Stats 545: Midterm exam

This is a 75-minute exam for 32 points. Write your name and PUID on each sheet, and also include the number of answer sheets. Attempt all questions.

1 Clustering [4 pts]

1. X_1 and X_2 are $n \times d$ matrices, each row being a d-dimensional observation. Recall that centroid-linkage clustering defines the distance between X_1 and X_2 as the distance between the mean of the observations of X_1 from the mean of observation in X_2. Complete linkage clustering defines the distance as the largest distance between an observation in X_1 and one in X_2. Write a few lines of `R` to do both of these. [3pts]
2. Explain why you might use complete-linkage clustering vs centroid-linkage. [1pts]

2 Matrix operations [11 pts]

Assume multiplying two $N \times N$ matrices requires N^3 operations while an $N \times N$ matrix times an $N \times 1$ vector requires N^2 operations. Let A be an $N \times N$ matrix, and b an $N \times 1$ vector. Define $A^4 = A \cdot A \cdot A \cdot A$.

1. How many operations are needed to calculate A^4? How many operations are needed to calculate $A^4 \cdot b$? [3pts]
2. What is the condition number $\kappa(A)$? Show that $\kappa(A) \geq 1$. Give a matrix where $\kappa(A) = 1$. [2pts]
3. Consider the system of equations $Ax = b$. With A fixed, for a small change δb in b, let the change in x be δx. Show that $\frac{\|\delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\delta b\|}{\|b\|}$. [2pts]
4. What is the QR decomposition of a matrix? Given the QR decomposition of A, how would you simulate an N-dim Gaussian with mean m and covariance A? You only can generate standard normals using `rnorm`. [2pts]
5. Let L be an $N \times N$ lower triangular matrix Write an expression for its determinant $|L|$. What is the structure of L^{-1}? (lower/upper triang., diagonal, unstructured) [2pts]

3 Dynamic programming [7 pts]

1. For a binary heap with N nodes, what
   (a) is the cost of **removing** the **largest** element from the binary heap? [1pts]
   (b) are the costs of **finding** the **largest, fifth largest** and **smallest** elements in the binary heap? [3pts]
2. Briefly explain the problem that the Needleman-Wunsch solves, and the forward pass of the dynamic program. Explain what the cost of the forward pass is. [3pts]

4 Exponential family distributions [3 pts]

1. Let $(X_1, X_2, \cdots, X_T)$ be an N-state Markov chain, with $p(X_1 = i) = \pi_i$, and $p(X_{t+1} = j | X_t = i) = A_{ij}$. Show that $p(X_1, \cdots, X_T)$ is exponential family and write down its sufficient statistics and natural parameters. [2pts]
2. Suppose we only observe X_2 and X_5, with all other observations missing. Is $p(X_2, X_5)$ exponential family? Explain your answer. [1pts]

5 The EM algorithm [7 pts]

The number of people who swipe into a building on weekdays is Poisson distributed with unknown mean λ (recall that the Poisson distribution has the form $p(x|\lambda) = \lambda^x \exp(-\lambda)/x!$). On weekends, only the security person might enter (with unknown probability π), else no one enters. We observe counts $X = (x_1, \ldots, x_T)$ of the number of people who entered for some T days (not in sequence). Unfortunately, we did not record the day of the week of each observation, all we know is that is was a weekend with probability $2/7$, and weekday with probability $5/7$. Write $W = (w_1, \ldots, w_T)$ for the set of missing indentifiers, with $w_i = 1$ indicating the ith observation was a weekday.

1. Write down the log joint-probability $\log p(X, W|\lambda, \pi)$. [2pts]
2. Write down the EM lower-bound $\mathcal{F}(q, \pi, \lambda)$. [2pts]
3. For λ and π given, write the $q_i(w_i)$ that maximizes $\mathcal{F}$ for observation i. Given the q's how would you update π and λ? If you don't derive the latter, then explain the intuition. [3pts]