Supervised Learning

Uzmabanu Kapadia

EMLIPD: 23864352

Ukapadi000@citymail.cuny.edu

Abstract:

Supervised learning can be applied in a wide array of industries from finance and healthcare to image recognition and natural language processing. This paper provides a comprehensive overview of supervised learning, a foundational paradigm in machine learning where algorithms are trained on labeled datasets to make predictions or decisions. Supervised learning consists of various algorithms each designed to perform a specific task.

The paper begins by explaining the fundamental principles of supervised learning, emphasizing the role of labelled training data in enabling algorithms to generalize and make accurate predictions on unseen examples. It examines some of the prominent supervised learning algorithms, including linear regression, logistic regression, k-nearest neighbors, decision trees, and random forest. For each algorithm, the paper discusses an overview of the given algorithm, potential applications as well as strength. It then goes into discussing a little about code and results of two of the supervised learning algorithms being discussed in this paper, the decision trees and the random forests.

Through a comparative analysis, the paper explores the strengths of different algorithms, aiding in algorithm selection based on the characteristics of their datasets and the nature of their prediction tasks. This study helps in creating a better understanding towards various algorithms involved in supervised learning in context to their usage.

What is Supervised Learning?

Supervised learning is a type of machine learning algorithm that trains a model based on labeled data where the correct answers are already provided [1]. Data labelling is a part of the preprocessing stage when designing a machine learning model. It requires raw data to be identified and have labels added to that data to specify its context for the model to allow for more accurate predictions [5]. The model learns from this labeled data and is then used to make predictions on new, unlabeled data [2].

Supervised learning had a diverse array of uses in tasks such as classification, regression, and anomaly detection [3]. Classification involves assigning a discrete category to each input. For instance, a classification model could predict whether an email is spam or not, or whether a customer will churn or not [4]. Regression involves predicting a continuous numerical value for each input. For example, a regression model could predict the price of a house or the amount of rain that will fall on a given day [4]. Anomaly detection involves identifying data points that significantly deviate from the rest of the data. For instance, an anomaly detection model could identify fraudulent credit card transactions [1]. To conduct a supervised learning process, it is necessary to explore the different supervised learning algorithms some of which include linear regression, logistic regression, k-nearest neighbors, decision trees, and random forest.

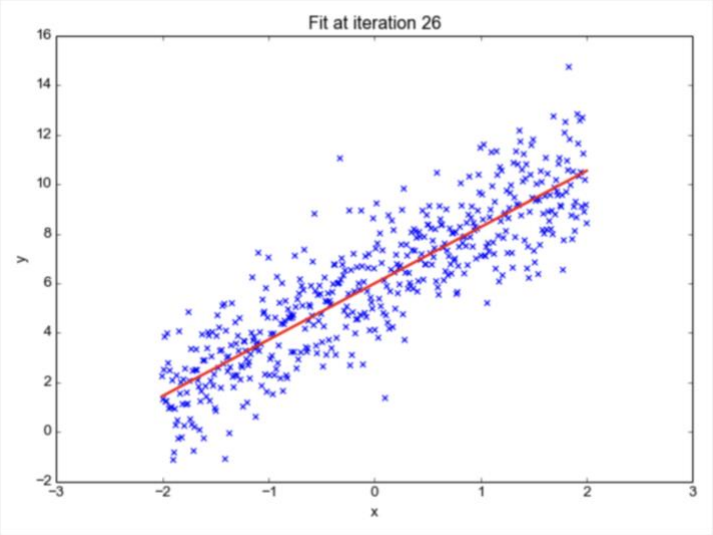
Linear Regression:

One of the supervised learning algorithms is linear regression. The linear regression algorithm is used to recognize the correlation between a dependent variable and one or more independent variables and is usually leveraged to make predictions regarding future outcomes [6]. In instances where there exists just one independent variable and one dependent variable, the approach is referred to as simple linear regression. An example of this is predicting house prices

based on the footage of the house. This is an example of a simple linear regression as it is basing the house prices on solely on one variable, the house footage.

However, as there is an increase in the number of independent variables, it becomes known as multiple linear regression [6]. An example of a multiple linear regression would be predicting a students' GPA based on multiple variables such as their high school GPA, SAT scores, and number of extracurricular activities. Regardless of the type of linear regression, the goal is to establish a line of best fit, which is computed through the method of least squares but, unlike other regression models, the resulting line in linear regression is straight when visualized on a graph [6].

