EECS 491 Assignment 4

Due Sat Apr 25 before midnight. 100 points total.

Submitting assignments to Canvas

- For jupyter notebooks, submit the .ipynb file and a pdf export of the notebook.
- Make sure you check that the pdf export represents the latest state of your notebook and that the equations and figures are properly rendered.
- If your are not using notebooks, writeup your assignment using latex and submit a pdf with your code. The writeup should include relevant code with description if it can fit on a page.
- Use the following format for filenames:
  - `EECS491-A4-yourcaseid.ipynb`
  - `EECS491-A4-yourcaseid.pdf`
- If you have more than these two files, put any additional files in a directory named `EECS491-A4-yourcaseid` . Do not include binaries or large data files. Then zip this directory and submit it with the name `EECS491-A4-yourcaseid.zip` . Do not use other compression formats. The `.ipynb` file can be included in the zipped directory, but make sure you submit the `.pdf` file along with the `.zip` file. This is so it appears at the top level on canvas, which allows for easier grading.

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. Plot samples from these distributions to show these properties. Use a different mean for each. Make sure your plots show the density.

1.2 (5 pts) Compute the principal axes for each of these distributions, i.e. the eigenvectors of the covariance matrices. Use can use a linear algebra package. Plot the samples again, but this time overlay the 1, 2, and 3-sigma contours and with the scaled eigenvectors.

Exercise 2. Linear Gaussian Models (20 pts)

Consider two independent multi-dimensional Gaussian random vector variables

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

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 to illustrate the result in Q2.1. Show both the components of $\mathbf{y} = \mathbf{x} + \mathbf{z}$ and that the sampling from the analytic result is the same as adding two samples.

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)

Like in previous assignments, in this exercise you have more lattiude and are meant to do creative exploration. The intention is for you to teach yourself about a topic beyond what's been covered above. Please consult the rubric below for what is expected.

Exploration Grading Rubric

Exploration problems will be graded according the elements in the table below. The scores in the column headers indicate the number of points possible for each rubric element (given in the rows). A score of zero for an element is possible if it is missing entirely.

	Substandard (+1)	Basic (+2)	Good (+3)	Excellent (+5)
Pedagogical Value	No clear statement of idea or concept being explored or explained; lack of motivating questions.	Simple problem with adequate motivation; still could be a useful addition to an assignment.	Good choice of problem with effective illustrations of concept(s). Demonstrates a deeper level of understanding.	Problem also illustrates or clarifies common conceptual difficulties or misconceptions.
Novelty of Ideas	Copies existing problem or makes only a trivial modification; lack of citation(s) for source of inspiration.	Concepts are similar to those covered in the assignment but with some modifications of an existing exericse.	Ideas have clear pedagogical motivation; creates different type of problem or exercise to explore related or foundational concepts more deeply.	Applies a technique or explores concept not covered in the assignment or not discussed at length in lecture.
Clarity of Explanation	Little or confusing explanation; figures lack labels or useful captions; no explanation of motivations.	Explanations are present, but unclear, unfocused, wordy or contain too much technical detail.	Clear and concise explanations of key ideas and motivations.	Also clear and concise, but includes illustrative figures; could be read and understood by students from a variety of backgrounds.
Depth of Exploration	Content is obvious or closely imitates assignment problems.	Uses existing problem for different data.	Applies a variation of a technique to solve a problem with an interesting motivation; explores a concept in a series of related problems.	Applies several concepts or techniques; has clear focus of inquiry that is approached from multiple directions.

In [ ]: