DOWNLOAD PDF MACHINE LEARNING MODELS AND ALGORITHMS FOR BIG DATA CLASSIFICATION

Chapter 1 : A Tour of The Top 10 Algorithms for Machine Learning Newbies

This book presents machine learning models and algorithms to address big data classification problems. Existing machine learning techniques like the decision tree (a hierarchical approach), random forest (an ensemble hierarchical approach), and deep learning (a layered approach) are highly suitable.

As Big Data is the hottest trend in the tech industry at the moment, machine learning is incredibly powerful to make predictions or calculated suggestions based on large amounts of data. So if you want to learn more about machine learning, how do you start? For me, my first introduction is when I took an Artificial Intelligence class when I was studying abroad in Copenhagen. My lecturer is a full-time Applied Math and CS professor at the Technical University of Denmark, in which his research areas are logic and artificial, focusing primarily on the use of logic to model human-like planning, reasoning and problem solving. The textbook that we used is one of the AI classics: At the end of the class, in a team of 3, we implemented simple search-based agents solving transportation tasks in a virtual environment as a programming project. I have learned a tremendous amount of knowledge thanks to that class, and decided to keep learning about this specialized topic. In the last few weeks, I have been multiple tech talks in San Francisco on deep learning, neural networks, data architecture â€" and a Machine Learning conference with a lot of well-known professionals in the field. In this post, I want to share some of the most common machine learning algorithms that I learned from the course. Machine learning algorithms can be divided into 3 broad categories â€" supervised learning, unsupervised learning, and reinforcement learning. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance-event outcomes, resource costs, and utility. Take a look at the image to get a sense of how it looks like. As a method, it allows you to approach the problem in a structured and systematic way to arrive at a logical conclusion. Naive Bayes Classification Some of real world examples are: To mark an email as spam or not spam Classify a news article about technology, politics, or sports Check a piece of text expressing positive emotions, or negative emotions? Used for face recognition software. Ordinary Least Squares Regression: If you know statistics, you probably have heard of linear regression before. Least squares is a method for performing linear regression. You can think of linear regression as the task of fitting a straight line through a set of points. Ordinary Least Squares Regression Linear refers the kind of model you are using to fit the data, while least squares refers to the kind of error metric you are minimizing over. Logistic regression is a powerful statistical way of modeling a binomial outcome with one or more explanatory variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Logistic Regression In general, regressions can be used in real-world applications such as: Credit Scoring Measuring the success rates of marketing campaigns Predicting the revenues of a certain product Is there going to be an earthquake on a particular day? SVM is binary classification algorithm. Given a set of points of 2 types in N dimensional place, SVM generates a N â€" 1 dimensional hyperplane to separate those points into 2 groups. Say you have some points of 2 types in a paper which are linearly separable. SVM will find a straight line which separates those points into 2 types and situated as far as possible from all those points. Support Vector Machine In terms of scale, some of the biggest problems that have been solved using SVMs with suitably modified implementations are display advertising, human splice site recognition, image-based gender detection, large-scale image classification

DOWNLOAD PDF MACHINE LEARNING MODELS AND ALGORITHMS FOR BIG DATA CLASSIFICATION

Chapter 2 : Machine Learning: What it is and why it matters | SAS

This book presents machine learning models and algorithms to address big data classification problems. Existing machine learning techniques like the decision tree (a hierarchical approach), random forest (an ensemble hierarchical approach), and deep learning (a layered approach) are highly suitable for the system that can handle such problems.

Get started with Amazon SageMaker Amazon SageMaker is a fully-managed platform that enables developers and data scientists to quickly and easily build, train, and deploy machine learning models at any scale. Amazon SageMaker removes all the barriers that typically slow down developers who want to use machine learning. Machine learning often feels a lot harder than it should be to most developers because the process to build and train models, and then deploy them into production is too complicated and too slow. First, you need to collect and prepare your training data to discover which elements of your data set are important. After deciding on your approach, you need to teach the model how to make predictions by training, which requires a lot of compute. Then, you need to tune the model so it delivers the best possible predictions, which is often a tedious and manual effort. All of this takes a lot of specialized expertise, access to large amounts of compute and storage, and a lot of time to experiment and optimize every part of the process. Amazon SageMaker removes the complexity that holds back developer success with each of these steps. Amazon SageMaker includes modules that can be used together or independently to build, train, and deploy your machine learning models. Introducing Amazon SageMaker How It Works Build Amazon SageMaker makes it easy to build ML models and get them ready for training by providing everything you need to quickly connect to your training data, and to select and optimize the best algorithm and framework for your application. Amazon SageMaker includes hosted Jupyter notebooks that make it easy to explore and visualize your training data stored in Amazon S3. You can also download these open source containers to your local environment and use the Amazon SageMaker Python SDK to test your scripts in local mode before using Amazon SageMaker for training or hosting your model in production. You also have the option of using your own framework. Train You can begin training your model with a single click in the Amazon SageMaker console. Amazon SageMaker manages all of the underlying infrastructure for you and can easily scale to train models at petabyte scale. To make the training process even faster and easier, Amazon SageMaker can automatically tune your model to achieve the highest possible accuracy. Deploy Once your model is trained and tuned, Amazon SageMaker makes it easy to deploy in production so you can start generating predictions a process called inference for real-time or batch data. Amazon SageMaker deploys your model on auto-scaling clusters of Amazon SageMaker ML instances that are spread across multiple availability zones to deliver both high performance and high availability. Amazon SageMaker takes away the heavy lifting of machine learning, so you can build, train, and deploy machine learning models quickly and easily. Benefits Get to Production with Machine Learning Quickly Amazon SageMaker significantly reduces the amount of time needed to train, tune, and deploy machine learning models. Amazon SageMaker manages and automates all the sophisticated training and tuning techniques so you can get models into production quickly. Choose Any Framework or Algorithm Amazon SageMaker supports all machine algorithms and frameworks so you can use the technology you are already familiar with. If you want to train with an alternative framework or algorithm, you can bring your own in a Docker container. Easily Integrate With Your Existing Workflow Amazon SageMaker is designed in three modules that can be used together or independently as part of any existing ML workflow you might already have in place. Featured SageMaker Customers "Harnessing data and analytics across hardware, software and biotech, GE Healthcare is transforming healthcare by delivering better outcomes for providers and patients. The scalability of Amazon SageMaker, and its ability to integrate with native AWS services, adds enormous value for us. We are excited about how our continued collaboration between the GE Health Cloud and Amazon SageMaker will drive better outcomes for our healthcare provider partners and deliver improved patient care. Machine Learning and other computations that could take months

