Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Data Management for Digital Health
Revision of Exercise III

Borchert, Rasheed, Dr. Bayat, Dr. Schapranow

Data Management for Digital Health
Winter 2022

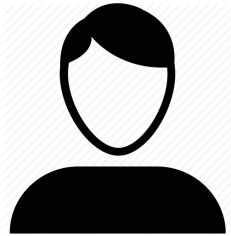
Exercise III
Topics

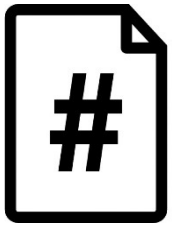
- Medical Use Case Oncology
- Bio Recap
- Genome Data Acquisition and Processing
- NLP & Text Data
- Text Data and Machine Learning

Exercise III
Key Stats

Hasso Plattner Institut

25 Questions 50 Points	42 Submissions 42 Passed	Average score 43.5 / 87%	Average time 67 min





<< 3h

Q11: What are examples of feature selection algorithms, as discussed in class?

Hasso Plattner Institut

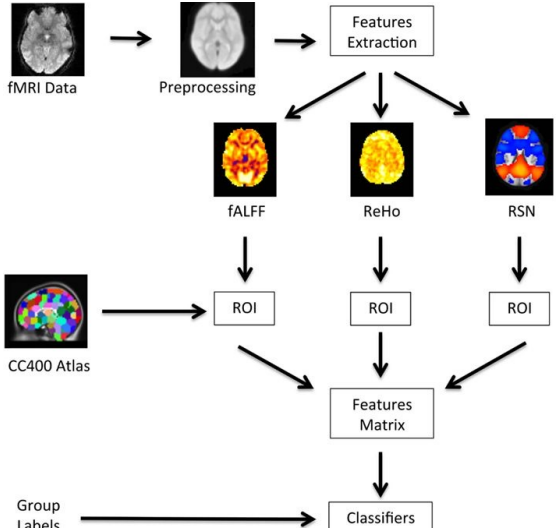
- ✗ Deep feature synthesis
- ✗ Word embeddings
- ✓ Boruta
- ✗ Graph neural networks

Frequently missed

Frequent incorrect answer

3. Data Preparation
Feature Extraction

- Aims to reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features)
- New reduced set of features should then be able to summarize most of the information contained in the original set
- Create some interaction (e.g., multiply or divide) between each pair of variables → lengthy process
- Deep feature synthesis (DFS) is an algorithm which enables you to quickly create new variables with varying depth



https://matlab1.com/feature-extraction-image-processing/

3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

Evaluation Exercise III

Data Management for Digital Health, Winter 2022

Q25: Which models among the followings are designed to account for censored data in the model?

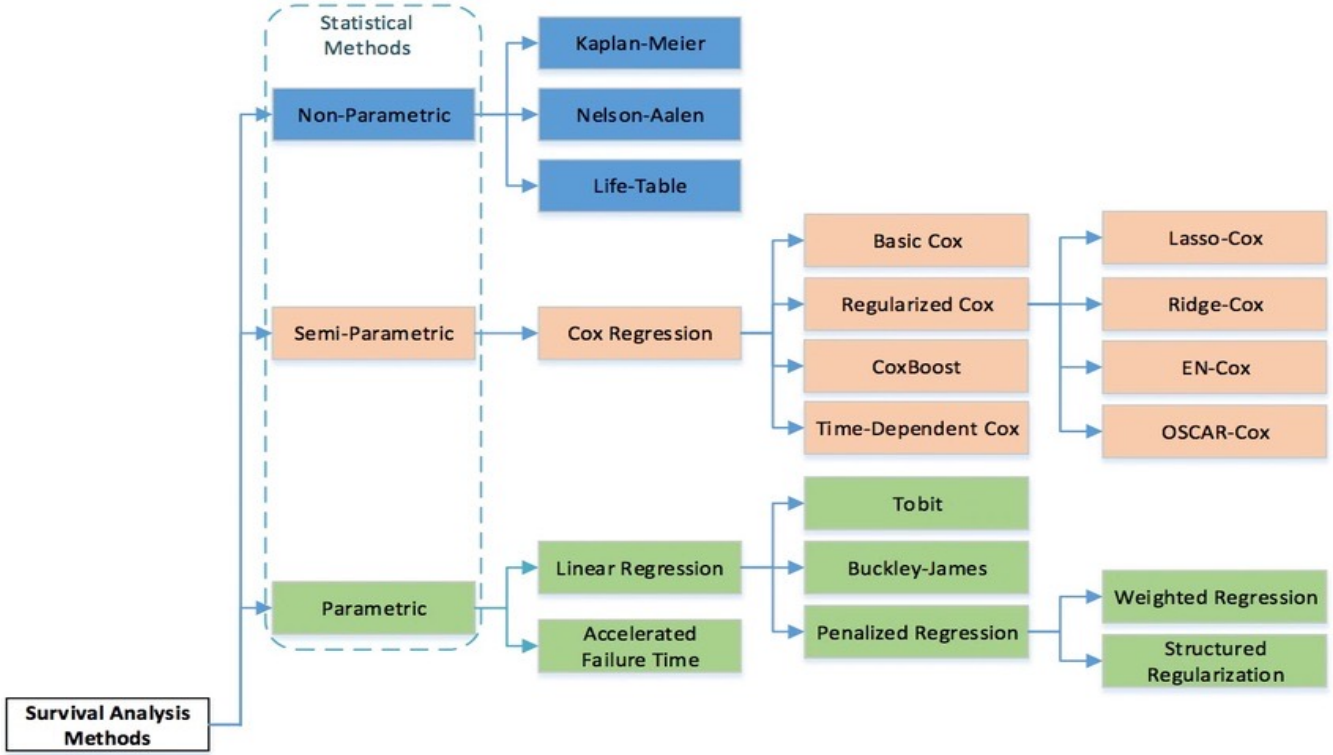
Hasso Plattner Institut

- ✓ Cox Regression
- ✗ Support Vector Machine
- ✗ Linear regression
- ✓ Accelerated Failure Time Model

Frequently missed

Frequent incorrect answer

Methods to work with censored data



https://humboldt-wi.github.io/blog/research/information_systems_1920/group2_survivalanalysis/

Q14: Please select all correct statements about feature selection.

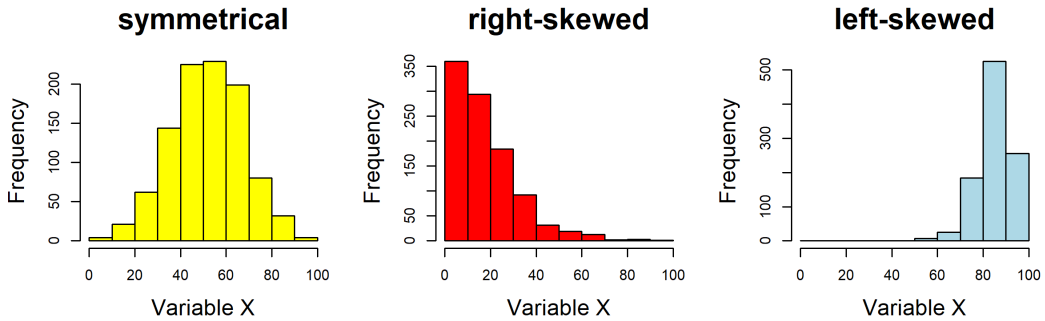
- ✗ Feature selection always decreases model performance.
- ✓ Correlation is used to identify the relationships between two continuous variables.
- ✗ Autocorrelation is used to identify the relationship between two continuous variables.
- ✗ Box-Cox is a feature selection method.

Frequently missed

Frequent incorrect answer

3. Data Preparation: Transformed Attributes Box-Cox

- Box-Cox Remove skewness and to normalize the data
- Log transformation → for strongly right-skewed data
- Sqrt transformation → for slightly right-skewed data
- Power transformation →for left-skewed data



https://www.davidzeleny.net/anadat-r/doku.php/en:data_preparation

Q19: Which of the following are ensemble algorithms?

- Random Forest

- XGBoost

- Logistic Regression

- Support Vector Machines

Frequently missed

Frequent incorrect answer

Simple Binary Classification: Logistic Regression

	x_1	x_2	y
	Age	Systolic Blood Pressure	Coronary Heart Disease
$x^{(1)}$	17	118	0
$x^{(2)}$	46	117	0
$x^{(3)}$	53	146	1
$x^{(4)}$	62	158	1
$x^{(5)}$	20	106	1
.	...	...	...
	20	124	0
.	48	144	1
.	42	154	0
.	51	124	1
$x^{(n)}$	58	214	0

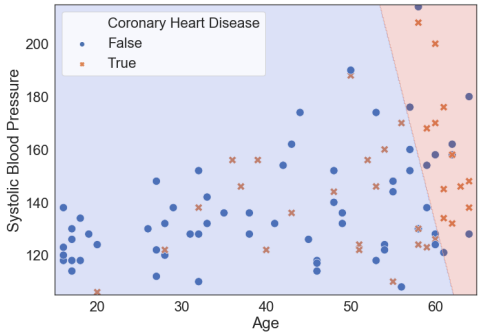
Model $P\left(y^{(i)} = 1 \,\middle|\, \boldsymbol{x_i}\right) = \sigma(\boldsymbol{w^T x^{(i)}} + b)$

