COMPSCI762: Foundations of Machine Learning

Data Preprocessing

Katerina Taskova and Jörg Simon Wicker

The University of Auckland

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

SCIENCE
SCHOOL OF COMPUTER SCIENCE

Week 5-8

- In weeks 5-8, we will cover:
  - Data Preprocessing

$$\text{Data} \longrightarrow \boxed{\text{Preprocessing}} \longrightarrow \boxed{\text{Learning Algorithm}}$$

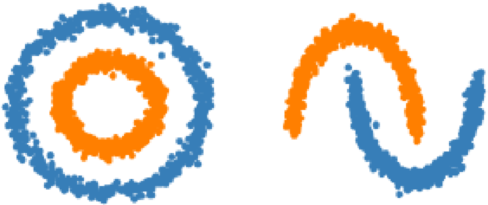
Week 5-8

- In weeks 5-8, we will cover:
  - Bayes Learning

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Week 5-8

- In weeks 5-8, we will cover:
  - Clustering



Week 5-8

- In weeks 5-8, we will cover:
  - Association Rules

If X buys *bread*, then X buys *milk* [support 50 %, confidence = 100 %]

Bread	Eggs	Milk	Oranges
1	1	1	0
0	0	1	0
1	0	1	0
0	1	0	1

—

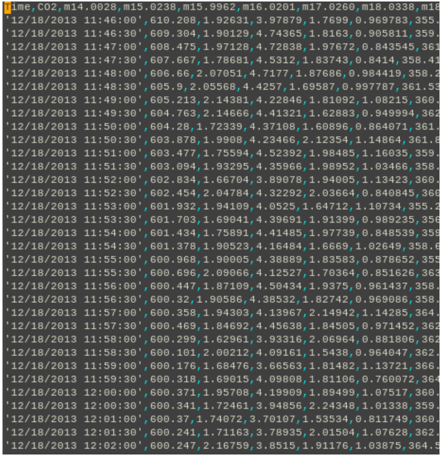
Data Preprocessing

This week we will cover

Data Preprocessing
- Data Cleaning
- Missing Data
- Preprocessing and Evaluation
- Data Reduction
- Noisy Data
- Data Transformation and Data Discretization
- Imbalanced Data

Why preprocess?

- Preprocessing means to transform the data before we feed it to a learning algorithm
- Why would we do that?
- What would we for example do?



```
Time,CO2,m14.0028,m15.0238,m15.9962,m16.0201,m17.0260,m18.0338,m18
'12/18/2013 11:46:00',610.208,1.92631,3.97879,1.7699,0.969783,355.
'12/18/2013 11:46:30',609.304,1.90129,4.74365,1.8163,0.905811,359.
'12/18/2013 11:47:00',608.475,1.97128,4.72838,1.97672,0.843545,361
'12/18/2013 11:47:30',607.667,1.78681,4.5312,1.83743,0.8414,358.41
'12/18/2013 11:48:00',606.66,2.07051,4.7177,1.87686,0.984419,358.2
'12/18/2013 11:48:30',605.9,2.05568,4.4257,1.69587,0.997787,361.53
'12/18/2013 11:49:00',605.213,2.14381,4.22846,1.81092,1.08215,360.
'12/18/2013 11:49:30',604.763,2.14666,4.41321,1.62883,0.949994,362
'12/18/2013 11:50:00',604.28,1.72339,4.37108,1.60896,0.864071,361.
'12/18/2013 11:50:30',603.878,1.9908,4.23466,2.12354,1.14864,361.8
'12/18/2013 11:51:00',603.477,1.75594,4.52392,1.98485,1.16035,359.
'12/18/2013 11:51:30',603.094,1.93295,4.35966,1.98952,1.03466,358.
'12/18/2013 11:52:00',602.834,1.66704,3.89078,1.94005,1.13423,360.
'12/18/2013 11:52:30',602.454,2.04784,4.32292,2.03664,0.840845,360
'12/18/2013 11:53:00',601.932,1.94109,4.0525,1.64712,1.10734,355.2
'12/18/2013 11:53:30',601.703,1.69041,4.39691,1.91399,0.989235,356
'12/18/2013 11:54:00',601.434,1.75891,4.41485,1.97739,0.848539,359
'12/18/2013 11:54:30',601.378,1.90523,4.16484,1.6669,1.02649,358.6
'12/18/2013 11:55:00',600.968,1.90005,4.38889,1.83583,0.878652,355
'12/18/2013 11:55:30',600.696,2.09066,4.12527,1.70364,0.851626,363
'12/18/2013 11:56:00',600.447,1.87109,4.50434,1.9375,0.961437,358.
'12/18/2013 11:56:30',600.32,1.90586,4.38532,1.82742,0.969086,358.
'12/18/2013 11:57:00',600.358,1.94303,4.13967,2.14942,1.14285,364.
'12/18/2013 11:57:30',600.469,1.84692,4.45638,1.84505,0.971452,362
'12/18/2013 11:58:00',600.299,1.62961,3.93316,2.06964,0.881806,362
'12/18/2013 11:58:30',600.101,2.00212,4.09161,1.5438,0.964047,362.
'12/18/2013 11:59:00',600.176,1.68476,3.66563,1.81482,1.13721,366.
'12/18/2013 11:59:30',600.318,1.69015,4.09808,1.81106,0.760072,364
'12/18/2013 12:00:00',600.371,1.95708,4.19909,1.89499,1.07517,360.
'12/18/2013 12:00:30',600.341,1.72461,3.94856,2.24348,1.01338,359.
'12/18/2013 12:01:00',600.37,1.74072,3.70107,1.53534,0.811749,360.
'12/18/2013 12:01:30',600.241,1.71163,3.78935,2.01504,1.07628,362.
'12/18/2013 12:02:00',600.247,2.16759,3.8515,1.91176,1.03875,364.5
```

This week we will...