to refine now take weeks or days, allowing us to engage, inform and excite the fan in new and unique ways. Amazon SageMaker simplifies machine learning, helping our development teams to build models for predictions that create new connections that otherwise might have never been possible. We will create novel large-scale machine learning and AI algorithms and deploy them on this platform to solve complex problems that can power prosperity for our customers. Working with Amazon SageMaker enabled us to design a natural language processing capability in the context of a question answering application. With Amazon SageMaker, the distributed training, optimized algorithms, and built-in hyperparameter features should allow my team to quickly build more accurate models on our largest data sets, reducing the considerable time it takes us to move a model to production. It is simply an API call. Amazon SageMaker will significantly reduce the complexity of machine learning, enabling us to create a better experience for our customers, fast. We chose AWS because of their strength, depth, and proven expertise in delivering machine learning services to create new opportunities to share never-before-seen metrics. Amazon SageMaker can easily train and deploy machine learning models which can more effectively target online ads, providing better customer engagement and conversion. Once built, models can be hosted easily in low-latency, auto-scaling endpoints, or passed to other real-time bidding systems. Credit Default Prediction Amazon SageMaker makes it easier to predict the likelihood of credit default, a common machine learning problem. Amazon SageMaker integrates tightly with existing analytical frameworks like Amazon Redshift, Amazon EMR, and AWS Glue, allowing you to publish large, diverse datasets into an Amazon S3 data lake, then transform them quickly, build machine learning models, and immediately host them for online prediction. Industrial IoT and Machine Learning Industrial IoT and machine learning can enable real-time predictions to anticipate machinery failure or maintenance scheduling, to achieve higher levels of efficiency. A digital twin, or replica, of physical assets, processes, or systems, can be generated as models to predict preventive maintenance or to optimize output of complex machines or industrial processes. Supply Chain and Demand Forecasting Amazon SageMaker provides the infrastructure and algorithms needed to develop individual sales forecasts for every product in the largest ecommerce settings. With time series and product category data alone, Amazon SageMaker picks up on seasonality, trends, and product similarities to deliver accurate forecasts, even for new items. Click-through Prediction Amazon SageMaker provides both single machine and distributed CPU implementations of XGboost algorithms, which are useful in multiple classification, regression, and ranking use-cases, such as ad click-through rate prediction. Click prediction systems are central to most online advertising systems, since it is crucial to predict the most accurate click-through rate CTR as possible to ensure consumers have the best experience. Using the XGBoost algorithm, you can run a real-time predictor and return a scored prediction result. You can then determine whether or not to serve ads from a particular advertiser and improve your CTR prediction in display ads. Predicting Quality of Content Amazon SageMaker has a number of tools for pre-processing and finding structures within text, using that information to make predictions about content quality. You can generate word embeddings to find similar semantic and syntactic words in large text volumes, and group together similar words to avoid sparsity. Finally, build independent classification models by cluster on the reduced dimensional grouped word data to determine whether documents need to be moderated. Learn more about Amazon SageMaker.

Chapter 3 : The 10 Algorithms Machine Learning Engineers Need to Know

Algorithms Grouped by Learning Style. There are different ways an algorithm can model a problem based on its interaction with the experience or environment or whatever we want to call the input data.

