Data Mining

Nicolas PASQUIER

Laboratoire I3S (UMR-7271 UNS/CNRS)

Université Nice Sophia-Antipolis

http://www.i3s.unice.fr/~pasquier

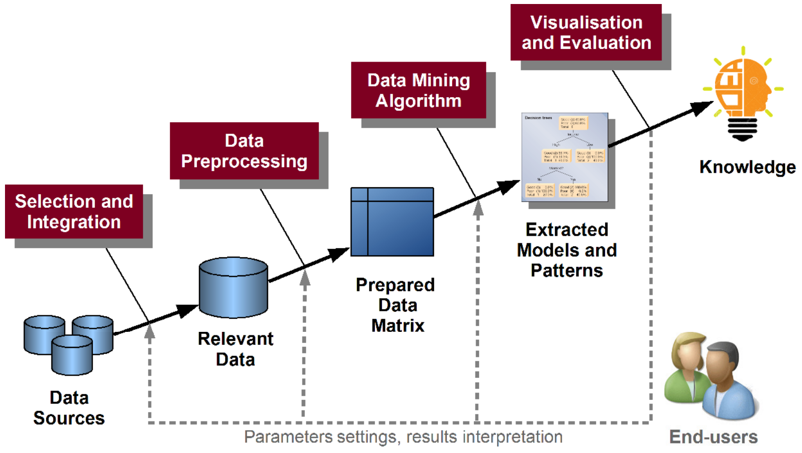
mailto:nicolas.pasquier@unice.fr

Definition

- Computing process of discovering information from very large heterogeneous datasets using algorithmic methods involving machine learning, statistics and database principles
- Extraction of knowledge models and patterns for:
  - Data space understanding (descriptors, groups, etc.)
  - Data relationships identification (links, sequences, cycles, etc.)
  - Information extrapolation for prediction (predictive models)
- Model and pattern representations
  - Implication rules, graphs, trees, partitions, sequences, time series, functions, etc.
  - Statistical indicators are associated to each model or pattern to assess its relevance and usefulness
- Knowledge Discovery from Data (KDD)

Knowledge Discovery from Data

- Interactive and iterative process



Selection and Integration

Data Preprocessing

Data Mining Algorithm

Visualisation and Evaluation

Data Sources

Relevant Data

Prepared Data Matrix

Extracted Models and Patterns

Knowledge

End-users

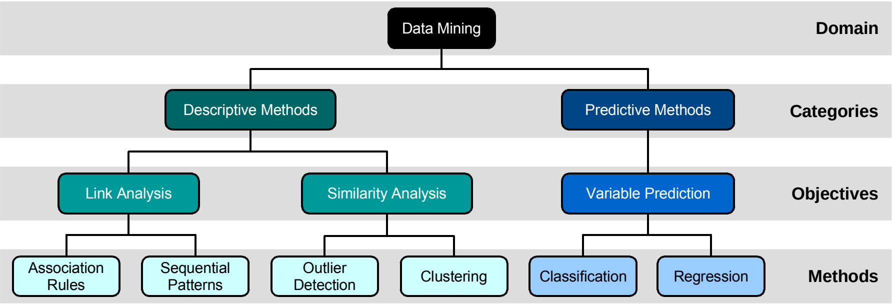
Parameters settings, results interpretation

Data Mining Objectives

- Two main categories of methods corresponding to distinct objectives
- Descriptive methods: Understand data space structure and properties
  - Frequent patterns (itemsets, closed sets, association rules)
  - Instance groups and partitions (clusters)
  - Ordered recurrent patterns (sequential patterns, chronological patterns)
  - Exceptions and deviations analysis (outlier detection)
- Predictive methods: Learn from past examples to predict future values (predictive models)
  - Predict categorical variables (supervised classification)
  - Predict numerical variables (regression)

Data Mining Methods

- Hierarchical categorization of data mining methods



- Different algorithms for each method, each one relying on a peculiar theoretical framework
- Different properties: Computation complexity in time and space, scalability, secondary memory accesses, etc.

What is Data?