Here we can see a plot for a linear regression model, it basically represents the line of best fit:



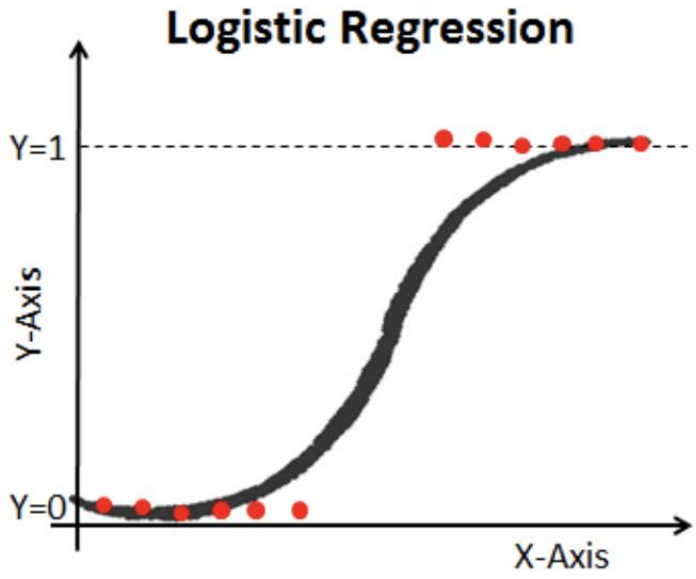
Linear regression has wide applications across various domains in supervised learning. Applications can be seen in economics and finance for predicting stock prices based on historical data [7]. Uses of linear regression can also be seen applicable in the healthcare field for predicting patient outcomes or in marketing to forecasting sales based on advertising as well as

many other fields such as education or science. Some of the major strengths of a linear regression model include interpretability meaning linear regression allows for a straightforward interpretation of the relationship between the independent and dependent variable [8]. Additionally, linear regression is simplistic making it easy to understand, implement, and efficient for variables with linear relationship. Linear regression also allows for the identification of the most influential features affecting the dependent variable [3].

Logistic Regression:

Logistic Regression algorithm is used when the dependent variable is categorical. This means they have binary outputs, such as true and false or yes and no [6]. Though both, linear and logistic regression try to understand relationships between inputted data, logistic regression is mostly used to solve binary classification problems [6]. Logistic regression uses a logit function to make predictions about this binary problem [10].

Here we can see an example of a logistic regression plot:



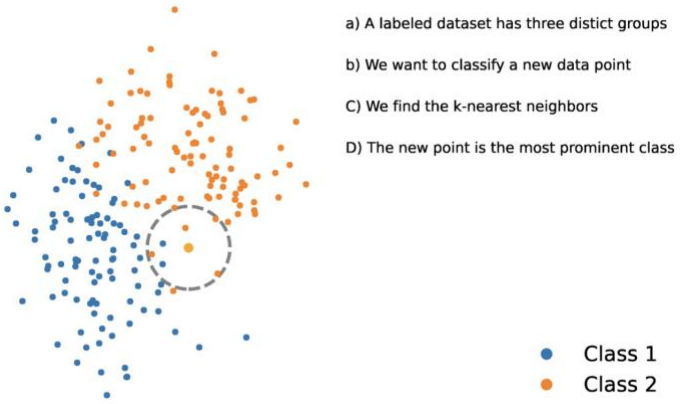
Some applications of logistic regression include spam detection, medical research such as Predicting the likelihood of a patient having a particular medical condition based on various diagnostic features. Also, when looking at credit scoring to assess the probability of a customer defaulting a loan aiding in credit risk evaluation [9]. Another application of logistic regression predicting election outcomes or political preferences based on demographic and polling data. When considering its applications, it's also important to look at the strengths of the algorithm to see where it can be best used overall.

Logistic regression similarly to linear regression has good interpretability meaning logistic regression models provide interpretable coefficients that indicate the impact of each predictor on the log-odds of the outcome. Some other strengths include its efficiency with binary outcomes meaning when dealing with binary data classification, as previously established, the best algorithm to use would most likely be logistic regression. Also, logistic regression allows assessment through various good-to-fit tests to check for model performance and regularization through L1 and L2 can be used to prevent overfitting [3].

