Feature Engineering and Data Preprocessing

Pabitra Mitra

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: $=$ $\neq$
  - Order: $<$ $>$
  - Addition: $+$ $-$
  - Multiplication: $*$ $/$

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects (< >).	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes

- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

Types of data sets

- **Record**
  - **Data Matrix**
  - **Document Data**
  - **Transaction Data**
- **Graph**
  - **World Wide Web**
  - **Molecular Structures**
- **Ordered**
  - **Spatial Data**
  - **Temporal Data**
  - **Sequential Data**
  - **Genetic Sequence Data**

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

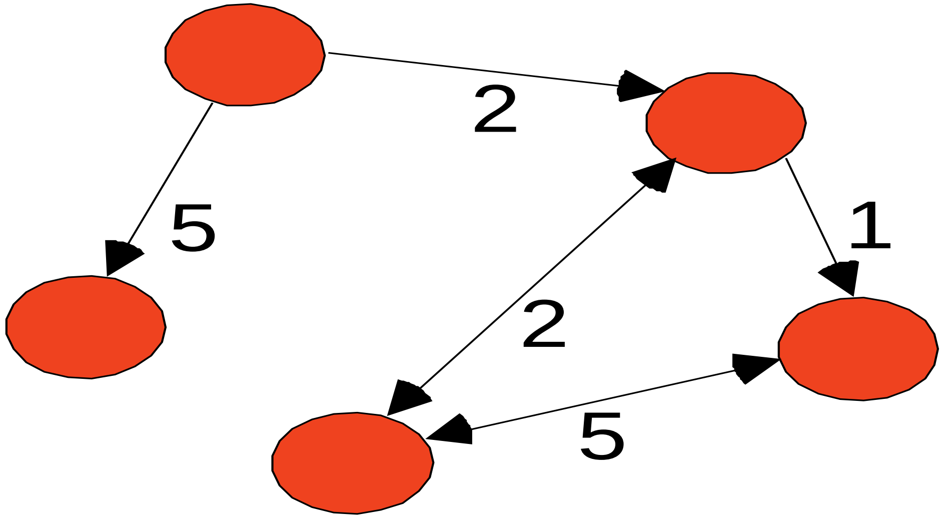
Transaction Data

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data

- Examples: Facebook graph and HTML Links



Ordered Data

- Genomic sequence data

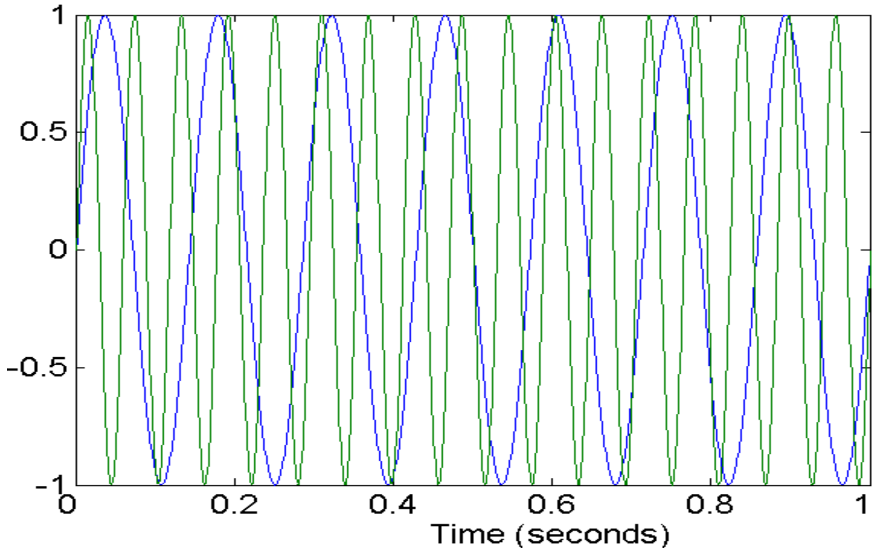
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Data Quality

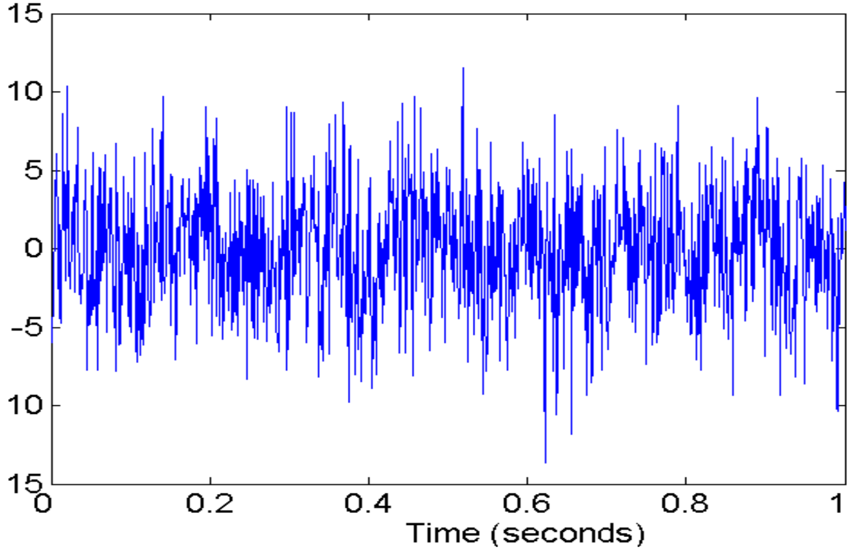
- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?
- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data