- A collection of data objects and their attributes
- Attribute: Property or characteristic of an object
  - Ex: Eye color of a person, duration of a movie
  - Attribute is also known as variable, field, feature, characteristic, or column
- Object: A set of attributes values
  - Object is also known as instance, record, case, sample, tuple, or row

Attributes

Objects

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	70	96	False	Yes
Rainy	68	80	False	Yes
Rainy	65	70	True	No
Overcast	64	65	True	Yes
Sunny	72	95	False	No
Sunny	69	70	False	Yes
Rainy	75	80	False	Yes
Sunny	75	70	True	Yes
Overcast	72	90	True	Yes
Overcast	81	75	False	Yes
Rainy	71	91	True	No

Descriptive Methods: Association Rules

- Directed relationships depicting frequent co-occurrences of variable values in data instances
- Association rule: X → Y, support (%), confidence (%)
  - X and Y are terms with the form Variable = Value
  - Support: Measure of frequency of the terms in the dataset
  - Confidence: Measure of accuracy of the rule in the dataset
- Application examples
  - Market Basket Data analysis: Identify the most frequently associated items in transactions (item placement, promotional offer definition)

    Ex : Buy = Cereals ∧ Buy = Sugar → Buy = Milk,
    support = 11%, confidence = 62%
  - E-commerce: Shopping cart associations and conversions analysis, item suggestions for cross-sales (e.g. most frequently items consulted or bought together)

Association Rules Example

Transactionnal Dataset (Market Basket Data)

Customer ID	Amount	Coffee	Fruits & Vegetables	Fish	Sodas	Sugar	Fruit Juice	…
39808	427.12	True	False	False	False	True	True	…
67362	253.56	False	True	True	False	False	False	…
10872	206.17	True	False	False	True	True	False	…
26748	236.88	False	True	True	False	False	True	…
91609	188.13	True	False	False	False	True	True	…
…	…	…	…	…	…	…	…	…

Analysis of Association between Item Sales

Association Rule	Support (frequency)	Confidence (accuracy)
Coffee → Sugar, Fruit Juice	12 %	40 %
Fruits & Vegetables → Fish	9 %	61 %
Amount = [0..50] → Sodas	15 %	39 %
Coffee, Sugar → Fruit Juice	12 %	43 %
…	…	…

Descriptive Methods: Clustering

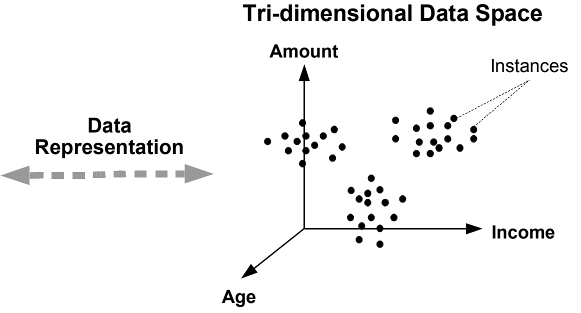
- Clustering: Identify groups of instances that are as much as possible
  - Similar among themselves within the group
  - Different from one group to another
- Unsupervised process: No target variable or classes
  - No priori knowledge of the number and type of “natural” clusters in the data space
- Instances are compared using a similarity measure (distance) to form data groups (clusters)
  - Maximize intra-cluster similarity
  - Minimize inter-cluster similarity
- Terminology: Unsupervised learning, unsupervised classification, segmentation

Clustering: Example

- Example tri-dimensional dataset: Dimensions are Age, Income and Amount

Dataset

ID	Age	Income	Amount
1241	33	1412.24	124.49
4827	42	2515.30	301.70
7204	21	1734.02	119.63
4729	58	1102.54	92.45
2948	34	2056.92	354.51
1086	22	1094.73	102.72
8293	60	3456.91	427.33
3275	51	2003.65	289.95
5678	49	3860.28	334.82
9356	31	1389.57	169.45
7221	48	1292.46	98.23
3959	23	1906.32	225.01
6576	36	2158.04	298.56
…	…	…	…

