection $\vec{\lambda}^*$ (part 2)

$$\Leftrightarrow \left[\vec{\lambda}^T S_W \vec{\lambda}\right] 2S_B\vec{\lambda} - \left[\vec{\lambda}^T S_B \vec{\lambda}\right] 2S_W\vec{\lambda} = 0$$

- Divide by $\vec{\lambda}^T S_W \vec{\lambda}$:

$$\left[\frac{\vec{\lambda}^T S_W \vec{\lambda}}{\vec{\lambda}^T S_W \vec{\lambda}}\right] S_B\vec{\lambda} - \left[\frac{\vec{\lambda}^T S_B \vec{\lambda}}{\vec{\lambda}^T S_W \vec{\lambda}}\right] S_W\vec{\lambda} = 0$$

$$\Leftrightarrow S_B\vec{\lambda} - DS_W\vec{\lambda} = 0$$

$$\Leftrightarrow S_W^{-1} S_B\vec{\lambda} = D\vec{\lambda}$$

- Solution of eigenvalue problem $S_W^{-1} S_B\vec{\lambda} = D\vec{\lambda}$

Linear Fisher Discriminant Analysis

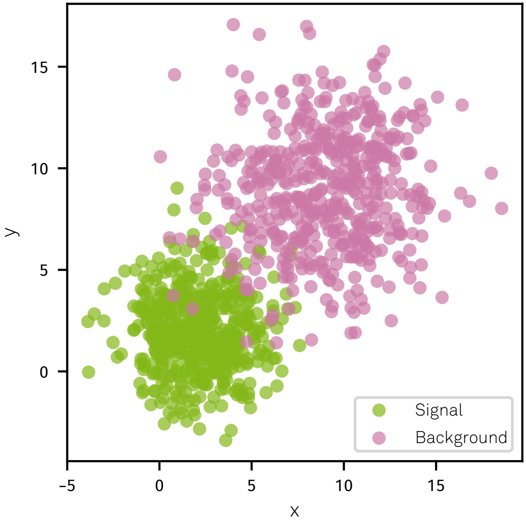
- Optimal separation of two classes with n observables each by $(n-1)$-dimensional hyperplane
- The projection $\vec{\lambda}$ that maximizes $D(\vec{\lambda})$ is searched for
    1. Calculation of n-dimensional mean vectors
    2. Calculation of scattering matrices
    3. Calculation of projection $\vec{\lambda}^*$
    4. Define cut onto projection axis

