STAT 479: Study Guide for Quiz

Quiz will be 3 HW questions + 2 of the following:

1. **MLE for Exponential Distribution**

   Suppose we observe data $x_1, x_2, \ldots, x_n$ drawn from an exponential distribution with PDF:

   $$f(x; \lambda) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

   (a) **Log-Likelihood Function**: Which of the following correctly represents the log-likelihood function $\ell(\lambda)$ for this dataset?

   A. $\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$
   B. $\ell(\lambda) = n \log \lambda + \lambda \sum_{i=1}^{n} x_i$
   C. $\ell(\lambda) = n \log \lambda - \lambda \prod_{i=1}^{n} x_i$
   D. $\ell(\lambda) = n\lambda - \lambda \sum_{i=1}^{n} x_i$

   (b) **Gradient of the Log-Likelihood**: What is the gradient of the log-likelihood $\ell(\lambda)$ with respect to λ?

   A. $\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$
   B. $\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} + \sum_{i=1}^{n} x_i$
   C. $\frac{\partial \ell}{\partial \lambda} = n\lambda - \sum_{i=1}^{n} x_i$
   D. $\frac{\partial \ell}{\partial \lambda} = \lambda \prod_{i=1}^{n} x_i$

   (c) **MLE for** λ: Which of the following is the Maximum Likelihood Estimator (MLE) for λ?

   A. $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$
   B. $\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$
   C. $\hat{\lambda} = \frac{n}{\prod_{i=1}^{n} x_i}$
   D. $\hat{\lambda} = \frac{1}{\sum_{i=1}^{n} x_i}$

2. **HMMs at the Dishonest Casino**

Background You are playing a game against a casino that secretly switches between two types of dice:

- A **fair die (F)** where each face appears with equal probability.
- A **loaded die (L)** that is biased toward rolling a six.

Each round, you and the casino both roll a die:

- You always roll a **fair six-sided die**.
- The casino rolls either a **fair** or a **loaded** die.
- The player with the **higher number wins**. If there is a tie, the casino wins.

The casino follows a **Hidden Markov Model (HMM)**, meaning that while you see the sequence of rolls, you do not know when the casino switches between the fair and loaded dice.

Transition and Emission Probabilities The casino follows these transition and emission probabilities:

- **State Transitions**:

$$P(F \to F) = s_{FF}, \qquad P(F \to L) = s_{FL} = 1 - s_{FF}$$
$$P(L \to L) = s_{LL}, \qquad P(L \to F) = s_{LF} = 1 - s_{LL}$$

- **Emission Probabilities**:
  - **Fair die (F)**: $P(X = k|F) = \frac{1}{6}$ for all $k \in \{1, 2, 3, 4, 5, 6\}$.
  - **Loaded die (L)**: $P(X = 6|L) = \frac{1}{2}$, while all other faces appear with probability $\frac{1}{10}$.
- **Initial Probabilities**:

$$P(S_1 = F) = \pi_F, \qquad P(S_1 = L) = \pi_L = 1 - \pi_F$$

Your goal is to ultimately develop an HMM-based strategy to decide when to play (bet) and when to skip a round to maximize your expected winnings.

(a) Draw the graphical representation of the HMM for this problem. Your diagram should include:
- Hidden states (S_t) representing whether the fair or loaded die is in use.
- Observations (X_t) representing the dice rolls.

(b) Using the **Markov assumption** (i.e., the probability of a state depends only on the previous state), write the probability expression for a sequence of two states and two observations, $P(X_1, X_2, S_1, S_2)$.

(c) Express $P(S_1 = F|X_1 = 6, X_2 = 6)$ in terms of the given initial, transition, and emission probabilities.

(d) **Bonus:** Suppose at time $t = 1$ the casino demonstrates to you that the dice is fair (i.e. $P(S_1 = F) = 1, P(S_1 = L) = 0$). Starting from $t = 2$, develop a strategy to decide when to play and when to skip a round. Consider the following payoffs when designing your strategy:

- If you win the roll, you receive **3 times** your bet.
- If the casino wins, you lose your bet.
- You may choose to **skip a round** and neither gain nor lose anything.

3. **Bayesian Network Reasoning**

Consider the following Bayesian Network, where A, B, C, and D are random variables:

$$A \to B \to D, \quad A \to C \to D.$$

The conditional probability tables below describe the relationships in the network, but their exact numerical values are hidden.

