SUPERVISED MACHINE LEARNING

Najmiddinov Shakhzodbek Shukhrat ugli
Odiljonov Umidjon Odiljon ugli
Tashkent University of Information Technologies
named afterMuhammad al- Kharezmy

ABSRTACT. The term "supervised learning," which is also used to refer to supervised machine learning, refers to the process of teaching algorithms to correctly classify data or predict outcomes using labeled datasets. The model modifies its weights as input data is fed into it until it is well fitted. This happens as part of the cross validation procedure to make sure the model does not fit too well or too poorly. A common example of how supervised learning aids companies is by classifying spam in a distinct folder from your email. Neural networks, naive bayes, linear regression, logistic regression, random forests, and support vector machines (SVM) are a few techniques used in supervised learning.

KEY WORDS. Supervised learning, labeled datasets, distinct folder, neural networks, naïve bayes.

INTRODUCTION. Computers can learn through machine learning even when they are not expressly trained for the task at hand. Learning occurs when data and mathematical models[1] are coupled, for instance by determining optimal values for the model's unobserved variables. Fitting a straight line to data could be the simplest example of learning, however machine learning typically deals with models that are much more flexible than straight lines. The purpose of doing this is to provide a result that can be used to make judgments about fresh data that wasn't used in the model's learning. Using a data set of 1000 photographs of puppies, we may train a model that, if carefully chosen, may be able to determine whether another image is of a puppy or not.

The process of supervised machine learning typically involves the following steps:

1. Data collection: The first step in any machine learning project is to collect and prepare the data. This involves collecting a dataset that contains both input data and corresponding output data[2], or labels.
2. Data preprocessing: The dataset must be cleaned and preprocessed to remove any errors, inconsistencies, or missing values. This may involve techniques such as normalization, feature scaling, and feature engineering.

3. Model selection: Once the data has been preprocessed, the next step is to select an appropriate model for the task at hand. This will depend on the nature of the problem, the size and complexity of the dataset, and the desired level of accuracy.

4. Model training: The model is then trained on the labeled dataset using an optimization algorithm such as gradient descent. During training, the algorithm adjusts its internal parameters to minimize the difference between its predictions and the actual labels in the training data.

5. Model evaluation: Once the model has been trained, it is evaluated on a separate dataset, known as the validation dataset, to measure its performance. This helps to identify any issues with overfitting or underfitting, and to fine-tune the model's hyperparameters[3] if necessary.

6. Model deployment: Finally, the trained model can be deployed to make predictions on new, unseen data.

One of the key advantages of supervised machine learning is that it can be used to make highly accurate predictions on a wide range of tasks. However, it requires a large amount of labeled data to train the model, which can be time-consuming and expensive to obtain. Additionally, the model's performance may degrade over time if the underlying data distribution changes, which can require retraining the model on updated data.

ENSEMBLE METHODS. *Bagging* - The trade-off between bias and variance is a key idea in machine learning. In general, a model's bias will decrease the more flexible it is. In other words, a flexible model has the ability to depict intricate input-output interactions. Naturally, if the real relationship between inputs and outputs is complex, as it frequently is in machine learning applications, then this is advantageous.

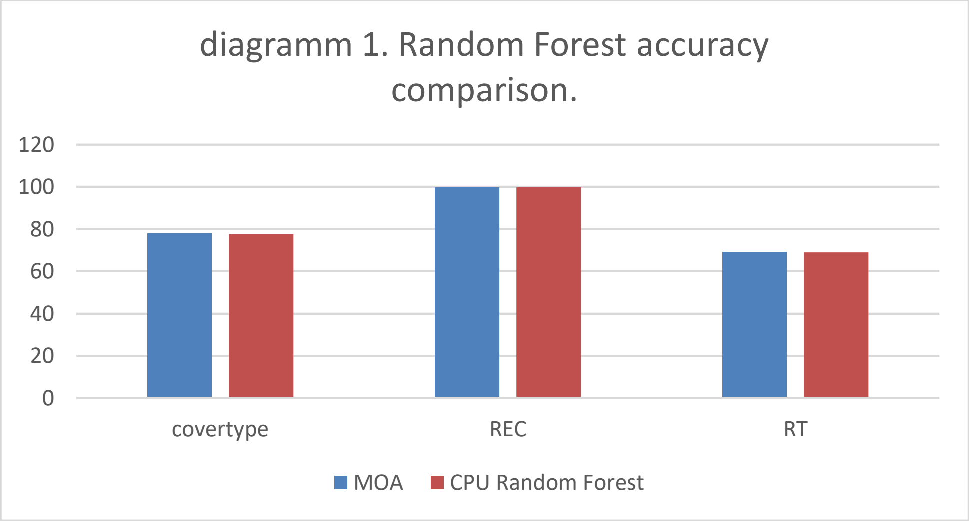
The bootstrap - A strong and extensively used statistical method is the bootstrap. Its primary application is, in fact, estimating the uncertainty in statistical estimators, but in this case, we'll utilize it to imitate the variance reduction strategy described above.

Random forests - The variance reduction obtained by averaging is limited by the correlation between the individual ensemble members[4]. A natural question to ask is therefore if it is possible to reduce this correlation. One simple trick for accomplishing this is a method known as random forests

Random forests presume that these base models are provided by classification or regression trees[5], but bagging is a universal technique that can theoretically be used to reduce the variance of any base model. The goal is to lessen the association between each tree by adding extra randomness as it is being built. This may at first appear like a ridiculous idea because, according to logic, arbitrarily altering a model's training should have a negative impact on its performance. However, there is a justification for this perturbation, which we will go over later, but first, we'll go over the specifics of the technique.

	MOA	CPU Random Forest
covertype	77.93	77.55
REC	99.8	99.8
RT	69.13	69

Table 1. Random Forest accuracy comparison.



CONCLUSION. In conclusion, supervised machine learning is an effective method for correctly classifying and predicting a variety of tasks. It is feasible to obtain high levels of accuracy and generalize well to new, unknown data by carefully choosing and preparing the data, choosing a good model, and adjusting the model's hyperparameters.

REFERENCES.

1. "The Mathematics of Machine Learning" by Ali Rahimi and Ben Recht, available at: https://arxiv.org/abs/1803.08375
2. "Data Output and Analysis" by Richard J. Larsen and Morris L. Marx.
3. "Hyperparameter optimization: A review of algorithms and applications" by James Bergstra and Yoshua Bengio.
4. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
5. "Regression Trees" by Leo Breiman, Jerome Friedman, Charles Stone, and Richard Olshen.