MUSA650_Spring2022_HW1

February 2, 2022

0.1 HW1 - Basics of ML

Include your code in the relevant cells below. Subparts labeled as questions (Q1.1, Q1.2, etc.) should have their answers filled in or plots placed prominently, as appropriate.

Important note: On this and future homeworks, depending on the data size and your hardware configuration, experiments may take too long if you use the complete dataset. This would be counter-productive, as you will need to run multiple experiments. Accordingly, please start first with a smaller sample that will allow you to run your code in a reasonable time.

Once you complete all tasks, before the final submission, you can allow longer run times and run your code with the complete set. However, if this is still taking too much time or causing your computer to freeze it will be OK to submit experiments using a sample size that is feasible for your setting.

Grading of the homework will not be affected from this type of variations in the design of your experiments.

0.1.1 S1: Understanding the data

- Load MNIST Fashion training and testing datasets with reduced size (n6000 for training and n1000 for testing) https://drive.google.com/drive/folders/1ytbYCba9LUU_8L2V8Bks7pIfu4x1sXQy

Q1.1: What is the number of features in this dataset: _____

Q1.2: What is the number of samples in this dataset: _____

Q1.3: What is the dimensionality of each data sample: _____

0.1.2 S2: Viewing the data

- Select one random example from each category from the training set. Convert the feature vector for the selected example to a 2D image. Display the image with the name of the category

Q2.1: Show the example images

0.1.3 S3: Exploring the dataset

- Select all images in category "Dress" in the training set. Create and display a pixel-wise "average image" for this category.

- Create and display a pixel-wise "standard deviation image" for this category.
- Repeat the items above for the the category "Dress" in the testing set. Compare the average and standard deviation images.
- Repeat the items above for a different category you select.

Q3.1: Plot the 2D mean and std images for dresses in training and testing sets

Q3.2: Plot the 2D mean and std images for the category you selected in training and testing sets

Q3.3: Comment on differences between the mean and std images from training and testing datasets. What do you notice, and what might it mean?

0.1.4 S4: Image distances

- In the training set, find the dress image that is most dissimilar to the average dress image. Show it as a 2D image
- In the training set, find the dress image most similar to mean image. Show it as a 2D image

Hint: You can use the "euclidian distance" as your similarity metric. Given that an image i in category dress is represented with a flattened feature vector v_i , and the mean image for category dress with the feature vector v_m, the distance between these two images can be calculated using the vector norm of their differences (| v_i - v_m |)

Q4.1: What is the index of most dissimilar dress image: _____

Q4.2: What is the index of most average looking dress image: _____

Q4.3: Plot the most dissimilar dress image in 2D: _____

Q4.4: Plot the most similar dress image in 2D: _____

0.1.5 S5: Image distances, part 2

- Repeat questions S3 and S4 after binarizing the images first

Q5.1: What is the index of most dissimilar dress image: _____

Q5.2: What is the index of most similar dress image: _____

Q5.3: Did the answer change after binarization? How do you interprete this finding?

0.1.6 S6: Binary classification between dresses and sandals

- Select images from these two categories (dresses and sandals) in the training dataset
- Split them into two sets (Set1, Set2) with 70% to 30% random split
- Replace category labels as 0 (dress) and 1 (sandal)
- Use Set1 to train a linear SVM classifier with default parameters and predict the class labels for Set2
- Use Set2 to train a linear SVM classifier with default parameters and predict the class labels for Set1

Q6.1: What is the prediction accuracy using the model trained on Set1: _____

Q6.2: What is the prediction accuracy using the model trained on Set2: _____

0.1.7 S7: Binary classification between dresses and sandals, part 2

- Select images from these two categories (dresses and sandals) in the training dataset
- Split them into two sets (Set1, Set2) with 20% to 80% random split
- Select images from these two categories in the testing dataset
- Replace category labels as 0 (dress) and 1 (sandal)
- Use Set1 to train a linear SVM classifier with default parameters and predict the class labels for testing images
- Use Set2 to train a linear SVM classifier with default parameters and predict the class labels for testing images

Q7.1: What is the prediction accuracy using the model trained on Set1: _____

Q7.2: What is the prediction accuracy using the model trained on Set2: _____

Q7.3: Comment on the differences in the accuracy of the two models. If there is a difference, why do you think that is?

0.1.8 S8: k-NN Error Analysis

- In training and testing datasets select the images with labels: Dress, Coat, Sandal, Shirt or Sneaker
- Train a k-NN classifier using 4 to 40 nearest neighbors with a step size of 4
- Calculate and plot overall testing accuracy for each experiment

Q8.1: For k=4 what is the label that was predicted with lowest accuracy: _____

Q8.2: For k=20 what is the label that was predicted with lowest accuracy: _____

Q8.3: What is the label pair that was confused most often (i.e. class A is labeled as B, and vice versa): _____

Q8.4: Visualize 5 mislabeled samples with their actual and predicted labels

0.1.9 S9: Feature extraction

- We describe each image by using a reduced set of features (compared to n=784 initial features for each pixel value) as follows:

  1. Binarize the image (background=0, foreground=1)
  2. For each row i, find m_i, the index of the first non-zero pixel (m_i will be a value from 0 to 28 - if all pixels in a row are zero then m_i=28)

  Example image: 0 0 0 0 1 1 1 … 0 0 1 1 1 0 0 … 0 0 0 0 0 1 0 … … Extracted features: [4, 2, 5, …]

  This strategy gives a feature vector with n=28 features (for each row)

  Repeat classification experiments in Q6 using this reduced feature set.

Q9.1: What is the prediction accuracy using the model trained on Set1: _____

Q9.2: What is the prediction accuracy using the model trained on Set2: _____

0.1.10 BONUS:

Repeat S9 by extracting 28x4 features this time by applying the same rule in four different directions and concatenating them (left->right, right->left, top->bottom, bottom->top)