(a) **Joint Probability Structure**: Suppose we want to compute $P(A = 1, B = 1, C = 1, D = 1)$. Which of the following expressions correctly represents how this probability should be computed?

A. $P(A)P(B|A)P(C|A)P(D|B, C)$
B. $P(A)P(B)P(C)P(D|B, C)$
C. $P(A|B, C)P(B|D)P(C|D)P(D)$
D. $P(A|B)P(B|C)P(C|D)P(D)$

(b) **Effect of Marginalization**: To compute the marginal probability $P(D = 1)$, which of the following steps is required?

A. Summing over all possible values of A, B, and C in $P(A, B, C, D = 1)$.
B. Multiplying $P(A)$, $P(B)$, and $P(C)$ and then summing over A.
C. Directly using $P(D)$, since it does not depend on other variables.
D. Integrating the probability function over all possible states of A, B, and C.

(c) **Inference and Conditional Probability**: If we observe $D = 1$, which of the following would most likely increase the probability that $A = 1$?

A. If $P(B = 1|A = 1)$ and $P(C = 1|A = 1)$ are both high.
B. If $P(D = 1|B = 1, C = 1)$ is very low.
C. If $P(A = 1)$ is independent of $P(D)$.
D. If $P(B = 1|A = 1)$ and $P(C = 1|A = 1)$ are both low.

4. **Conditional vs Joint Models**

Let's consider two different probabilistic models for a categorical outcome Y given feature variables $X = (X_1, X_2, \ldots, X_n)$.

Model 1 (Conditional Model) The probability of Y given X is modeled as:

$$P(Y = k|X) = \frac{\exp(\theta_k^T X)}{\sum_j \exp(\theta_j^T X)}$$

Model 2 (Joint Model) A joint probability distribution over (Y, X) is defined using an unnormalized score:

$$P(Y, X) = \frac{\psi(Y, X)}{Z}, \quad \text{where } \psi(Y, X) = \exp(\theta_Y^T X)$$

and Z is the normalization constant.

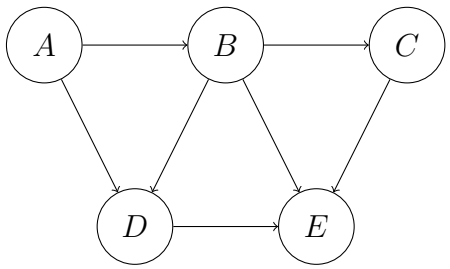
(a) Derive an explicit expression for Z in Model 2. Over which variables does the sum run?

(b) Express $P(Y|X)$ in Model 2.

(c) Show that Model 1 can be derived from Model 2 by choosing an appropriate form of Z.

5. **Conditional Independencies in Directed Graphical Models**

Consider the following Bayesian Network:



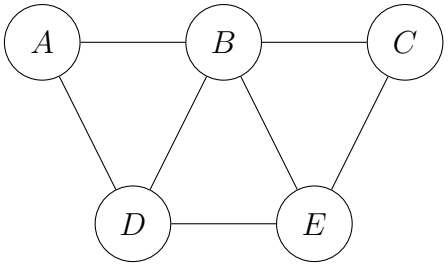
(a) Write the joint probability distribution $P(A, B, C, D, E)$ using the factorization implied by this Bayesian network.

(b) Below are conditional probability tables (CPTs) that specify distributions for some of the variables. Determine whether these conditional probability tables (CPTs) are consistent with the structure of the directed graph. If the tables do **not** match the graph, explain which independence assumptions are violated.

A	B	$P(B\|A)$	$P(A)$
0	0	0.6	0.5
0	1	0.4	0.5
1	0	0.3	0.5
1	1	0.7	0.5

B	C	E	$P(E\|B, C, D)$
0	0	0	0.8
0	0	1	0.2
0	1	0	0.5
0	1	1	0.5
1	0	0	0.4
1	0	1	0.6
1	1	0	0.1
1	1	1	0.9

6. **Identifying Cliques and Matching to Data Distributions** Consider the following undirected probabilistic graphical model, where nodes represent random variables, and edges represent direct dependencies between variables:



(a) Identify all **maximal cliques** in this graph. Write in the form (x, y, z) if nodes x, y, z for a maximal clique.