With the rapid growth of big data and availability of programming tools like Python and R â€”machine learning is gaining mainstream presence for data scientists. Machine learning applications are highly automated and self-modifying which continue to improve over time with minimal human intervention as they learn with more data. To address the complex nature of various real world data problems, specialized machine learning algorithms have been developed that solve these problems perfectly. Machine Learning algorithms are classified as â€" 1 Supervised Machine Learning Algorithms Machine learning algorithms that make predictions on given set of samples. Supervised machine learning algorithm searches for patterns within the value labels assigned to data points. These machine learning algorithms organize the data into a group of clusters to describe its structure and make complex data look simple and organized for analysis. Over time, the algorithm changes its strategy to learn better and achieve the best reward. What other machine learning algorithms do you think should have been on the list? Would you like to be updated when other readers reply to this question? It is a simple classification of words based on Bayes Probability Theorem for subjective analysis of content. If you have a moderate or large training data set. If the instances have several attributes. Given the classification parameter, attributes which describe the instances should be conditionally independent. Document Categorization- Google uses document classification to index documents and find relevancy scores i. PageRank mechanism considers the pages marked as important in the databases that were parsed and classified using a document classification technique. A good bet for multi class predictions as well. K-Means is a non-deterministic and iterative method. The algorithm operates on a given data set through pre-defined number of clusters, k. The output of K Means algorithm is k clusters with input data partitioned among the clusters. K Means clustering algorithm can be applied to group the webpages that talk about similar concepts. So, the algorithm will group all web pages that talk about Jaguar as an Animal into one cluster, Jaguar as a Car into another cluster and so on. Given a smaller value of K, K-Means clustering computes faster than hierarchical clustering for large number of variables. This helps search engines reduce the computational time for the users. It works by classifying the data into different classes by finding a line hyperplane which separates the training data set into classes. As there are many such linear hyperplanes, SVM algorithm tries to maximize the distance between the various classes that are involved and this is referred as margin maximization. If the line that maximizes the distance between the classes is identified, the probability to generalize well to unseen data is increased. For example, the training data for Face detection consists of group of images that are faces and another group of images that are not faces in other words all other images in the world except faces. Under such conditions, the training data is too complex that it is impossible to find a representation for every feature vector. Separating the set of faces linearly from the set of non-face is a complex task. SVM renders more efficiency for correct classification of the future data. The best thing about SVM is that it does not make any strong assumptions on data. It does not over-fit the data. Applications of Support Vector Machine SVM is commonly used for stock market forecasting by various financial institutions. For instance, it can be used to compare the relative performance of the stocks when compared to performance of other stocks in the same sector. The relative comparison of stocks helps manage investment making decisions based on the classifications made by the SVM learning algorithm. Association rule implies that if an item A occurs, then item B also occurs with a certain probability. For the algorithm to derive such conclusions, it first observes the number of people who bought an iPad case while purchasing an iPad. This way a ratio is derived like out of the people who purchased an iPad, 85 people also purchased an iPad case. Basic principle on which Apriori Machine Learning Algorithm works: If an item set occurs frequently then all the subsets of the item set, also occur frequently. If an item set occurs infrequently then all the supersets of the

item set have infrequent occurrence. Advantages of Apriori Algorithm It is easy to implement and can be parallelized easily. Apriori implementation makes use of large item set properties. Applications of Apriori Algorithm Detecting Adverse Drug Reactions Apriori algorithm is used for association analysis on healthcare data like-the drugs taken by patients, characteristics of each patient, adverse ill-effects patients experience, initial diagnosis, etc. This analysis produces association rules that help identify the combination of patient characteristics and medications that lead to adverse side effects of the drugs. Market Basket Analysis Many e-commerce giants like Amazon use Apriori to draw data insights on which products are likely to be purchased together and which are most responsive to promotion. For example, a retailer might use Apriori to predict that people who buy sugar and flour are likely to buy eggs to bake a cake. Auto-Complete Applications Google auto-complete is another popular application of Apriori wherein - when the user types a word, the search engine looks for other associated words that people usually type after a specific word. The algorithm shows the impact on the dependent variable on changing the independent variable. The independent variables are referred as explanatory variables, as they explain the factors the impact the dependent variable. Dependent variable is often referred to as the factor of interest or predictor. Advantages of Linear Regression Machine Learning Algorithm It is one of the most interpretable machine learning algorithms, making it easy to explain to others. It is easy of use as it requires minimal tuning. It is the mostly widely used machine learning technique that runs fast. Applications of Linear Regression Estimating Sales Linear Regression finds great use in business, for sales forecasting based on the trends. If a company observes steady increase in sales every month - a linear regression analysis of the monthly sales data helps the company forecast sales in upcoming months. A health insurance company can do a linear regression analysis on the number of claims per customer against age. This analysis helps insurance companies find, that older customers tend to make more insurance claims. Such analysis results play a vital role in important business decisions and are made to account for risk. Data Science Libraries in Python to implement Linear Regression â€" statsmodel and SciKit Data Science Libraries in R to implement Linear Regression â€" stats Explanations about the top machine learning algorithms will continue, as it is a work in progress. Stay tuned to our blog to learn more about the popular machine learning algorithms and their applications!!! Whenever you want to visit a restaurant you ask your friend Tyrion if he thinks you will like a particular place. To answer your question, Tyrion first has to find out, the kind of restaurants you like. You give him a list of restaurants that you have visited and tell him whether you liked each restaurant or not giving a labelled training dataset. Tyrion asks you several informative questions to maximize the information gain and gives you YES or NO answer based on your answers to the questionnaire. Here Tyrion is a decision tree for your favourite restaurant preferences. A decision tree is a graphical representation that makes use of branching methodology to exemplify all possible outcomes of a decision, based on certain conditions. In a decision tree, the internal node represents a test on the attribute, each branch of the tree represents the outcome of the test and the leaf node represents a particular class label i. The classification rules are represented through the path from root to the leaf node. Types of Decision Trees Classification Trees- These are considered as the default kind of decision trees used to separate a dataset into different classes, based on the response variable. These are generally used when the response variable is categorical in nature. Regression Trees-When the response or target variable is continuous or numerical, regression trees are used. These are generally used in predictive type of problems when compared to classification. Decision trees can also be classified into two types, based on the type of target variable-Continuous Variable Decision Trees and Binary Variable Decision Trees. It is the target variable that helps decide what kind of decision tree would be required for a particular problem. Why should you use Decision Tree Machine Learning algorithm? These machine learning algorithms help make decisions under uncertainty and help you improve communication, as they present a visual representation of a decision situation. Decision tree machine learning algorithms help a data scientist capture the idea that if a different decision was taken, then how the operational nature of a situation or model would have changed intensely. Decision tree algorithms help make optimal decisions by allowing a data scientist to traverse through forward and backward

calculation paths. When to use Decision Tree Machine Learning Algorithm Decision trees are robust to errors and if the training data contains errors- decision tree algorithms will be best suited to address such problems. Decision trees are best suited for problems where instances are represented by attribute value pairs. If the training data has missing value then decision trees can be used, as they can handle missing values nicely by looking at the data in other columns. Decision trees are best suited when the target function has discrete output values. Advantages of Using Decision Tree Machine Learning Algorithms Decision trees are very instinctual and can be explained to anyone with ease. People from a non-technical background, can also decipher the hypothesis drawn from a decision tree, as they are self-explanatory. When using decision tree machine learning algorithms, data type is not a constraint as they can handle both categorical and numerical variables. Decision tree machine learning algorithms do not require making any assumption on the linearity in the data and hence can be used in circumstances where the parameters are non-linearly related. These machine learning algorithms do not make any assumptions on the classifier structure and space distribution. These algorithms are useful in data exploration. Decision trees implicitly perform feature selection which is very important in predictive analytics. When a decision tree is fit to a training dataset, the nodes at the top on which the decision tree is split, are considered as important variables within a given dataset and feature selection is completed by default. Decision trees help save data preparation time, as they are not sensitive to missing values and outliers. Missing values will not stop you from splitting the data for building a decision tree. Outliers will also not affect the decision trees as data splitting happens based on some samples within the split range and not on exact absolute values. Drawbacks of Using Decision Tree Machine Learning Algorithms The more the number of decisions in a tree, less is the accuracy of any expected outcome. A major drawback of decision tree machine learning algorithms, is that the outcomes may be based on expectations. When decisions are made in real-time, the payoffs and resulting outcomes might not be the same as expected or planned. There are chances that this could lead to unrealistic decision trees leading to bad decision making. Any irrational expectations could lead to major errors and flaws in decision tree analysis, as it is not always possible to plan for all eventualities that can arise from a decision.

Chapter 4 : Predictive analytics - Wikipedia

As Big Data is the hottest trend in the tech industry at the moment, machine learning is incredibly powerful to make predictions or calculated suggestions based on large amounts of data.

Overview[edit] Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field: Machine learning tasks[edit] Machine learning tasks are typically classified into several broad categories: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. As special cases, the input signal can be only partially available, or restricted to special feedback. The computer is given only an incomplete training signal: The computer can only obtain training labels for a limited set of instances based on a budget , and also has to optimize its choice of objects to acquire labels for. When used interactively, these can be presented to the user for labeling. No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself discovering hidden patterns in data or a means towards an end feature learning. Here, it has learned to distinguish black and white circles. Another categorization of machine learning tasks arises when one considers the desired output of a machine-learned system: This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email or other messages and the classes are "spam" and "not spam". In regression , also a supervised problem, the outputs are continuous rather than discrete. In clustering , a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task. Density estimation finds the distribution of inputs in some space. Dimensionality reduction simplifies inputs by mapping them into a lower-dimensional space. Topic modeling is a related problem, where a program is given a list of human language documents and is tasked to find out which documents cover similar topics. Among other categories of machine learning problems, learning to learn learns its own inductive bias based on previous experience. Developmental learning , elaborated for robot learning , generates its own sequences also called curriculum of learning situations to cumulatively acquire repertoires of novel skills through autonomous self-exploration and social interaction with human teachers and using guidance mechanisms such as active learning, maturation, motor synergies, and imitation. History and relationships to other fields[edit] See also: Timeline of machine learning Arthur Samuel , an American pioneer in the field of computer gaming and artificial intelligence , coined the term "Machine Learning" in while at IBM [11]. As a scientific endeavour, machine learning grew out of the quest for artificial intelligence. Already in the early days of AI as an academic discipline, some researchers were interested in having machines learn from data. They attempted to approach the problem with various symbolic methods, as well as what were then termed " neural networks "; these were mostly perceptrons and other models that were later found to be reinventions of the generalized linear models of statistics. Probabilistic systems were plagued by theoretical and practical problems of data acquisition and representation. Their main success came in the mids with the reinvention of backpropagation. The field changed its goal from achieving artificial intelligence to tackling solvable problems of a practical nature. It shifted focus away from the symbolic approaches it had inherited from AI, and toward methods and models borrowed from statistics and probability theory. Relation to data mining[edit] Machine learning and data mining often employ the same methods and overlap significantly, but while machine learning focuses on prediction, based on known properties learned from the training data, data mining focuses on the discovery of previously unknown properties in the data this is the analysis step of knowledge discovery in databases. Data mining uses many machine learning methods, but with different goals; on the other hand, machine learning also employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy. Much of the confusion between these two research communities which do often have separate conferences and separate journals, ECML PKDD being a major exception comes from the basic assumptions they work with: Evaluated with respect to known knowledge, an

