1-Cross-Validation Test

Cross-validation is a model validation technique for evaluating how the results of a statistical analysis will generalize to an independent data set. It is mostly used in settings where the goal is the prediction. Knowing how to calculate a test such as this enables you to validate the application of your model on real-world, i.e., independent data sets.

1.1-Univariate Analysis

Univariate analysis is the simplest form of analyzing data. *Uni* means "one," so your data has only one variable. It doesn't deal with causes or relationships, and its main purpose is to describe. It takes data, summarizes that data, and finds patterns in the data.

The patterns found in univariate data include central tendency (mean, mode, and median) and dispersion, range, variance, maximum, minimum, quartiles (including the interquartile range), and standard deviation.

Example:

How many students are graduating with a data science degree? You have several options for describing data using a univariate approach. You can use frequency distribution tables, frequency polygons, histograms, bar charts, or pie charts.

1.2-Bivariate Analysis

Bivariate analysis is when two variables are analyzed together for any possible association or empirical relationship, such as, for example, the correlation between gender and graduation with a data science degree? Canonical correlation in the experimental context is to take two sets of variables and see what is common between the two sets.

Graphs that are appropriate for bivariate analysis depend on the type of variable. For two continuous variables, a scatterplot is a common graph. When one variable is categorical and the other continuous, a box plot is common, and when both are categorical, a mosaic plot is common.

1.3-Multivariate Analysis

Multivariate data analysis refers to any statistical technique used to analyze data that arises from more than one variable. This essentially models reality, in which each situation, product, or decision involves more than a single variable.

More than two variables are analyzed together for any possible association or interactions.

Example:

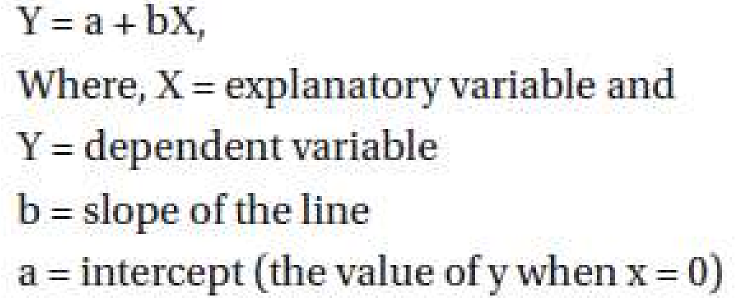
What is the correlation between gender, country of residence, and graduation with a data science degree? Any statistical modeling exercise, such as regression, decision tree, SVM, and clustering are multivariate in nature. The analysis is used when more than two variables determine the final outcome.

2-Linear Regression

Linear regression is a statistical modeling technique that endeavors to model the relationship between an explanatory variable and a dependent variable, by fitting the observed data points on a linear equation, for example, modeling the body mass index (BMI) of individuals by using their weight.

2.1-Simple Linear Regression

Linear regression is used if there is a relationship or significant association between the variables. This can be checked by scatterplots. If no linear association appears between the variables, fitting a linear regression model to the data will not provide a useful model. A linear regression line has equations in the following form:



$Y = a + bX$,
Where, X = explanatory variable and
Y = dependent variable
b = slope of the line
a = intercept (the value of y when $x = 0$)

2.1-RANSAC Linear Regression

RANSAC (RANdom SAmple Consensus) is an iterative algorithm for the robust estimation of parameters from a subset of inliers from the complete data set. An advantage of RANSAC is its ability to do robust estimation of the model parameters, i.e., it can estimate the parameters with a high degree of accuracy, even when a significant number of outliers is present in the data set. The process will find a solution, because it is so robust.

Generally, this technique is used when dealing with image processing, owing to noise in the domain.

2.2-Hough Transform

The Hough transform is a feature extraction technique used in image analysis, computer vision, and digital image processing. The purpose of the technique is to find imperfect instances of objects within a certain class of shapes, by a voting procedure. This voting procedure is carried out in a parameter space, from which object candidates are obtained as local maxima in a so-called accumulator space that is explicitly constructed by the algorithm for computing the Hough transform.

3-Logistic Regression

Logistic regression is the technique to find relationships between a set of input variables and an output variable (just like any regression), but the output variable, in this case, is a binary outcome (think of 0/1 or yes/no).

3.1-Simple Logistic Regression

I will guide you through a simple logistic regression that only compares two values. A real-word business example would be the study of a traffic jam at a certain location in London, using a binary variable. The output is a categorical: yes or no. Hence, is there a traffic jam? Yes or no?

The probability of occurrence of traffic jams can be dependent on attributes such as weather condition, day of the week and month, time of day, number of vehicles, etc. Using logistic regression, you can find the best-fitting model that explains the relationship between independent attributes and traffic jam occurrence rates and predicts probability of jam occurrence.

