Machine Learning MC886

University of Campinas (UNICAMP), Institute of Computing (IC)
Assignment #3, 2019s2, Prof. Sandra Avila

Objective

Exploring unsupervised learning techniques by using dimensionality reduction techniques.

Activities

1. (0.5 pts) Baseline: Explore Neural Networks with Fashion-MNIST. What is the accuracy? Describe your Neural Network architecture.
2. (1 pts) Using PCA: Re-do the first experiment considering the PCA dimensionality reduction. Consider three different energies (variance) for reducing the image dimensionality. What are the conclusions when using PCA in this problem? Does the accuracy improve?
3. (1 pts) Using Autoencoders: Re-do the first experiment considering Autoencoders for reducing the image dimensionality. Consider two different latent vector size for reducing the image dimensionality. What are the conclusions when using Autoencoders in this problem? Does accuracy improve?

   Autoencoders are a branch of neural network which attempt to compress the information of the input variables into a reduced dimensional space and then recreate the input data set. Typically the autoencoder is trained over a number of iterations using gradient descent, minimizing the mean squared error. The key component is the "bottleneck" hidden layer. This is where the information from the input has been compressed. By extracting this layer from the model, each node can now be treated as a variable in the same way each chosen principal component is used as a variable in following models.
4. (2.5 pts) Using clustering techniques: Explore two clustering algorithms using the reduced features (PCA or Autoencoders). Do the clusters make sense? Check the validity/quality of your clusters.
5. (4 pts) Prepare a 4-page (max.) report with all your findings. It is UP TO YOU to convince the reader that you are proficient on unsupervised learning techniques, and the choices it entails.
6. (1 pts) You should provide a single Jupyter notebook with your solution (in Python 3 code). You can use the Keras API :-)

Dataset

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 28×28 grayscale image, associated with a label from 10 classes.

Dataset Information:

- You should respect the following traininig/test split: $60,000$ training examples, and $10,000$ test examples. Avoid overfitting.
- The data is available at:
  `https://www.dropbox.com/s/qawunrav8ri0sp4/fashion-mnist-dataset.zip`:
  'train' folder (fashion-mnist_train.csv.zip) + 'test' folder (fashion-mnist_test.csv.zip).

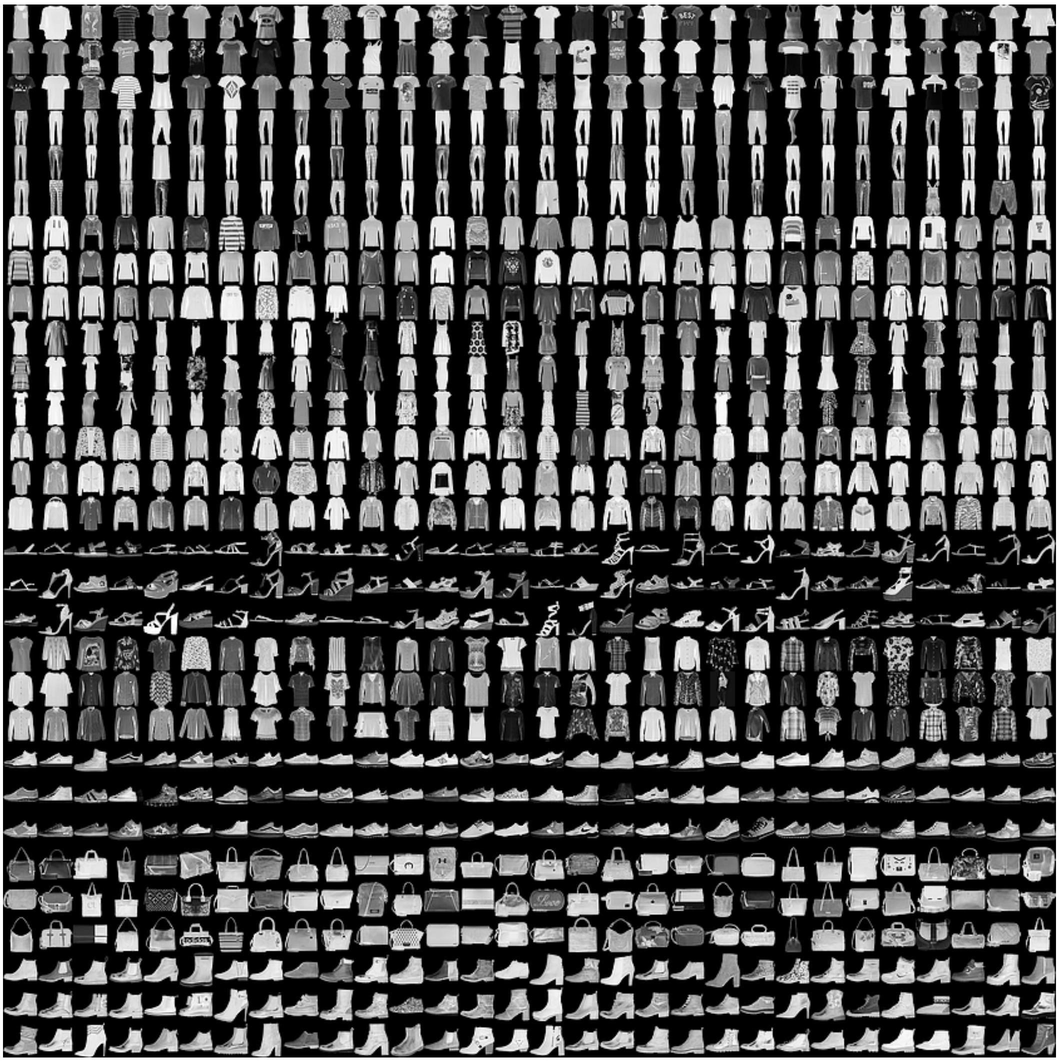

Figure 1: Classes in the dataset, each class takes three-rows. Figure reproduced from `https://github.com/zalandoresearch/fashion-mnist`.

- Each training and test example is assigned to one of the following labels: 0 t-shirt/top, 1 trouser, 2 pullover, 3 dress, 4 coat, 5 sandal, 6 shirt, 7 sneaker, 8 bag, 9 ankle boot.
- Each row is a separate image. Column 1 is the class label. The remaining columns are pixel numbers (784 total). Each value is the darkness of the pixel (1 to 255). Dataset was converted to CSV with this script: `https://pjreddie.com/projects/mnist-in-csv`.

Deadline

Sunday, **November 3rd**, 11h59 pm.

Penalty policy for late submission: You are not encouraged to submit your assignment after due date. However, in case you did, your grade will be penalized as follows:

- November 4th 11h59 pm : grade * 0.75
- November 5th 11h59 pm : grade * 0.5
- November 6th 11h59 pm : grade * 0.25

Submission

On November 4th (Monday), bring your 4-page printed report. The template for report is available at `https://www.dropbox.com/s/nc6d89otr8ekvjd/report-model.zip`. Please, print on both sides of the page. The report should be written in Portuguese or English.

Submit a zip file, with the code and the report (PDF file), via Moodle.

This activity is NOT individual, it must be done in pairs (two-person group).