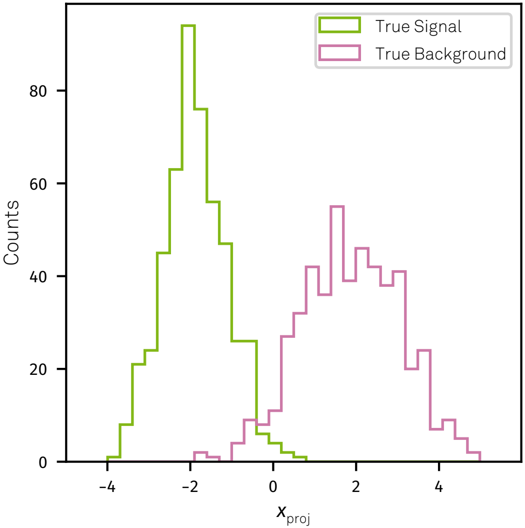
Linear Fisher Discriminant Analysis

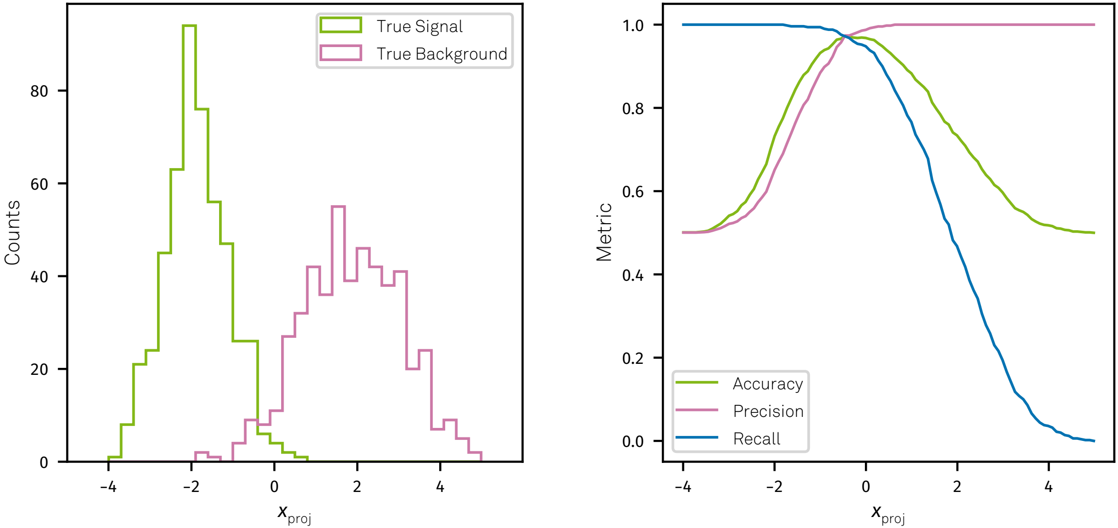
4. Define cut onto projection axis
   - Each n-dimensional point is projected into one dimension
   - A cut onto the projection axis is wanted which divides between both populations

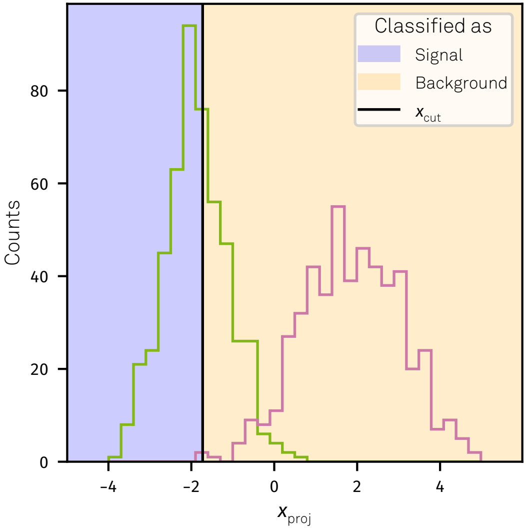
Linear Fisher Discriminant Analysis

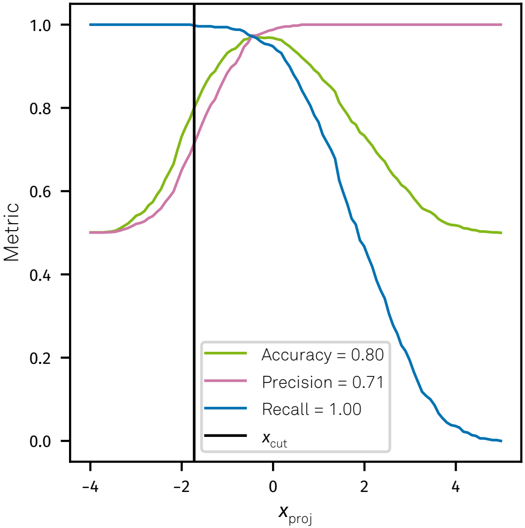
4. Define cut onto projection axis
   - Each n-dimensional point is projected into one dimension
   - A cut onto the projection axis is wanted which divides between both populations
   - No generalized best cut can be given
   - Must be motivated for each specific problem
     - Trade-off between recall and precision