This process is called binary logistic regression. The state of the traffic changes for No = Zero to Yes =

One, by moving along a curve modeled by the following code, is illustrated in Figure 10-10.

```
for x in range(-10,10,1): print(math.sin(x/10))
```

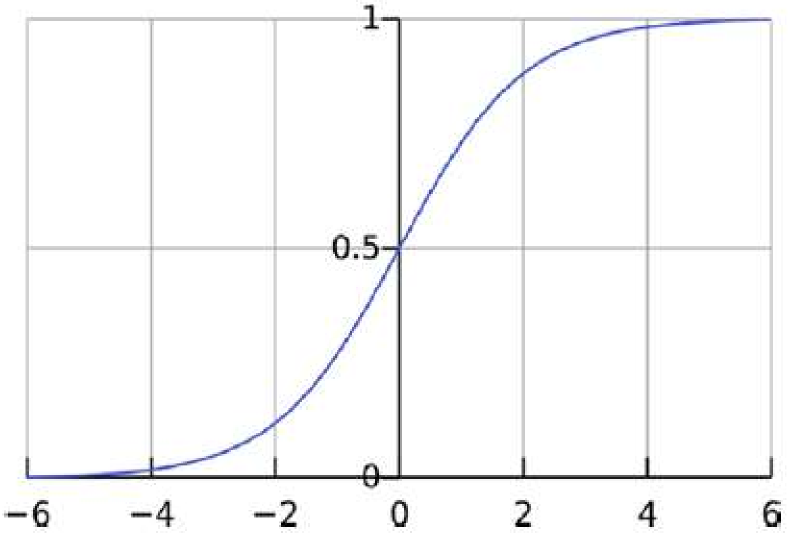


Figure 10-10. Binary logistic regression

3.2-Multinomial Logistic Regression

Multinomial logistic regression (MLR) is a form of linear regression analysis conducted when the dependent variable is nominal with more than two levels. It is used to describe data and to explain the relationship between one dependent nominal variable and one or more continuous-level (interval or ratio scale) independent variables. You can consider the nominal variable as a variable that has no intrinsic ordering.

This type of data is most common in the business world, as it generally covers most data entries within the data sources and directly indicates what you could expect in the average data lake. The data has no intrinsic order or relationship.

Ordinal Logistic Regression

Ordinal logistic regression is a type of binomial logistics regression. Ordinal regression is used to predict the dependent variable with ordered multiple categories and independent variables. In other words, it is used to facilitate the interaction of dependent variables (having multiple ordered levels) with one or more independent variables.

This data type is an extremely good data set to process, as you already have a relationship between the data entries that is known. Deploying your Transform step's algorithms will give you insights into how strongly or weakly this relationship supports the data discovery process.

4-Clustering Techniques

Clustering (or segmentation) is a kind of unsupervised learning algorithm, in which a data set is grouped into unique, differentiated clusters. Let's say we have customer data spanning 1000 rows. Using clustering, we can group the customers into separate clusters or segments, based on the variables. In the case of customers' data, the variables can be demographic information or purchasing activities.

Clustering is an unsupervised learning algorithm, because the input is unknown to the data scientist as no training set is available to pre-train a model of the solution.

You do not train the algorithm on any past input-output information, but you let the algorithm define the output for itself. Therefore, there is no right solution to a clustering algorithm, only a reasonably best-fit solution, based on business usability. Clustering is also known as unsupervised classification.

There are two basic types of clustering techniques.

I. Hierarchical clustering

II. Partitional clustering

4.1-Hierarchical Clustering

Hierarchical clustering is a method of cluster analysis whereby you build a hierarchy of clusters. This works well for data sets that are complex and have distinct characteristics for separated clusters of data.

The following would be an example. People on a budget are more attracted to your sale items and multi-buy combinations, while more prosperous shoppers are more brand-orientated. These are two clearly different clusters, with poles-apart driving forces to buy an item.

There are two design styles

- **Agglomerative:** This is a bottom-up approach. Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive:** This is a top-down approach. All observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. I will take you through the transform process to generate a hierarchical cluster.

4.2-Partitional Clustering

A partitional clustering is simply a division of the set of data objects into nonoverlapping subsets (clusters), such that each data object is in exactly one subset. Remember when you were at school? During breaks, when you played games, you could only belong to either the blue team or the red team. If you forgot which team was yours, the game normally ended in disaster.

5-ANOVA

The one-way analysis of variance (ANOVA) test is used to determine whether the mean of more than two groups of data sets is significantly different from each data set.