Noise

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



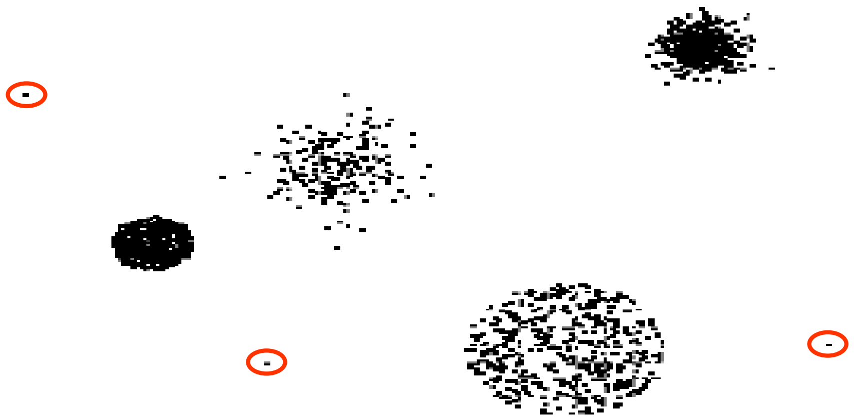
Two Sine Waves



Two Sine Waves + Noise

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Missing Values

- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogenous sources

- Examples:
  - Same person with multiple email addresses

- Data cleaning
  - Process of dealing with duplicate data issues

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

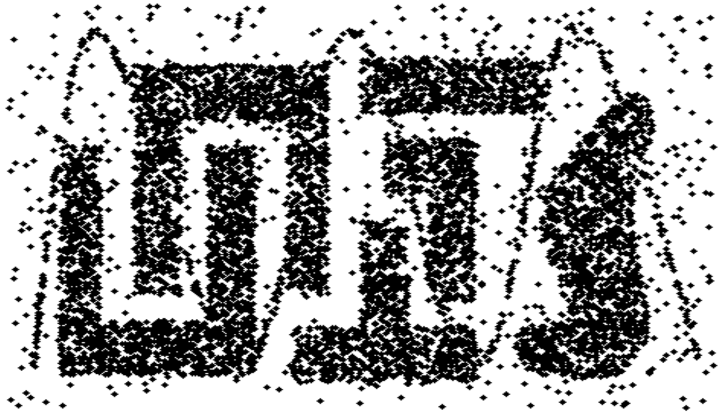
Sampling

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

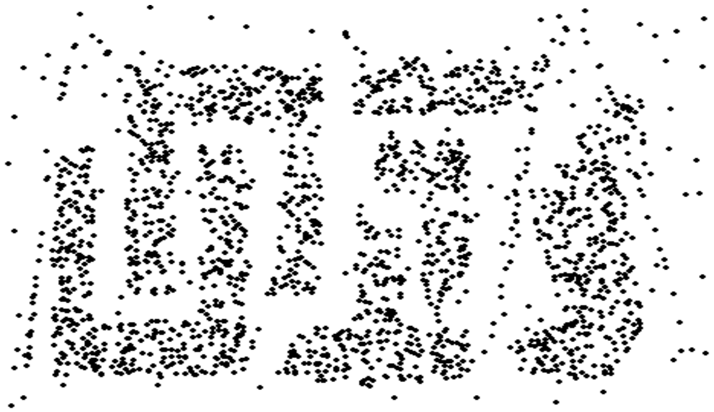
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.

- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

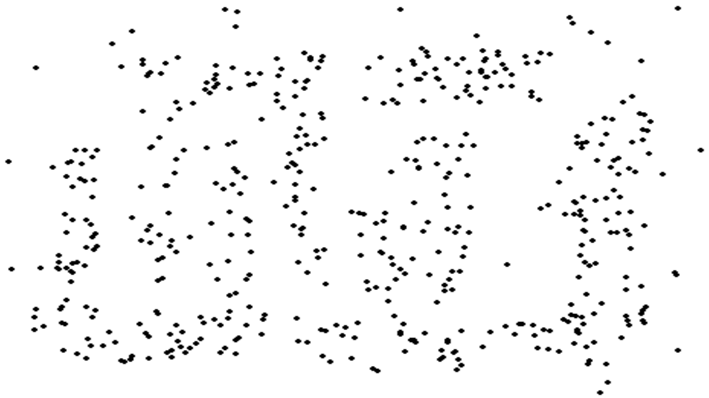
Sample Size



8000 points



2000 Points



500 Points

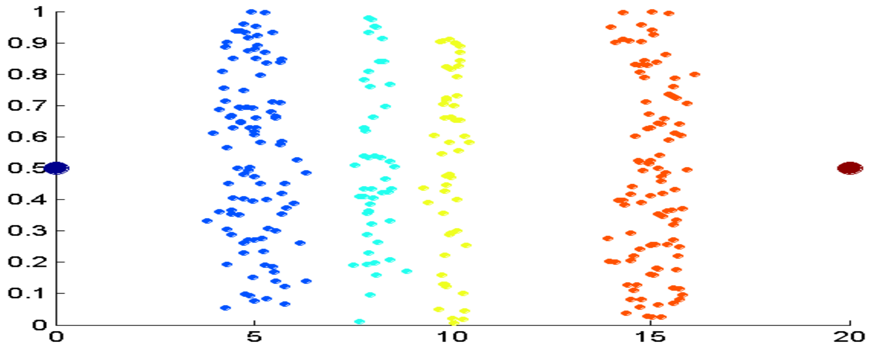
Sampling …

- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

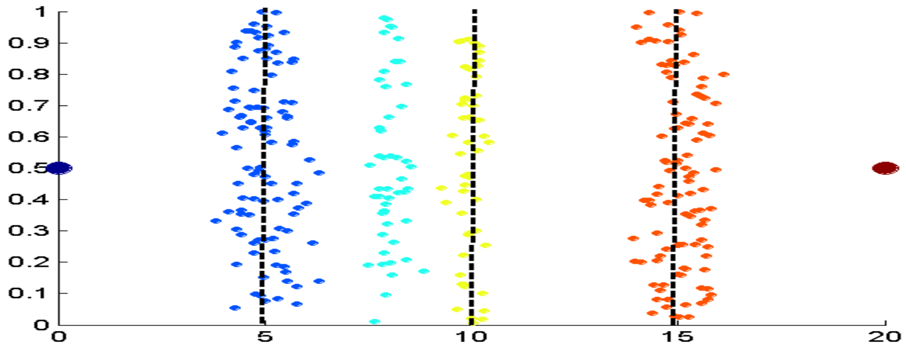
Types of Sampling

- Simple Random Sampling
  - There is an equal probability of selecting any particular item
- Sampling without replacement
  - As each item is selected, it is removed from the population
- Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

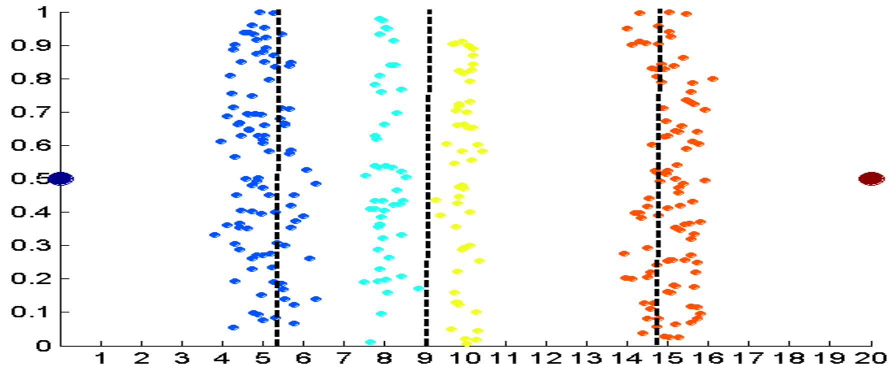
Discretization



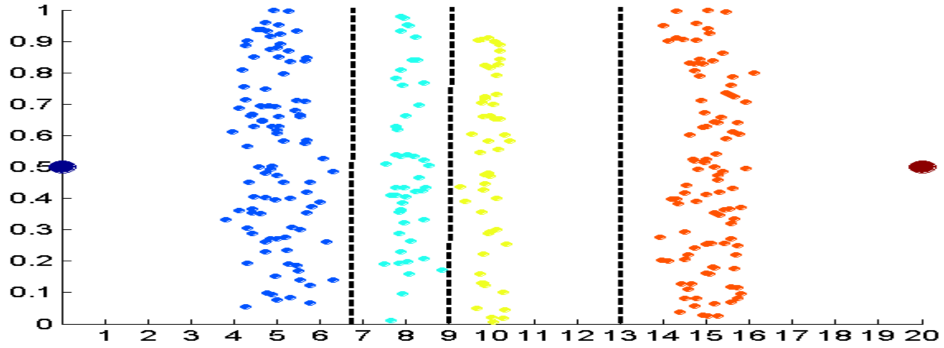
Data



Equal interval width



Equal frequency



K-means

Attribute Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, $\log(x)$, e^x, $|x|$
  - Standardization and Normalization

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies

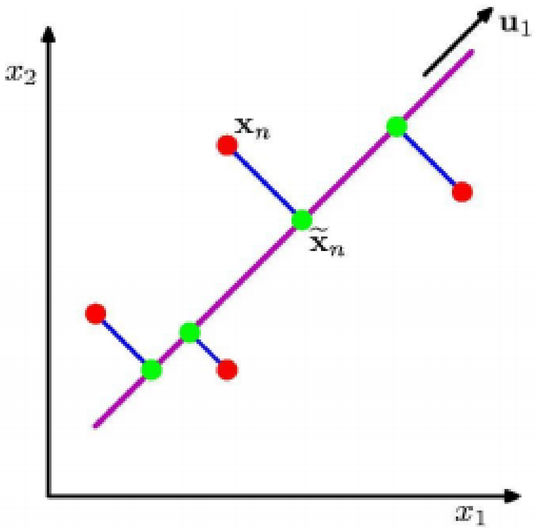
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful

Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- Techniques
  - Principle Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

Principal Component Analysis



PCA:

Orthogonal projection of the data onto a lower-dimension linear space that...

