ENRICHMENT FOR ALL

Digital Adventure Ride to the Future

7 – 18 January, 2024

Utrecht University

1

ENRICHMENT FOR ALL

Regression and Demand Forecasting

Hakim Qahtan

Department of Information and Computing Sciences
Utrecht University

Utrecht University

2

Today

- Classification
- Model and Data Bias
- Fairness

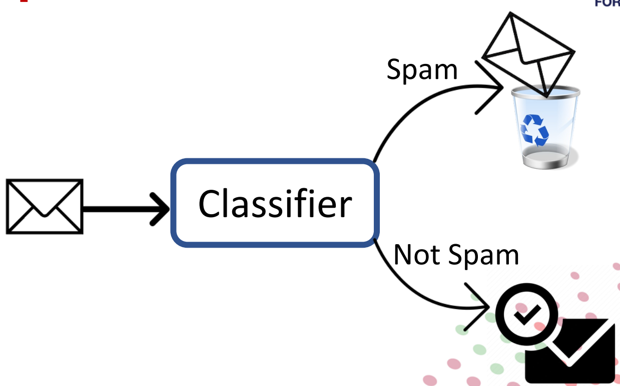
Classification

Topics

- Classification
  - Train-Test Split
  - Classification Algorithms
  - Evaluation
- Evaluation

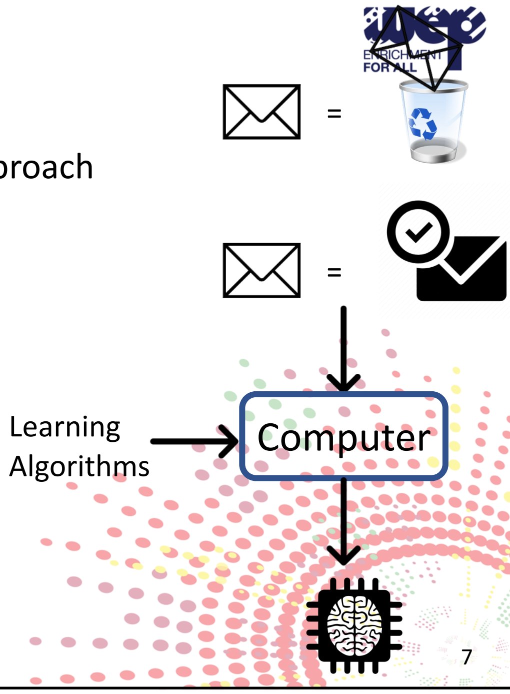
Real World Examples

- Credit/loan approval
- Medical diagnosis
- Fraud detection
- Web page categorization
- Spam detection in emails
- Image classification
- Text sentiment



How Classifiers Work?

- Classification is a supervised learning approach
  - Labeled dataset is needed
  - Features/Examples: X
  - Outcomes/Labels: Y
  - Learn the patterns
  - Use them to predict the labels



Utrecht University

7

7

Classification techniques

- Include:
  - Logistic Regression
  - Learning from the neighbors
  - Decision tree-based methods
  - Support vector machines
  - Neural networks
  - Bayesian classification