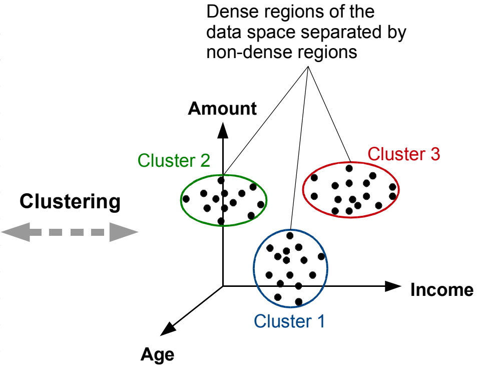


Clusters: Data Space Patterns

- Clusters: Groups of instances that are close in the data space

Dataset

ID	Age	Income	Amount	Cluster
1241	33	1412.24	124.49	1
4827	42	2515.30	301.70	3
7204	21	1734.02	119.63	1
4729	58	1102.54	92.45	1
2948	34	2056.92	354.51	2
1086	22	1094.73	102.72	1
8293	60	3456.91	427.33	3
3275	51	2003.65	289.95	2
5678	49	3860.28	334.82	3
9356	31	1389.57	169.45	2
7221	48	1292.46	98.23	1
3959	23	1906.32	225.01	2
6576	36	2158.04	298.56	3
…	…	…	…	**…**

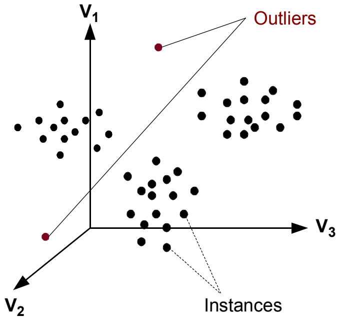

Dense regions of the data space separated by non-dense regions
Clustering
Amount
Cluster 2
Cluster 3
Cluster 1
Income
Age

Clustering: Application Examples

- Customer segmentation
  - Identify groups of customers with shared needs, common interests, similar lifestyles or similar demographic profiles
  - Determine segments that are likely to be the most profitable or have growth potential, define target markets and appropriate promotional products and services
- Social media network analysis
  - Identify groups of users with close interests or communities of individuals (friendship patterns, similar opinions, etc.)
- Bio-informatics
  - Identify groups of genes and proteins that are co-expressed, i.e. with linked variations in activities for specific biological conditions
  - Determine which one intervene jointly in biological functions and processes

Descriptive Methods: Outlier Detection

- Deviation analysis
  - Outlier: Instance which variable values are well outside of the expected range of values
  - Data noise, exceptions, rare events
  - Points that are widely separated from other points in the data space
- Application examples
  - Detection of fraud in credit card transactions
  - Sensor fault detection and filtering
  - Data quality evaluation



Descriptive Methods: Sequential Patterns

- Discovery of patterns from data that are structured as sequences
- Sequential patterns: Ordered sets of discrete values that are frequent in the data
- Application examples
  - Identification of similarities between DNA sequences (nucleotides), e.g. TTCAGTTGTG AATGAATGGA CGTCAGTTAC CATGCCAGTT…
  - Webpage click-stream analysis for the optimisation and personalisation of website's navigation
  - Text analysis: Sets of sentences from texts are processed as sequence databases to find sub-sequences of words frequently occurring in the texts
  - Time series analysis (e.g. stock data), when discretization is performed as a pre-processing step
  - Analysis of seasonal factors in sales data

Sequential Patterns: Example

Sequence Dataset

TID	Customer	Date	Book Title
1201	723	07/01/16	The Fellowship of the Ring
1202	927	09/01/16	Foundation
1203	209	10/01/16	Nine Princes in Amber
1204	723	14/01/16	The Two Towers
1205	518	14/01/16	The Fellowship of the Ring
1206	209	15/01/16	Foundation
1207	723	21/01/16	The Return of the King
1208	465	22/01/16	The Guns of Avalon
1209	927	24/01/16	Foundation and Empire
1210	518	27/01/16	The Two Towers
1211	305	30/01/16	Nine Princes in Amber
1212	209	01/02/16	Foundation and Empire
1213	518	04/02/16	The Return of the King
…	…	…	…

