Handwriting digit prediction using ML.

Mustapha Hadrous

EC Utbildning

Examensarbete Data Science

2024–02

Abstract:

This project is focused on how to learn my computer to predict handwritten digits in correct way. We use MNIST dataset which contains small images of handwritten digits. There are many classification algorithms (LogisticRegression, KNeighborsClassifier , RandomForest, etc) which can be trained on this dataset including learning algorithms and evaluate. We can improve accuracy by trying a different algorithm together. We use Confusion Matrix to minimize false.

Keywords:

statistical analysis, machine learning, KNN, Scaling and classification.

Acknowledgement:

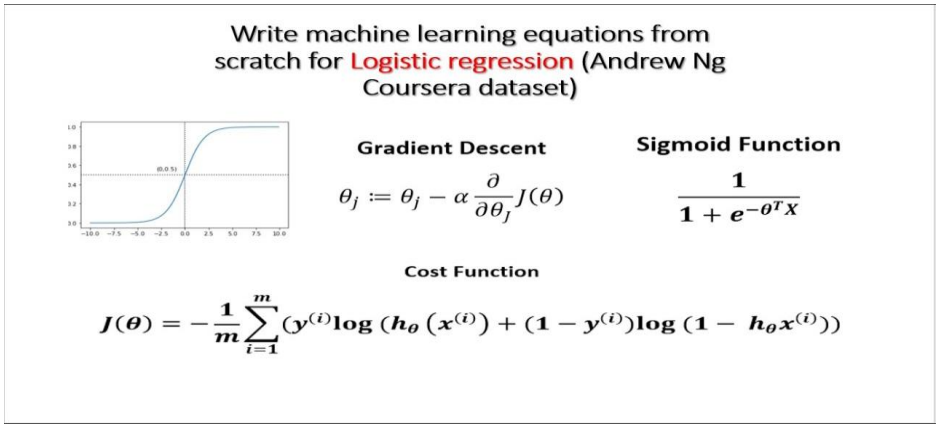
I would like to thank Antonio Prgomet (My teacher) for his information, helping, advice, ideas, discussions and supporting to do this project.

I would like to thank my classmates which we discussed and helping each other to solve problems I faced. Especially Robert 😊

Skapas automatiskt i Word genom att gå till Referenser> Innehållsförteckning.

Innehållsförteckning

Logistic regression: is a data analysis technique that uses mathematics to find the relationships between two data factors. It then uses this relationship to predict the value of one of those factors based on the other.



Formula

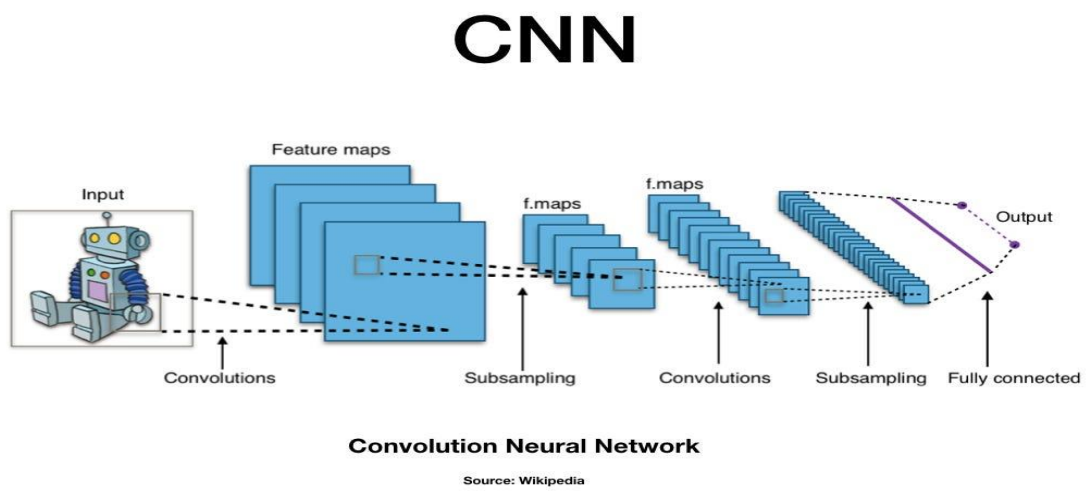
$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values
$y_1, y_2, \ldots, y_n$ are observed values
n is the number of observations