uninformed unsupervised method will easily be outperformed by other supervised methods, while in a typical KDD task, supervised methods cannot be used due to the unavailability of training data. Relation to optimization[edit] Machine learning also has intimate ties to optimization: Loss functions express the discrepancy between the predictions of the model being trained and the actual problem instances for example, in classification, one wants to assign a label to instances, and models are trained to correctly predict the pre-assigned labels of a set of examples. The difference between the two fields arises from the goal of generalization: According to Michael I. Jordan , the ideas of machine learning, from methodological principles to theoretical tools, have had a long pre-history in statistics. Some statisticians have adopted methods from machine learning, leading to a combined field that they call statistical learning. Computational learning theory A core objective of a learner is to generalize from its experience. The training examples come from some generally unknown probability distribution considered representative of the space of occurrences and the learner has to build a general model about this space that enables it to produce sufficiently accurate predictions in new cases. The computational analysis of machine learning algorithms and their performance is a branch of theoretical computer science known as computational learning theory. Because training sets are finite and the future is uncertain, learning theory usually does not yield guarantees of the performance of algorithms. Instead, probabilistic bounds on the performance are quite common. The biasâ€”variance decomposition is one way to quantify generalization error. For the best performance in the context of generalization, the complexity of the hypothesis should match the complexity of the function underlying the data. If the hypothesis is less complex than the function, then the model has underfit the data. If the complexity of the model is increased in response, then the training error decreases. But if the hypothesis is too complex, then the model is subject to overfitting and generalization will be poorer. In computational learning theory, a computation is considered feasible if it can be done in polynomial time. There are two kinds of time complexity results. Positive results show that a certain class of functions can be learned in polynomial time. Negative results show that certain classes cannot be learned in polynomial time.

Chapter 5 : Machine learning - Wikipedia

With machine learning and artificial intelligence, a data scientist can make his or her work of process Big Data easily. Considering the volume of data sets, software models and conventional databases turn out to be less effective.

There are many factors at play, such as the size and structure of your dataset. Of course, the algorithms you try must be appropriate for your problem, which is where picking the right machine learning task comes in. The Big Principle However, there is a common principle that underlies all supervised machine learning algorithms for predictive modeling. Machine learning algorithms are described as learning a target function f that best maps input variables X to an output variable Y: If we did, we would use it directly and we would not need to learn it from data using machine learning algorithms. This is called predictive modeling or predictive analytics and our goal is to make the most accurate predictions possible. For machine learning newbies who are eager to understand the basic of machine learning, here is a quick tour on the top 10 machine learning algorithms used by data scientists. Predictive modeling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explainability. We will borrow, reuse and steal algorithms from many different fields, including statistics and use them towards these ends. The representation of linear regression is an equation that describes a line that best fits the relationship between the input variables x and the output variables y , by finding specific weightings for the input variables called coefficients B. Linear Regression For example: Different techniques can be used to learn the linear regression model from data, such as a linear algebra solution for ordinary least squares and gradient descent optimization. Linear regression has been around for more than years and has been extensively studied. Some good rules of thumb when using this technique are to remove variables that are very similar correlated and to remove noise from your data, if possible. It is a fast and simple technique and good first algorithm to try. It is the go-to method for binary classification problems problems with two class values. Logistic regression is like linear regression in that the goal is to find the values for the coefficients that weight each input variable. Unlike linear regression, the prediction for the output is transformed using a non-linear function called the logistic function. The logistic function looks like a big S and will transform any value into the range 0 to 1. This is useful because we can apply a rule to the output of the logistic function to snap values to 0 and 1 e. IF less than 0. Logistic Regression Because of the way that the model is learned, the predictions made by logistic regression can also be used as the probability of a given data instance belonging to class 0 or class 1. This can be useful for problems where you need to give more rationale for a prediction. Like linear regression, logistic regression does work better when you remove attributes that are unrelated to the output variable as well as attributes that are very similar correlated to each other. If you have more than two classes then the Linear Discriminant Analysis algorithm is the preferred linear classification technique. The representation of LDA is pretty straight forward. It consists of statistical properties of your data, calculated for each class. For a single input variable this includes: The mean value for each class. The variance calculated across all classes. Linear Discriminant Analysis Predictions are made by calculating a discriminate value for each class and making a prediction for the class with the largest value. The technique assumes that the data has a Gaussian distribution bell curve , so it is a good idea to remove outliers from your data before hand. The representation of the decision tree model is a binary tree. This is your binary tree from algorithms and data structures, nothing too fancy. Each node represents a single input variable x and a split point on that variable assuming the variable is numeric. Decision Tree The leaf nodes of the tree contain an output variable y which is used to make a prediction. Predictions are made by walking the splits of the tree until arriving at a leaf node and output the class value at that leaf node. Trees are fast to learn and very fast for making predictions. They are also often accurate for a broad range of problems and do not require any special preparation for your data. The model is comprised of two types of probabilities that can be calculated directly from your training data: Once calculated, the probability model can be used to make predictions for new data using Bayes Theorem. When

