

**Homework 1. Due midnight Sunday Feb 5 (a little less than 2 weeks from now).  
To be submitted electronically as a pdf to a dropbox in Angel.**

1. Consider the (unnormalized) 2D bivariate distribution  $f(x,y)$

	$x = 1$	$x = 2$	$x = 3$	$x = 4$
$y = 0$	10	10	10	10
$y = 1$	10	20	20	0
$y = 2$	0	10	0	0

For each of the following, give your answer as a NORMALIZED distribution (sums to 1)

- a) What is the marginal distribution  $g(x)$  ?
- b) What is the marginal distribution  $h(y)$  ?
- c) What is the conditional distribution  $p(x | y=0)$ ?
- d) What is the conditional distribution  $p(y | x=2)$ ?
- e) Are the random variables  $x$  and  $y$  are independent?
- f) What is the expected value of  $x$ ?
- g) What is the expected value of  $\text{sqrt}(y)$ ?

**2a.** Let  $f(x,y)$  be a general bivariate distribution, either discrete or continuous, and let  $g(x)$  be its marginal distribution with respect to  $x$ . Prove that the expected value of  $x$  with respect to  $f(x,y)$  is equal to the expected value of  $x$  with respect to  $g(x)$ . That is, we want to show that the joint mean of  $x$  is equal to the marginal mean of  $x$ .

**2b.** Come up with a simple counterexample of a discrete distribution  $f(x,y)$  and its marginal distribution  $g(x)$  to show that the  $x$  coordinate of the MODE of  $f(x,y)$  is not necessarily the same as the mode of the marginal  $g(x)$ . Recall that the mode of a discrete distribution is the location where the highest probability value  $p_{\max}$  occurs, assuming that value is unique (there are alternate definitions if the distribution has more than one mode... choose your example so that  $f(x,y)$  and  $g(x)$  each have only one mode).

**3.** Recall that we briefly discussed in class taking the derivative of a scalar function with respect to a vector (for example, we showed that  $\partial x^T A x / \partial x = 2x$ ). Show that

$$\frac{\partial x^T A x}{\partial x} = (A + A^T)x$$

**4.** In class, we reviewed the principle of maximum likelihood estimation for estimating the parameters of a distribution from a set of independent and identically distributed (i.i.d.) samples, and derived estimates of mean and variance for a circularly symmetric Gaussian distribution. Consider instead a set of  $N$  binary samples  $X = \{x_1, x_2, \dots, x_N\}$  where each  $x_i \in \{0, 1\}$ , drawn independently from a Bernoulli distribution with parameter  $\mu$ . The samples might represent coin flips (tails=0, heads=1), or might represent black and white pixel values (black=0, white=1), or whatever. Recall that a Bernoulli distribution has the form

$$P(x_i | \mu) = \mu^{x_i} (1 - \mu)^{(1-x_i)} .$$

Derive the maximum likelihood estimate for  $\mu$  given a set of  $N$  i.i.d. samples from this distribution.

**5.** Implement the k-means algorithm, described in class, and apply it generate  $K$  clusters (or codewords) using sample data from a problem of interest to you. Come up with a way to visualize the clusters, so we can see if it is doing a good job. The visualization will probably be domain dependent, so will have to be designed to make sense for the problem you are trying to solve. For example, if the vector sizes are small enough (say 1D or 2D points), you could plot the estimated cluster centroids, and plot each data point in a color according to which cluster it is estimated to belong to. By the way, I know that k-means is already built into the statistics toolbox in matlab, and there are a million implementations floating around on the web, but I want you to implement it yourself (it really is not hard at all) to get a better understanding of how the algorithm works. If this is too easy for you, you could try implementing and testing a version where you also adjust/estimate the variance of each cluster along with the cluster center.