

Machine Learning Project

Havva Nilsu Öz

havvanilsu.oz@studenti.unipd.it

1. Introduction

In this project I discussed the image classification problem on a clothing dataset. Several machine learning models have been created and tried to improve their prediction accuracy. So we can compare the models. The goal is to be able to predict the new dataset as accurately as possible, taking into account the parameters such as the time it takes. The model I got the best result in line with the analysis I made is Artificial Neural Network with one input layer, six hidden layer and one output layer.

2. Dataset

The Fashion-MNIST dataset of Zalando's article images is comprised of 28 x 28 grayscale images of 70,000 fashion products, from 10 categories, with 7,000 images per category.

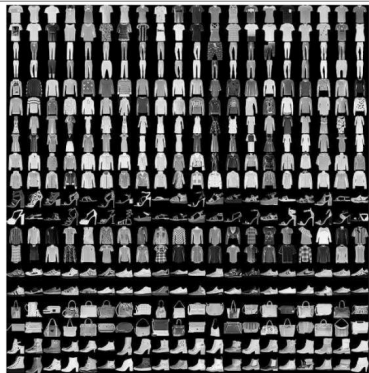
Label	Description	Examples
0	T-Shirt/Top	
1	Trouser	
2	Pullover	
3	Dress	
4	Coat	
5	Sandals	
6	Shirt	
7	Sneaker	
8	Bag	
9	Ankle boots	

Figure 1. Caption

2.1. Re-shaping the data

The given data is 28x28 array which is 2-dimensional. Since the machine learning algorithms we work on are expecting a 1-dimensional input, I reshaped 28x28 array into a vector of 784 pixels.

2.2. Splitting the data

I split the dataset into 48,000 training set, 12,000 validation set and 10,000 test set in order to prevent overfitting.

I obtained this result by turning 20% of 60,000 training set into the validation set.

2.3. Normalizing the data

The pixels values range from 0 (black) to 255 (white) so the variation of the variables is quite a lot. In order to improve my efficiency on models, I use standardization to the data which makes the values centered around the mean with a unit standard deviation. The pixel values are scaled from a range of [0, 255] to a range [0, 1].

3. Methods

3.1. Artificial Neural Network

The main problem with image classification is the difficulty to find useful features. It is not easy to create features from images like shapes, edges, regions. However a neural network, along with learning a model for classification is able to create and select automatically useful features.

In my project a model with one input layer and one output layer is applied to achieve our goal. However a single-layer neural network can only be used to represent linearly separable functions. This means very simple problems where, the two classes in a classification problem can be nearly separated by a line. But our classification problem is more complicated so I decided to add hidden layers. I tried the same model with 1 and 6 hidden layers which makes the model multi-layer perceptron. Each hidden layer consists of 128 neuron.

After finding the right model with sufficient number of hidden layers I applied different epochs and batch-size with the help of the EarlyStopping class.

3.2. Logistic Regression

Logistic regression is a linear model which can be subjected for nonlinear transforms. Since this is multi-class problem I used Softmax Regression, which is nothing but Logistic Regression for multi-class classification problems

3.3. K-Nearest Neighbors

KNN algorithm stores all the available cases and classifies the new data based on a similarity measure. So it classifies based on the classification of neighbors. 'k' in KNN is

a parameter that refers to the number of nearest neighbours to include in the majority of the voting process. I used $k = 5$ in my project.

3.4. Decision Tree

Decision tree is one of the most powerful and popular tool for classification and prediction. Decision trees are a non-statistical approach to pattern classification. I used several max depths in order to find the most accurate one. Grid-search is used to find the optimal hyperparameters in order to find the most accurate predictions for the new dataset. None, 5, 10 and 20 max depths are tried.

3.5. Random Forest

Random forest algorithms are a machine learning technique that is being used increasingly for image classification. It is formulated through the combination of several decision trees. In most cases, a forest of several decision trees can be assumed to be a more accurate model than to a single individual tree. To select the best parameters for estimation, again I performed grid search with given number of estimators and maximum depth.

3.6. Support Vector Machine

Support Vector Machine is a type of supervised learning which can be used for both classification and regression models. Two different penalty parameter of the error term, which corresponds to C in the model, is used to improve the prediction accuracy. However it takes considerably more time than other models. However SVM's apply multiclass classification in a one versus rest of the classes and this makes harder to create separating hyperplane. So they are more useful with binary classification.

