Artificial Intelligence

Neural Networks

Deise Origuella

Exercise 1.

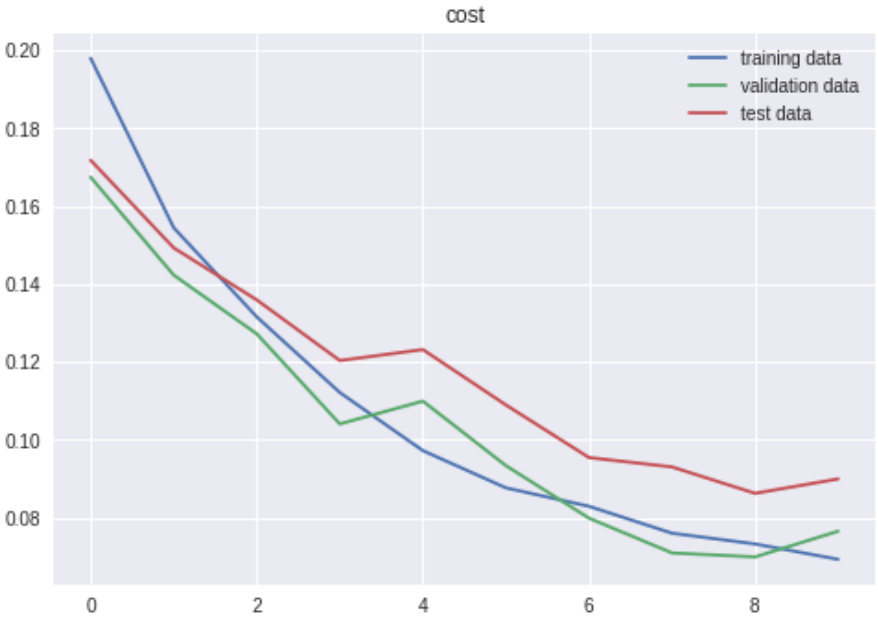
a. Training samples: 9268
   Test samples: 3867

b. There are 17 different classes:
```
['apple', 'apricot', 'avocado', 'cherry', 'grape',
'grapefruit', 'kiwi', 'lemon', 'nectarine', 'orange',
'papaya', 'peach', 'peach flat', 'pear', 'plum',
'pomegranate', 'strawberry']
```

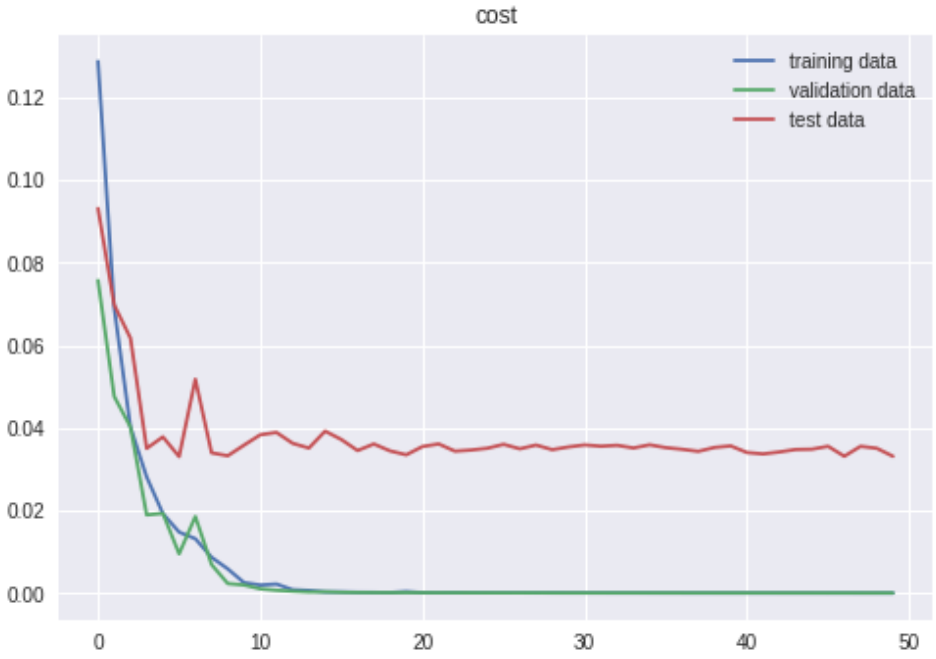
c. Labeling the data according to its class, adjust pixel values range and resampling (all images must have the same spatial characteristics), before adding to the dataset.

