|  |
| --- |
|  |
|  |
|  | **How to use random forest for prediction?**  **https://www.codementor.io/@agarrahul01/multiclass-classification-using-random-forest-on-scikit-learn-library-hkk4lwawu** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  | **Categorize by the input**: If it is a labeled data, it’s a supervised learning problem. |
|  | Supervised learning requires that the algorithm’s possible outputs are already known and that |
|  | the data used to train the algorithm is already labeled with correct answers. |
|  |  |
|  | **Commonly used machine learning algorithms** |
|  |  |
|  | **Linear regression** uses one independent variable X to explain or predict the outcome of the dependent variable y |
|  |  |
|  | **Logistic regression** performs binary classification, so the label outputs are binary. |
|  |  |
|  | **K-means clustering** is a clustering algorithm used to automatically divide a large group into smaller groups. |
|  |  |
|  | **K-nearest neighbors** is a classification algorithm, which is a subset of supervised learning |
|  |  |
|  | **K-means** is a clustering algorithm, which is a subset of unsupervised learning |
|  |  |
|  |  |
|  | **Random Forest** |
|  |  |
|  | So we take a random set of measures and a random sample of our training set and we build a decision tree. Then we do the same many |
|  | times using a different random set of measurements and a random sample of data each time. At the end we have many decision trees, we use |
|  | each of them to predict |
|  |  |
|  | **Choosing the Best Algorithm for your Classification Model.** |
|  | In machine learning, there’s something called the “No Free Lunch” theorem which means no one algorithm works well for every problem. |
|  | This is widely applicable in Prediction Models where we train our dataset on an algorithm and later use the trained model for |
|  | predictions on new data. |
|  |  |
|  | **As a result, you should try many different algorithms for your problem, while using a hold-out “test set” of data to evaluate** |
|  | **performance and select the winner** |
|  |  |
|  | Let’s list down the tasks we are going to perform to achieve our goal |
|  | •Read the Data |
|  | • Create Dependent and Independent Datasets based on our Dependent and Independent features |
|  | •Split the Data into Training and Testing sets |
|  | •Train our Model for different Classification Algorithms namely Naive bayes, Logistic Regression, Random Forest Classifier. |
|  | •Select the Best Algorithm |
|  |  |
|  | **What is Naive Bayes algorithm?** |
|  | It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, |
|  | a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other |
|  | feature. |
|  |  |
|  |  |
|  | **Pros:** |
|  |  |
|  | It is easy and fast to predict class of test data set. It also perform well in multi class prediction |
|  | When assumption of independence holds, a Naive Bayes classifier performs better compare to other |
|  | It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption). |
|  | **Cons:** |
|  |  |
|  | If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation. |
|  | Pros and Cons of Random ForestPros The following are the advantages of Random Forest algorithm −   * It overcomes the problem of overfitting by averaging or combining the results of different decision trees. * Random forests work well for a large range of data items than a single decision tree does. * Random forest has less variance then single decision tree. * Random forests are very flexible and possess very high accuracy. * Scaling of data does not require in random forest algorithm. It maintains good accuracy even after providing data without scaling. * Random Forest algorithms maintains good accuracy even a large proportion of the data is missing.  Cons The following are the disadvantages of Random Forest algorithm −   * Complexity is the main disadvantage of Random forest algorithms. * Construction of Random forests are much harder and time-consuming than decision trees. * More computational resources are required to implement Random Forest algorithm. * It is less intuitive in case when we have a large collection of decision trees. * The prediction process using random forests is very time-consuming in comparison with other algorithms |
|  |  |
|  |  |