Sub-sequences of books occurring repeatedly for different customers (sequences)

Sequential Patterns

#	Book Title	Time Window
A1	The Fellowship of the Ring	T_0
A2	The Two Towers	T_0+13
A3	The Return of the King	T_0+21
B1	Foundation	T_0
B2	Foundation and Empire	T_0+17

Predictive Methods: Classification

- Supervised learning: Learn a model that predicts the class of an instance (value of class variable) according to values of other variables
- The model, called classifier, is learning from the training set (set of instances whose class is known)
- The classifier will then be applied to predict the class of new instances of unknown class
- Different types of classifiers: Decision trees, classification rules, neural networks, random forests, support vector machines, etc.
- Application examples
  - Risk factor analysis: Diagnose patient risk to develop a disease according to medical analyses (blood pressure, etc.), age, gender, etc.
  - Credit scoring: Categorize credit applications into risky, safe or requires human intervention according to income, age, etc.

Classification: Example

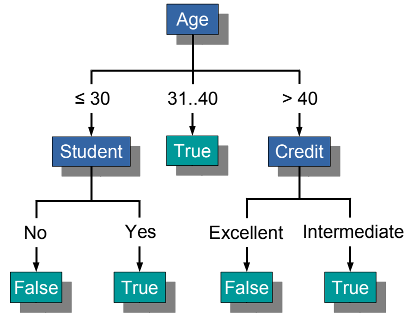
- Learning: Predictive model construction

Training Dataset (customers)

Age	Income	Student	Credit	Buyer
28	High	No	Intermediate	False
24	High	No	Excellent	False
39	High	No	Intermediate	True
47	Medium	No	Intermediate	True
41	Low	Yes	Intermediate	True
52	Low	Yes	Excellent	False
35	Low	Yes	Excellent	True
19	Medium	No	Intermediate	False
22	Low	Yes	Intermediate	True
54	Medium	Yes	Intermediate	True
23	Medium	Yes	Excellent	True
34	Medium	No	Excellent	True
37	High	Yes	Intermediate	True
…	…	…	…	…

Analyse co-occurrences of variable values for the two classes (Buyer = Yes and Buyer = No) to learn class prediction criteria

Decision Tree Classifier



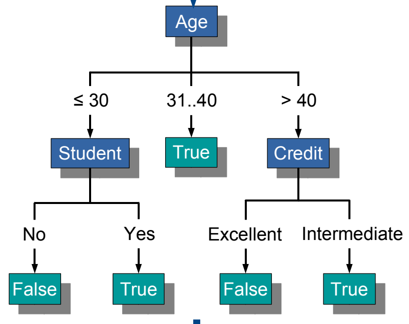
Classification: Example

- Prediction: Predictive model application

Prospect Dataset (new customers)

Age	Income	Student	Credit	Buyer
44	High	No	Intermediate	?
32	Medium	No	Excellent	?
19	Low	No	Intermediate	?
47	Medium	No	Excellent	?
24	Low	Yes	Intermediate	?

Process each instance: Identify the corresponding path in the tree



Prospect Dataset (new customers)

Age	Income	Student	Credit	Buyer
44	High	No	Intermediate	True
32	Medium	No	Excellent	True
19	Low	No	Intermediate	False
47	Medium	No	Excellent	False
24	Low	Yes	Intermediate	True

Assign to the instance the class in the leaf node

Predictive Methods: Regression

- Learn a model that predicts the value of a continuous variable according to values of other variables
- Estimate the relationship between a dependent variable (target) and the explanatory (predictive) variables
- Different types of regression: Linear regression, non-linear regression, logistic regression, non-parametric regression, etc.
- Application examples
  - Linear regression is used in business to evaluate trends and make estimates or forecasts, e.g. analysing monthly sales data to forecast sales in future months
  - In finance, the capital asset pricing model uses regression to analyse and quantify the risk of an investment
  - In epidemiology, regression is used to analyse the environmental factors affecting the health and illness of populations

Regression: Example

- Example regression model: Churn scoring (propensity to leave)

Training Dataset (customers)