- maximizes variance of projected data (purple line)

- minimizes mean squared distance between
  - data point and
  - projections (sum of blue lines)

slide by Barnabás Póczos and Aarti Singh

Principal Component Analysis

Idea:

- Given data points in a d-dimensional space, project them into a lower dimensional space while preserving as much information as possible.
  - Find best planar approximation to 3D data
  - Find best 12-D approximation to 10^4-D data

- In particular, choose projection that minimizes squared error in reconstructing the original data.

slide by Barnabás Póczos and Aarti Singh

Principal Component Analysis

- **PCA Vectors** originate from the center of mass.

- Principal component #1: points in the direction of the **largest variance.**

- Each subsequent principal component
  - is **orthogonal** to the previous ones, and
  - points in the directions of the **largest variance of the residual subspace**

slide by Barnabás Póczos and Aarti Singh

Classifier Evaluation

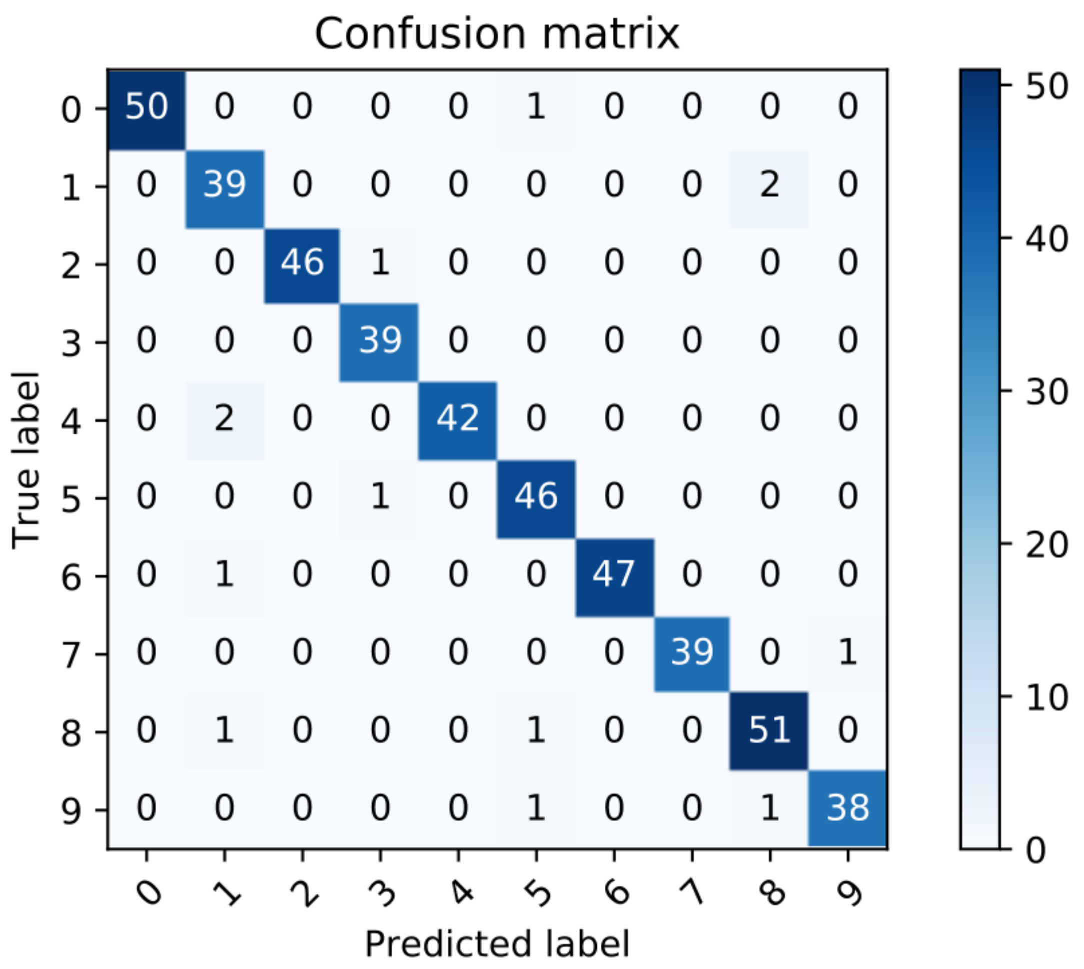
- Is the model good enough for use?
- What is the best hyper-parameter value?
- How do we compare various models?

ML Evaluation Measures

- Classification
  - Confusion matrix
  - Precision, Recall, F-Score
  - AUROC
- Regression
  - Mean squared error, RMSE
  - Mean absolute error
- Unsupervised clustering
  - Silhouette coefficient
  - Davis-Bouldin index
- Application Independent Measures
- Application Dependent Measures

Classifier Evaluation: Confusion Matrix

Digit Recognition



Spam Filter!

	Predicted	
Actual	Inbox	Spam
Inbox	25	0
Spam	5	60

Robot 1

	Predicted	
Actual	Inbox	Spam
Inbox	20	5
Spam	0	65

Robot 2

COVID Test!