Utrecht University

8

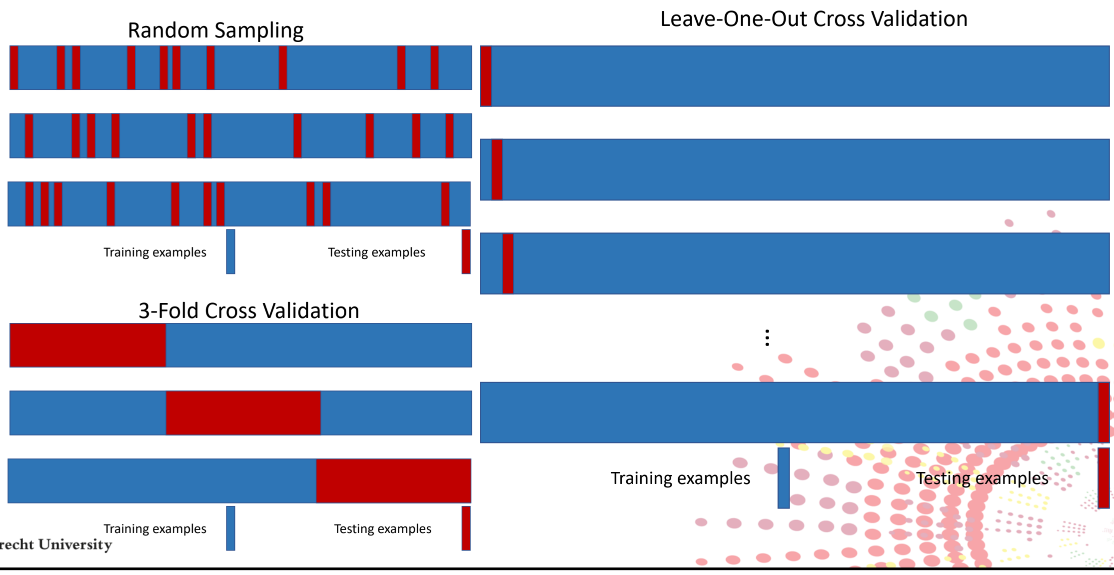
8

Building a Classification Model

- Split the data into training and testing
- Train the classifier on the training data
- Select a set of performance measures
  - Accuracy, Balanced Accuracy, Precision, Recall, F-Score, …
- Evaluate the classifier on the test set

NOTE: NEVER USE THE TEST SET FOR TRAINING!

Train-Test Split



Train-Test Split

- Problem
  - The training/testing set may not include examples from one of the classes
- Solution
  - Split with stratification
    - Ensures that each class is represented with approximately equal proportions in both subsets

Classification – A Three-Step Process

- Model construction
  - describing a set of predetermined classes
- Model evaluation
  - testing if the model will perform well on unseen data
  - labels are available but not provided to the model (classifier)
- Model usage
  - for classifying future or unknown objects

Model Construction

Training Data

Classification Algorithms

Classifier (Model)

Name	Profession	Years	Buys Computer
Mike	Student	1	No
Mary	Professor	7	Yes
Bill	Professor	2	Yes
Jim	Student	7	Yes
Anne	Student	2	No
Dave	Student	3	Yes

IF ‘profession’ = ‘professor’
OR ‘years’ > 2
THEN ‘Buys Computer’ = ‘yes’

Utrecht University 13

13

Step (2): Evaluating the Model

Testing Data

Classifier (Model)

Name	Rank	Years
Tom	Student	2
Joane	Professor	7
Greg	Professor	1
Sally	Student	4

prediction

Name	Rank	Years	Buys Computer
Tom	Student	2	No
Joane	Professor	7	Yes
Greg	Professor	1	Yes
Sally	Student	4	Yes

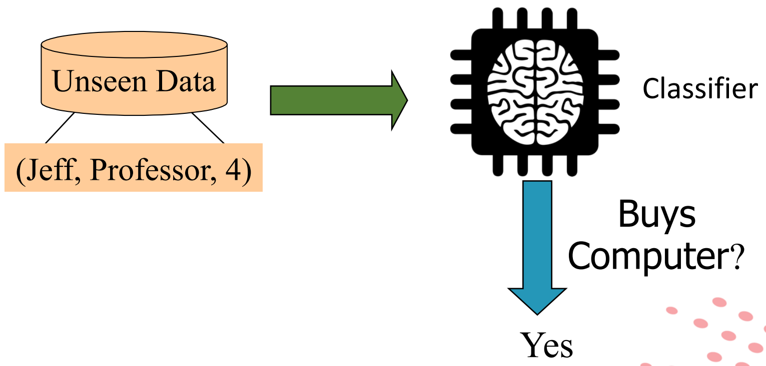
Actual Labels

Buys Computer
No
No
Yes
Yes

Utrecht University 14

14

Step (3): Using the Model in Prediction



Utrecht University

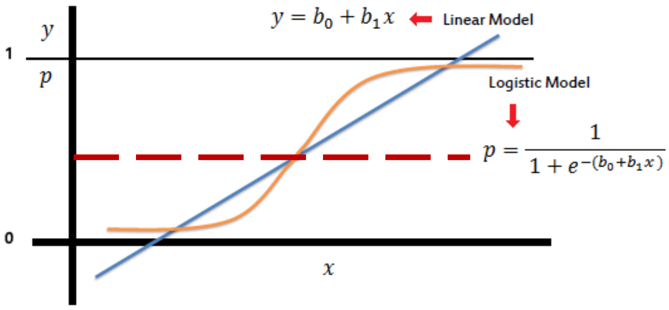
15

Classification Techniques

Utrecht University

16

Logistic Regression



A sigmoid function that assumes values in the range [0,1]

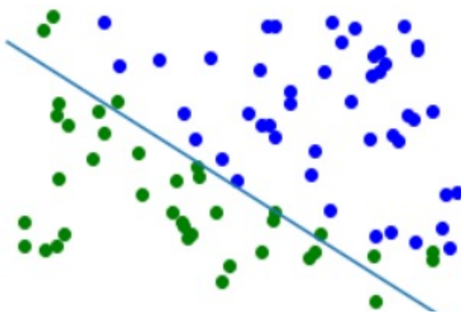
$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