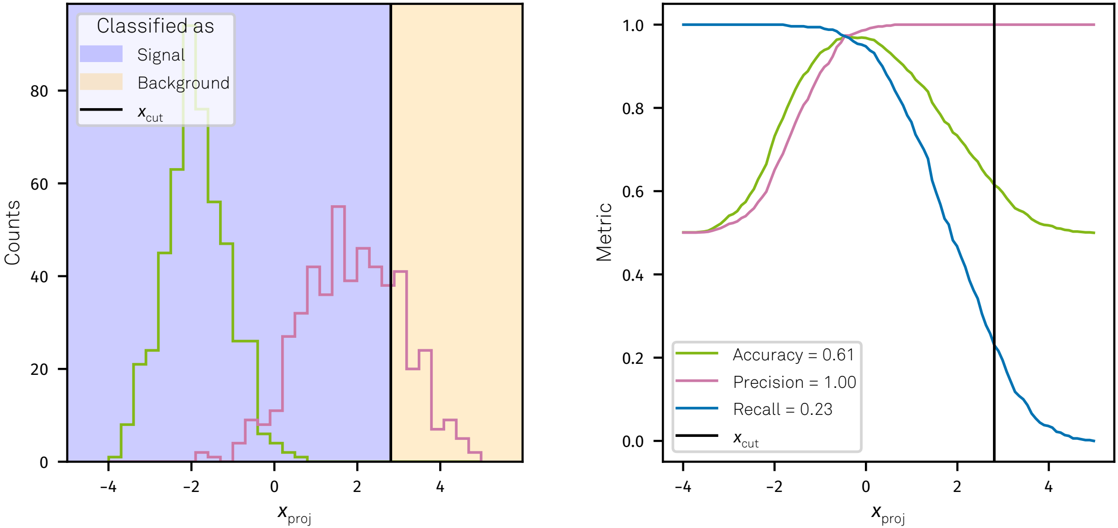

Signal
Background
y
x
15
10
5
0
-5
0
5
10
15


True Signal
True Background
Counts
x_{proj}
80
60
40
20
0
-4
-2
0
2
4


True Signal
True Background
Counts
80
60
40
20
0
−4
−2
0
2
4
x_{proj}
Accuracy
Precision
Recall
Metric
1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
2
4
x_{proj}


Classified as
Signal
Background
x_{cut}
Counts
80
60
40
20
0
-4
-2
0
2
4
x_{proj}


Metric
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy = 0.80
Precision = 0.71
Recall = 1.00
x_{cut}
-4
-2
0
2
4
x_{proj}


Classified as
Signal
Background
x_{cut}
Counts
x_{proj}
Accuracy = 0.61
Precision = 1.00
Recall = 0.23
x_{cut}
Metric
x_{proj}

Data Mining

Fayyad et al. 2002

"The capacity of digital data storage worldwide has doubled every nine months for at least a decade, at twice the rate predicted by Moore's Law for the growth of computing power during the same period."

Overview

Data Mining

Typical Exercises in Data Mining

Data Mining

- Was originally a step of so-called “Knowledge Discovery in Databases” processes — nowadays equal to KDD
- Data Mining often means application of machine learning algorithms
  - “[Machine learning is a] field of study that gives computers the ability to learn without being explicitly programmed.” (Arthus Smith, 1959)
  - “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.” (Tom M. Mitchell, 1997)
- This part of the lecture gives an **insight** into the large field of data mining

Data Mining Process According to Fayyad et al.

1. Goal definition of knowledge discovery
2. Provision of background knowledge for the respective area of expertise
3. Data selection
4. Data cleaning/preprocessing
5. Data reduction and transformation
6. Model selection
7. Data mining
8. Interpretation

$\Longrightarrow$ **KDD/Data mining processes are iterative and interactive**

$\Longrightarrow$ In practice, some steps are inseparable and the order can be slightly different