Example:

A BOGOF (buy-one-get-one-free) campaign is executed on 5 groups of 100 customers each. Each group is different in terms of its demographic attributes. We would like to determine whether these five respond differently to the campaign. This would help us optimize the right campaign for the right demographic group, increase the response rate, and reduce the cost of the campaign.

The analysis of variance works by comparing the variance between the groups to that within the group. The core of this technique lies in assessing whether all the groups are in fact part of one larger population or a completely different population with different characteristics.

6-Principal Component Analysis (PCA)

Dimension (variable) reduction techniques aim to reduce a data set with higher dimension to one of lower dimension, without the loss of features of information that are conveyed by the data set. The dimension here can be conceived as the number of variables that data sets contain.

Two commonly used variable reduction techniques follow.

- Factor Analysis
- Conjoint Analysis

6.1-Factor Analysis

The crux of PCA lies in measuring the data from the perspective of a principal component. A principal component of a data set is the direction with the largest variance. A PCA analysis involves rotating the axis of each variable to the highest Eigen vector/Eigen value pair and defining the principal components, i.e., the highest variance axis or, in other words, the direction that most defines the data. Principal components are uncorrelated and orthogonal.

PCA is fundamentally a dimensionality reduction algorithm, but it is just as useful as a tool for visualization, for noise filtering, for feature extraction, and engineering.

6.2-Conjoint Analysis

Conjoint analysis is widely used in market research to identify customers' preference for various attributes that make up a product. The attributes can be various features, such as size, color, usability, price, etc.

Using conjoint (trade-off) analysis, brand managers can identify which features customers would trade off for a certain price point. Thus, such analysis is a highly used technique in new product design or pricing strategies.

Example:

The data is a ranking of three different features (TV size, TV type, TV color).

TV size options are 42", 47", or 60".

TV type options are LCD or Plasma.

TV color options are red, blue, or pink.

The data rates the different stimuli types for each customer. You are tasked with determining which TVs to display on Krennwallner's billboards.

7-Decision Trees

Decision trees, as the name suggests, are a tree-shaped visual representation of routes you can follow to reach a particular decision, by laying down all options and their probability of occurrence. Decision trees are exceptionally easy to understand and interpret. At each node of the tree, one can interpret what would be the consequence of selecting that node or option. The series of decisions leads you to the end result, as shown in Figure 10-13.

Before you start the example, I must discuss a common add-on algorithm to decision trees called AdaBoost. AdaBoost, short for "adaptive boosting," is a machine-learning meta-algorithm.

The classifier is a meta-estimator, because it begins by fitting a classifier on the original data set and then fits additional copies of the classifier on the same data set, but where the weights of incorrectly classified instances are adjusted, such that subsequent classifiers focus more on difficult cases. It boosts the learning impact of less clear differences in the specific variable, by adding a progressive weight to boost the impact.

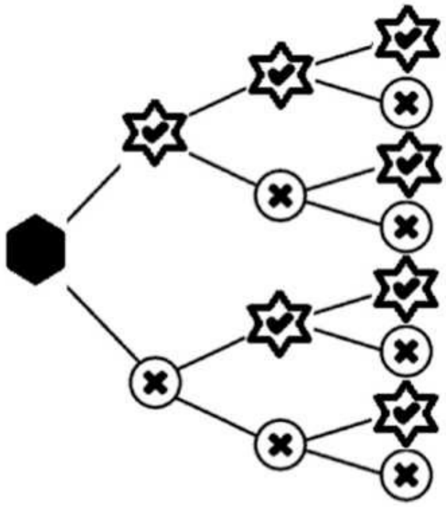


Figure 10-13. Simple decision tree

8-Support Vector Machines, Networks, Clusters, and Grids

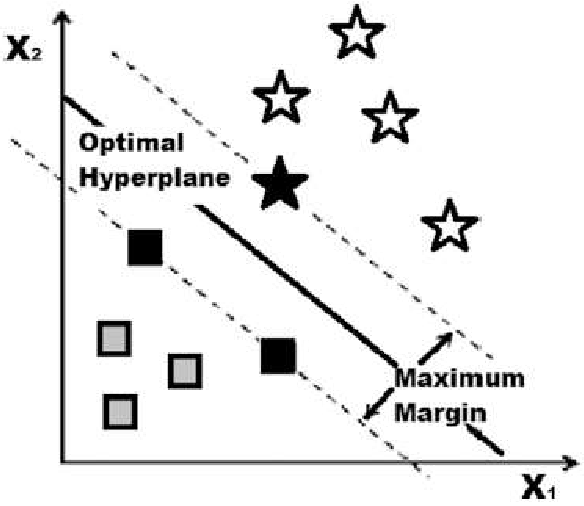
The support vector machine (SVM) constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification and regression. The support vector network (SVN) daisy chains more than one SVM together to form a network. All the data flows through the same series of SVMs.