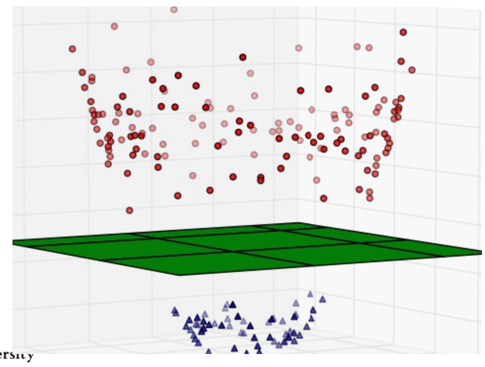
Utrecht University

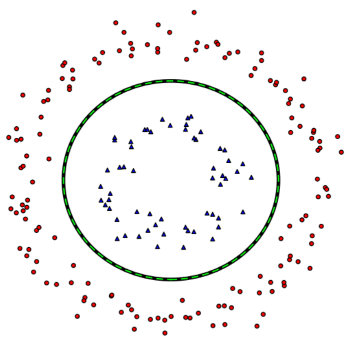
17

17

Decision Boundaries







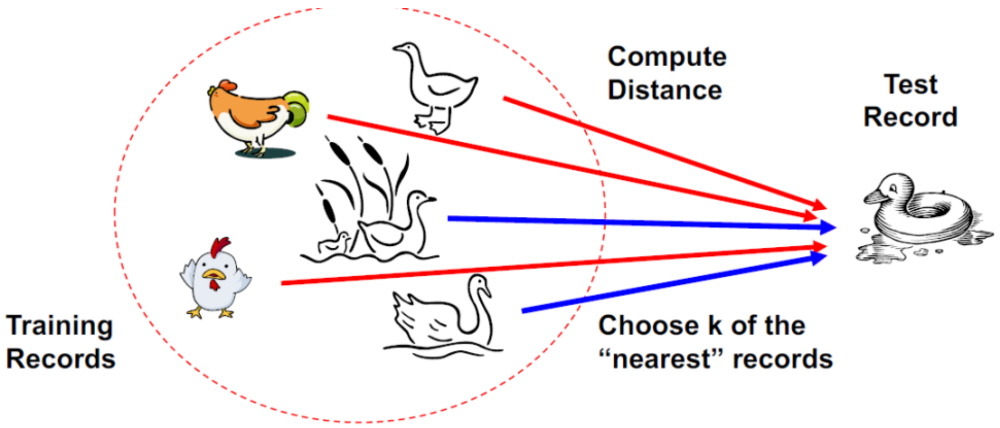
Utrecht University

18

18

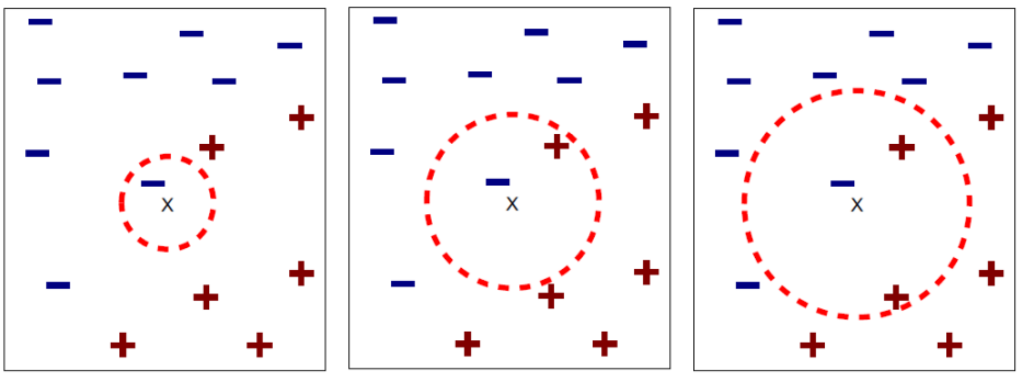
Nearest Neighbors Classifiers

Basic idea: if it walks like a duck, quacks like a duck, then it is probably a duck



19

Nearest Neighbors Classifiers (Cont.)