Data Mining Dictionary

- Feature: attribute, observable, measured variable, property
- Label: target quantity → labeled/unlabeled : target value known/unknown
- Classes: values of discrete target quantity (classes/labels are often used equivalently)
- Supervised learning: “Supervised learning is the machine learning task of inferring a function from labeled training data.” (Foundations of Machine Learning, 2012)
- Unsupervised learning: structure recognition independent of target quantities, optimization criteria, feedback signal or other information which goes beyond actual data
- Warning: meanings of the terms are rarely universally defined

Overview

Data Mining

Typical Exercises in Data Mining

Signal-to-Background Separation of Muon Neutrinos in IceCube

- All cascades are considered as background
- For tracks, a distinction must be made between atmospheric muons and neutrino-induced muons

Typical Excersises in Data Mining

- Measurement of muon neutrino spectrum wth IceCube needs a signal-to-background separation
  - ⇒ Discrete target quantity (class affiliation: signal/background)

Estimating the Age of Abalones

- “Tasmanian Aquaculture and Fisheries Institute” wants to avoid overfishing of abalones
- Overview of numbers and age of current stock are required
- Estimation of age
    1. Cut shell
    2. Polish shell
    3. Dye shell
    4. Count rings under a microscope
- Estimation is very costly and lengthy
- Fast method for age estimation on using external characteristics is needed

“Abalones (Haliotis) are a genus of of large snails [...].”
(Wikipedia)

Typical Excersises in Data Mining

- Measurement of muon neutrino spectrum wth IceCube needs a signal-to-background separation
  - ⇒ Discrete target quantity (class affiliation: signal/background)
- Protection of the abalone stock off the Tasmanian coast requires age estimation abalones
  - ⇒ Estimation of a continuous quantity

“Teekesselchen” - Game Involving Homonyms

- A task that is becoming increasingly important in “machine”-human communication is the understanding of the content of language
  (research area: natural language processing)
- Current approaches dispense with the explicit implementation of grammar and word meanings
- Algorithms learn language by processing many millions of texts
- One task is to find out whether a word is a homonym (a word with fundamentally different meanings depending on context)
- Example:
  - “Nocturnal flying mammals” — “An implement use to hit a ball” ⇒ bat

Typical Excersises in Data Mining

- Measurement of muon neutrino spectrum wth IceCube needs a signal-to-background separation
  - ⇒ Discrete target quantity (class affiliation: signal/background)
- Protection of the abalone stock off the Tasmanian coast requires age estimation abalones
  - ⇒ Estimation of a continuous quantity
- Understanding texts of a language requires knowledge of homonyms
  - ⇒ Search for object groups with similar properties (occurrence of word in similar contexts)

Data Selection

- Data must represent the question
- Example
  - Low-energy neutrinos in DeepCore are irrelevant for the measurement of high-energy muon neutrino spectrum
  - Measurement of bred abalone are (probably) not usable for age estimation abalones off the Tasmanian cost
  - Homonyms are not equal in all languages and dialects; differences between spoken and written language
  - ...
- Selection must be made and motivated context-dependent

Data Cleaning/Preprocessing

- Type and formatting of entries must be adapted to following operations:
  - Time/date formatting
  - Attribute types and scaling measures
    - Nominal: = / ≠
    - Ordinal: = / ≠ / < / >
    - Metrical: = / ≠ / < / > / + / ...
  - ...
- All entries must be unambiguous to process; i. e. gaps in the data, NaNs and infinite entries must be replaced logically
- If the data consists of multiple data sets, it must be ensured that all parts fit together and can be combined

Data Cleaning/Preprocessing

- Attributes that do not contain any or misunderstandable information shall are to be eliminated
  - IDs
  - Constant quantities
  - Attributes with too many missing entries
- In case of simulations, it must be ensured that the simulations do not contain information that is not present in the actual data.