	Predicted	
Actual	Negative	Positive
Negative	25	0
Positive	5	60

Robot 1

	Predicted	
Actual	Negative	Positive
Negative	20	5
Positive	0	65

Robot 2

Only Accuracy not Enough!

- Unequal cost of decision
  - *Medical diagnosis*
  - *Spam Filtering*

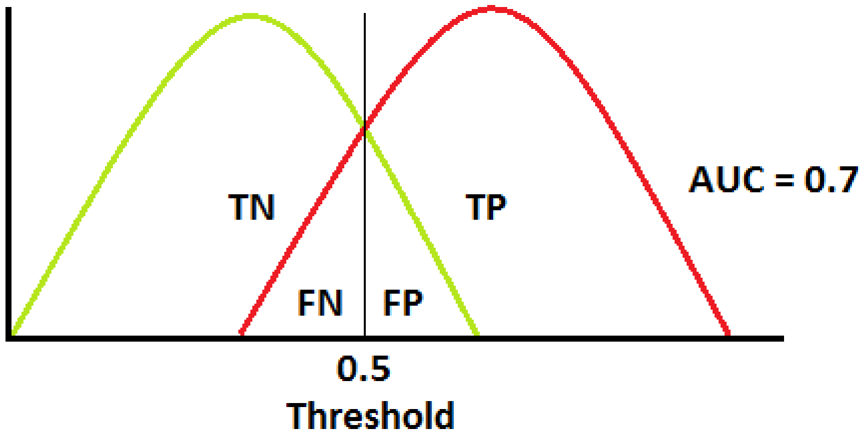
- Unbalanced Classes
  - *Medical diagnosis: 95 % healthy, 5% disease.*
  - *e-Commerce: 99 % do not buy, 1 % buy*

Multiple Scores

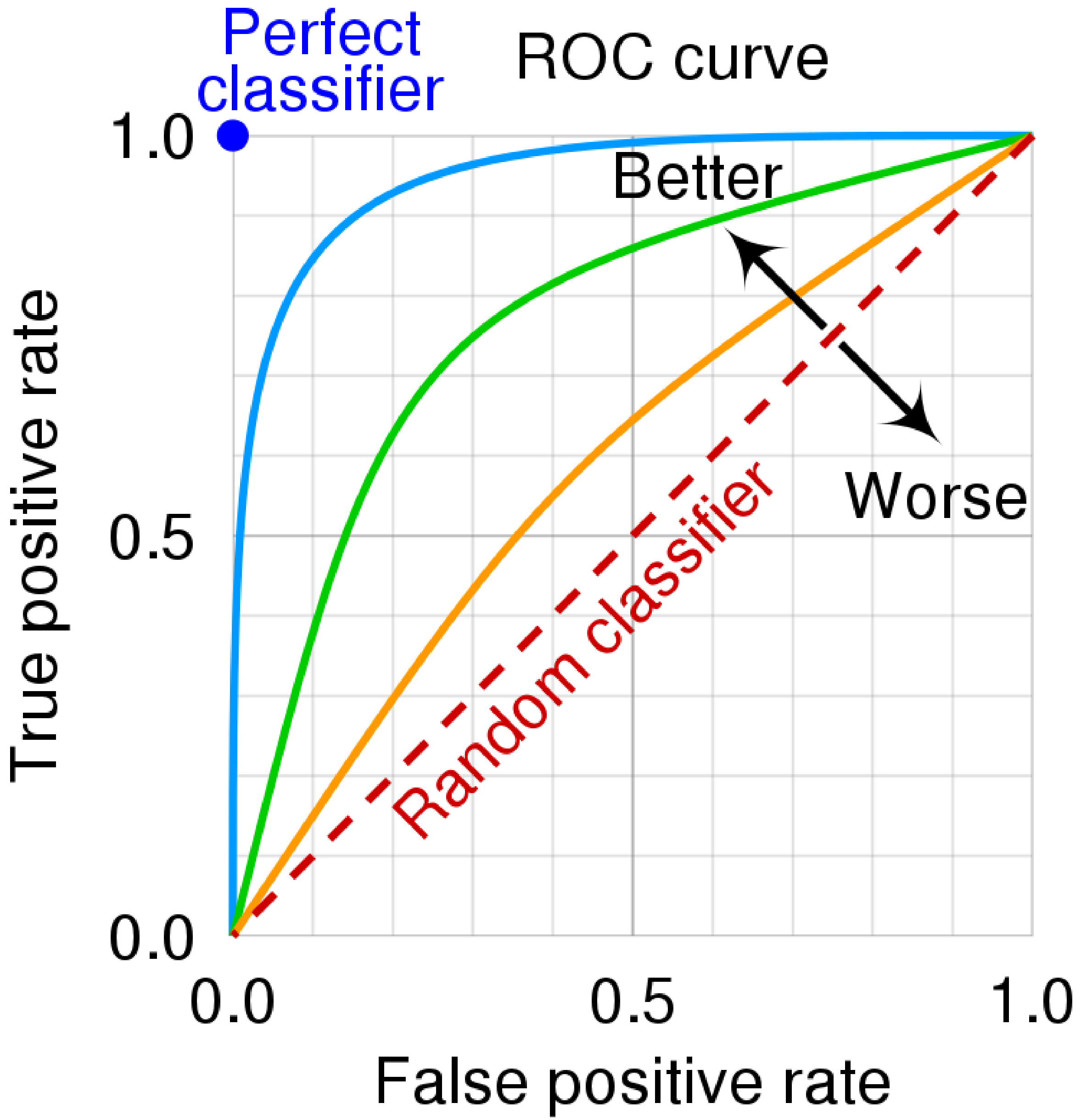
		predicted	
		negative	positive
actual examples	negative	a TN - True Negative correct rejections	b FP - False Positive false alarms type I error
	positive	c FN - False Negative misses, type II error overlooked danger	d TP - True Positive hits