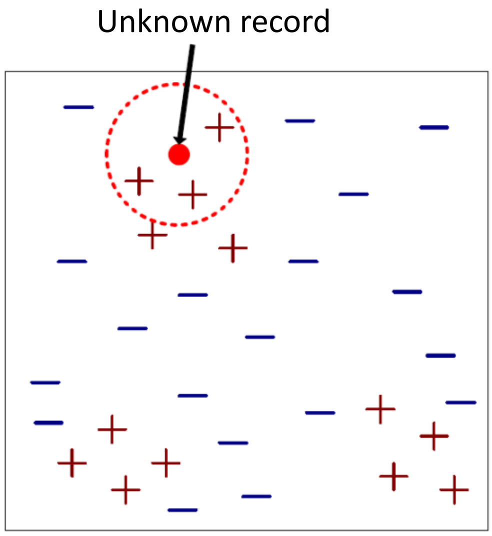


1- nearest neighbor   2- nearest neighbors   3- nearest neighbors

20

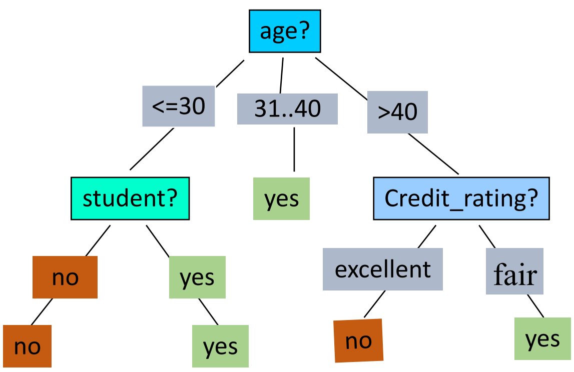
Nearest Neighbors Classifiers (Cont.)

- Three requirements:
  - Set of records (training set)
  - Distance metric
  - The number of neighbors to be considered *k*
- Classifying unknown record x:
  - Compute the distance from x to the other training records
  - Identify the k-Nearest Neighbors (kNN(x))
  - Use class labels of the kNN records to determine the class of x .. How?



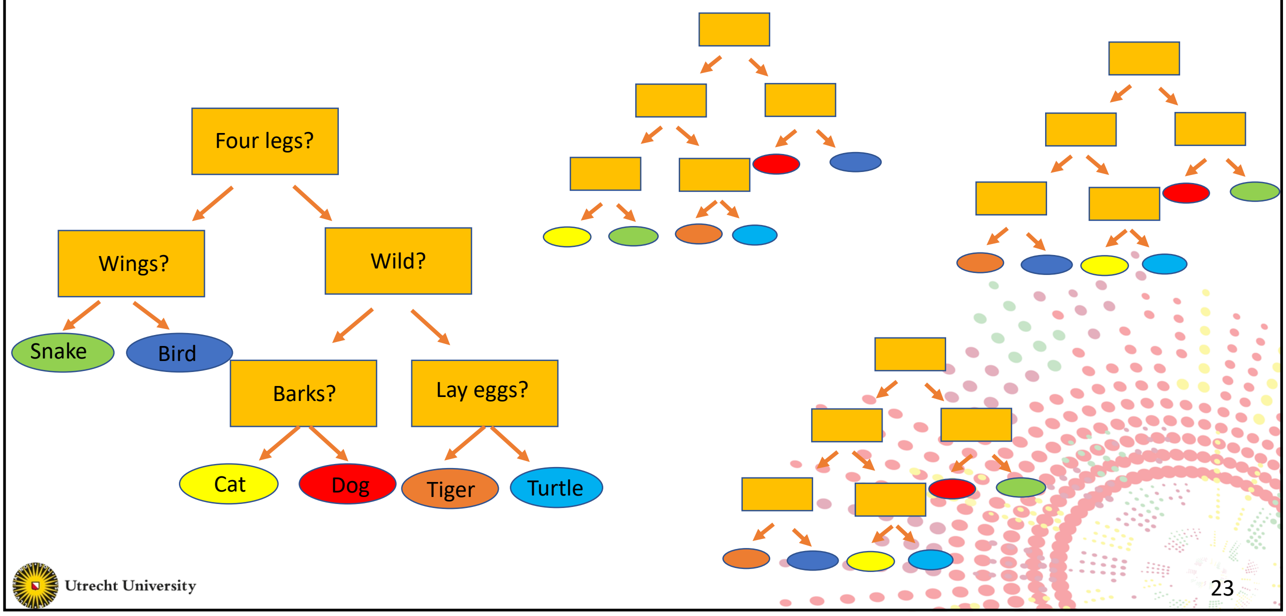
Decision Trees

- Outcome: Buys_computer
- Resulting tree:



age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Random Forests

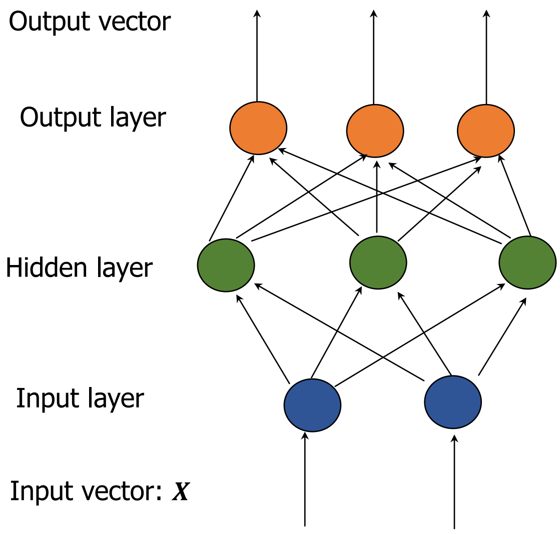


Utrecht University

23

23

Neural Network



$$w_j^{(k+1)} = w_j^{(k)} + \lambda \left(y_i - \hat{y}_i^{(k)} \right) x_{ij}$$

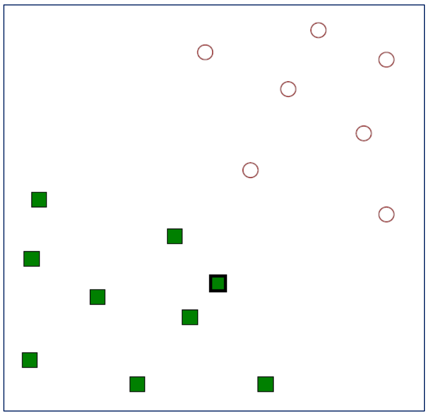
Utrecht University

24

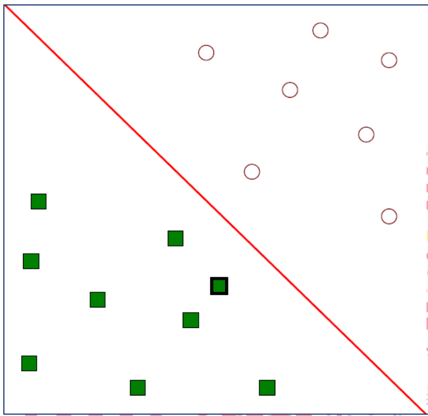
24

Support Vector Machines (SVM)

Find linear hyperplane (decision boundary) that will separate the data

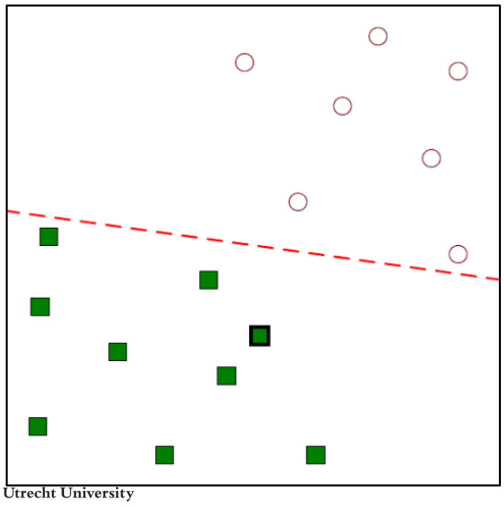


One possible separators

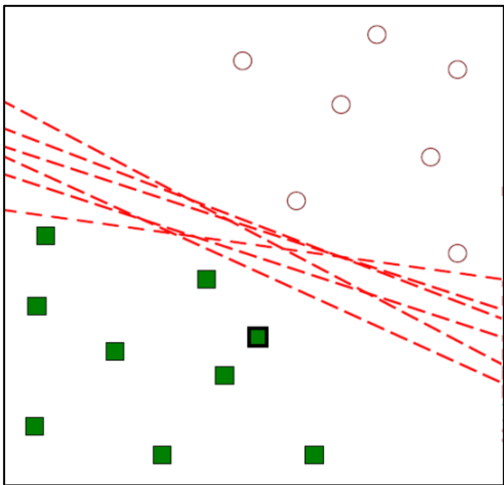


Support Vector Machines (SVM) (Cont.)

