**Measurements of data representation:**

• Nominal – categorical. • Ordinal – ranked

• Interval – difference is meaningful.

• Ratio – ratio with a unit

**Privacy and Sensitive Data:** Information that is protected against unwarranted disclosure  
• Health-related data. • Genetic data, biometric data

• Personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, sexual orientation

**Data “Born Digital”:** Originally recorded or created in digital form.

**Data lifecycle:** Define Questions -> Collect/find Data -> Store Data -> Extract Data -> Pre-process Data -> Analyze Data -> Present Results -> Publish Data

**Collect/find Data:**

**• Frequency:** Interval of collection. **• Granularity:** Range for each group.

• **Cost:** Takes many forms: money, time, storage, processing, effort, etc.

• **Utility**

**Extract Data:** Data queries to extract useful subsets or slices of the data.

**Preprocessing:** Reformatting**,** Conversion**,** Cleaning**,** Imputation, Integration, Feature generation, Feature construction, and Feature selection.

**Possible Pre-processing Steps:**

• **Data reformatting:** changing the format or encoding of the data: Changing an image from JPG to PDF

• **Data conversion:** changing the unit of measurement or representation: Average temperatures in different countries, data is in C and F

**• Data cleaning:** detecting and correcting errors: Temperature time series: 70, 68...

**• Data imputation:** hypothesizing missing values: Temperature time series: 9am 50, 10am --, 11am 60.

**• Data integration:** mapping objects across datasets, merging them: Use data from two separate social networks

**Programming language:** High-level Programming Language -> Compiler -> Low-level Programming Language

**Compiled vs Interpreted languages.**

**Algorithms:** An algorithm is a mechanical procedure that describes how to carry out an explanation of some data the logic (like a recipe). Algorithm describes a process and rules to execute the process with a machine.

**Algorithms vs Programs:** An algorithm is a mechanical procedure that describes how to carry out a computation on some data (the logic), like a recipe. Programmers design algorithms and then turn them into programs that can be executed.

**Computational Workflows:** Workflow is represented as a graph of connected nodes.

• Nodes represent programs and data (alternatively)

• Links represent how data flows from program to program (output to input).

• **No user interaction during execution. • No cycles/loops or iterations allowed !!!!!**

**Repeatability:** Same lab, same data, same Analysis.

**Reproducibility:** Different lab, same data, same Analysis.

**Replication:** Different lab, different data, same method (re-run study).

**Provenance:** Documenting How Results Are Obtained. what workflow was used, what its components were, what the input data was, and what values were assigned to the parameters.

**Observational Studies:** Collect data in a way that does not directly interfere with how the data arise (observe). Only allows to associate variables.

**Experiments:** Randomly assign subject to treatment. Establish causal connections.

**Types of variables:**

图示

描述已自动生成

**Independent Variable:** Variable that you think affects the other variable. **Also called**: Exposure variable, Control variable, Explanatory variable, Manipulated variable. Manipulated in an experiment. Measured in Observation.

**Dependent Variable:** The outcome variable (you think Inferential Tests and Statistical Significance it has an effect on). **Also called**: Outcome variable, Controlled variable, Explained variable, Response variable.

**Confounding Variables:** Variables that affect both the independent and dependent variable, and that make it seem like there is a relationship between them.

**Dependent Variable:** The outcome variable (you think it has an effect on). Measured in an experiment and observation.

|  |  |  |
| --- | --- | --- |
|  | Observational | Experimental |
| IV | Measure | Manipulate |
| DV | Measure | Measure |

**Correlation Does Not Imply Causation**: directionality problem & third-variable problem.

**Observational vs. Experimental:**

**Observational research: (1)** Important, “hard to manipulate” real-life outcomes. **(2)** Could not ethically or practically manipulate.

**Experimental research: (1)** Make causal claims**. (2)** Can manipulate cleanly

**Measures of center:** mean, median, mode.

**Measures of spread:** variance, standard deviation, range, inter-quartile range.

**Law of large numbers:** In any probability space the average of the results obtained from a large number of trials converge to the expected value.

**Independence:** Two processes are independent if knowing the outcome of one provides no useful information about the outcome of the other.

**Conditional Probability:**

**68-95-99.7% rule:**

图表, 直方图

描述已自动生成

**Descriptive vs Inferential Statistics:**

**• Descriptive statistics:** Summarizes data.

**• Inferential Statistics:** Inferences from the data.

**Inferential Tests and Statistical Significance:** If is less than designated cutoff, then we say the result is statistically significant. **Other way of representing value:** incorrectly rejecting the null, false positive, type I error.

**Effect Size:** is a measure of the strength of the relationship between two variables. Most commonly reported effect sizes is Cohen’s .

**Types of statistical tests:**

|  |  |  |
| --- | --- | --- |
|  | Continuous predictor | Categorical predictor |
| Continuous outcome | correlation or regression | t-test or ANOVA |
| Categorical outcome | Logistical regression | Chi-square test or loglinear. |

**Independent Samples t-test:** Compares two independent groups.

**Paired-Samples t-test:** Compares one group that has been tested twice, Before and after; under two different conditions.

