Government Statistical Service

Best Practice and Impact

GSS Methodology Advisory Committee

Glossary

Privacy and Data Confidentiality

National Statistician’s Quality Review

December 2018

Term	Definition
Active learning	A branch of semi-supervised learning where at each iteration the system identifies the unlabeled items that are most likely to be informative and asks a user to provide labels/responses from which it can learn.
Active Server Pages	A Microsoft technology for creating web pages with embedded scripts or programs that perform some computation before returning results to the user.
Aggregate data	Record level data summed to create a table. Aggregate data includes frequency tables and magnitude tables.
Anonymisation	A process that transforms personal and/or identifiable data into unidentifiable data.
Approved researcher	Researchers with permission to access personal information to assist in statistical research. They must meet criteria under the Statistics and Registration Service Act (2007) to be accredited as an approved researcher.
Attribute disclosure	A type of disclosure where an intruder can discover new information about a specific individual, household or business. This form of disclosure usually occurs in tabular data releases and arises from the presence of empty cells either in a released table or linkable set of tables after any subtraction has taken place.
Base classifiers	The individual models within an ensemble of machine learning models (see ensemble learning).
Bayesian latent model	A Bayesian model where the parameter space includes latent variables; used in classical record linkage.
Binary decision	A decision in machine learning that can take two possible values, such as true/false or linked/not linked.
Confidentiality	The right or expectation of an individual or organisation to not have information about them disclosed.
Comparison vector	Given a set of F common features present in two datasets A and B, and two records a (from A), and b (from B), the comparison vector C is a list of length F, containing the differences between the values of a and b for each feature.
Data divergence	The sum of all differences between two data sets (in format or granularity, or due to variations in coding practice, errors on one dataset or the other etc.).

Distributed access	A set of protocols that allow organisations to control the extent to which remote users have access to their systems and data, in order to preserve confidentiality.
Distributed data	Datasets which are physically distributed over several locations.
Dominance rule	See (n, k) rule.
End User Licence (EUL)	A licence used for data, normally microdata, that have been de-identified and partially anonymised. The user cannot attempt to identify an individual, nor claim to have (inadvertently) identified an individual.
Ensemble learning	A machine learning approach which creates several different models (base classifiers) and then combines their predictions by voting etc. Different models will make different errors, therefore combining the votes of several models should improve accuracy.
Euclidean distance	The straight-line distance between two points calculated as the square root of the sum of the squared distances for each feature/dimension. Best known in 2 dimensions via Pythagoras' theorem relating the lengths of the sides of a right-angled triangle.
Expectation Maximisation (EM)	A statistical method for estimating the most likely value for parameters in a machine learning model that represents a dataset.
Feasibility intervals / region	In an optimisation problem, interval or region containing all values of the variables that satisfy the constraints of the problem. The optimal solution must be in the feasibility interval/region.
Five Safes	A protocol to ensure that personal information stored at the ONS (and other organisations) is secure. The safes are Safe People, Safe Projects, Safe Settings, Safe Outputs, Safe Data.
Frequency tables	Tables of counts often used to display data collected in surveys and censuses. Each cell in a table represents the frequency or count of the defined combination of categories.
Group disclosure	A type of disclosure usually seen in tabular data when information about a small group can be determined. If all respondents in a group fall into a sensitive category then group disclosure is possible.

Global data environment	Theoretically, the sum of all data in the world. Pragmatically, the set of all data that might be linked to a given dataset. The concept is most relevant for open data releases.
Herfindahl concentration index	A measure of the relative size of different firms in a market. Calculated as the sum of the squares of their market share.
Heuristic parameter choice	Choice of a parameter by trial and error or a rule of thumb.
Hypergraph partitioning	A hypergraph is a set of nodes (points of intersection), with a set of edges (hyperedges) connecting two or more nodes. For example, a simple 2-D table with row and column totals. Hypergraph partitioning is the problem of sub-dividing a hypergraph whilst minimising the number of hyperedges that link partitions.
Identification dataset	Data set containing identifier attributes that is linked to an anonymised dataset with the goal of reidentifying the subjects to whom the records in the latter dataset correspond.
Identification disclosure	The act of identifying a person or statistical unit in the table. This identification could lead to the disclosure of potentially sensitive information about the respondent.
Identification key / key variables	A small number of key variables which can be linked to determine whether a record is a sample unique. These variables are usually visible and can assist in identifying respondents in the dataset.
K-anonymisation	A measure used to assess whether there is sufficient uncertainty within a microdata set. There must be least K records within the de-identified microdata set that have the same combination of indirect identifiers.
L-diversity	An extension of k-anonymisation. Each sensitive variable contains at least L categories with one or more records.
Linear programming	A set of mathematical optimisation methods to determine the best result of a mathematical model given a set of constraints represented by linear relationships.
Machine learning	An application of artificial intelligence that allows systems to learn automatically and improve from experience without being explicitly programmed. Machine learning algorithms receive input data, using statistical analysis to look for