Another possible separator



Other possible separators



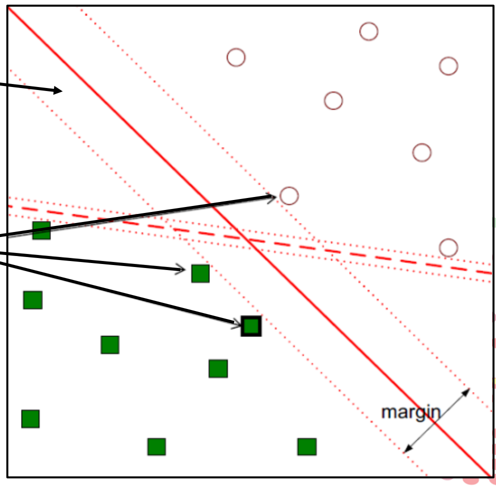
Support Vector Machines (SVM) (Cont.)

Find the hyperplane that maximizes the margin

Better separator

Support Vectors

are the points that the margin is pushed up against



Evaluation: Confusion Matrix

- Focus on the predictive capability of a model

Confusion Matrix:

		Predicted Label	
		Class = Y	Class = N
Actual Label	Class = Y	TP	FN
	Class = N	FP	TN

TP = True Positive
FP = False Positive
FN = False Negative
TN = Ture Negative

- The basic used metric:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

Evaluation Measures

Measure	Formula
Accuracy (acc.)	$\frac{TP+TN}{\text{TP+FP+TN+FN}}$
Precision (P)	$\frac{TP}{\text{TP+FP}}$
Recall (R)	$\frac{TP}{\text{TP+FN}}$
F1-Score	$2\frac{P \times R}{\text{P+R}}$
Sensitivity (TPR)	Recall (TPR)
Specificity (TNR)	$\frac{TN}{\text{TN+FP}}$
Balanced Accuracy (BA)	$\frac{\text{Sensitivity} + \text{Specificity}}{2}$

Utrecht University

29

29

Accuracy

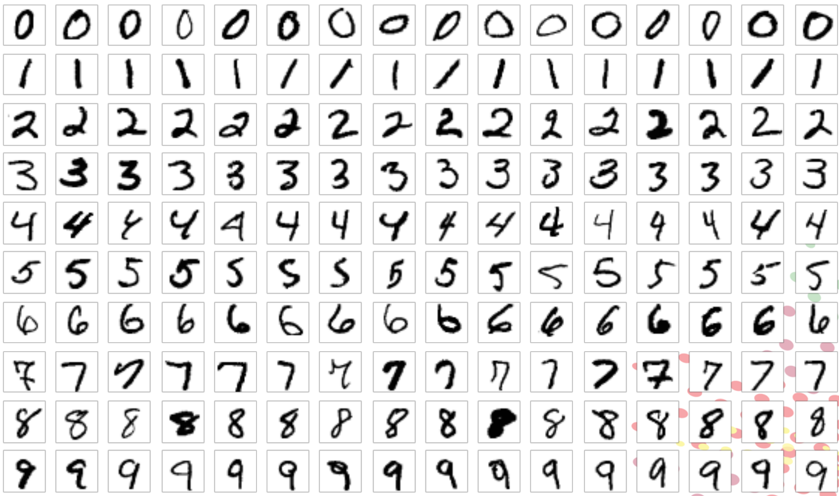
- Limitation of accuracy metric:
  - The problem of unbalanced classes
  - Consider a 2-class problem
    - Number of class 1 examples = 9990
    - Number of class 0 examples = 10
  - If the model predict everything as class 1
    - $Accuracy = \frac{9990}{10000} = 99.9\%$
- Accuracy is misleading because the classifier didn't predict any class 0 examples
- $Weighted\ Accuracy = \frac{w_1TP + w_4TN}{w_1TP + w_2FN + w_3FP + w_4TN}$

Utrecht University

30

30

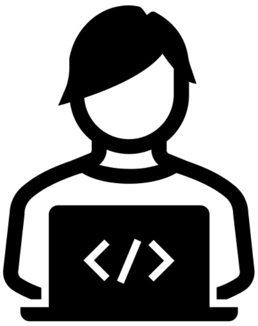
Example: Classifying hand-written digits



