CS559 HW4

Dimitrios Haralampopoulos

2023-04-08

You shall submit a zip file named Assignment4_LastName_FirstName.zip which contains:

- python files (.ipynb or .py) including all the code, comments, plots, and result analysis. You need to provide detailed comments in English.
- pdf file including explanations and answers for non-coding questions.

1. (10 points) Implement a basic k-NN model on the yeast dataset (data is provided as an attachment in Canvas). The task is to predict the compartment in a cell that a yeast protein will localize to based on the properties of its sequence. More details can be found in the dataset description. Apply cross-validation and report your best performance.

- (7 pts) Implementation of k-NN model.
- (3 pts) Apply cross-validation and report the best results.

2. (5 points) Suppose we clustered a set of N data points using two different clustering algorithms: k-means and Gaussian mixtures. In both cases we obtained 5 clusters and in both cases the centers of the clusters are exactly the same. Can a few (say 3) points that are assigned to different clusters in the kmeans solution be assigned to the same cluster in the Gaussian mixture solution? If no, explain. If so, sketch an example or explain in 1-2 sentences.

It is possible that 3 points that were classified differently in the kmeans solution can be classified in the same cluster in the gaussian mixture solution. Just because the cluster centers and number of clusters are the same, it doesn't imply that the clusters will be of the same size in both models. Additionally, both models converge on a local minimum, which may not be the same local minimum or a local minimum that is optimal for the dataset.

3. (10 points) Do the following statements hold in each of the above networks ? Please explain your reasoning

- (5pts) $A \perp C | B, D$ in Graph (a)

$A \not\perp C | B, D$ since C is dependent on B and D which are both dependent on A.

- (5pts) $A \perp C | B, D$ in Graph (b)

$A \perp C | B, D$ since D and B are dependent on both A and C, but A and C do not depend on one another through D and B, hence they are marginally independent with one another.

- (5pts) $B \perp D | A, C$ in Graph (a)

$B \perp D | A, C$ by D-Separation. D and B are dependent on A, but are marginally independent of C.

- (5pts) $B \perp D | A, C$ in Graph (b)

$B \not\perp D | A, C$ by D-Separation. D and B both marginally depend on A and C.

4. (30 points) Given the matrix X whose rows represent different data points, you are asked to perform a k-means clustering on this dataset using the Euclidean distance as the distance function. Here k is chosen as 3. The Euclidean distance d between a vector x and a vector y both in $\mathcal{R}^d$ is defined as $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$. All data in X were plotted in Figure1. The centers of 3 clusters were initialized as $\mu_1 = (6.2, 3.2)$(red), $\mu_2 = (6.6, 3.7)$(green), $\mu_3 = (6.5, 3.0)$(blue).

(a). (24 pts) Implementation of K-Means (submit your code)

(b). (1.5 pts) What's the center of the first cluster (red) after one iteration? (Answer in the format of $[x_1, x_2]$, round your results to three decimal places)

(c). (1.5 pts) What's the center of the second cluster (green) after two iterations?

(d). (1.5 pts) What's the center of the third cluster (blue) when the clustering converges?

(e). (1.5 pts) How many iterations are required for the clusters to converge?

5. (25 points) In this question you will implement the EM algorithm for Gaussian Mixture Models. A good read on gaussian mixture EM can be found at this link. A sample dataset for this problem can be downloaded in canvas files. For this problem:

- n is the number of training points
- f is the number of features
- k is the number of gaussians
- X is an $n \times f$ matrix of training data
- w is an $n \times k$ matrix of membership weights. $w(i, j)$ is the probability that x_i was generated by gaussian j
- π is a $k \times 1$ vector of mixture weights (gaussian prior probabilities). π_i is the prior probability that any point belongs to cluster i
- μ is a $k \times f$ matrix containing the means of each gaussian
- Σ is an $f \times f \times k$ tensor of covariance matrices. $\Sigma(:, :, i)$ is the covariance of gaussian i

6. (20 points) Develop a Convolutional Neural Network (CNN) model to predict a handwritten digit images into 0 to 9 (You can use Keras or other packages). The pickled file represents a tuple of 3 lists: the training set, the validation set and the testing set. Each of the three lists is a pair formed from a list of images and a list of class labels for each of the images. An image is represented as numpy 1-dimensional array of 784 (28 x 28) float values between 0 and 1 (0 stands for black, 1 for white). The labels are numbers between 0 and 9 indicating which digit the image represents. The code block below shows how to load the dataset.

   import pickle, gzip, numpy

   # Load the dataset f = gzip.open('mnist.pkl.gz', 'rb') u = pickle._Unpickler(f) u.encoding = 'latin1' train_set, valid_set, test_set = u.load() f.close()

- (10 pts) Implementation of CNN.
- (1 pts) Choose the proper activation and loss function.
- (2 pts) Plot the train, validation, and test errors as a function of the epochs.
- (2.5 pts) Report the best accuracy on the validation and test data sets. Discuss the parameter choices such as the filter size, number of filters etc.
- (1 pts) Apply early stopping using the validation set to avoid overfitting.
- (2 pts) Give a brief description of your observations.
- (1.5 pts) Does pooling make the model more or less sensitive to small changes in the input images? Why? By small changes, we mean moving the input images to the left or right, rotating them slightly etc.