Assignment 6:

Due Nov 16, 2016, 4pm by email to dalhousieml2016@gmail.com with subject line A6.

CSCI4155 (undergraduate students only):

1. Write a program to demonstrate **k-mean clustering** as in the example shown in class. The example points should be sampled from Normal distributions with unit variance and means $\mu_1=(1,1)$ and $\mu_2=(5,5)$. The starting value for the means should be $\mu_1^{start}=(1,5)$ and $\mu_2^{\,start}=(5,1)$. Are the classes always consistently labeled? Explain why or why not. Include you program in the submission.
   Note that this is an individual assignment and that you should not share your code with other students in the course.

2. Explain with your own word in not more than ½ page what is **unsupervised learning.**

3. Explain in your own words in not more than ½ page what is a **generative model**.

CSCI6505 (graduate students only)

1. **Naïve Bayes**:
   This is a small project to test the Naïve Bayes on the 20newsgroups dataset. For this you should go through the tutorial
      http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
   This tutorial shows you how to read the data and how to 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 and attach your program to the email submission.

2. Write an **autoencoder** for the MNIST data set and explore how the reconstruction accuracy is influenced by the compression rate. Discuss your implementation and results in less than one page.