The support vector cluster (SVC) runs SVM on different clusters of the data in parallel. Hence, not all data flows through all the SVMs.

The support vector grid (SVG) is an SVC of an SVN or an SVN of an SVC. This solution is the most likely configuration you will develop at a customer site. It uses SVMs to handle smaller clusters of the data, to apply specific transform steps. As a beginner data scientist, you only need to note that they exist.

8.1-Support Vector Machines

A support vector machine is a discriminative classifier formally defined by a separating hyperplane. The method calculates an optimal hyperplane (Figure 10-14) with a maximum margin, to ensure it classifies the data set into separate clusters of data points.



***e 10-14.** Simple support vector machine's optimal hyperplane*

8.2-Support Vector Networks

The support vector network is the ensemble of a network of support vector machines that together classify the same data set, by using different parameters or even different kernels. This a common method of creating feature engineering, by creating a chain of SVMs.

8.3-Support Vector Clustering

Support vector clustering is used were the data points are classified into clusters, with support vector machines performing the classification at the cluster level. This is commonly used in highly dimensional data sets, where the clustering creates a grouping that can then be exploited by the SVM to subdivide the data points, using different kernels and other parameters.

I have seen SVC, SVN, SVM, and SVG process in many of the deep-learning algorithms that I work with every day. The volume, variety, and velocity of the data require that the deep learning do multistage classifications, to enable the more detailed analysis of the data points to occur after a primary result is published.

9-Data Mining

Data mining is processing data to pinpoint patterns and establish relationships between data entities. Here are a small number of critical data mining theories you need to understand about data patterns, to be successful with data mining.

9.1-Association Patterns

This involves detecting patterns in which one occasion is associated to another. If, for example, a loading bay door is opened, it is fair to assume that a truck is loading goods. Pattern associations simply discover the correlation of occasions in the data. You will use some core statistical skills for this processing.

Correlation is only a relationship or indication of behavior between two data sets. The relationship is not a cause-driven action.

Example:

If you discover a relationship between hot weather and ice cream sales, it does not mean high ice cream sales cause hot weather, or vice versa. It is only an observed relationship.
This is commonly used in retail basket analysis and recommender systems.

9.2-Classification Patterns

This technique discovers new patterns in the data, to enhance the quality of the complete data set. Data classification is the process of consolidating data into categories, for its most effective and efficient use by the data processing. For example, if the data is related to the shipping department, you must then augment a label on the data that states that fact.

A carefully planned data-classification system creates vital data structures that are easy to find and retrieve. You do not want to scour your complete data lake to find data every time you want to analyze a new data pattern.

9.2.1-Clustering Patterns

Clustering is the discovery and labeling of groups of specifics not previously known.

An example of clustering is if, when your customers buy bread and milk together on a Monday night, you group, or cluster, these customers as “start-of-the-week small-size shoppers,” by simply looking at their typical basic shopping basket.

Any combination of variables that you can use to cluster data entries into a specific group can be viewed as some form of clustering. For data scientists, the following clustering types are beneficial to master.

9.2.2-Connectivity-Based Clustering

You can discover the interaction between data items by studying the connections between them. This process is sometimes also described as hierarchical clustering.

9.2.3-Centroid-Based Clustering (K-Means Clustering)

"Centroid-based” describes the cluster as a relationship between data entries and a virtual center point in the data set. K-means clustering is the most popular centroid- based clustering algorithm.
Distribution-Based Clustering

This type of clustering model relates data sets with statistics onto distribution models. The most widespread density-based clustering technique is DBSCAN.

9.2.4-Density-Based Clustering

In density-based clustering, an area of higher density is separated from the remainder of the data set. Data entries in sparse areas are placed in separate clusters. These clusters are considered to be noise, outliers, and border data entries.

9.2.5-Grid-Based Method

Grid-based approaches are common for mining large multidimensional space clusters having denser regions than their surroundings. The grid-based clustering approach differs from the conventional clustering algorithms in that it does not use the data points but a value space that surrounds the data points.

9.3-Bayesian Classification

Naive Bayes (NB) classifiers are a group of probabilistic classifiers established by applying Bayes's theorem with strong independence assumptions between the features of the data set. There is one more specific Bayesian classification you must take note of, and it is called tree augmented naive Bayes (TAN).

Tree augmented naive Bayes is a semi-naive Bayesian learning method. It relaxes the naive Bayes attribute independence assumption by assigning a tree structure, in which each attribute only depends on the class and one other attribute. A maximum weighted spanning tree that maximizes the likelihood of the training data is used to perform classification.

