Name: ________________________________

STAT161/261: Quiz

1. Basic PCA. Given 3 data points in 2-d space, $(1,1)$, $(2,2)$ and $(3,3)$:

   (a) What is the first principle component vector? (List the eigenvector and eigenvalue.)

   (b) What is the proportion of variance if we project onto that first principal component? Remember the POV for projecting onto K of p components is:

   $$\text{POV} = \frac{\lambda_1 + \cdots + \lambda_K}{\lambda_1 + \cdots + \lambda_p}.$$

   (c) If we use the first principal component vector to represent our data, and then we go represent those vectors in the original space, what will be the reconstruction error? Explain.

2. Basic k-means: Suppose you are given the 3 data points in Fig. 1. Starting with two cluster centers at $(0,0)$ and $(0,2)$, what does k-means converge to?

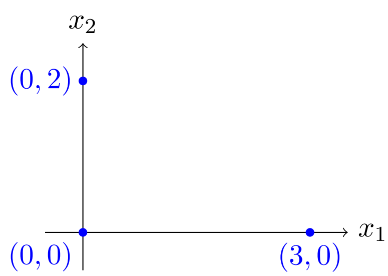

Figure 1: Data points for Problem 2.

3. You have N measurements from data (x_i, y_i) that can be modelled as

   $$y = w_0 + w_1 x + w_2 x^2 + d, \quad d \sim \mathcal{N}(\mu, 1),$$

   where d is Gaussian with *non-zero* mean of μ and variance 1.

   (a) Find the maximum likelihood estimate of $[w_0, w_1, w_2]$.

   (b) What is the bias of the estimator for w_2?

4. Nonparametric: Given data $x = \{1, 2, 4, 8\}$:

   (a) Draw the density estimate:

   $$\hat{p}_1(x) = \frac{1}{Nh} \sum_i w\left(\frac{x - x_i}{h}\right).$$

   where $w(x)$ is the rectrangular window and $h = 1$.

(b) Draw the 2-nearest neighbor estimated pdf

$$\hat{p}_2(x) = \sum_i \frac{1}{Nd_2(x_i)} w\left(\frac{x - x_i}{d_2(x_i)}\right).$$

(c) What is the estimate for $P(X >= 2)$ in (a) and (b).

5. True / False: PCA is always good tool to discriminate between classes.

6. True / False: The log-likelihood of the data will always be non-decreasing through successive iterations of the expectation maximization algorithm. [SR: Changed increasing to non-decreasing]

7. True / False: A method of assessing reliability and validity of your nonparametric clustering algorithm is to use methods based on different h's and k's, and compare the results. [SR]

8. You take N independent measurements from an exponential distribution where each measurement x_i has a density:

$$p(x_i|\theta) = \theta \exp(-x_i\theta), \quad x_i, \theta \in (0, \infty).$$

(a) Find the maximum likelihood estimate of θ given data x_i, $i = 1, \ldots, N$, when x_i are drawn independently N times from the distribution. Use steps i and ii:

(i) Write down the log likelihood function of theta given the N measurements.

(ii) Maximize it by taking the derivative. Solve for θ.

(b) Write down the log likelihood equation you would have for take the MAP estimate of *theta* above if θ had a prior of $\mathcal{N}(0, 1)$. DO NOT differentiate and find the MAP estimate. Just write down the function you would differentiate.