$= \sigma(w_1\ x_1^{(i)} + w_2\ x_2^{(i)} + b)$

Features $\boldsymbol{x_i} = (x_i^{(1)}, x_i^{(2)})$

Weights $\boldsymbol{w} = (w_1\ , w_2\ )$, bias b

Logistic function $\sigma(z) = \frac{1}{1+e^{-z}}$



→ Result of running logistic regression:

$w_1 = 0.057 \quad w_2 = 0.0044 \quad b = -3.857$

$P(y = 1 \mid \boldsymbol{x^{(1)}}) = \sigma(0.057 \cdot 17 + 0.0044 \cdot 118 - 3.857) = \mathbf{0.08}$

$P(y = 1 \mid \boldsymbol{x^{(4)}}) = \sigma(0.057 \cdot 62 + 0.0044 \cdot 158 - 3.857) = \mathbf{0.56}$

Hasso Plattner Institut

4 Predictive Modeling

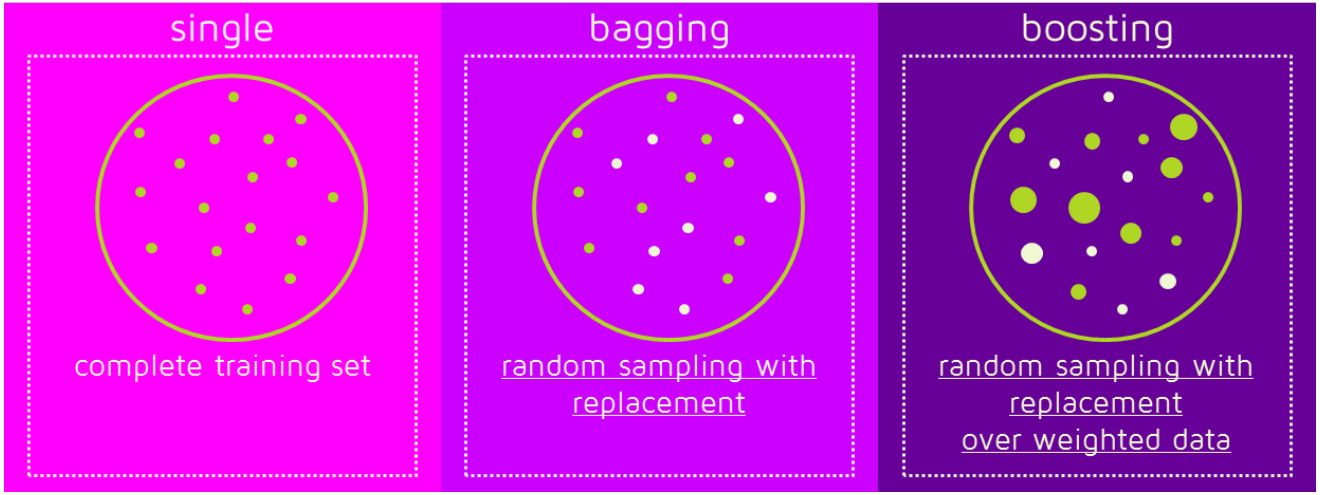
- Algorithm selection
- Feature selection
- Model training
- Hyperparameter tuning
- Model selection

Evaluation Exercise III

Data Management for Digital Health, Winter 2022

Ensemble methods: Bagging vs. Boosting

- In Bagging each element has the same probability of appearing in the training data set.
- In Boosting, the observations are weighted and therefore, some of them will take part in the new sets more often



https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/

Linear Models and Ensembel

- Logistic regression and SVM are linear Models
- Ensembel methods are non linear methods combining many models.
- Often Trees. Gradient boositng and Bagging
- Examples: Random Forest(Bagging) and Xgboost (Boosting)

Q18: Which of the following statement is true about Matthiews Correlation Coeffecient MCC.

- ✗ MCC and Pearson's correlation coefficients are the same metrics.
- ✓ MCC values lie between [-1,1]
- ✓ MCC is a symmetric metric for the evaluation of supervised machine learning.
- ✗ Accuracy is the preferred metric for evaluating a supervised machine learning problem with imbalanced classes compared to MCC.

Frequently missed

Frequent incorrect answer

Evaluation in Imbalanced class: F1 and MCC

Hasso Plattner Institut

- F1 score takes precision and recall into account.
- The disadvantages of F1 score:
  - It is not normalized
  - It is not symmetric (when swapping positive and negative classes)
- Matthew's Correlation Coefficient (MCC) is normalized and symmetric.

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

- Similarly interpretable as Pearson's correlation coefficient [-1,1]
- 1: perfect prediction. 0 : random prediction. -1: negative prediction