Utrecht University

31

31



Coding Time

Utrecht University

32

Coffee Break

Utrecht University

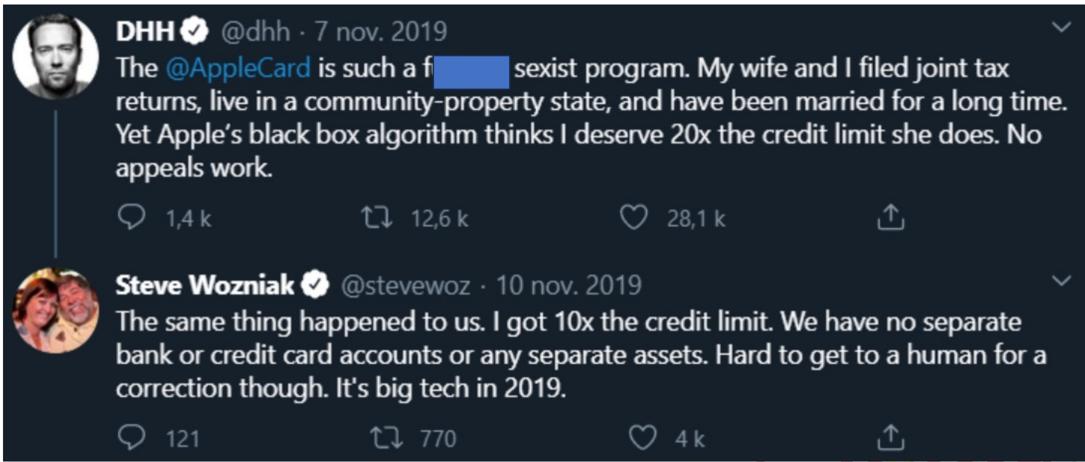
33

Model and Data Bias

Utrecht University

34

Unfair Algorithms – Apple Card

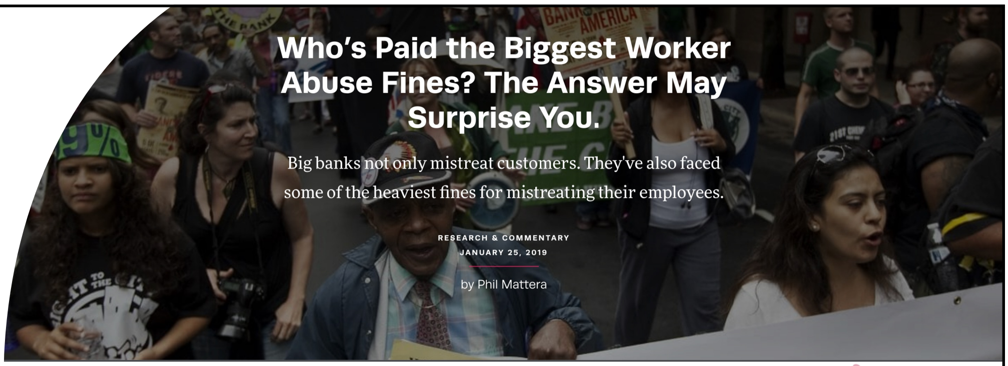


https://www.bloomberg.com/opinion/articles/2019-11-11/is-the-apple-and-goldman-sachs-credit-card-sexist,
https://www.theguardian.com/technology/2019/nov/10/apple-card-issuer-investigated-after-claims-of-sexist-credit-checks

Utrecht University

35

35



Parent Companies with the Highest Disclosed Discrimination Penalties

Rank	Parent	Penalty Total	Cases
1	Bank of America	$210,296,593	8
2	Coca-Cola	$200,616,000	9
3	Novartis	$183,000,000	2
4	Morgan Stanley	$150,385,000	6
5	Abercrombie & Fitch	$90,115,600	4
6	FedEx	$80,035,138	15
7	Boeing	$79,935,059	7
8	Verizon Communications	$71,504,891	6
9	Wells Fargo	$68,099,000	5
10	SoftBank (parent of Sprint)	$62,852,756	3

https://inequality.org/research/penalties-workplace-abuse/

Utrecht University

36

36

Reasons

- Historical bias in the decision variable
- Limited / less informative features
- Biased data collection
- Imbalanced representation of different demographic groups