9.4-Sequence or Path Analysis

This identifies patterns in which one event leads to another, later resulting in insights into the business. Path analysis is a chain of consecutive events that a given business entity performs during a set period, to understand behavior, in order to gain actionable insights into the data.

I suggest you use a combination of tools to handle this type of analysis. I normally model the sequence or path with the help of a graph database, or, for smaller projects, I use a library called **networkx** in Python.

Example:

Your local telecommunications company is interested in understanding the reasons or flow of events that resulted in people churning their telephone plans to their competitor.

9.5-Forecasting

This technique is used to discover patterns in data that result in practical predictions about a future result, as indicated, by predictive analytics of future probabilities and trends.

10-Pattern Recognition

Pattern recognition identifies regularities and irregularities in data sets. The most common application of this is in text analysis, to find complex patterns in the data.

11-Machine Learning

The business world is bursting with activities and philosophies about machine learning and its application to various business environments. Machine learning is the capability of systems to learn without explicit software development. It evolved from the study of pattern recognition and computational learning theory.

The impact is that with the appropriate processing and skills, you can amplify your own data capabilities, by training a processing environment to accomplish massive amounts of discovery of data into actionable knowledge, while you have a cup of coffee, for example. This skill is an essential part of achieving major gains in shortening the data-to-knowledge cycle.

I will cover a limited rudimentary theory, but machine learning encompasses a wide area of expertise that merits a book by itself. So, I will introduce you only to the core theories.

11.1-Supervised Learning

Supervised learning is the machine-learning task of inferring a function from labeled training data. The training data consists of a set of training examples. In supervised learning, each example is a pair consisting of an input object and a desired output value. You use this when you know the required outcome for a set of input features.

Example:

If you buy bread and jam, you can make a jam sandwich. Without either, you have no jam sandwich.

11.2-Unsupervised Learning

Unsupervised learning is the machine-learning task of inferring a function to describe hidden structures from unlabeled data. This encompasses many other techniques that seek to summarize and explain key features of the data.

Example:

You can take a bag of marble with different colors, sizes, and materials and split them into three equal groups, by applying a set of features and a model. You do not know up front what the criteria are for splitting the marbles.

11.2-Reinforcement Learning

Reinforcement learning (RL) is an area of machine learning inspired by behavioral psychology that is concerned with how software agents should take actions in an environment, so as to maximize, more or less, a notion of cumulative reward. This is used in several different areas, such as game theory, swarm intelligence, control theory, operations research, simulation-based optimization, multi-agent systems, statistics, and genetic algorithms (Figure 10-15).

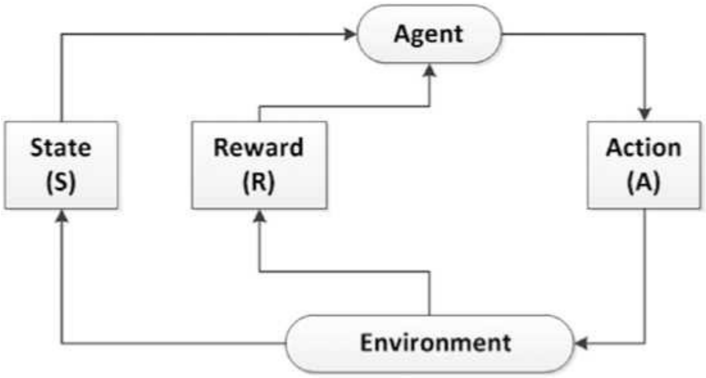


Figure 10-15. Reinforced learning diagram

The process is simple. Your agent extracts features from the environment that are either “state” or “reward" State features indicate that something has happened. Reward features indicate that something happened that has improved or worsened to the perceived gain in reward. The agent uses the state and reward to determine actions to change the environment.

This process of extracting state and reward, plus responding with action, will continue until a pre-agreed end reward is achieved. The real-world application for these types of reinforced learning is endless. You can apply reinforced learning to any environment which you can control with an agent.

I build many RL systems that monitor processes, such as a sorting system of purchases or assembly of products. It is also the core of most robot projects, as robots are physical agents that can interact with the environment. I also build many “soft-robots” that take decisions on such data processing as approval of loans, payments of money, and fixing of data errors.

12-Bagging Data

Bootstrap aggregating, also called bagging, is a machine-learning ensemble metaalgorithm that aims to advance the stability and accuracy of machine-learning algorithms used in statistical classification and regression. It decreases variance and supports systems to avoid overfitting.