Minutes	Invoice	Professional	Seniority	Income	Score
276,46	48,43	28,11	3,50	68,86	64,98
189,01	61,93	22,57	2,42	77,31	52,65
197,49	47,90	27,48	2,42	56,89	63,72
256,77	66,92	44,84	2,34	75,23	72,11
274,82	72,78	37,56	3,38	87,60	83,45
…	…	…	…	…	…

Regression parameters

- Target: Score
- Input: Minutes, Invoice, Professional, Seniority, Income
- Type: Simple linear regression

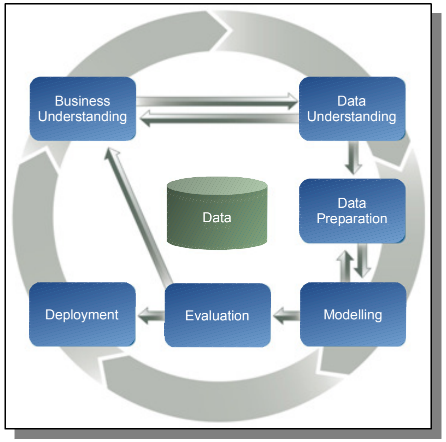
Score value will be estimated by a linear function of input variable values

```
Linear regression model

  Minutes       *  0.1747
+ Invoice       *  0.05427
+ Professional  * -0.1204
+ Seniority     * -2.369
+ Income        *  0.07443
+                 15.46
```

CRISP-DM Methodology

- Cross-Industry Standard Process for Data Mining
- Data mining project management methodology
- Six phases process
- Edges represent the most frequent relationships
- Non-strict sequence: The actual needs of the project define the sequence



CRISP-DM Methodology

- Business Understanding: Comprehend project objectives and requirements, convert them into data mining objectives and tasks
- Data Understanding: Use data exploration techniques (data visualisation, statistics and queries) to identify data quality problems and discover first insights in data (structures, relationships, etc.)
- Data Preparation: Construct the dataset by data selection (variables and instances), and data transformations and cleaning
- Modelling: Apply different algorithmic configurations to discover knowledge models and patterns, optimize parametrizations to improve their relevance and usefulness
- Evaluation: Evaluate the models and patterns, and their construction processes, assess their adequacy to the business objectives
- Deployment: Put in practice the models and patterns, organize and present them to the end-users according to the requirements

Data Exploration

- Comprehend the multi-dimensional data space structure and identify its main properties
- Identify data quality issues: Detect noise or exceptions in the data, standardize unknown values representation and units of measurement
- Determine required data transformations, depending on planned data mining tasks and algorithms
  - Normalisation of continuous numerical variables for computations based on distance measures between instances: Ensure measurements are independent from the amplitude of value domains
  - Discretization of continuous numerical variables for sets-based algorithms: Variable value domains are divided into intervals
  - Dimension reduction: Summarize the initial variables by a smaller number of new uncorrelated variables, while minimizing information loss

Data Exploration

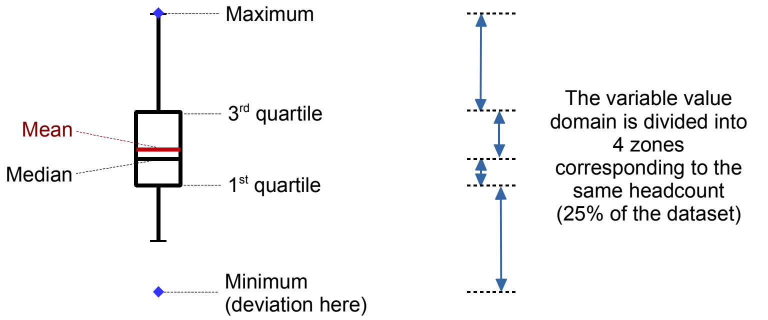
- Discover insights on data, such as significant groups or relationships
  - Distribution of instances regarding variable values
  - Linkages between variables and between variable values
- Some prominent tools for data exploration
  - Data queries: Selections of data corresponding to different criteria (viewpoints)
  - Statistics: Measures such as min, max, mean, quartiles, standard deviation, variance, correlation, covariance, etc.
  - Mono-dimensional visualizations: Distribution histograms and curves, boxplots
  - Multi-dimensional visualizations: Scatter plots, heat-maps, parallel coordinates, radar charts