your data is real-valued it is common to assume a Gaussian distribution bell curve so that you can easily estimate these probabilities. Bayes Theorem Naive Bayes is called naive because it assumes that each input variable is independent. This is a strong assumption and unrealistic for real data, nevertheless, the technique is very effective on a large range of complex problems. The model representation for KNN is the entire training dataset. Predictions are made for a new data point by searching through the entire training set for the K most similar instances the neighbors and summarizing the output variable for those K instances. For regression problems, this might be the mean output variable, for classification problems this might be the mode or most common class value. The trick is in how to determine the similarity between the data instances. The simplest technique if your attributes are all of the same scale all in inches for example is to use the Euclidean distance, a number you can calculate directly based on the differences between each input variable. K-Nearest Neighbors KNN can require a lot of memory or space to store all of the data, but only performs a calculation or learn when a prediction is needed, just in time. You can also update and curate your training instances over time to keep predictions accurate. The idea of distance or closeness can break down in very high dimensions lots of input variables which can negatively affect the performance of the algorithm on your problem. This is called the curse of dimensionality. It suggests you only use those input variables that are most relevant to predicting the output variable. The Learning Vector Quantization algorithm or LVQ for short is an artificial neural network algorithm that allows you to choose how many training instances to hang onto and learns exactly what those instances should look like. These are selected randomly in the beginning and adapted to best summarize the training dataset over a number of iterations of the learning algorithm. After learned, the codebook vectors can be used to make predictions just like K-Nearest Neighbors. The most similar neighbor best matching codebook vector is found by calculating the distance between each codebook vector and the new data instance. The class value or real value in the case of regression for the best matching unit is then returned as the prediction. Best results are achieved if you rescale your data to have the same range, such as between 0 and 1. If you discover that KNN gives good results on your dataset try using LVQ to reduce the memory requirements of storing the entire training dataset. A hyperplane is a line that splits the input variable space. In SVM, a hyperplane is selected to best separate the points in the input variable space by their class, either class 0 or class 1. The SVM learning algorithm finds the coefficients that results in the best separation of the classes by the hyperplane. The best or optimal hyperplane that can separate the two classes is the line that has the largest margin. Only these points are relevant in defining the hyperplane and in the construction of the classifier. These points are called the support vectors. They support or define the hyperplane. In practice, an optimization algorithm is used to find the values for the coefficients that maximizes the margin. SVM might be one of the most powerful out-of-the-box classifiers and worth trying on your dataset. It is a type of ensemble machine learning algorithm called Bootstrap Aggregation or bagging. The bootstrap is a powerful statistical method for estimating a quantity from a data sample. Such as a mean. You take lots of samples of your data, calculate the mean, then average all of your mean values to give you a better estimation of the true mean value. In bagging, the same approach is used, but instead for estimating entire statistical models, most commonly decision trees. Multiple samples of your training data are taken then models are constructed for each data sample. When you need to make a prediction for new data, each model makes a prediction and the predictions are averaged to give a better estimate of the true output value. Random Forest Random forest is a tweak on this approach where decision trees are created so that rather than selecting optimal split points, suboptimal splits are made by introducing randomness. The models created for each sample of the data are therefore more different than they otherwise would be, but still accurate in their unique and different ways. Combining their predictions results in a better estimate of the true underlying output value. If you get good results with an algorithm with high variance like decision trees , you can often get better results by bagging that algorithm. This is done by building a model from the training data, then creating a second model that attempts to correct the errors from the first model. Models are added until the training set is predicted perfectly or a maximum number of models are added. AdaBoost was the first really successful boosting algorithm

developed for binary classification. It is the best starting point for understanding boosting. Modern boosting methods build on AdaBoost, most notably stochastic gradient boosting machines. AdaBoost AdaBoost is used with short decision trees. After the first tree is created, the performance of the tree on each training instance is used to weight how much attention the next tree that is created should pay attention to each training instance. Training data that is hard to predict is given more weight, whereas easy to predict instances are given less weight. Models are created sequentially one after the other, each updating the weights on the training instances that affect the learning performed by the next tree in the sequence. After all the trees are built, predictions are made for new data, and the performance of each tree is weighted by how accurate it was on training data. Because so much attention is put on correcting mistakes by the algorithm it is important that you have clean data with outliers removed. Even an experienced data scientist cannot tell which algorithm will perform the best before trying different algorithms. Although there are many other Machine Learning algorithms, these are the most popular ones. You can find my own code on GitHub , and more of my writing and projects at https: You can also follow me on Twitter , email me directly or find me on LinkedIn. Sign up for my newsletter to receive my latest thoughts on data science, machine learning, and artificial intelligence right at your inbox!