Exercise 2.

a. Train and test loss for each epoch:

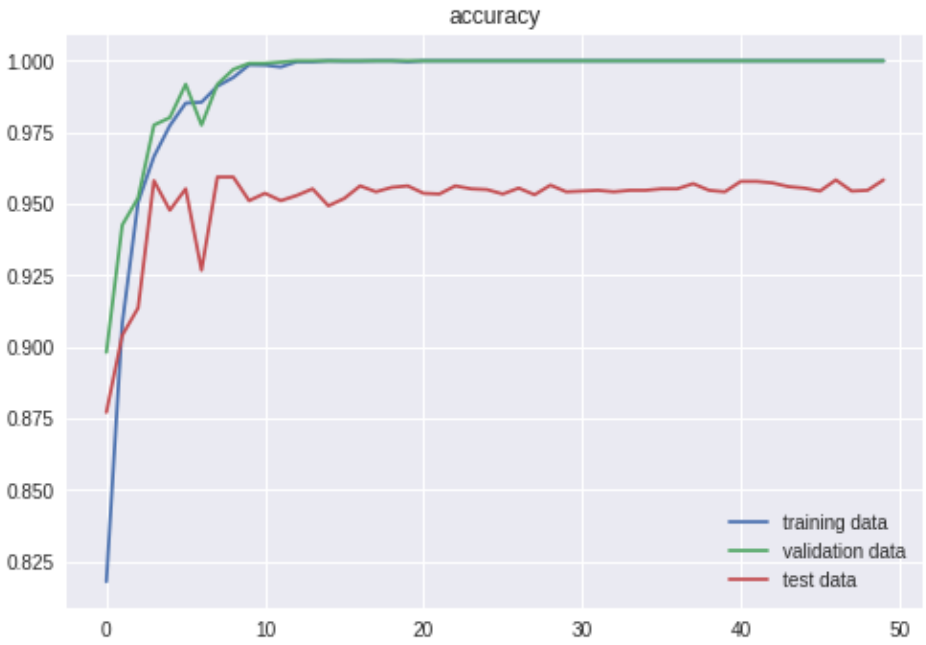


Analyzing the cost function decrease, we noticed that the algorithm has not yet fully converged, since the cost function did not stabilize and it was still decreasing in a high rate. We increased the epochs from 10 to 50 and the learning rate (step size) from 1e-2 to 1e-1 and, as seen by the graph below, the cost function has stabilized.



b. Specialization. By definition, it is a top-down approach in which one higher level entity can be broken down into two other levels. On this case, the fruits database was broken down into apples and non-apples. Generalization is defined as the opposite of specialization.

c. In our database, the accuracy is given by the graph below. It means the ratio of correctly predicted cases to the total number of cases. It could not be used as a loss function, since they calculate the probabilities of a prediction being correct. You want to preserve the information in the probabilities up until the decision is made, when all (other) information is known and the costs of the different types of errors can be guessed.



Exercise 3.

