EECS 491 Assignment 4

Due Fri Apr 12 before midnight. 100 points total.

Submitting assignments to Canvas

- For jupyter notebooks, submit the .ipynb file and an export of the notebook in both html and pdf formats.
  - Check that the exports accurately represent the latest state of your notebook.
  - If you use interactive plots, make sure the exports for the static file is representative of the points you wish to make.
- If you are not using a notebook, write up your assignment in latex and submit a pdf with your code. The writeup should include relevant code snippets with description if it can fit on a page. Do not include binaries or large data files.
- Use the following filename format:

  `EECS491-A4-yourcaseid.ipynb` (or `.pdf` or `.html`)

- If your code uses extra files, put them in a directory, zip it, and use the name

  `EECS491-A4-yourcaseid-files.zip`

  Do not use other compression formats. You can include your notebook or code in this directory, but be sure to also submit these separately to Canvas as above.

Exercise 1. Multivariate Gaussians (10 points)

1.1 (5 pts) Consider the 2D normal distribution

$$p(x, y) \sim \mathcal{N}(\mu, \mathbf{\Sigma})$$

Define three separate 2D covariance matrices $\mathbf{\Sigma}$ for each of the following cases: x and y are uncorrelated; x and y are correlated; and x and y are anti-correlated.

1.2 (5 pts) Compute the principal axes for each of these distributions, i.e. the eigenvectors of the covariance matrices.

Exercise 2. Linear Gaussian Models (20 pts)

Consider two multi-dimensional Gaussian random vector variables

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu_{\mathbf{x}}, \mathbf{\Sigma}_x)$$
$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mu_{\mathbf{z}}, \mathbf{\Sigma}_z)$$

Now consider a third variable that is the sum of the first two:

$$\mathbf{y} = \mathbf{x} + \mathbf{z}$$

2.1 (5 pts) What is the expression for the distribution $p(\mathbf{y})$?

2.2 (5 pts) What is the expression for the condidtional distribution $p(\mathbf{y}|\mathbf{x})$?

2.3 (10 pts) Write code that simulates this data and illustrate the results.

Exercise 3. Dimensionality Reduction and PCA (25 pts)

In this quesiton you will use principal component analysis to reduce the dimensionality of your data and analyze the results.

3.1 (5 pts) Find a set of high dimensional data. It should be continuous and have at least 6 dimensions, e.g. stats for sports teams, small sound segments or images patches also work. Note that if the dimensionality of the data is too large, you might run into computational efficiency problems using standard methods. Describe the data and illustrate it, if appropriate.

3.2 (5 pts) Compute the principal components of the data. Plot a few of the largest eigenvectors and interpret them in terms of how there are modeling the structure of the data.

3.3 (5 pts) Plot, in decreasing order, the cumulative percentage of variance each eigenvector accounts for as a function of the eigenvector number. These values should be in decreasing order of the eigenvalues. Interpret these results.

3.4 (10 pts) Plot the original data projected into the space of the two principal eigenvectors (i.e. the eigenvectors with the largest two eigenvalues). Be sure to either plot relative to the mean, or subtract the mean when you do this. Interpret your results. What insights can you draw? Interpret the dimensions of the two largest principal components. Which dimensions of the data are correlated? Or anti-correlated?

Exercise 4. Gaussian Mixture Models (25 pts)

4.1 (10 pts) Use the EM equations for multivariate Gaussian mixture model to write a program that implements the Gaussian Mixture Model to estimates from an ensemble of data the means, covariance matrices, and class probabilities. Choose reasonable values for your initial values and a reasonable stopping criterion. Explain your code and the steps of the algorithm. Do not assume a diagonal or isotropic covariance matrices.

4.2 (5 pts) Write code to plot the 3-sigma contours of each Gaussian overlayed on the data (try to find a library function to plot ellipses). Illustrate with an example.

4.3 (5 pts) Define a two-model Gaussian mixture test case, synthesize the data, and verify that your algorithm infers the (approximately) correct values based on training data sampled from the model and plotting the results.

4.4 (5 pts) Apply your model to the Old Faithful dataset (supplied with the assignment files). Run the algorithm for the cases $K = 1$, $K = 2$, and $K = 3$. For each case, plot the progression of the solutions at the beginning, middle, and final steps in the learning. For each your plots (you should have 9 total), you should also print out the corresponding values of the mean, covariance, and class probabilities.

Exploration (20 points)

Select a topic related to the problems above or the recent lectures and write your own exercise. It should aim to teach or explore a concept you don't understand or found interesting.

Grading rubric:

- Clarity of explanation. Could another student read and do this? (5 pts)
- Novelty or distinctness. Does it complement or go beyond what was covered above? (5 pts)
- Does the exercise teach something about the concept(s)? (5 pts)
- How deeply does it explore the concept(s)? (5 pts)

In [ ]:

```
1
```