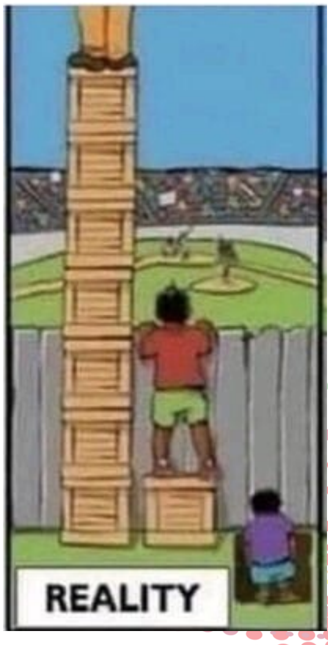
Utrecht University

37

37

Consequences

- One gets more than is needed
- Huge disparity is created



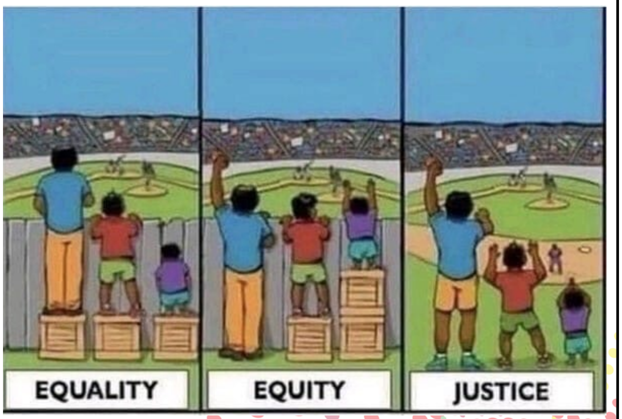
Utrecht University

38

38

What Should be Done?

- Equality: everyone benefits from the same support
- Equity: everyone gets the needed support
- Justice: remove the causes of inequity



Utrecht University

39

39

Algorithmic Fairness

Utrecht University

40

What do you think is fair?

"Fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the decision-making context"

Utrecht University

41

41

Fairness Through Unawareness

- What if you do not even collect sensitive data?
  - Useful to have the sensitive features to check for fairness
- Removing the protected attributes does not always work
  - Proxy attributes

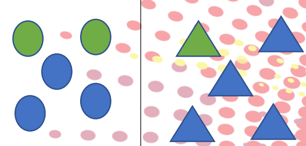
Utrecht University

42

42

Statistical Parity

- People of different groups should have the same probability of getting the positive outcome
- Disparate Impact Ratio
  - Base rate unprivileged / base rate privileged
  - Should be between 0.8 and 1.25



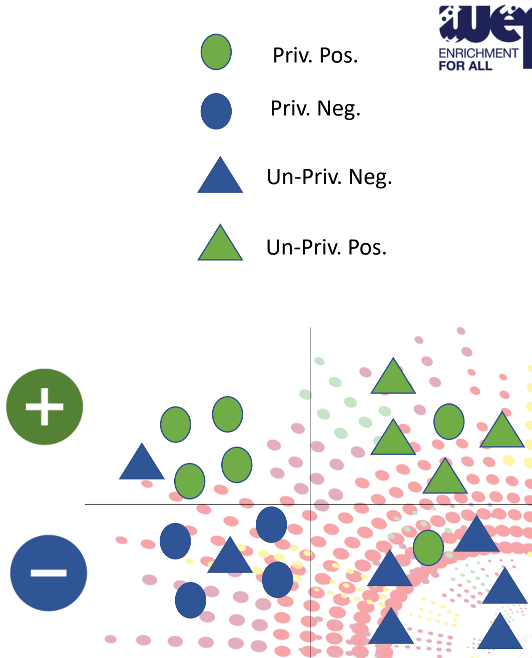
Utrecht University

43

43

Consistency

- An individual fairness measure
- Determines how similar the labels are for the similar instances in a dataset based on the k-neighbors of the instance
- This Metric: takes values between 0 and 1 with 1 is the optimal



Utrecht University

44

44

Bias Mitigation Algorithms

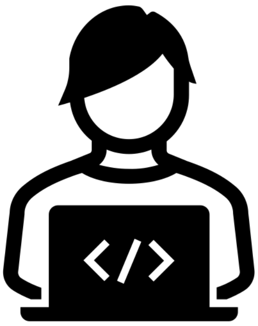
ENRICHMENT FOR ALL

- Pre-Processing Algorithms
- In-Processing Algorithms
- Post-Processing Algorithms

Utrecht University

45

45



ENRICHMENT FOR ALL

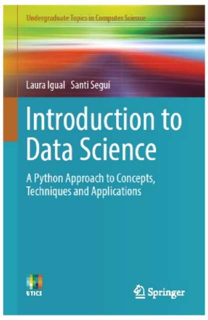
Coding Time

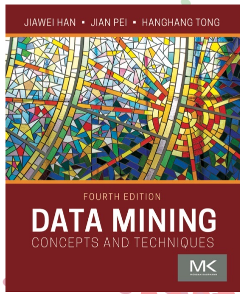
Utrecht University

46

Reading Material for Interested Students

- Introduction to Data Science, Ch 5.
  [Supervised Learning]
- Data Mining: Concepts-and-Techniques
  Ch 6. Classification





Acknowledgement: parts of the material were prepared by Xiangliang Zhang, Fenna Woudstra and Begum Hattatoglu

ENRICHMENT FOR ALL

Utrecht University

47

47

Thank You

Utrecht University

48

Utrecht University

DISCLAIMER

The information in this presentation has been compiled with the utmost care,
but no rights can be derived from its contents.

49