- Accuracy $= (a+d)/(a+b+c+d) =$ $(TN + TP)/total$
- **True positive rate**, recall, sensitivity $= d/(c+d) =$ $TP/actual\ positive$
- Specificity, true negative rate $= a/(a+b) =$ $TN/actual\ negative$
- Precision, predicted positive value $= d/(b+d) =$ $TP/predicted\ positive$
- **False positive rate**, false alarm $= b/(a+b) =$ $FP/actual\ negative = 1$ - $specificity$
- False negative rate $= c/(c+d) = FN/actual\ positive$

ROC Curve



TN
TP
AUC = 0.7
FN
FP
0.5
Threshold



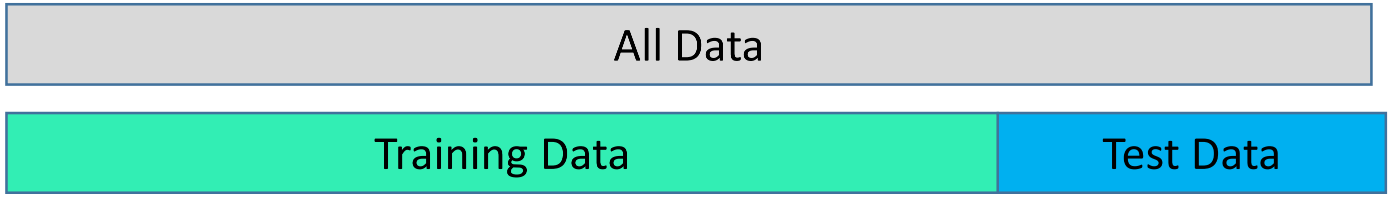
Perfect classifier
ROC curve
1.0
Better
Worse
0.5
0.0
True positive rate
Random classifier
0.0
0.5
1.0
False positive rate

Estimation of Generalization Performance

- A classifier should perform well on <u>unseen</u> examples drawn from the underlying data distribution
  - Underlying distribution unknown
- We only have a sample from the data distribution!
- How to estimate true generalization error?
  - Robust estimation using the sample