The KNN classifier: is one of the simplest and widely used classification algorithms, where a new data point is classified based on its similarity to a specific group of neighboring data points. This tutorial provides an overview of the KNN algorithm, its implementation in Python, and its applications.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain. The Convolutional Neural Network is a subtype of Neural Networks that is mainly used for applications in image and speech recognition. Its built-in convolutional layer reduces the high dimensionality of images without losing its information.



Metod:

To better understand the MNIST dataset, we need to analyses and visualize it by the following steps:

- We started by examining the training and testing parts of the dataset to see if the data is balanced in terms of understanding the number of classes and their attributes.
- we checked the dimensionality of each part of the dataset
- we plotted some samples from the training dataset to visualize the digits included in the MNIST dataset.
- Cleaning data by (Simple imputer).
- Standard Scaler: will normalize the features.
- Calculating Confusion Matrix.
- Check Classification Report.
- Select the models and fit it.
- Check the score and accuracy.

Resultat och Diskussion

	Score
Logistic Regression	93.4 %
KNeighborsClassifier	98.2 %

```
lrr = lr.score(X_train,y_train)
knnn = knn.score(X_train,y_train)
if lrr > knnn:
    print ('logistice regretion is Best Modul with score', lrr)
else:
    print('KNeighborsClassifier is the best Modul with score :' , knnn)
```

executed in 43.3s, finished 01:12:06 2024-03-20

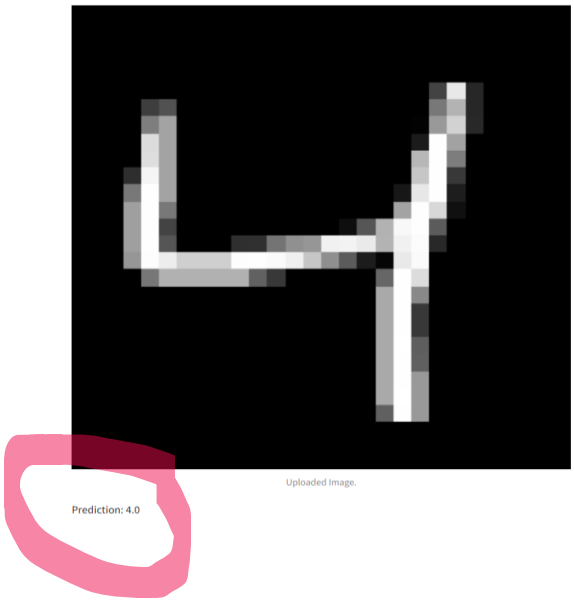
```
KNeighborsClassifier is the best Modul with score : 0.9829682539682539
```

After checking the score between the two models, we will notice that K neighbours Classifiers score is better than logistic regression. So will use KNN model to prediction.

Slutsatser

As we see we use statistical techniques, to analyse the dataset which effective for dimensionality reduction. In addition, we used both logistic regression and K neighbours Classifiers models to classify the dataset. These models achieved accuracies of 93% and 98%, respectively. So, when doing scaling and imputer together with KNN can enhance the efficiency and accuracy of the classifier. It will perform dimensionality reduction by transforming features into a set of principal components, thus reducing the number of features in the dataset. Subsequently, training the KNN on this reduced dataset will result in quicker training times and improved performance.