I have seen many data science solutions over the last years that suffer from overfitting, because they were trained with a known data set that eventually became the only data set they could process. Thanks to inefficient processing and algorithms, we naturally had a lead way for variance in the data.

The new GPU (graphics processing unit)-based systems are so accurate that they overfit easily, if the training data is a consistent set, with little or no major changes in the patterns within the data set.

13-Random Forests

Random forests, or random decision forests, are an ensemble learning method for classification and regression that works by constructing a multitude of decision trees at training time and outputting the results the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

The result is an aggregation of all the trees' results, by performing a majority vote against the range of results. So, if five trees return three yeses and two nos, it passes a yes out of the Transform step.

Sometimes, this is also called tree bagging, as you take a bagging concept to the next level by not only training the model on a range of samples from the data population but by actually performing the complete process with the data bag and then aggregating the data results.

This is simply the daisy-chaining of a series of random forests to create a solution. I have found these to become more popular over the last two years, as solutions become more demanding and data sets become larger. The same principles apply; you are simply repeating them several times in a chain.

14-Computer Vision (CV)

Computer vision is a complex feature extraction area, but once you have the features exposed, it simply becomes a matrix of values.

15-Natural Language Processing (NLP)

Natural language processing is the area in data science that investigates the process we as humans use to communicate with each other. This covers mainly written and spoken words that form bigger concepts. Your data science is aimed at intercepting or interacting with humans, to react to the natural language.

There are two clear requirements in natural language processing. First is the direct interaction with humans, such as when you speak to your smartphone, and it responds with an appropriate answer. For example, you request “phone home," and the phone calls the number set as “home."

The second type of interaction is taking the detailed content of the interaction and understanding its context and relationship with other text or recorded information. Examples of these are news reports that are examined, and common trends are found among different news reports. This a study of the natural language's meaning, not simply a response to a human interaction.

- **Text-Based:** If you want to process text, you must set up an ecosystem to perform the basic text processing.
- **Speech-Based:** There is a major demand for speech-to-text conversion, to extract features.

16- Neural Networks

Neural networks (also known as artificial neural networks) are inspired by the human nervous system. They simulate how complex information is absorbed and processed by the human system. Just like humans, neural networks learn by example and are configured to a specific application.

Neural networks are used to find patterns in extremely complex data and, thus, deliver forecasts and classify data points. Neural networks are usually organized in layers. Layers are made up of a number of interconnected “nodes" Patterns (features) are presented to the network via the “input layer," which communicates to one or more “hidden layers," where the actual processing is done. The hidden layers then link to an “output layer," where the answer is output, as shown in Figure 10-16.

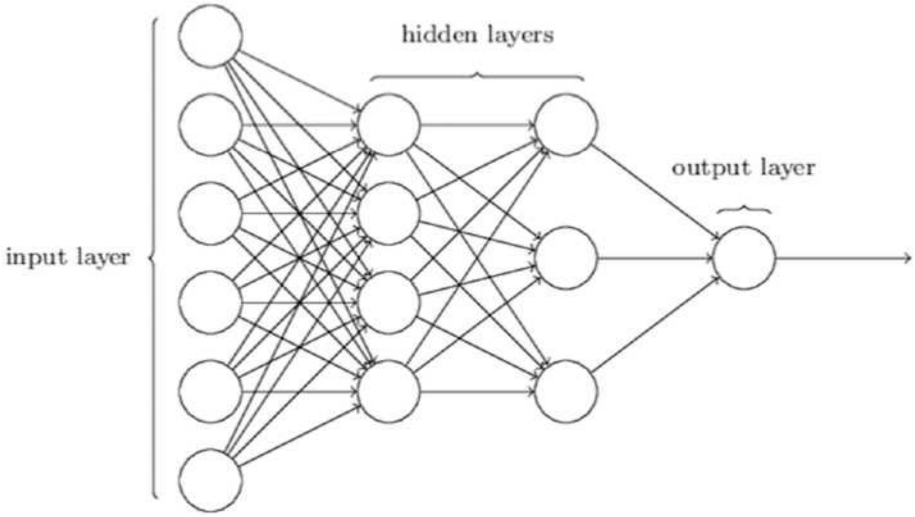


Figure 10-16 General artificial neural network

When you feed these to the neural network, it will use them as simple 0 or 1 values, and this is what neural networks really excel at solving. Unfortunately, the most important feature when buying a ball for a dog is “Does the dog fit under the house?” It took me two hours to retrieve one dog and one black ball from a space I did not fit!

The lesson: You must change criteria as you develop the neural network. If you keep the question simple, you can just add more questions, or even remove questions, that result in features.