- Talk about problems that can appear in data
- Introduce strategies to solve these problems
- Talk about feature selection, a very important technique in machine learning

Major Tasks in Data Preprocessing

- Data cleaning
  - Missing values
  - Noisy data
  - Outliers
- Data reduction
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- Transformation and discretization
  - Normalization
  - Hierarchy generation

Data Cleaning

- Basic assumption in machine learning?

- But, real-world data are, in most cases, dirty
- This can lead to problems, if data are

  Incomplete lacking attribute values, certain attributes, or containing only aggregate data

  Noisy containing noise, errors, or outliers

  Inconsistent containing discrepancies in codes or names

  Intentially wrong for example, there are a lot of pictures with a GPS location just a bit west of Africa

Incomplete (Missing) Data

- Data are not always available
  - Many tuples have no recorded value for several attributes
  - E.g. customer income in sales data
- Missing data may be due to
  - Equipment malfunction
  - Inconsistent with other recorded data and thus deleted
  - Data not entered due to misunderstanding
  - Certain data may not be considered important at the time of entry
  - Data history or changes of the data not recorded
- Missing data may need to be inferred
  - When, for example?

What to Consider When Handling Missing Data?

- Missing completely at random (MCAR)
  - Completely unrelated to the data

Name	Country	Income
Jane	NZ	$50k
Kate	NZ	$75k
Tom	US	$53k
George	UK	$64k
Mark	UK	$77k
Philippe	US	$80k

MCAR →

Name	Country	Income
Jane	NZ	
	NZ	$75k
Tom	US	
George		$64k
	UK	$77k
Philippe	US	$80k

—

  - Potential problem? Small sample size

What to Consider When Handling Missing Data?

- Missing at random (MAR)
  - The fact the data are missing is related not to the missing attribute, but to some other data in the data set

Name	Country	Income
Jane	NZ	$50k
Kate	NZ	$75k
Tom	US	$53k
George	UK	$64k
Mark	UK	$77k
Philippe	US	$80k

MAR →

Name	Country	Income
Jane	NZ	$50k
Kate	NZ	$75k
Tom	US	$53k
George	UK	
Mark	UK	
Philippe	US	$80k

—

  - Potential problem? Bias due to row-wise deletion

What to Consider When Handling Missing Data?

- Missing not at random (MNAR)
  - There is a reason the data are missing and it is related to the attribute itself

Name	Country	Income
Jane	NZ	$50k
Kate	NZ	$75k
Tom	US	$53k
George	UK	$64k
Mark	UK	$77k
Philippe	US	$80k

MNAR

Name	Country	Income
Jane	NZ	
Kate	NZ	$75k
Tom	US	
George	UK	
Mark	UK	$77k
Philippe	US	$80k

  - Potential problem? Bias due to row-wise deletion

How to Handle Missing Data – Imputation

- Ignore the tuple

$$
\overset{X}{\boxed{\begin{matrix} 0 & 1 & 1 & 1 & \ldots \\ ? & ? & ? & 1 & \ldots \\ 1 & 0 & ? & ? & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & 1 & 0 & \ldots \end{matrix}}}
\rightarrow
\overset{X'}{\boxed{\begin{matrix} 0 & 1 & 1 & 1 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & 1 & 0 & \ldots \end{matrix}}}
$$

  - Usually done when the class label is missing (classification)
  - Not effective when the fraction of missing values varies considerably

How to Handle Missing Data – Imputation

- Fill in the missing data manually

$$X = \begin{bmatrix} 0 & 1 & 1 & 1 & \dots \\ ? & ? & ? & 1 & \dots \\ 1 & 0 & ? & ? & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 1 & 0 & \dots \end{bmatrix} \rightarrow X' = \begin{bmatrix} 0 & 1 & 1 & 1 & \dots \\ 1 & 0 & 0 & 1 & \dots \\ 1 & 0 & 1 & 1 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 1 & 0 & \dots \end{bmatrix}$$