After checking in Streamlet , it working well now as picture below



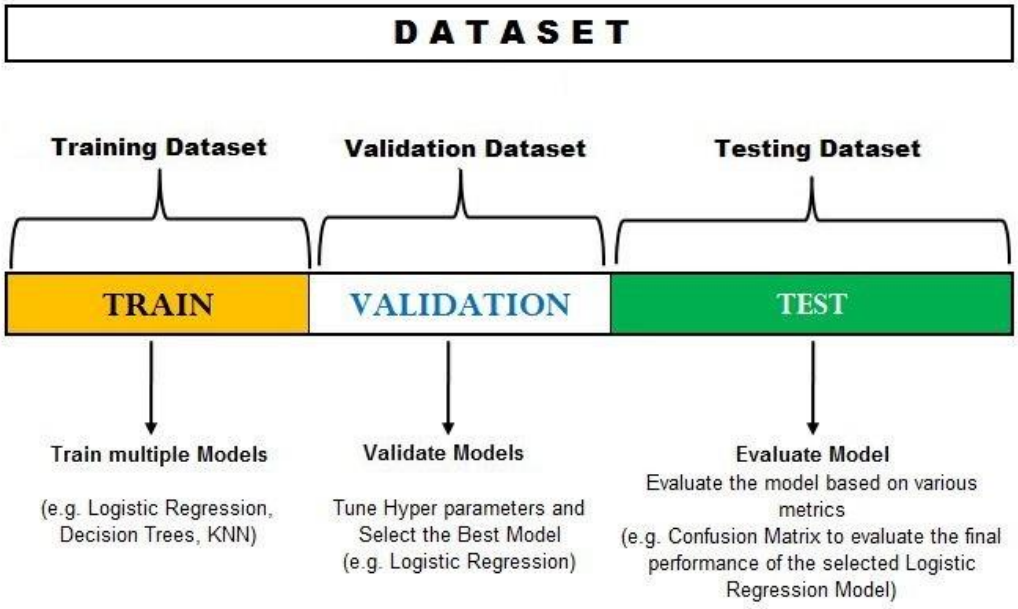
Teoretiska frågor

1. Kalle delar upp sin data i ”Träning”, ”Validering” och ”Test”, vad används respektive del för?

    - **Training Data**: is the data you use to train an algorithm or machine learning model to predict the outcome you design your model to predict. The quality of the training data directly impacts the accuracy and reliability of the machine learning algorithm

    - **Test Data :** Once your machine learning model is built (with your training data), you need unseen data to test your model. This data is called testing data, and you can use it to evaluate the performance and progress of your algorithms' training and adjust or optimize it for improved results.

    - **Validation Data**: provides an initial check that the model can return useful predictions in a real-world setting, which training data cannot do. The ML algorithm can assess training data and validation data at the same time.



2. Julia delar upp sin data i träning och test. På träningsdatan så tränar hon tre modeller; ”Linjär Regression”, ”Lasso regression” och en ”Random Forest modell”. Hur skall hon välja vilken av de tre modellerna hon skall fortsätta använda när hon inte skapat ett explicit ”valideringsdatasats” :

    - In the absence of a dedicated validation dataset Julia will attempt to select the best model among linear regression, lasso regression, and random forest. By analysing their performance and score on training data. This involves evaluating (MSE), (RMSE) to measure how well each model fits the training data. In addition, considerations such as the complexity of problems, the interpretability of the models, and the potential for over- or under-fitting play critical roles in making the decision. By carefully weighing these factors, Julia can choose the model that has the optimal balance between performance and simplicity.

3. Vad är ”regressionsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden?

    - **Regressions problem:** Regression describes how to numerically relate an independent variable to the dependent variable. To represent a linear relationship between two variables. **Regression problem** is how to model one or several dependent variables/responses, Y, by means of a set of predictor variables X and Y For Linear Regression problems usually we use :

        i. **MAE** measures the average squared difference between an observation's actual and predicted values. The output is a single number representing the cost, or score, associated with our current set of weights. Our goal is to minimize MSE to improve the accuracy of our model.