Let me offer an example of feature development for neural networks. Suppose you have to select three colors for your dog's new ball: blue, yellow, or pink. The features would be

- Is the ball blue? Yes/No
- Is the ball yellow? Yes/No
- Is the ball pink? Yes/No

16.1-Gradient Descent

Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient of the function at the current point.

16.2-Regularization Strength

Regularization strength is the parameter that prevents overfitting of the neural network. The parameter enables the neural network to match the best set of weights for a general data set. The common name for this setting is the epsilon parameter, also known as the learning rate.

17- TensorFlow

TensorFlow is an open source software library for numerical computation using dataflow graphs. Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) communicated between them. The flexible architecture allows you to deploy computation to one or more CPUs or GPUs.

TensorFlow was originally developed by researchers and engineers working on the Google Brain Team within Google's machine intelligence research organization, for the purposes of conducting machine learning and deep neural networks research, but the system is general enough to be applicable in a wide variety of other domains.

The next big advantage is the Cloud Tensor Processing Unit (TPU) (https://cloud. google.com/tpu/) hardware product, which was specifically designed to calculate tensor processing at better performance levels than standard CPU or GPU hardware. The TPU supports the TensorFlow process with an extremely effective hardware ecosystem. The TensorFlow Research Cloud provides users with access to these second-generation cloud TPUs, each of which provides 180 teraflops of machine-learning acceleration.

Did you notice how with minor changes TensorFlow handles larger volumes of data with ease? The advantage of TensorFlow is the simplicity of the basic building block you use to create it, and the natural graph nature of the data pipelines, which enable you to easily convert data flows from the real world into complex simulations within the TensorFlow ecosystem.

TensorFlow process used mostly for the following three models:

- *Basket analysis:* What do people buy? What do they buy together?
- *Forex trading:* Providing recommendations on when to purchase forex for company requirements.
- *Commodities trading:* Buying and selling futures.

1-Organize Superstep

The Organize superstep takes the complete data warehouse you built at the end of the Transform superstep and subsections it into business-specific data marts. A data mart is the access layer of the data warehouse environment built to expose data to the users. The data mart is a subset of the data warehouse and is generally oriented to a specific business group.

1.1-Vertical Style

Performing vertical-style slicing or subsetting of the data warehouse is achieved by applying a filter technique that forces the data warehouse to show only the data for specific preselected filtered outcomes against the data population. The vertical-style slicing selects the subset of columns from the population, while preserving the rows. That is, the data science tool can see only the preselected columns from a record for all the records in the population.

1.2-Island Style

Performing island-style slicing or subsetting of the data warehouse is achieved by applying a combination of horizontal- and vertical-style slicing. This generates a subset of specific rows and specific columns reduced at the same time.

1.3-Secure Vault Style

The secure vault is a version of one of the horizontal, vertical, or island slicing techniques, but the outcome is also attached to the person who performs the query.

This is common in multi-security environments, where different users are allowed to see different data sets.

This process works well, if you use a role-based access control (RBAC) approach to restricting system access to authorized users. The security is applied against the “role," and a person can then, by the security system, simply be added or removed from the role, to enable or disable access.

The security in most data lakes I deal with is driven by an RBAC model that is an approach to restricting system access to authorized users by allocating them to a layer of roles that the data lake is organized into to support security access.

It is also possible to use a time-bound RBAC that has different access rights during office hours than after hours.

1.4-Association Rule Mining

Association rule learning is a rule-based machine-learning method for discovering interesting relations between variables in large databases, similar to the data you will find in a data lake. The technique enables you to investigate the interaction between data within the same population.

This example I will discuss is also called “market basket analysis." It will investigate the analysis of a customer's purchases during a period of time.

The new measure you need to understand is called “lift." Lift is simply estimated by the ratio of the joint probability of two items x and y, divided by the product of their individual probabilities:

$$Lift = \frac{\mathrm{P}(x,y)}{\mathrm{P}(x)P(y)}$$

If the two items are statistically independent, then P(x,y) = P(x)P(y), corresponding to Lift = 1, in that case. Note that anti-correlation yields lift values less than 1, which is also an interesting discovery, corresponding to mutually exclusive items that rarely cooccur.

The general algorithm used for this is the Apriori algorithm for frequent item set mining and association rule learning over the content of the data lake. It proceeds by identifying the frequent individual items in the data lake and extends them to larger and larger item sets, as long as those item sets appear satisfactorily frequently in the data lake. The frequent item sets determined by Apriori can be used to determine association rules that highlight common trends in the overall data lake.

2-Report Superstep

The Report superstep is the step in the ecosystem that enhances the data science findings with the art of storytelling and data visualization. You can perform the best data science, but if you cannot execute a respectable and trustworthy Report step by turning your data science into actionable business insights, you have achieved no advantage for your business.