	patterns in the data whilst constantly updating outputs as new data become available.
Magnitude data	A variable in a dataset where the cell entries are continuous variables such as trade with a particular country, number of employees etc.
Magnitude tables	Summed (or averaged) magnitude data. These are often used to display data from business surveys. Each cell would represent a total value for the businesses which are contributors to that cell.
Manhattan distance	Also known as city-block distance. Calculated as the sum of the absolute differences in values for each feature/dimension.
Markov matrix	Also known as a stochastic matrix, transition matrix or probability matrix, it is a square matrix containing the probabilities of transition between several states.
Meta-heuristic feature selection approaches	The use of artificial intelligence optimisation methods such as evolutionary algorithms to find ‘good’ subsets of features from a dataset. Often used to improve the performance of machine learning algorithms when datasets contain irrelevant and/or highly correlated features.
Microdata	Microdata (also known as record-level data or row-level data) are data on the characteristics of units of a population, such as individuals, households, or establishments, collected by a census, sample survey, or from administrative data. The data are in the form of tables where each row corresponds to an individual person, business, other statistical unit or event.
Microdata Review Board	Microdata Review Boards are situated within organisations releasing statistical data to inform decisions about releasing microdata and the mode of access.
Multiple imputation	Multiple imputation is a method for assigning values to missing data in a sample, based on a model of the data that are available. Although it was initially proposed to deal with missing data, multiple imputation has also been used to generate synthetic data.
Naïve Bayes learning	A form of machine learning that creates classifiers that give the probability of different labels for an observation. They apply Bayes’ rule under the assumption that the values of descriptive features are independent.

(n,k) rule	Applied to each cell in a magnitude table. Under this rule a cell is regarded as unsafe if the n largest units contribute more than k % to the cell total.
Non-parametric methods	These methods seek to make inferences on a data sample without making any hypothesis on the distribution of that sample (as opposed to parametric modelling).
Non-perturbative methods	The appearance of the data (but not the data itself) is changed. This includes methods such as table redesign and suppression.
NP-hard	In computer science, NP-hardness (non-deterministic polynomial-time hardness) denotes a problem that is at least as hard as the hardest problem in NP. In other words, NP-hard problems are those that are likely to be unsolvable in polynomial time.
Ontology	A set of concepts and categories in a subject area or domain that shows their properties and the relations between them. It can be roughly understood as a taxonomy or a classification of concepts.
Parametric modelling	Consists of adjusting to a data sample a family of distributions that is hypothesised to fit the data. The adjustment process entails finding the distribution parameters that best fit the data.
Perturbative methods	Changes have been made to some values of the data. This includes methods such as rounding and the addition of noise.
Privacy	Privacy is breached if a unit in the data can be identified through unique or rare combinations of variables. Privacy is applicable to data subjects whereas confidentiality applies to data.
Privacy preserving record linkage	Methods for efficiently linking records from different datasets without compromising the confidentiality included in either.
Pseudo-F metric	Measure of the quality of clustering found by an unsupervised learning method. It describes the ratio of between-cluster variance to within cluster variance.
Pseudonymisation	The initial step when protecting microdata. This is the act of removing direct identifiers from a record and replacing with another code (such as a row number).

Quasi-identifying variables	Variables which identify individuals indirectly such as age, gender, occupation, place of residence.
Random noise	Random numbers that are generated to be added to raw data values. Noise may be positive or negative. It is often chosen from a fixed statistical distribution centered at 0, e.g. the Laplace distribution.
Record-level data	See microdata.
Relational databases	A standard way of storing data in a set of tables, designed to minimise duplication of information.
Remote Access	Online access to disclosive microdata. Analysis can be carried out remotely by the researcher but no data are downloaded to the researcher's computer and outputs will be checked by the organisation with responsibility for the data prior to release.
Row-level data	See microdata.
R-U (risk-utility) map	A graphical representation of the trade-off between disclosure risk and data utility. The aim is to use a method of disclosure control which maintains as much utility as possible.
Secure Research Service	The Secure Research Service (SRS), formerly the Virtual Microdata Laboratory (VML), is an ONS facility for providing secure access to sensitive detailed data, typically microdata.
Seeds	The initial set of items with labels in semi-supervised machine learning.
Self-learning	A branch of semi-supervised learning where at each iteration the system makes predictions for the unlabeled items. It then accepts those predictions for which it has greatest confidence, to expand the training set for the next iteration of learning.
Semi-supervised learning	A form of model learning that makes use of a set of training examples, where the correct response is only available for some of them.
Server-side web page delivery	Technologies that perform some processing on the web-server before returning results to a user on a client machine. Found in many online business platforms.
Sufficient statistic	A statistic that can be used to describe another, where the estimation of the latter would not be improved with more knowledge of the sample.

Supervised learning	A form of model learning that makes use of a set of training examples, alongside the correct response for each. The differences between the system's outputs and the true outputs are used to adapt the learned model.
Synthetic data	Data that do not relate to real statistical units but have the look and structure of real data. They will have been generated from one or more population models, designed to be non-disclosive, and used either for teaching purposes, for testing code, or for use in developing methodology.
Topology	Mathematical description of the shape and structure of a landscape or space.
Threshold rule	A cell in a table of frequencies is defined to be sensitive if the number of respondents is less than some specified number (between 3 and 5 are commonly used).
Uniform Resource Identifier	A Uniform Resource Identifier (URI) is a string of characters that unambiguously identifies a particular resource. The most common form of URI is the Uniform Resource Locator (URL), frequently referred to informally as a web address (W3C definition).
Univariate distribution	A probability distribution of a one-dimensional random variable.
Unsupervised learning	A form of model learning that makes use of a set of training examples where there are no labels or responses. Typically used for clustering. Model adaptation is driven by statistical measures describing the 'quality' of the clustering.
Weak learners	Simple machine learning models whose performance is better than random guessing, but may not be extremely accurate.
Within group disclosure	Occurs in tables when there is one respondent in a single category, with all other respondents in a different category. This would enable the single respondent to be an intruder and find out additional information about those members of the other category.