        $$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i |$$

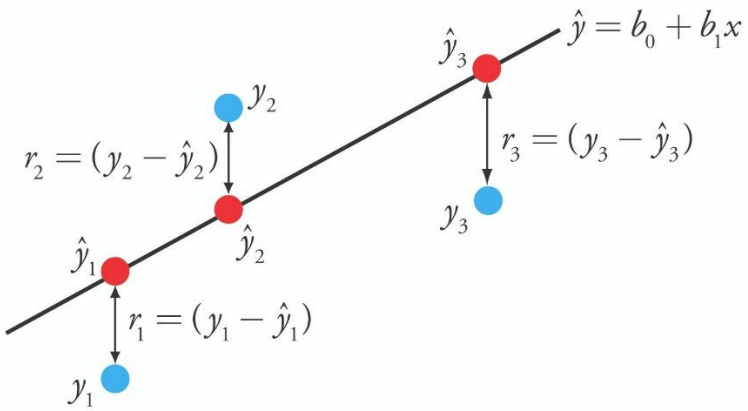
        ii. **Root Mean Square Error (RMSE):** It gives an idea of how much error the system typically makes in its predictions, with a higher weight for large errors.

4. Hur kan du tolka RMSE och vad används det till:

   - **Root Mean Square Error (RMSE)** is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

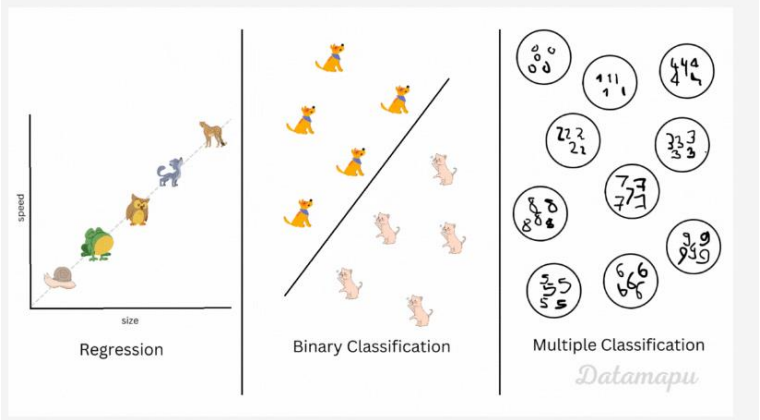
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

   - Y : are observed values
   - Y^ : are predicted value
   - N : is the number of observations



5. Vad är ”klassificieringsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden? Vad är en ”Confusion Matrix”?

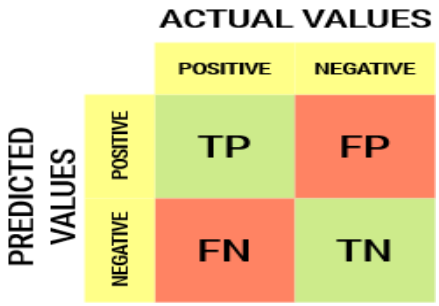
   - **Classification problems** are the problems in which an object is to be classified in one of the n classes based on the similarity index of its features with that of each class. By classes, we mean a collection of similar objects. **Classification problem** can be thought of as two separate problems – binary classification and multiclass classification. In binary classification, a better understood task, only two classes are involved, whereas multiclass classification involves assigning an object to one of several classes. For classification we can use Confusion Matrix , accuracy , precision, recall ….



Models we can use for classification problems:

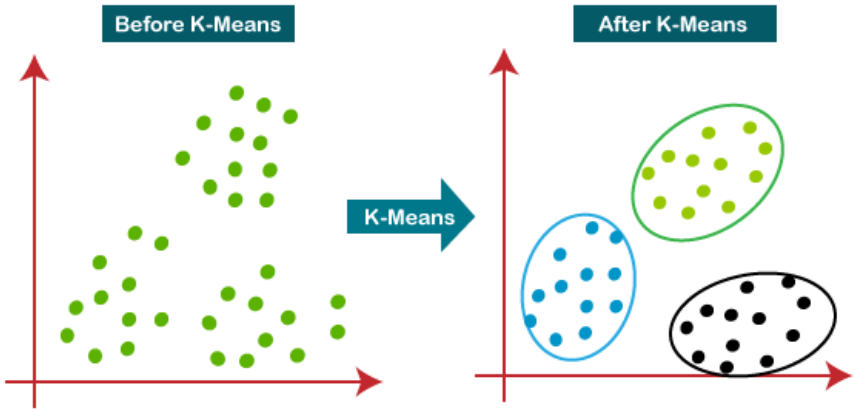
- Logistic Regression.
- K-Nearest Neighbours.
- Decision Tree.