2.1-Summary of the Results

The most important step in any analysis is the summary of the results. Your data science techniques and algorithms can produce the most methodically, most advanced mathematical or most specific statistical results to the requirements, but if you cannot summarize those into a good story, you have not achieved your requirements.

2.2-Understand the Context

What differentiates good data scientists from the best data scientists are not the algorithms or data engineering; it is the ability of the data scientist to apply the context of his findings to the customer.

Example:

The bar has served last rounds at 23:45 and there is one person left. The person drinking has less than 5% of the beer still in his glass at 23:50.

From the context of the drinker, he will have to drink slower to stay till midnight or drink the rest immediately, if he has to catch the midnight bus. From the context of the bar staff, they want to close at 23:45, and they have to get this last patron out the door to lock up, as they, too, want to catch the midnight bus.

It is clear that if you can determine if both parties want to be on the midnight bus, you will have the context of the amount of beer in the drinker's glass. The sole data science measure of the 5% left in the glass is of no value.

I had a junior data scientist try to use image processing from a brewery's CCTV system to determine the levels of people's beers, to enable his data science to notify the bar staff what their rate of beer consumption was during the night. The intention was to generate 3% more profit on the late shift.

The issue was more sinister, as every night, several local regulars would stay after midnight, to catch the new subsidized “Drive-Safe” night bus from the bus terminal across the street home after 12:30, hence not having to pay the normal bus fares. So, they got beer for less than their fare home and traveled at no cost, because they smelled like beer.

The brewery ended paying extra hours of overtime at double rates to six staff members in their three bars, owing to it being after midnight, and staff are not allowed to be in the bar alone with customers, for security reasons.

The staff was always late returning home, as, not qualifying for the late charities' bus, they had to use taxis. It was only after the brewery supplied the data scientist with the external CCTV that we found the real issue. The solution was to supply a coupon for a normal bus trip if customers took it before 23:30 and had beer at the brewery. The total profit equaled +7.3% on the late shift. No complex data science needs only some context.

I have seen too many data scientists spend hours producing great results using the most complex algorithms but being unable to articulate their findings to the business in a manner it understood. Or even worse, not able to get their results into production, because at a bigger scale, they simply did not work as designed by the data scientist— which immediately made the whole exercise pointless in the first place.

2.3-Appropriate Visualization

It is true that a picture tells a thousand words. But in data science, you only want your visualizations to tell one story: the findings of the data science you prepared. It is absolutely necessity to ensure that your audience will get your most important message clearly and without any other meanings.

2.4-Eliminate Clutter

Have you ever attended a presentation where the person has painstakingly prepared 50 slides to feedback his data science results? The most painful image is the faces of the people suffering through such a presentation for over two hours.

The biggest task of a data scientist is to eliminate clutter in the data sets. There are various algorithms, such as principal component analysis (PCA), multicollinearity using the variance inflation factor to eliminate dimensions and impute or eliminate missing values, decision trees to subdivide, and backward feature elimination, but the biggest contributor to eliminating clutter is good and solid feature engineering.

2.5-Draw Attention Where You Want It

Remember: Your purpose as a data scientist is to deliver insights to your customer, so that they can implement solutions to resolve a problem they may not even know about. You must place the attention on the insight and not the process. However, you must ensure that your process is verified and can support an accredited algorithm or technique.

2.6-Telling a Story (Freytag’s Pyramid)

Under Freytag's pyramid, the plot of a story consists of five parts: exposition, rising action, climax, falling action, and resolution. This is used by writers of books and screenplays as the basic framework of any story. In the same way, you must take your business through the data science process.

Exposition is the portion of a story that introduces important background information to the audience. In data science, you tell the background of the investigation you performed.

Rising action refers to a series of events that build toward the point of greatest interest. In data science, you point out the important findings or results. Keep it simple and to the point.

The climax is the turning point that determines a good or bad outcome for the story's characters. In data science, you show how your solution or findings will change the outcome of the work you performed.

During the falling action, the conflict between what occurred before and after the climax takes place. In data science, you prove that after your suggestion has been implemented in a pilot, the same techniques can be used to find the issues now proving that the issues can inevitably be resolved.

Resolution is the outcome of the story. In data science, you produce the solution and make the improvements permanent.

Decide what repeatable sound bite you can use to help your core message stick with your audience. People remember the core things by your repeating them. You must ensure one thing: that the customer remembers your core message.

Make sure you deliver that message clearly and without any confusion. Drive the core message home by repeating sound bites throughout the meeting.

The art of a good data scientist lies in the ability to tell a story about the data and get people to remember its baseline.