K-nearest Neighbors:

The K-nearest neighbors also known as the KNN algorithm is a non-parametric algorithm that evaluates the proximity of one data point to another to be able to decide whether the two points can be grouped together [10]. This proximity between data points shows the degree to which they can be compared to one another. For example, if there was a graph with group A and group B, each of these groups would be represented by a point on the graph, so when new data would be added to the graph, the group of this new point will depend on which group its closer to [10].

Here we can see a plot for a K-nearest neighbor algorithm:



K-nearest has various application in supervised learning. For example, in retail recommender system can suggest products based on the purchasing behavior of similar customers. Another application of K-nearest can be in image recognition to classify images based on visual feature as well as in finance for credit scoring and risk assessment for loan approval [9]. Overall, K-nearest neighbor algorithm is particularly useful when the decision boundary is complex or when dealing with non-linear relationships in the data. Some strengths of the K-nearest neighbor include its simplicity and intuitiveness allowing it to be easy to understand and implement. K-nearest neighbor algorithms is non-parametric in nature meaning it does not rely on assumptions related to shape or parameters of the data distribution and since it makes no assumptions when it comes to the underlying data distribution, it is versatile to be used for various types of datasets [3]. This algorithm can also easily extend to become a multiclass classification problem and it robust to outliers as the predictions the algorithm makes is based on the majority class within the k-nearest neighbor.

Decision Trees:

One important algorithm used in supervised learning is decision trees. They are a type of probability tree like structure model, continuously separating data to categorize or make

predictions based on the results of the previous set of questions [10]. The decision tree model evaluates the given data, providing responses to the questions to provide assistance in formulating more informed choices [10]. An image for a decision tree created for this paper can be seen in the code result discussion below.

Decision trees can be used in which the answers produced are yes or no and are used as a way of reaching an end decision on predicting what a specific thing may be based on given data. The decision tree algorithm is used in both classification and regression tasks. Some applications of a decision tree can be for fraud detection where such fraud activities can be detected in financial transactions through analyzing patterns and anomalies. Supervised learning with decision trees are also used for prediction of equipment failure or maintenance based previous historical performance data.

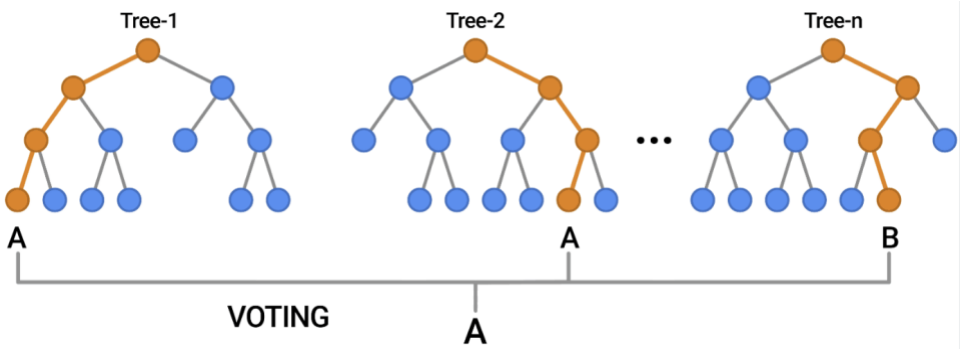
When using decision trees some strengths that they have are the ability to handle non-linearity meaning they can model complex non-linear relationships in data without needing transformations [3]. Another strength is they do not make assumptions about its underlying data distribution, allowing them to be versatile as well as measuring variable importance to aid in feature selection and understand predictor impacts. Decision trees can also handle missing values by either assigning predicted values or bypassing splits that may involve the missing value.

Random Forests:

The last supervised learning algorithm discussed here is random forests. The random forest algorithm is built on trees. Unlike of decision trees which use a single tree, random forests use multiple decision trees to be able to make judgement which basically ends up in a forest as it is using a collection of uncorrelated decision trees [10]. These trees are merged together to

decrease variance and establish more accurate data predictions [6]. Random forests can be used for both classification and regression purposes.