Descriptive Statistics

- Quartiles: Instances are ordered according to the variable value and the dataset is divided into four partitions of the same size
- Ex: Variable Temperature

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
64.00	69.25	72.00	73.57	78.75	85.00

- They are graphically displayed using boxplots

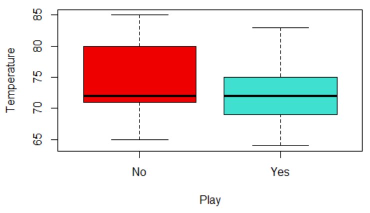


The variable value domain is divided into 4 zones corresponding to the same headcount (25% of the dataset)

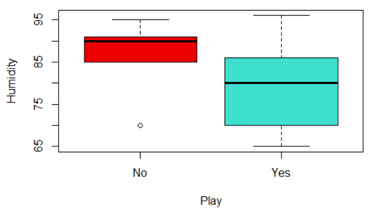
Boxplots

- The lengths of the box and the whiskers (vertical lines extending from the box) show the dispersion of the values
- The median value is outlined inside the box
- Deviant values are identified by adjacency analysis of values
- Outliers are plotted as individual points
- Some implementations include the representation of the mean of the values

Temperature Boxplots for each Class

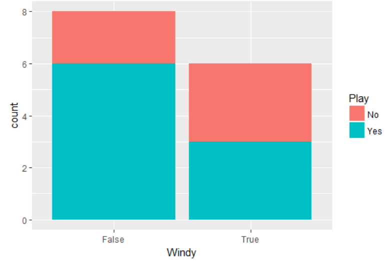


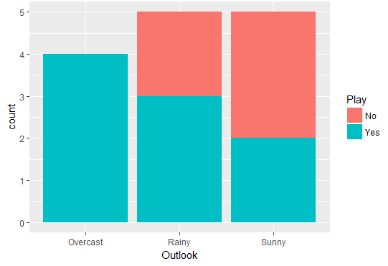
Humidity Boxplots for each Class



Distribution Histograms

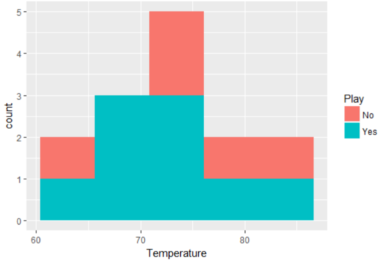
- For each value of a discrete variable (e.g. Boolean or categorical), a bar represents the number of instances with the value
- Colors distinguish instances of classes Play=Yes (green) and Play=No (red)
- We note that the headcounts of variable values are balanced
- We can see that all instances with the Overcast value for the Outlook variable are of class Play=Yes

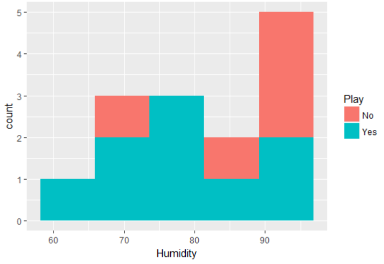




Distribution Histograms

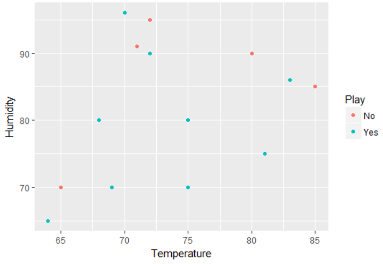
- For each interval of values of a continuous variable (e.g. Temperature in F°), a bar represents the number of instances with a value in the interval
- We note that medium temperatures are more frequent than others
- We can see that the majority of instances with low Humidity values are of class Play=Yes, whereas the majority of instances with high Humidity values are of class Play=No

