COGSCI 109: Homework Assignment 2 (3 pages), due: Monday October 25, 10:30 am (before class)

Note: Please note this assignment is 3 pages. Check the course web page and wiki for information and announcements:

(A) (6 points) Are the following functions proper probability density functions? If no, why not?

$$f_1(x) = \begin{cases} 1 & : & A \leq x \leq B \\ 0 & : & \text{else} \end{cases} \quad . \tag{1}$$

$$f_2(x) = \begin{cases} 0 & : & |x| > 1 \\ x^2 & : & |x| \leq 1 \end{cases} \quad . \tag{2}$$

$$f_3(x) = \begin{cases} \lambda \exp(-\lambda x) & : & x \geq 0 \\ 0 & : & \text{else} \end{cases} \quad , \text{where } \lambda > 0 \ . \tag{3}$$

(B) **You should not integrate to answer these questions. Review the properties of the normal density function and expectation.**

What is

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-5)^2}{2}} =$$

[2 points]

What is

$$\int_{-\infty}^{\infty} e^{\frac{-(x-5)^2}{2}} =$$

[2 points]

What is

$$\int_{-\infty}^{5} e^{\frac{-(x-5)^2}{2}} =$$

[2 points]

What is

$$\int_{-\infty}^{\infty} x e^{\frac{-(x-5)^2}{2}} =$$

[2 points] NOTE THAT THERE IS AN X IN THIS INTEGRAL.

(C) **Mean and Variance.**(4 points) For expected values the following relations hold: E[aX]=aE[X] and E[X+Y] = E[X] + E[Y], where X and Y are random variables and a is a constant. Consider the scalar random variable X with mean μ and variance σ^2. Consider constructing the following new random variables from X (a and b are constant numbers): P=X+ a, Q=bX. Use the definitions of mean and variance and the relations above to derive expressions for the mean and variance of P and Q expressed as functions of μ and σ.

(D) **Expected values.** (4 points) For expected values the following relations hold: E[aX]=aE[X] and E[X+Y] = E[X] + E[Y], where X and Y are random variables and a is a constant. Using these results, now prove the following formula for the calculation of the variance of a scalar random variable X:

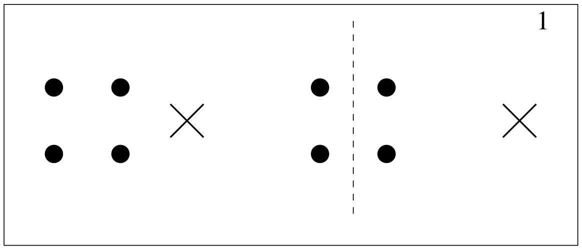
$$var(X) = E[X^2] - (E[X])^2$$

Hint: start from the definition: $var(X) = E[(X - \mu)^2]$, where $\mu = E[X]$ is the mean of X, and apply the above rules to get it into the right form.

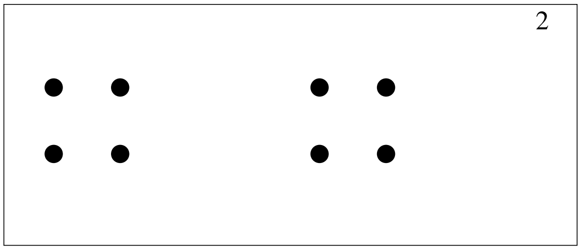
(E) **K-means** (6 points) Consider the example below. Cluster centers (or means) are depicted as X's, data points are dots. The box on the top (with 1 in the corner) represents a system with 8 data points and K=2 cluster centers. To help you answer the question, the dividing line midway between the two centers is shown by a dashed line.

a) In the box in the center (with 2 in the right corner), draw the next position of the means (in the next iteration) and **b)** in the box on the bottom (with 3 in the corner), draw the next position after that (in the next iteration) of the means. You may want to continue drawing dashed lines to keep track of how the space is being divided (and show us your work).

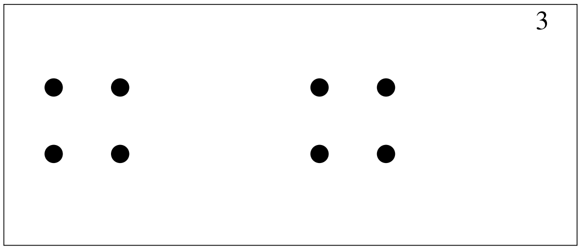
Initial State



a)



b)



c) In general (not just for the example above) when will the K-means algorithm stop? (what condition must be fulfilled? - be as specific as you can). You do not have to write it as an equation but you may.

(F) **Hypothesis Testing** (12 points) Consider the situation of a drug company testing a lot of different drugs to find one that reduces the length of cold symptoms. Let's assume that there are 10000 drugs to test and that for each drug, the company finds fifteen people that will use the drug. Assume that we know that the mean length of cold symptoms without any drugs is 7 days. The company will ask the people to report back with the number of days (which does not have to be a whole number) more or less than 7 for which they had cold symptoms (e.g. -2.3 for 4.7 days of cold symptoms).

In this question you will use the matlab command **ttest**. **h=ttest(datavec)** will return 0 if the null hypothesis that the data in datavec come from a distribution with mean 0 (e.g. the drug has no effect) can NOT be rejected at the 5% significance level. 1 will be returned if the null hypothesis can be rejected at the 5% significance level. Do a help on **ttest** in Matlab for more information.

a) Simulate the situation of performing 10000 random draws of 15 samples from a normal distribution with mean 0 and standard deviation 1 (use the **randn** command). Write down the command you used.

b)What does this simulate in terms of our drug test/ cold analogy?

c) For each of the 10000 random draws, test the hypothesis that the sample has mean 0 (using **ttest**). (save the combination of a) and c) in a file called testcheck.m) Hand in testcheck.m

d) How many tests rejected the null hypothesis? Why is this?

e) Rewrite your code to perform the above experiment without a loop (Hint: Note that ttest can be applied to a matrix – do help ttest on matlab). Save the code in a file called testcheck2.m. Hand in testcheck2.m

f) Using **tic** and **toc** report on the time to run the testcheck.m code and the testcheck2.m code.