Assignment 4– Probabilistic regression and classification, Generative Models, and Bayes Nets.

Assignment overview. This assignment is designed to introduce you the probabilistic regression and classification, and generative models as well as Bayes Nets. This assignment requires you apply EM with a Gaussian mixture model on the iris dataset, and also use lea.Lea for discrete probabilistic modeling. Follow the steps as indicated and complete the tasks. You are expected to figure out details of syntax by consulting Python's Help. Please answer each question by copying and commenting as needed on the material that you produce in the IPython console as well as all scripts you are asked to create, or use Jupyter Notebook and download it as a Python file.

Submission. Create a folder called ML_Assignment4 and put all the files inside the folder. Compress this folder to create either ML_Assignment4.zip or ML_Assignment4.rar. Submit this compressed folder as your assignment submission on Brightspace.

Submission deadline. Tuesday, Oct 12, 10:00 pm.

Late submission policy. If submitted after the due date, the penalty will be 10% per day.

Academic Integrity. Dalhousie academic integrity policy applies to all submissions in this course. You are expected to submit your own work. Please refer to and understand the academic integrity policy, available at https://www.dal.ca/academicintegrity

Python: We will be using Python for the programming exercises based on scientific Python libraries like

- **numpy** - mainly useful for its *N*-dimensional array objects.
- **matplotlib** - 2D plotting library producing publication quality figures.
- lea – mainly useful for working with discrete probability distributions.

If you have a question: Teaching Assistants (TAs) will be present during the labs to help you with any questions you may have. If you still have questions, feel free to email me at tt@cs.dal.ca.

Questions:

1. **[70 marks]** This Assignment requires you to write a Python script file called sol1.py to load the ***iris*** dataset from the first assignment and apply EM with a Gaussian mixture model on IRIS data. You are not allowed to use any Python public libraries like sklearn and scipy. You might compare the results of your program with sklearn models, but the whole exercise is to write the algorithm yourself.

   1.1. **[70 marks, 35 marks for Grads]** Write a program to implement EM with a Gaussian mixture model on the ***iris*** dataset and plot the Sepal data points based on the obtained clusters. Try different number of classes (k=2,3,4). Hint: use `numpy.linalg.pinv` and also use `numpy.copy` to temporary save a vector. You might plot the points using the obtained RGB colour values (i.e. if you have three clusters, there are three probability estimates of a data point belonging to each class).

1.2. **Graduate students only [20 marks]** Evaluate the prediction quality with different number of assumed classes (k=2,3,4). Explain briefly your evaluation method (maximum half a page).

2. **[30 marks, 15 marks for Grads]** This Assignment requires you to write a Python script file called `sol2.py` to calculate some inference of a simplified version of the Car repair example from the manuscript. Given is are the following probabilities:

   The marginal probability that the alternator is broken is 1/1000 and the marginal probability that the fan belt is broken is 2/100. The probability that the battery is charging when either the alternator or the fan belt is broken is zero. However, even if both are working there is a 5/1000 probability that the battery is not charging. When the battery is not charging then there is a 90% chance that the battery is flat, though even if the battery is charging then there is a 10% chance that the battery is flat. Finally, the car does not start if either the battery is flat, or there is no gas, or the starter is broken. However. Even if these three conditions don't hold there is a 5% chance that the car won't start.

   2.1. Draw the causal model of this system.
   2.2. What is the probability that the alternator is broken given that the car won't start?
   2.3. What is the probability that the fan belt is broken given that the car won't start?
   2.4. What is the probability that the fan belt is broken given that the car won't start and the alternator is broken?
   2.5. What is the probability that the alternator and the fan belt is broken given that the car won't start?

   Hint: You might use `lea.Lea` methods.

3. **Grads, only [30 marks]** Naïve Bayes:

   This Assignment requires you to write a Python script file called `sol3.py` to test the Naïve Bayes on the 20newsgroups dataset. Similar to the second question of assignment 2, you should read the data and work with sparse data in python. You should then write a Naïve Bayes program on your own (not using library function) to implement the binomial version of the Naïve Bayes rule outlined in the manuscript. Please provide the results in form of a confusion matrix.