**One-way ANOVA:** Used when the predictor variable has more than two levels.

**Chi-Square**: Used to analyze categorical (count or proportion) data.

**Correlation coefficient:** Used to represent the relationship between two continuous variables.

**Statistical tests evaluate a null hypothesis:** which states that there is NO relationship or effect (i.e., no difference between groups) Able to reject the null hypothesis if it is sufficiently unlikely (5/100 or .05) that your sample (and it’s test statistic) came from this sampling distribution (where there is NO relationship or effect; i.e., no differences between groups)

**Type I Errors:** Rejecting a true null hypothesis.

• Decide the probability α in advance for a given test

• The standard cutoff is .05, meaning there is a 5% chance of a Type I Error Sampling distribution

**Type II Errors:** Failing to reject a false null hypothesis.

• Statistical power is equal to 1 – β.

• Measure of the sensitivity of the test (increases with N).

• Influenced by experimental design and the size of the effect.

• Impossible to know precisely in advance, but you can estimate.

**“Fail to Reject” the Null Hypothesis:** Remember, only trying to disprove the null. When we do not reject, this is not the same as accepting the null! Could be support for experiment hypothesis

**Machine Learning algorithms** discover the relationships between the

variables of a system (input, output and hidden) from direct

samples/observations of the system.

**Supervised learning (Classification) & Unsupervised learning (clustering)**

**Reinforcement learning:** Learning through interaction with the environment by maximizing cumulative reward.

**• Discount factor** allows for calculating this for infinite horizon Markov Decision Processes.

**Training of Decision Tree:** **(1)** Start with the set of all instances in the

root node. **(2)** Select the attribute that splits the set best and create children nodes. **(3)** When a node has all instances in the same class, make it a leaf node. **(4)** Iterate until all nodes are leaves.

**of kNN**: **(1)** If is too small, sensitive to noise points. **(2)** If is too large, neighborhood may include points from other classes.

**Advantage**: Lazy learner. **Disadvantage:** Sensitive to Noise.

**Linear classification Main Assumptions: (1)** Linear weighted sum of attribute values. **(2)** Data is linearly separable. **(3)** Attributes are real valued.

**What model to choose**: Data scientists try different models, with different parameters, and check the accuracy to figure out which one works best for the

data at hand.

**Overfitting:** A model overfits the training data when it is very accurate with that data, and may not do so well with new test data.

Nonlinearity activation functions: Faster convergence.

• Hidden layers: Tanh, Sigmoid, RELU, GELU.

• Output layers: Sigmoid, Softmax.

Neural network training optimization does **not guarantee** reaching the global minimum (non-convex optimization).

**Bias-variance tradeoff**: Unfortunately, it is not always possible to minimize both variance and bias at **the same time**. In general, **bias** is reduced if we add more and more parameters to a model and make it more complex. However, **the more complex the model** becomes the more variance we introduce in the model. In its core, the problem alludes to over- and under- fitting.

• **Data: labeled instances**,: Training, Validation, Test

• **Training**

• Estimate parameters on training set.

• Tune hyperparameters on validation set.

• Anything short of this yields over-optimistic claims.

**• Evaluation**

• Ideally, the criteria used to train the classifier should be closely related to those used to evaluate the classifier.

**• Statistical issues**

• Error bars: want realistic (conservative) estimates of accuracy.

**Unseen test set provides an unbiased estimate of accuracy.**

**k-fold cross-validation:** The available data is partitioned into k equal-size disjoint subsets. Use each subset as the test set and combine the rest k-1 subsets as the training set to learn a classifier.

**Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted class | |
| Positive | Negative |
| True class | Positive | TP | FN |
| Negative | FP | TN |

**F1-Score:** It is hard to compare two classifiers using two measures. F1 score combines precision and recall into one measure.

**What affects the performance:** **(1)** Large amounts of features for simple models (high dimensionality). **(2)** Missing feature values for instances (sparse data). **(3)** Model capacity. **(4)** Errors in feature values for instances. **(5)** Errors in the labels of training instances (noisy or weak labels). **(6)** Too few instances for a complex classification task. **(7)** Uneven availability of instances in classes.

**Unsupervised Learning**: Learning the structure of the data. For example: PCA and Clustering.

**Syntax:** Sentence structure, its constituents and morphological presentation of a word.

**Semantics:** Meaning of text. It’s the fundamental take-away after you read a sentence.

**Stop words removal:** Common words with no or little value in helping with the task.

**Tokenizing:** Divide the sentences into words (or subwords).

**Optional task-dependent steps: (1)** Lowercase. **(2)** Removing stop words.

**Text Parsing**: Process of determining the syntactic structure of a text by analyzing its constituent words based on an underlying grammar (of the language).

**Named entity recognition:** Identify which components of a sentence are important for a task.

**Typical applications of NLP:** Text classification, Entity extraction, Question answering, Dialogue systems, Summarization, Information retrieval.

**NLP in Finance:** Use social media (e.g., Twitter) data to automatically measure public mood can be used rather than (expensive) traditional polls.

**Sentiment analysis is the detection of attitudes: (**1) Holder (source) of attitude. (2) Target (aspect) of attitude. (3) Type of attitude. (4) Text containing the attitude.

**Automatic summarization:** Reduce the amount of data (hence time) for other analysis tasks.

**Topic modeling:** Helps us identify “abstract” topics in documents.

**Word embeddings:** Compact representations (vectors) representing each word.

**Vector space models of words:** While learning these word representations, we are actually building a vector space in which all words reside with certain relationships between them.

**BERT:** Can give sentence and contextualized embedding.

**日程表

描述已自动生成**