Chapter 6 : Top 10 Machine Learning Algorithms

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify.

Definition[edit] Predictive analytics is an area of statistics that deals with extracting information from data and using it to predict trends and behavior patterns. The enhancement of predictive web analytics calculates statistical probabilities of future events online. Predictive analytics statistical techniques include data modeling, machine learning, AI, deep learning algorithms and data mining. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions. Predictive analytics is often defined as predicting at a more detailed level of granularity, i. This distinguishes it from forecasting. For example, "Predictive analyticsâ€”Technology that learns from experience data to predict the future behavior of individuals in order to drive better decisions. Define the project outcomes, deliverable, scope of the effort, business objectives, identify the data sets that are going to be used. Data mining for predictive analytics prepares data from multiple sources for analysis. This provides a complete view of customer interactions. Data Analysis is the process of inspecting, cleaning and modelling data with the objective of discovering useful information, arriving at conclusion Statistics: Statistical Analysis enables to validate the assumptions, hypothesis and test them using standard statistical models. Predictive modelling provides the ability to automatically create accurate predictive models about future. There are also options to choose the best solution with multi-modal evaluation. Predictive model deployment provides the option to deploy the analytical results into everyday decision making process to get results, reports and output by automating the decisions based on the modelling. Models are managed and monitored to review the model performance to ensure that it is providing the results expected. Types[edit] Generally, the term predictive analytics is used to mean predictive modeling , "scoring" data with predictive models, and forecasting. However, people are increasingly using the term to refer to related analytical disciplines, such as descriptive modeling and decision modeling or optimization. These disciplines also involve rigorous data analysis, and are widely used in business for segmentation and decision making, but have different purposes and the statistical techniques underlying them vary. Predictive models[edit] Predictive models are models of the relation between the specific performance of a unit in a sample and one or more known attributes or features of the unit. The objective of the model is to assess the likelihood that a similar unit in a different sample will exhibit the specific performance. This category encompasses models in many areas, such as marketing, where they seek out subtle data patterns to answer questions about customer performance, or fraud detection models. Predictive models often perform calculations during live transactions, for example, to evaluate the risk or opportunity of a given customer or transaction, in order to guide a decision. With advancements in computing speed, individual agent modeling systems have become capable of simulating human behaviour or reactions to given stimuli or scenarios. The available sample units with known attributes and known performances is referred to as the "training sample". The units in other samples, with known attributes but unknown performances, are referred to as "out of [training] sample" units. The out of sample units do not necessarily bear a chronological relation to the training sample units. For example, the training sample may consist of literary attributes of writings by Victorian authors, with known attribution, and the out-of sample unit may be newly found writing with unknown authorship; a predictive model may aid in attributing a work to a known author. Another example is given by analysis of blood splatter in simulated crime scenes in which the out of sample unit is the actual blood splatter pattern from a crime scene. The out of sample unit may be from the same time as the training units, from a previous time, or from a future time. Descriptive models[edit] Descriptive models quantify relationships in data in a way that is often used to classify customers or prospects into groups. Unlike predictive models that focus on predicting a single customer behavior such as credit risk ,

descriptive models identify many different relationships between customers or products. Descriptive models do not rank-order customers by their likelihood of taking a particular action the way predictive models do. Instead, descriptive models can be used, for example, to categorize customers by their product preferences and life stage. Descriptive modeling tools can be utilized to develop further models that can simulate large number of individualized agents and make predictions. Decision model Decision models describe the relationship between all the elements of a decisionâ€"the known data including results of predictive models , the decision, and the forecast results of the decisionâ€"in order to predict the results of decisions involving many variables. These models can be used in optimization, maximizing certain outcomes while minimizing others. Decision models are generally used to develop decision logic or a set of business rules that will produce the desired action for every customer or circumstance. Applications[edit] Although predictive analytics can be put to use in many applications, we outline a few examples where predictive analytics has shown positive impact in recent years. Analytical customer relationship management CRM [edit] Analytical customer relationship management CRM is a frequent commercial application of predictive analysis. Methods of predictive analysis are applied to customer data to pursue CRM objectives, which involve constructing a holistic view of the customer no matter where their information resides in the company or the department involved. CRM uses predictive analysis in applications for marketing campaigns, sales, and customer services to name a few. These tools are required in order for a company to posture and focus their efforts effectively across the breadth of their customer base. Several of the application areas described below direct marketing, cross-sell, customer retention are part of customer relationship management. Child protection[edit] Over the last 5 years, some child welfare agencies have started using predictive analytics to flag high risk cases. Additionally, sophisticated clinical decision support systems incorporate predictive analytics to support medical decision making at the point of care. A working definition has been proposed by Jerome A. It encompasses a variety of tools and interventions such as computerized alerts and reminders, clinical guidelines, order sets, patient data reports and dashboards, documentation templates, diagnostic support, and clinical workflow tools. They employed classical model-based and machine learning model-free methods to discriminate between different patient and control groups. Collection analytics[edit] Many portfolios have a set of delinquent customers who do not make their payments on time. The financial institution has to undertake collection activities on these customers to recover the amounts due. A lot of collection resources are wasted on customers who are difficult or impossible to recover. Predictive analytics can help optimize the allocation of collection resources by identifying the most effective collection agencies, contact strategies, legal actions and other strategies to each customer, thus significantly increasing recovery at the same time reducing collection costs. Cross-sell[edit] Often corporate organizations collect and maintain abundant data e. Customer retention[edit] With the number of competing services available, businesses need to focus efforts on maintaining continuous customer satisfaction , rewarding consumer loyalty and minimizing customer attrition. In addition, small increases in customer retention have been shown to increase profits disproportionately. Proper application of predictive analytics can lead to a more proactive retention strategy. Silent attrition, the behavior of a customer to slowly but steadily reduce usage, is another problem that many companies face. Predictive analytics can also predict this behavior, so that the company can take proper actions to increase customer activity. Direct marketing[edit] When marketing consumer products and services, there is the challenge of keeping up with competing products and consumer behavior. Apart from identifying prospects, predictive analytics can also help to identify the most effective combination of product versions, marketing material, communication channels and timing that should be used to target a given consumer. The goal of predictive analytics is typically to lower the cost per order or cost per action. Fraud detection[edit] Fraud is a big problem for many businesses and can be of various types: Some examples of likely victims are credit card issuers , insurance companies, [26] retail merchants, manufacturers, business-to-business suppliers and even services providers. Predictive modeling can also be used to identify high-risk fraud candidates in business or the public sector. Mark Nigrini developed a risk-scoring method to identify audit targets. He describes the use of this approach to detect fraud in the