Data Cleaning/Preprocessing

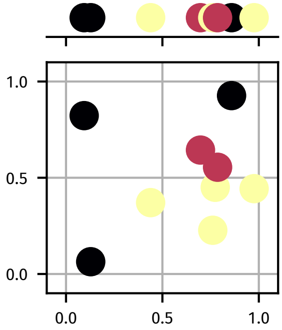
- Data cleaning using the example of signal-to-background separation in IceCube

Data Cleaning/Preprocessing

- Data cleaning using the example of signal-to-background separation in IceCube
- Files
  - signal.csv: simulation of muon neutrinos with energies from 10 GeV to 1 EeV from a solid angle of 4π
  - background.csv: simulation of air showers with primary energies from 600 GeV to 100 EeV
- All files are available for download in moodle

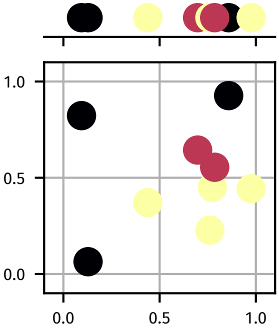
Curse of Dimensionality

- “Curse of Dimensionality” shaped by Bellmann 1961
- Number of measurements must increase exponentially
  - 1D: 3 bins →~ 3 measurements per bin
  - 2D: 3^2 bins →~ $\frac{1}{3}$ measurements per bin
  - 3D: 3^3 bins →~ $\frac{1}{9}$ measurements per bin



Curse of Dimensionality

- "Curse of Dimensionality" shaped by Bellmann 1961
- Number of measurements must increase exponentially
  - 1D: 3 bins →~ 3 measurements per bin
  - 2D: 3^2 bins →~ $\frac{1}{3}$ measurements per bin
  - 3D: 3^3 bins →~ $\frac{1}{9}$ measurements per bin
- Additional dimension yield additional information



Curse of Dimensionality

- Large amount of data is time and cost intensive in processing
- Redundant and for question irrelevant data parts should be eliminated soon

Data Reduction and Transformation

Goal: compact and meaningful presentation of data

Danger: useful information could be discarded or artifacts created in the data

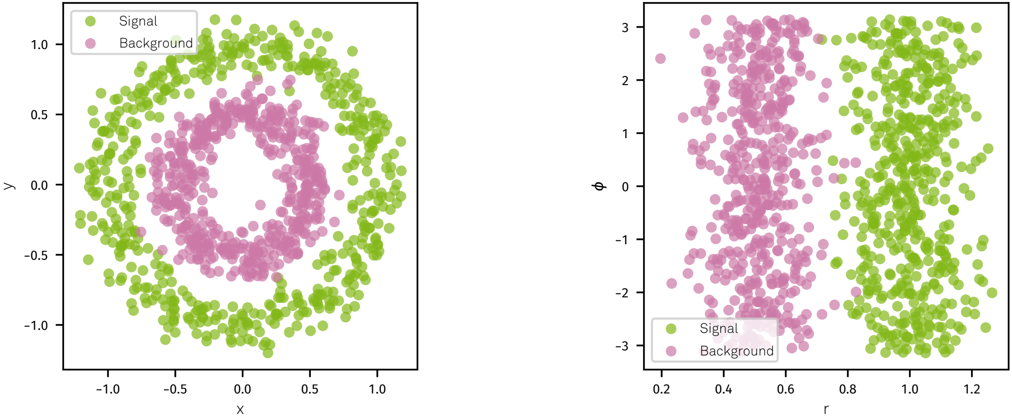
Two fundamental approaches:

- Feature extraction
  - Generation of new attributes which combine information of multiple attributes
    - Attributes by expert knowledge
      (e. g. exploitation of physical laws, boundary conditions, ...)
    - Attributes by combining existing
      (e. g. radius calculation of cylindrical detector from x and y coordinates)
  - Transformation of data space (Principal Component Analysis)
- Feature selection
  - Rejection of existing attributes

technische universität
dortmund

Feature Extraction

Simple example: Transformation from Cartesian into polar coordinates



Reject $\boldsymbol{\Phi}$ and keep $\boldsymbol{r} \implies$ Such transformations are only feasibly with expert knowledge and manual editing