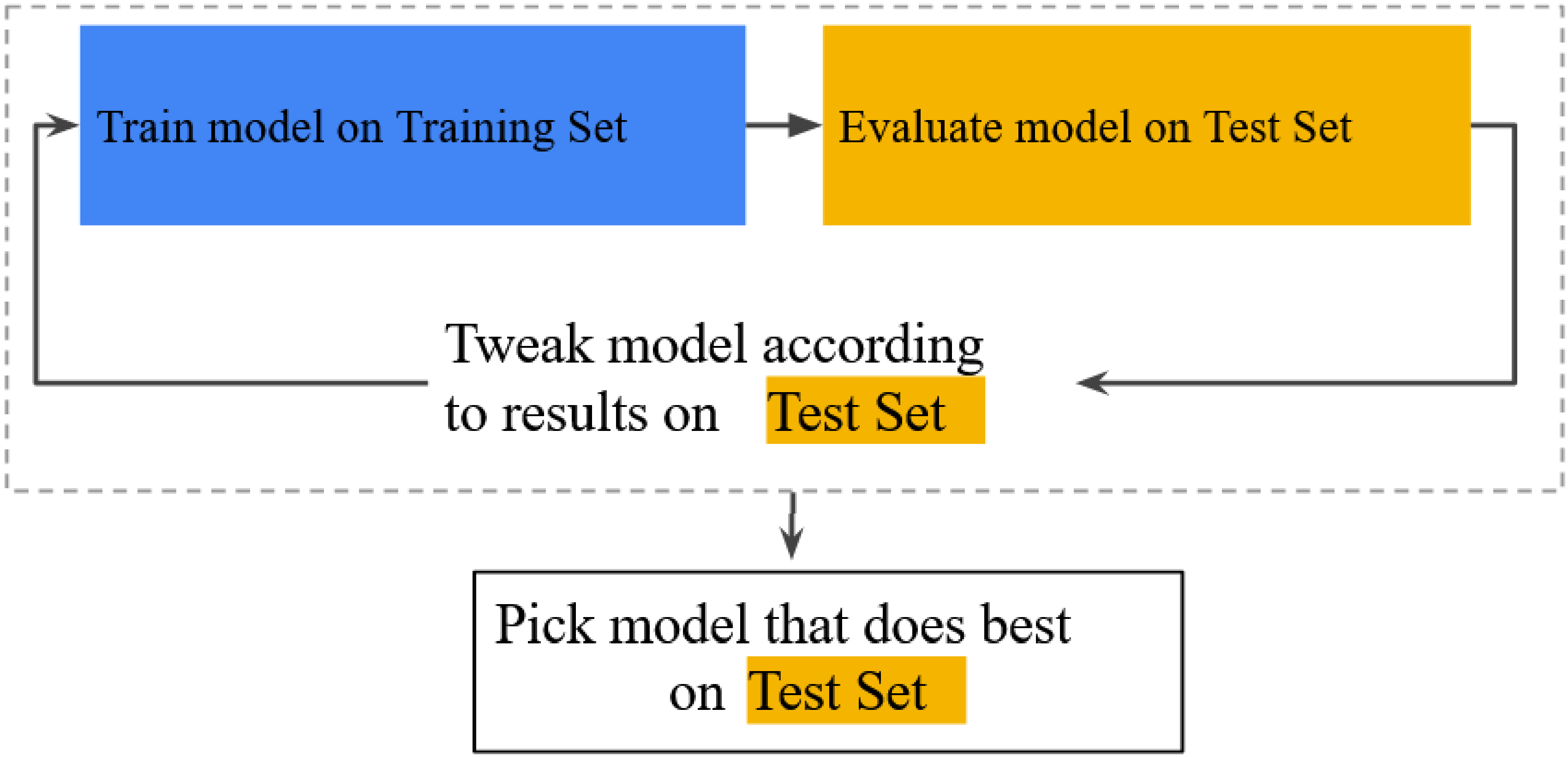
Hold-Out Set

- Randomly partition data into Train Set and Test Set



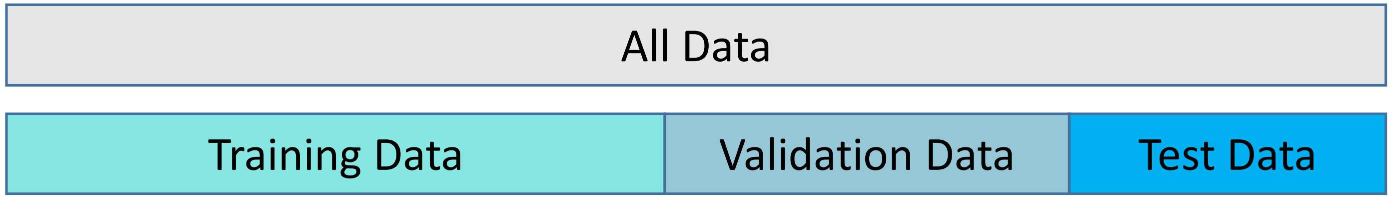
- Good performance on the test set is a useful indicator of good performance on the new data in general