A confusion matrix is a performance evaluation tool in machine learning, representing the accuracy of a classification model. It displays the number of true positives, true negatives, false positives, and false negatives A confusion matrix is a performance evaluation tool in machine learning, representing the accuracy of a classification model. It displays the number of true positives, true negatives, false positives, and false negatives.



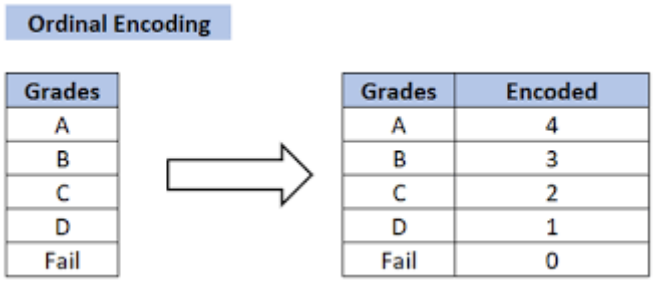
6. Vad är K-means modellen för något? Ge ett exempel på vad det kan tillämpas på

- **K-means clustering** is a type of unsupervised learning, which is used when you have unlabelled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. It finds the similarity between the items and groups them into the clusters. It is typically applied to data that has a smaller number of dimensions, is numeric, and is continuous examples include spam detection, sentiment analysis, scorecard prediction of exams.

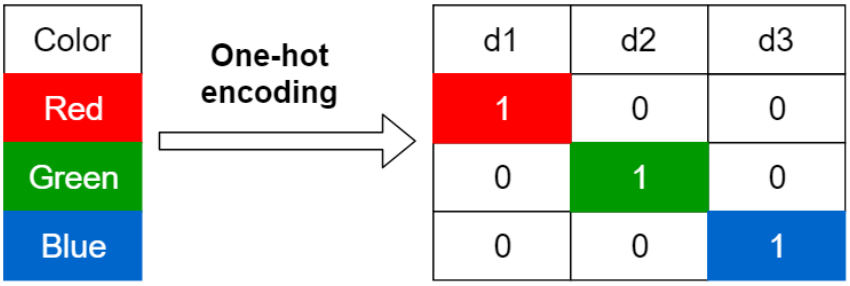


7. förklara (gärna med ett exempel): Ordinal encoding, one-hot encoding, dummy variable encoding. Se mappen ”l8” på GitHub om du behöver repetition.

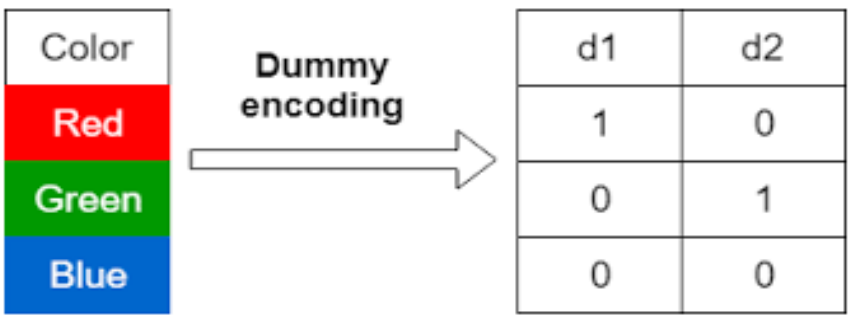
- **Ordinal encoding**: is a preprocessing technique used for converting categorical data into numeric values that preserve their inherent ordering.



- **One hot encoding:** is one method of converting data to prepare it for an algorithm and get a better prediction. With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector.



