Predictive Analytics Assignment — 2018

Submission

The assignment solution should be submitted electronically, and can be a combination of R code and PDF document, by email to `O.Obst@westernsydney.edu.au`. **Include the completed cover sheet that you can find at the end of the document**.

Submission is due on 18 Nov 2018, 11:59pm.

---

1. `Fashion-MNIST` is a dataset of Zalando's article images — consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes. Fashion-MNIST is a replacement for the original MNIST dataset (handwritten digits) for benchmarking machine learning algorithms. It shares the same image size and structure of training and testing splits. The web page `https://github.com/zalandoresearch/fashion-mnist` has more information.

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

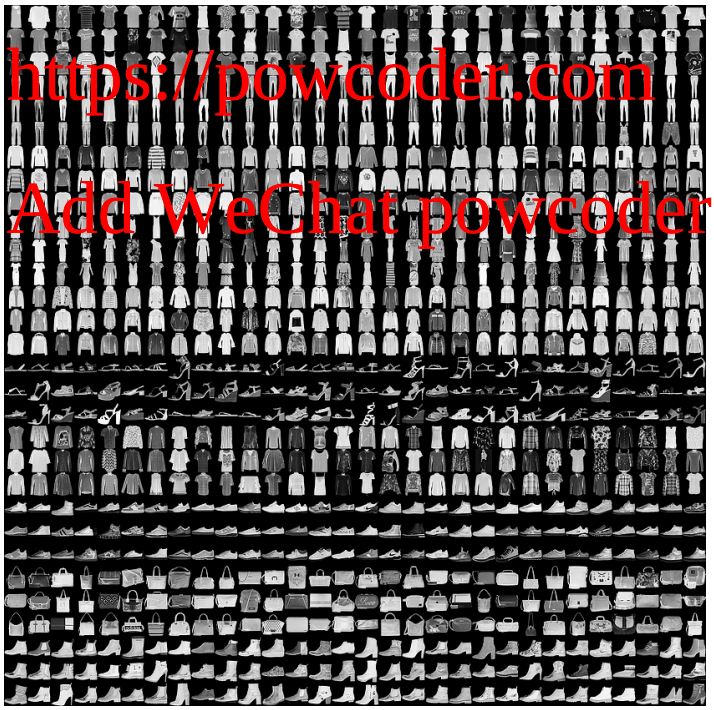

Figure 1: An example for how some the data looks like (each class takes three-rows).

In the vuws data directory for the assignment is a zip archive (`fashionmnist.zip`) that contains 4 data files (training and test sets, split into data and labels), plus some R code (`loader.R`).

You can use the functions in `loader.R` to load the unzipped files, and also to display individual images while you work on your code[1]. The function `load_mnist()` in this file loads all the data (must be in the same directory), and creates variables `train$n`, `train$x`, `train$y`, and `test$n`, `test$x`, `test$y`.

After loading, `train$x` is a 60000 x 784 matrix, each row is one digit (28x28). You can use the call `show_digit(train$x[5,])` to display the 5^{th} training example. The labels (0–9) for the data are in the `train$y` and `test$y` variables.

(a) Using the `nn2` function in the `RANN` library, implement a 3 nearest neighbour classification. **Submit your R code, and the achieved accuracy on the test set** (accuracy: proportion of correct results from total number of classifications).

(b) The `nn2` function also returns the distance to the k nearest neighbours (for details, see the `nn2` help text). Using the inverse of the distance between a data point i and its jth nearest neighbour, $c_{i,\text{NN}_j} = \frac{1}{\text{distance}(\mathbf{x}_i, \mathbf{x}_{\text{NN}_j})}$, implement a weighted version of 3 nearest neighbour classification, so that data points closer to the query point are weighted higher for a prediction. Consider the special case when data points are identical (i.e., when the distance to the nearest example is 0). **Submit your R code, and compare the achieved accuracy on the test set to the unweighted kNN from above**.

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Note: this is a fairly large data set. Dependent on your computer, expect loading of the data into R to take a moment, and 10-30min for kNN classification. To test your code, use only a part of the available data (e.g., the first 10% of the data), and only use the full data set once everything else is working as you expect.

2. Instead of $k = 3$, can you find a better value for k (implement in R), and how do you go about it (describe using your own words). Do not use any of the test data for finding a better k. How well does kNN predict the test data set with your best k (what is the accuracy)?

3. The `linear-train.csv` dataset on vuws contains one output column, and 4 input columns (2200 rows). The goal of this exercise is to predict the output,

[1] The images are fairly small, it's sometimes hard to identify them if displayed in large size

based on a simple linear regression. There is also a `linear-test.csv` data set (100 rows), with just the 4 input columns.

(a) Create a simple linear regression model from the first 2000 rows of training data. To estimate performance on new data, compute the mean square error of the prediction on the last 200 rows, $MSE = \frac{1}{200}\sum_{i=2001}^{2200}(\hat{y}^{(i)} - y^{(i)})^2$. Also create a prediction for the 100 test values. **Submit the MSE you computed, the 100 predictions for the test set, and R code to create the model** (one line is enough).

(b) Have a closer look at each of the inputs, and think about how the prediction could be improved by preprocessing the data. List possible improvements, and apply them to the data. How do you check your preprocessing does help, and which ones do you select? Briefly describe, why do you think your suggestion helps (what does it do to help)? Create a new model, on the modified data (rows 1-2000). Create a new prediction for the last 200 rows, and compute the MSE. **Submit a description of possible steps, and say which ones did you apply, and why. Submit the new MSE. Also submit the R code for your preprocessing steps.**

Assignment Project Exam Help

(c) Perform the same preprocessing step on the test data. Perform a prediction for the test data, using the model from the previous step. **Submit the 100 predictions for the test data.**

https://powcoder.com

Add WeChat powcoder

Assignment 1 Cover Sheet

School of Computing, Engineering and Mathematics

Student Name	
Student Number	
Unit Name and Number	301117: Predictive Analytics
Title of Assignment	Assignment 1
Due Date	18 Nov 2018
Date Submitted	

DECLARATION

Assignment Project Exam Help

I hold a copy of this assignment that I can produce if the original is lost or damaged.

https://powcoder.com

I hereby certify that no part of this assignment/product has been copied from any other students work or from any other source except where due acknowledgement is made in the assignment. No part of this assignment/product has been written/produced for me by another person except where such collaboration has been authorised by the subject lecturer/tutor concerned.

Add WeChat powcoder

Signature: . . . . . .

(Note: An examiner or lecturer/tutor has the right not to mark this assignment if the above declaration has not been signed)

	Question 1	Question 2	Question 3	Total
Mark				
Possible	15 + 20	20	10 + 20 + 15	100

This assignment is worth 40% of the unit assessment tasks.