franchisee sales reports of an international fast-food chain. Each location is scored using 10 predictors. The 10 scores are then weighted to give one final overall risk score for each location. The same scoring approach was also used to identify high-risk check kiting accounts, potentially fraudulent travel agents, and questionable vendors. A reasonably complex model was used to identify fraudulent monthly reports submitted by divisional controllers. This type of solution utilizes heuristics in order to study normal web user behavior and detect anomalies indicating fraud attempts. Portfolio, product or economy-level prediction[edit] Often the focus of analysis is not the consumer but the product, portfolio, firm, industry or even the economy. For example, a retailer might be interested in predicting store-level demand for inventory management purposes. Or the Federal Reserve Board might be interested in predicting the unemployment rate for the next year. These types of problems can be addressed by predictive analytics using time series techniques see below. They can also be addressed via machine learning approaches which transform the original time series into a feature vector space, where the learning algorithm finds patterns that have predictive power. Project risk management When employing risk management techniques, the results are always to predict and benefit from a future scenario. The capital asset pricing model CAP-M "predicts" the best portfolio to maximize return. Probabilistic risk assessment PRA when combined with mini- Delphi techniques and statistical approaches yields accurate forecasts. These are examples of approaches that can extend from project to market, and from near to long term. Underwriting see below and other business approaches identify risk management as a predictive method. Underwriting[edit] Many businesses have to account for risk exposure due to their different services and determine the cost needed to cover the risk. For example, auto insurance providers need to accurately determine the amount of premium to charge to cover each automobile and driver. For a health insurance provider, predictive analytics can analyze a few years of past medical claims data, as well as lab, pharmacy and other records where available, to predict how expensive an enrollee is likely to be in the future. Predictive analytics can help underwrite these quantities by predicting the chances of illness, default , bankruptcy , etc. Predictive analytics can streamline the process of customer acquisition by predicting the future risk behavior of a customer using application level data. Proper predictive analytics can lead to proper pricing decisions, which can help mitigate future risk of default. Technology and big data influences[edit] Big data is a collection of data sets that are so large and complex that they become awkward to work with using traditional database management tools. The volume, variety and velocity of big data have introduced challenges across the board for capture, storage, search, sharing, analysis, and visualization. Examples of big data sources include web logs , RFID , sensor data, social networks , Internet search indexing, call detail records, military surveillance, and complex data in astronomic, biogeochemical, genomics, and atmospheric sciences. Big Data is the core of most predictive analytic services offered by IT organizations. Regression techniques[edit] Regression models are the mainstay of predictive analytics. The focus lies on establishing a mathematical equation as a model to represent the interactions between the different variables in consideration. Depending on the situation, there are a wide variety of models that can be applied while performing predictive analytics. Some of them are briefly discussed below. Linear regression model[edit] The linear regression model analyzes the relationship between the response or dependent variable and a set of independent or predictor variables. This relationship is expressed as an equation that predicts the response variable as a linear function of the parameters.

Chapter 7 : classification and clustering algorithms

Support Vector Machine is a supervised machine learning algorithm for classification or regression problems where the dataset teaches SVM about the classes so that SVM can classify any new data. It works by classifying the data into different classes by finding a line (hyperplane) which separates the training data set into classes.

Chapter 8 : Machine Learning Models & Algorithms | Amazon SageMaker on AWS

DOWNLOAD PDF MACHINE LEARNING MODELS AND ALGORITHMS FOR BIG DATA CLASSIFICATION

Machine learning explores the study and construction of algorithms that can learn from and make predictions on data - such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions: 2 through building a model from sample inputs.