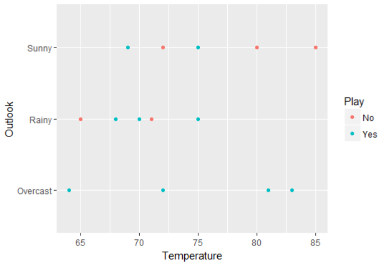




Scatter Plots

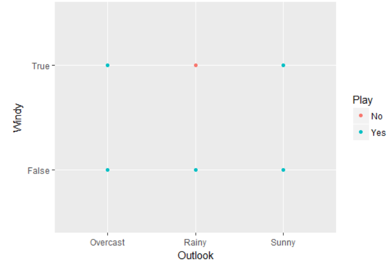
- Bi-dimensional visualization
- Each instance is represented as a point in the bi-dimensional data space where each dimension is a variable
- The points are positioned according to the instance values for the variables
- Colors differentiate instances of classes Play=Yes and Play=No
- For discrete variables, the points representing instances with the same value are on the same horizontal or vertical line

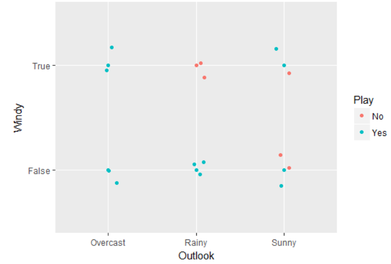




Scatter plots

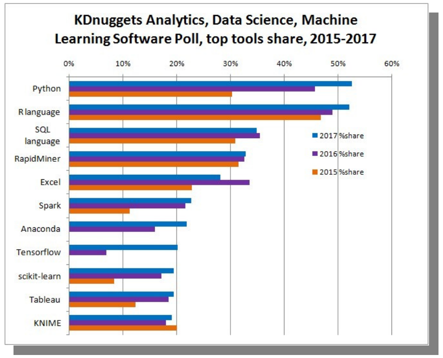
- Since the number of possible values for discrete variables is limited, several instances have the same value
- Several points can thus be superimposed in the plot (indistinguishable)
- A slight random displacement of the points, called “jitter”, is used to makes them distinguishable
- We can see that for some combinations of values, all instances are of the same class





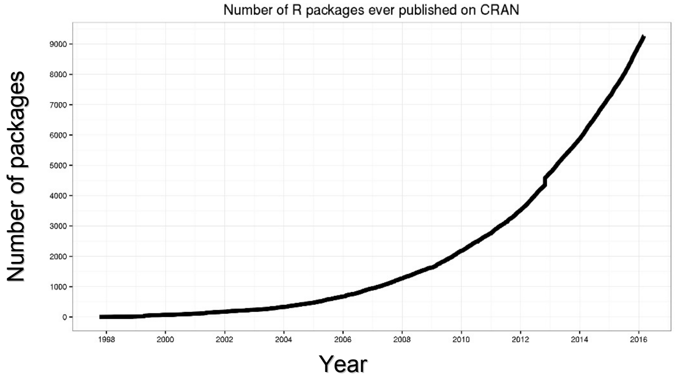
Data Mining Softwares

- KDNuggets poll: Which software did you use during the past 12 months for an analytical project?
- Data management, manipulation, visualization and analysis softwares
- R and Python are the dominating solutions
- R is the most complete solution, with several implementations of algorithms in all categories of methods
- https://www.kdnuggets.com/polls/



R Packages

- R provides an exhaustive collection of implementations (libraries)



- End 2017: More than 11 500 packages are available
- R command: `dim(available.packages())`

References

- Web sites
  - KDNuggets: Business Analytics, Big Data, Data Mining, Data Science, and Machine Learning. https://www.kdnuggets.com/
  - DataCamp: Learn Data Science Online. https://www.datacamp.com/
  - R and Data Mining: Documents, examples, tutorials and resources on R and data mining. http://www.rdatamining.com/
  - CRAN Task View: Machine Learning & Statistical Learning. https://cran.r-project.org/web/views/MachineLearning.html
- Bibliography

  - C. C. Aggarwal. *Data Mining: The Textbook*. Springer, 2015.
  - M. J. Zaki, W. Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.
  - J. Han, M. Kamber, Jian Pei. *Data Mining: Concepts and Techniques, Third Edition*. Morgan Kaufmann, 2012.