Using the Test Set



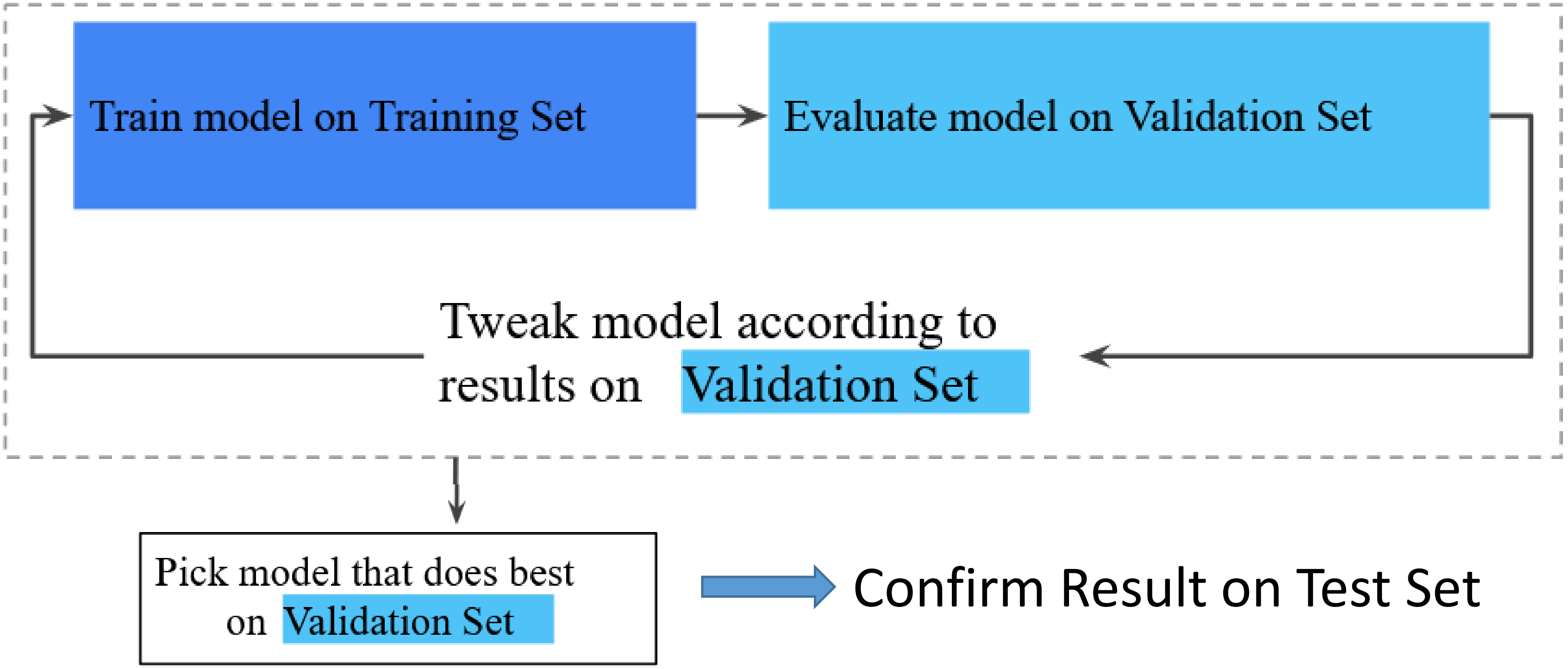
Validation Set

- Randomly partition data into Train, Validation, and Test Sets



- Motivation: One should never use test data during training.

Use of Validation Set



Training Set:
Practice problems to learn from

Validation Set:
Mock test to decide if more learning is necessary

Test Set:
Final Exam

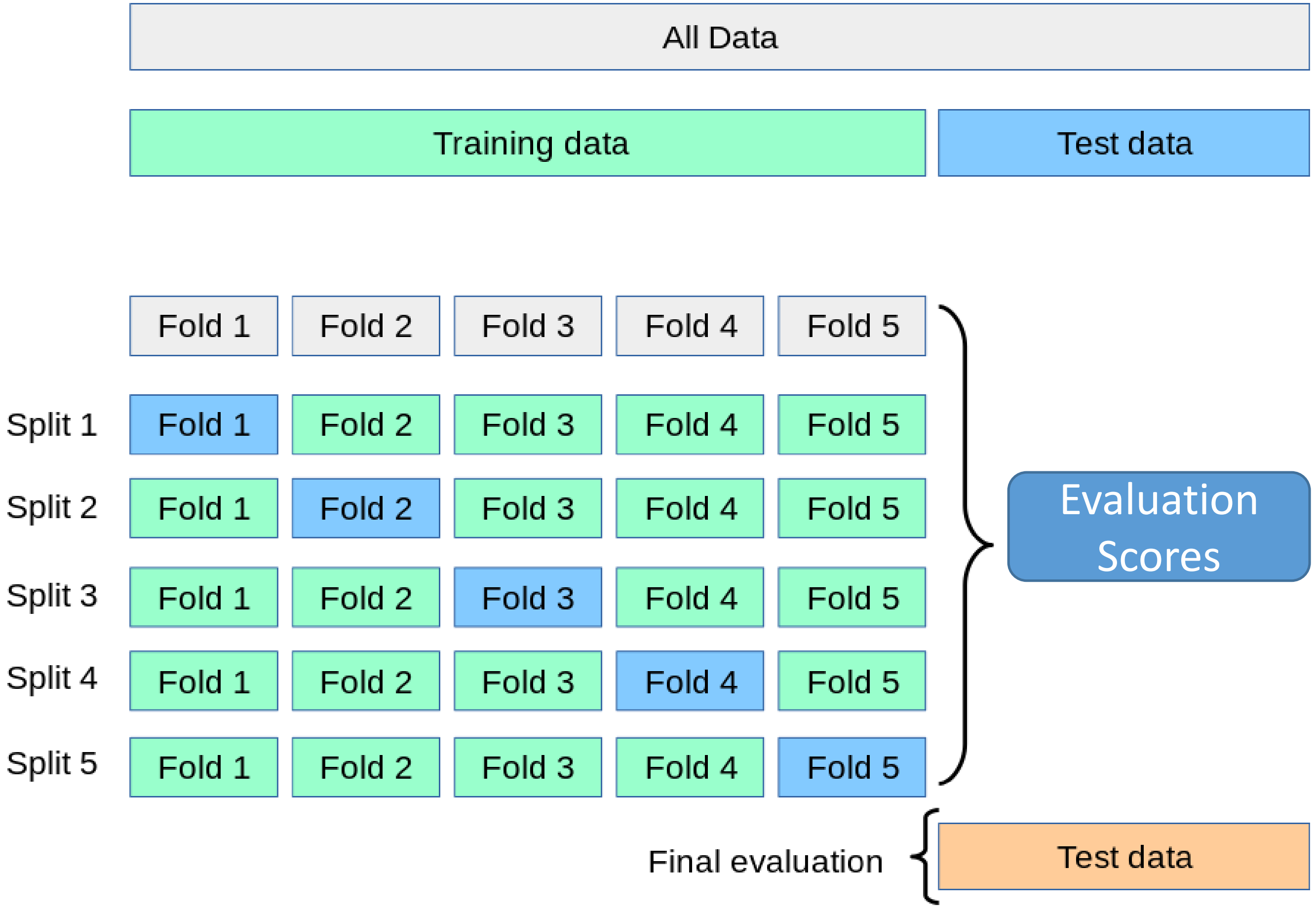
Bootstrap Estimates

Sampling with replacements
Data points may repeat

This work by Sebastian Raschka is licensed under a Creative Commons Attribution 4.0 International License.

Average of scores over each bootstrap sample

K-Fold Cross Validation



K = 5, 10

Leave-one-out: K = N,

N: size of data set

Acknowledgement