Below is a visual of how random forests work:

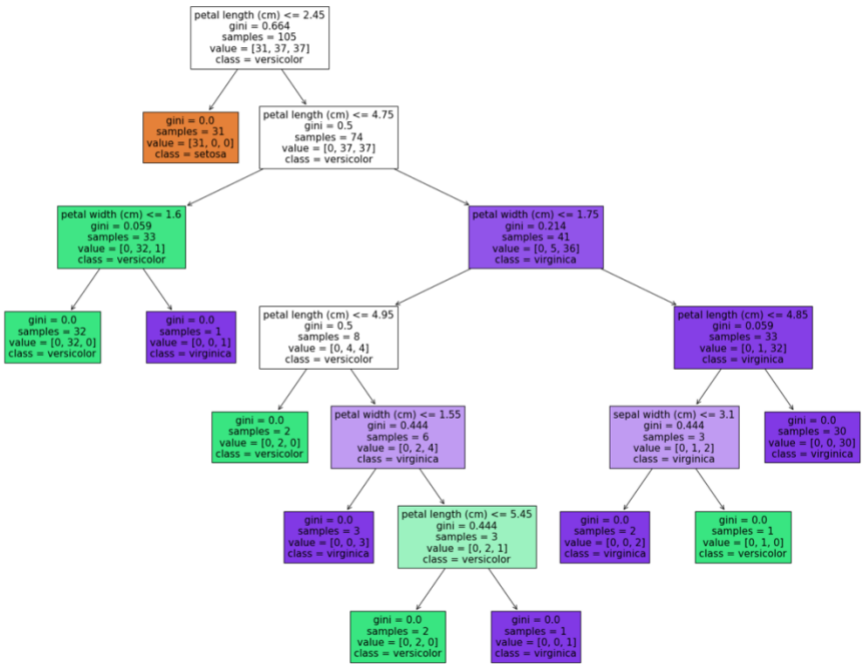


Random forests can be applied in various ways. For example, they can be used by ecologists for to model distribution of species and predicting ecological patterns or for remote sensing as land cover classification and mapping with satellite and aerial imagery as well as many other meaningful applications. The strengths of a random forest include high prediction accuracy, being less prone to overfitting as well as robust towards outliers in data as they use multiple trees reducing the impact of such data points [3]. Another strength for random forests in their effective handling of datasets with large number of features allowing them to suitable for higher dimension data. As random forests use decisions trees, all the strengths of decision trees would also apply to random forests.

Decision Tree and Random Forest Code Result Discussion:

For this paper the two algorithms, decision trees and random forest were coded. The code for this can be found in the midterm jupyter notebook. To begin, the iris dataset was used to create training sets for the models. After that, the training sets were used to test the models for accuracy. In the image below the decision tree created by the code for this iris dataset that it was trained on can be seen. Multiple of these decision trees were then used to create a random forest.

```
In [149]: # plot the decision tree
          fig = plt.figure(figsize=(25,20))
          _ = tree.plot_tree(model,
                             feature_names=irisFeatures,
                             class_names=iris.target_names,
                             filled=True)
```



The results of these algorithms can be viewed in the jupyter notebook. However, as expected, the random forest performed with greater accuracy considering it uses multiple decision trees to deal with data for better categorization.

Conclusion:

Overall, through the research of on supervised learning we were able to understand the supervised learning algorithms, see examples of their usage and their strengths. When conducting supervised learning it very important to analyze what kind of dataset it is being for to ensure that the right algorithm is used is. For the example coded for this paper an iris dataset was used for training and decision trees as well as random forest algorithms were used to be able to go through various categories and make decisions to come up with greatest accuracy. Hence, h this research allowed us to delve into the world of supervised learning and creating an understanding for the various algorithms involved.

Reference

1. Alpaydin, E. (2014). Introduction to machine learning (2nd ed.). MIT press.
2. Duda, R. O., & Hart, P. E. (1973). Pattern classification and scene analysis. Wiley.
3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
4. Mitchell, T. M. (1997). Machine learning. McGraw-Hill.
5. https://www.ibm.com/topics/data-labeling
6. https://www.ibm.com/topics/supervised-learning#:~:text=Supervised%20learning%2C%20also%20known%20as,data%20or%20predict%20outcomes%20accurately.
7. Brown, R. L., Durbin, J., & Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. Journal of the Royal Statistical Society: Series B (Methodological), 37(2), 149-192.
8. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning (Vol. 103). Springer.
9. Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). Credit scoring and its applications (Vol. 2). SIAM.
10. https://www.datacamp.com/blog/supervised-machine-learning