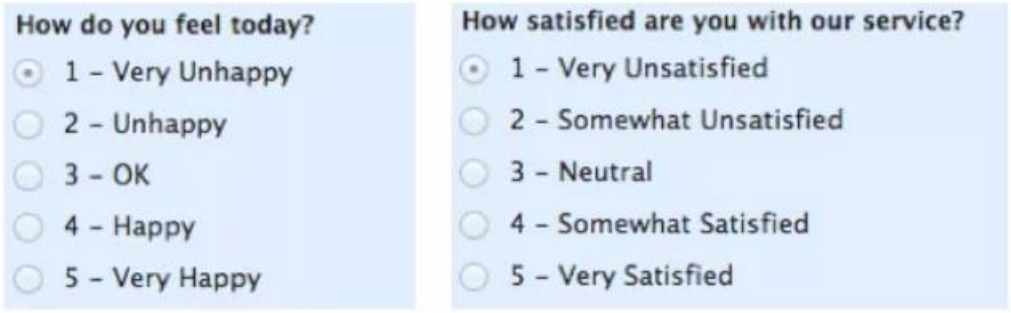
- **Dummy Encoding:** This categorical data encoding method transforms the categorical variable into a set of binary variables (also known as dummy variables). The dummy encoding is a small improvement over one-hot-encoding.



8. Göran påstår att datan antingen är ”ordinal” eller ”nominal”. Julia säger att detta måste tolkas. Hon ger ett exempel med att färger såsom {röd, grön, blå} generellt sett inte har någon inbördes ordning (nominal) men om du har en röd skjorta så är du vackrast på festen (ordinal) – vem har rätt?
    - Nominal: the data can only be categorized

Examples of Nominal Scales

    - Ordinal: the data can be categorized and ranked.



Example of Ordinal Scales

    - In this case, colours such as red, blue, green, yellow can be categorize into different groups. So its Nominal

9. Vad är Streamlit för något och vad kan det användas till?
    - **Streamlit**: is a free and open-source framework to transform, build and share Python, machine learning and data science and scripts into interactive web apps. Build dashboards, generate reports, or create chat apps. Once you've created an app, you can use our Community Cloud platform to deploy, manage, and share your app.

Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.
    - First challenge was understanding what machine learning mean is and do.
    - second challenge was How to read data.
    - Third challenge was to learn how to clean data, select features, select best models, Metrics models, data split, etc.
    - Last challenge was streamlit.
    - I faced these challenges by reading from book, teacher files, google, watch samples from YouTube, teacher videos, Datacamp, follow instructions and advice from teacher.

2. Vilket betyg du anser att du skall ha och varför.
    - Som du vill 😊

3. Något du vill lyfta fram till Antonio?
    - I would like to thank you for your explaining, patient, helping, solving the problems and your support to all.

Appendix

	precision	recall	f1-score	support
0.0	0.96	0.97	0.97	683
1.0	0.95	0.98	0.96	800
2.0	0.90	0.89	0.89	674
3.0	0.90	0.91	0.90	760
4.0	0.92	0.92	0.92	611
5.0	0.90	0.85	0.88	658
6.0	0.94	0.95	0.94	677
7.0	0.93	0.94	0.94	724
8.0	0.88	0.87	0.88	693
9.0	0.90	0.91	0.91	720
accuracy			0.92	7000
macro avg	0.92	0.92	0.92	7000
weighted avg	0.92	0.92	0.92	7000

Källförteckning

- Diego Lopez Yse. (2023). Introduction to K-Means Clustering. Pinecone. Retrieved from https://www.pinecone.io/learn/k-means-clustering/#K-means-clustering
- Jun Wu, Li Shi,Liping Yang, Xiaxia Niu ,2 Yuanyuan Li, Xiaodong Cui, Sang-Bing Tsai, Yunbo Zhang5. (den 3 May 2021). User Value Identification Based on Improved RFM Model and K-Means++ Algorithm for Complex Data Anlaysis. Hindawi, 2021, 8. doi:https://doi.org/10.1155/2021/99824