  - Tedious and sometimes infeasable

How to Handle Missing Data – Imputation

- Fill in automatically
  - A global constant

$$
X = \begin{bmatrix} \text{sunny} & \text{warm} & \text{Mon} & \text{May} & \ldots \\ \text{cloudy} & ? & ? & \text{July} & \ldots \\ \text{sunny} & \text{cold} & ? & ? & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \text{overcast} & \text{cold} & \text{Sat} & \text{June} & \ldots \end{bmatrix} \rightarrow X' = \begin{bmatrix} \text{sunny} & \text{warm} & \text{Mon} & \text{May} & \ldots \\ \text{cloudy} & \text{missing} & \text{missing} & \text{July} & \ldots \\ \text{sunny} & \text{cold} & \text{missing} & \text{missing} & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \text{overcast} & \text{cold} & \text{Sat} & \text{June} & \ldots \end{bmatrix}
$$

    - E.g. “missing”
    - A new class

How to Handle Missing Data – Imputation

- Fill in automatically
  - The attribute mean

$$X = \begin{bmatrix} 12 & 2 & 22 & 38 & \dots \\ 11 & ? & ? & 90 & \dots \\ 2 & 23 & ? & ? & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 9 & 11 & 54 & 23 & \dots \end{bmatrix} \rightarrow X' = \begin{bmatrix} 12 & 2 & 22 & 37 & \dots \\ 11 & 12 & 38 & 90 & \dots \\ 2 & 23 & 38 & 30 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 9 & 11 & 54 & 23 & \dots \end{bmatrix}$$

    - Done automatically by many implementations
    - Changes relationship with other variables $\Rightarrow$ bias in data

How to Handle Missing Data – Imputation

- Fill in automatically
  - The attribute mean of the samples belonging to the same class

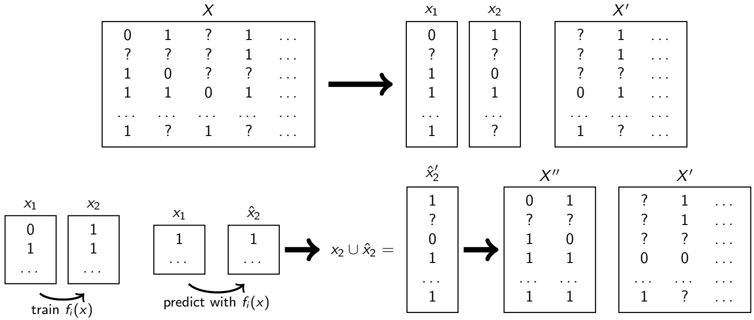
$$X|Y \qquad\qquad X'|Y$$

$$\left[\begin{array}{ccccc|c} 12 & 2 & 22 & 38 & \ldots & 1 \\ 11 & ? & ? & 90 & \ldots & 0 \\ 2 & 23 & ? & ? & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 9 & 11 & 54 & 23 & \ldots & 0 \end{array}\right] \rightarrow \left[\begin{array}{ccccc|c} 12 & 2 & 22 & 38 & \ldots & 1 \\ 11 & \color{red}{11} & \color{red}{54} & 90 & \ldots & 0 \\ 2 & 23 & \color{red}{22} & \color{red}{38} & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 9 & 11 & 54 & 23 & \ldots & 0 \end{array}\right]$$

    - Might change relationship with other variables other than class $\Rightarrow$ bias in data

How to Handle Missing Data – Imputation

- Fill in automatically
  - The most probable value



  - Inference-based such as Bayesian formula, decision tree, nearest neighbour,...

More on Imputation

- Matrix decomposition approaches
  - Decompose matrix using, e.g, Singular Value Decomposition
    - Decompose the data matrix X such that $X = U\Lambda V^T$
    - Create imputed matrix X' by multiplying $U \times \Lambda \times V^T$

$$\begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{bmatrix} \approx \begin{bmatrix} u_{11} & \cdots & u_{1k} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nk} \end{bmatrix} \begin{bmatrix} \lambda_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{nk} \end{bmatrix} \begin{bmatrix} v_{11} & \cdots & v_{1d} \\ \vdots & \ddots & \vdots \\ v_{k1} & \cdots & v_{kd} \end{bmatrix}$$

    - Minimize the sum of squared errors

$$\min_{U,\Lambda,V} \sum_{x_i j \in X} (x_{ij} - [U\Lambda V]_{ij})$$

Even More on Imputation

- EM imputation
  - Expectation Maximization
  - Use other variables to impute the values (Expectation)
  - Check if value is most probable (Maximization)
- Multiple imputation (e.g. MICE)
  1. Impute missing values using appropriate model (for example using classifier / regression model to predict the missing value)
  2. Repeat the step multiple times (3-5)
  3. Carry out required full analysis of data (e.g. build classifier and evaluate)
  4. Average the results (predictions or evaluation)
- So what is the best approach?

Preprocessing and Evaluation

- So now we know a preprocessing example
- Where would you put the preprocessing step in the evaluation?
- For example, for imputation:
    - Impute the values before splitting in train and test?
    - Impute the values in the training set – then how about the test set?

Conclusion

- Preprocessing is an important part in machine learning and data analysis
- Missing values can be caused by various reasons depending on what the reasons are, they must be addressed differently
- Various imputation approaches exist, they use the information of other instances and values to impute the missing values

Literature

- Material in Chapter 3 in Han's *Data Mining*

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

SCIENCE
SCHOOL OF COMPUTER SCIENCE

Thank you for your attention!

`https://ml.acukland.ac.nz`