4. Experiments

I imply a number of experiments were executed to reach the optimal set of parameters for the models. Now let's see the results.

4.1. Artificial Neural Network Results

First I start with one input and one output layer. Then after implying one and six layers, the test results is compared.

The model with 6 hidden layer has the best result compared to the others. After setting up the layers, I wanted to apply more epoch. However we must be careful when increasing the epoch because it may lead to overfitting. We can see this problem when the train accuracy is trained good but validation accuracy is not good. In other words when we look at the accuracy graph of train and validation, if we see that their curves are too far from each other we can say that

Layers	Test Accuracy
Input-Output	0.8690
Input-1 Hidden Layer-Output	0.8759
Input-6 Hidden Layer-Output	0.8844

Figure 2. Test accuracy of layers

the model overfits. To solve this problem EarlyStopping is used when analysing the epoch numbers affect to our model. In the end, the best parameters is found as epoch = 10 and batch size = 16 with the 6 hidden layer model. We can see from the train-validation graph that there is no overfitting problem.

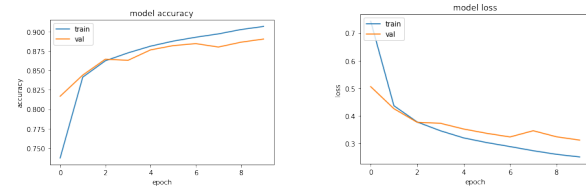


Figure 3. Accuracy and Loss graphs.

4.2. Logistic Regression Result

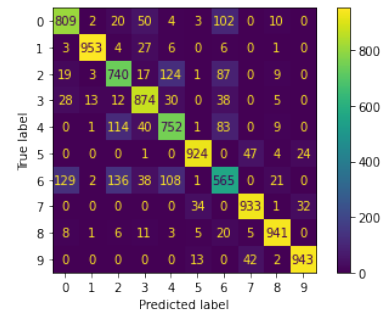


Figure 4. Confusion matrix

However when we analyse deeply with confusion matrix, we can see that the model fails a great amount of proportion when estimating the class 6.

4.3. K-Nearest Neighbors Result

4.4. Decision Tree Result

We must choose the max depth of the decision tree carefully in order to not face overfitting problem. So we must not set max depth too high. Therefore Gridsearch is used to determine the best max depth for our dataset. Among none, 5, 10 and 20; the best max depth is 10 with test accuracy 0.8057. However it is still below what we found with neural network.

4.5. Random Forest Result

As we expect that the random forest will perform better than the decision tree, it has 0.8636 test accuracy. Same logic we did with decision tree, Gridsearch is used to determine the best parameters for random tree. This time we also decide the n estimators. n estimators is the number of trees you want to build. So higher number of trees will give better performance but makes the code slower. The best parameters achieved as: max depth = 12 with n estimators= 25.

4.6. Support Vector Machine Result

For the SVM, both the model creating and data predicting process took the longest compared to others. Two different models is applied with $C = 0.01$ and $C = 1$. The model with $C = 0.01$ turns out perform better.

5. Conclusion

The final table for the models with best parameters is shown in Figure 5.

Model Name	Parameters	Execution Time	Test Accuracy
Neural Network - 6 hidden layer	epochs=10, batch_size= 16	2.75 s	0.8844
Logistic Regression	multi_class="multinomial", solver="lbfgs",max_iter= 100, n_jobs=-1	116 ms	0.8434
K-Nearest Neighbors	n_neighbors = 5, metric = 'minkowski', p=2, n_jobs=-1	1min 13s	0.8502
Decision Tree	criterion: 'entropy', 'max_depth': 10	24 ms	0.8057
Random Forest	criterion: entropy, max_depth: 12, n_estimators: 25	126 ms	0.8529
Support Vector Machine	kernel = 'linear', C = 0.01	2min 5s	0.8514

Figure 5. Test accuracy of best models

As we analyse through different models and different parameters to the same dataset, we can conclude that the most accurate model is artificial neural network. Both with test accuracy and execution time results are obtained best in ANN. Although I couldn't increase its accuracy a great extent, the model does a great image classification.

6. Formatting your paper

6.1. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [1]. Where appropriate, include the name(s) of editors of referenced books.

References

- [1] Authors. The frobnicatable foo filter, 2014. Face and Gesture submission ID 324. Supplied as additional material fg324.pdf.