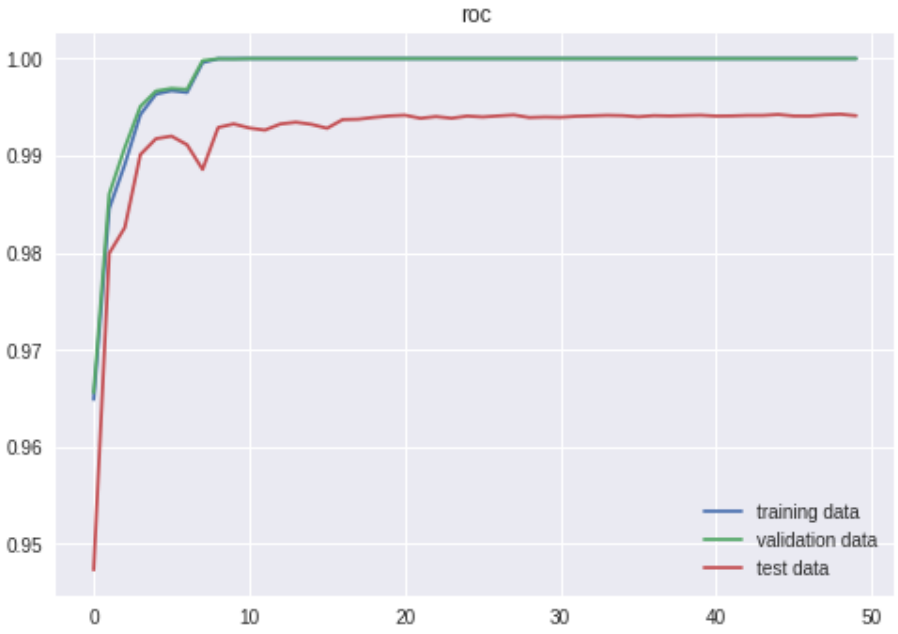
a. Through the output of the model, which is a single number range from 0~1 where the lower value indicates the image has more “apple” characteristics, and higher value if the model shows less “apple” characteristics.

c. Sensitivity measures the proportion of actual positives that are correctly identified as such. Specificity measures the proportion of actual negatives that are correctly identified as such.

d. F1-Score takes both false positives and false negatives into account. It is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost.

e. For our apple classifier, the advisable score would be

f. Below is the ROC curve for the apple model. ROC (Receiver Operating Characteristic) is a curve that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.



Exercise 4.

a. Learning rate, or step size, is a parameter that controls how much we are adjusting the weights of our network with respect the loss gradient. The lower the value, the slower we travel along the downward slope.

b. As explained in question 2a, a smaller learning rate (1e-1) was chosen because the algorithm was not reaching convergence.

c. Setting the learning rate too low will make the algorithm reach convergence, although at the expense of increased computational time.

d. Setting the learning rate too high will decrease computational time, but the algorithm may oscilate in the gradient descent curve, without reaching the local minima.

Exercise 5.

a.

b. Model architecture : (model_apple)

-Input layer

-Pooling layer, a downsampling by a factor 2

-Convolution layer with a kernel size 3 by 3

-Pooling layer with downsampling by factor 2

-Convolution layer with a kernel size 3 by 3

-Pooling layer with downsampling by factor 2

-Convolution layer with a kernel size 4 by 4

-Pooling layer with downsampling by factor 2

-Flatten layer compile the matrix into a vector

-Dense layers compile the number of outputs we want for the model.

c. Having a narrow layer offers the benefit of decreased computational time for the training and learning, although being too narrow would incorrectly train the network. Wider layers (higher number of feature maps) leads to an increase in computational time and training, increasing also the risk of overfitting.

Exercise 6.

a. The output layer in the model for fruit classification was updated to return 17 probability scores (from the categories 0 to 16), instead of 1 (set in the apple model).

b. By definition, the sigmoid function is used for a two-class logistic regression, whereas the softmax function is an extension of the sigmoid function and it is used for multiclass logistic regressions. On this case, the sigmoid function works best at predicting apples vs non-apples, while the softmax function is better suited for calculating probabilities for all of the 17 outputs.

c. Cross-entropy is preferred for classification, while mean squared error is one of the best choices for regression models. The reason is that looking at the problems themselves, in classification you work with very particular set of possible output values

thus MSE is badly defined (as it does not have this kind of knowledge thus penalizes errors in an incompatible way).

d.

Exercise 7.

a. True Positive means a correctly identified match, false Positive means an incorrectly identified match, true negative is a match that is correctly rejected and a false negative is a match that was incorrectly rejected.

b.