(b) Below is a conditional probability table (CPT) describing joint probabilities of some of the variables:

B	C	$P(B\|C)$	$P(C)$
0	0	0.7	0.5
0	1	0.4	0.5
1	0	0.3	0.5
1	1	0.6	0.5

B	C	E	$P(E\|B, C, D)$
0	0	0	0.8
0	0	1	0.2
0	1	0	0.5
0	1	1	0.5
1	0	0	0.4
1	0	1	0.6
1	1	0	0.1
1	1	1	0.9

Table 1: Conditional probability table for E.

Determine whether these conditional probability tables are consistent with the graphical model. Specifically:

(i) Does the factorization implied by the CPTs respect the independence assumptions of the graph?

(ii) If the tables do **not** match the graph, identify where the discrepancies occur and which independence assumptions are violated.

7. **Backpropagation with ℓ_2-Regularization**

Consider a neural network with one hidden layer. The network's output is:

$$\hat{y} = \sigma(w_2 \cdot h),$$

where:

- $h = \sigma(w_1 \cdot x)$,
- $\sigma(z)$ is the sigmoid activation function defined as $\sigma(z) = \frac{1}{1+e^{-z}}$,
- w_1 and w_2 are weight vectors,
- x is the input.

The network is trained using an ℓ_2-regularized loss function:

$$\mathcal{L}_{\text{reg}} = \mathcal{L} + \frac{\lambda}{2} \left(\|w_1\|_2^2 + \|w_2\|_2^2 \right),$$

where $\mathcal{L}$ is the negative log-likelihood for binary classification:

$$\mathcal{L} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}),$$

and $\lambda > 0$ is the regularization strength.

(a) **Gradient of $\mathcal{L}_{\text{reg}}$ w.r.t.** w_2: Derive the total gradient of $\mathcal{L}_{\text{reg}}$ with respect to w_2. Select the correct expression:

A. $(\hat{y} - y) \cdot h + \lambda w_2$
B. $(\hat{y} - y) \cdot h - \lambda w_2$
C. $(\hat{y} - y) \cdot h$
D. $(\hat{y} - y) + \lambda w_2$

(b) **Gradient of $\mathcal{L}_{\text{reg}}$ w.r.t.** w_1: Derive the total gradient of $\mathcal{L}_{\text{reg}}$ with respect to w_1. Select the correct expression:

A. $(\hat{y} - y) \cdot w_2 \cdot \sigma'(w_1 \cdot x) \cdot x + \lambda w_1$
B. $(\hat{y} - y) \cdot \sigma'(w_1 \cdot x) \cdot x + \lambda w_1$
C. $(\hat{y} - y) \cdot w_2 \cdot \sigma'(w_1 \cdot x) \cdot x$
D. $(\hat{y} - y) \cdot \sigma'(x) \cdot w_2 + \lambda w_1$

(c) **Impact of Regularization**: Suppose the regularization term is removed ($\lambda = 0$). Which of the following best describes the impact on the optimization process?

A. The model will fit the training data more closely, potentially overfitting.
B. The model will have higher training error but better generalization.

C. The gradients with respect to w_1 and w_2 will increase in magnitude.

D. The optimization will converge more slowly.

(d) **Regularization Strength**: If λ is increased, what effect will this have on the learned weights w_1 and w_2?

A. The weights will shrink, reducing overfitting.

B. The weights will increase, fitting the training data better.

C. The weights will remain unchanged, as λ does not affect the optimization.

D. The weights will oscillate during training.

8. **Deriving the Structure and Properties of a Restricted Boltzmann Machine**

Consider a probabilistic model with two types of binary random variables:

- **Visible variables** $V = \{V_1, \ldots, V_n\}$, which represent observed data.
- **Hidden variables** $H = \{H_1, \ldots, H_m\}$, which encode dependencies between visible variables.

After analyzing the model, you determine the following structural properties:

1. Each visible variable V_i is connected to some hidden variables H_j.
2. Each hidden variable H_j is connected to some visible variables V_i.
3. **No direct connections exist between visible variables**.
4. **No direct connections exist between hidden variables**.
5. The joint probability distribution is **defined using an energy function**, rather than conditional probabilities.

(a) **Graphical Model Structure**: Given the properties above, which of the following best describes the structure of this model?

A. A directed graphical model (Bayesian network) where each hidden node is a parent of multiple visible nodes.

B. A fully connected undirected graphical model, where every variable (visible or hidden) is connected to every other variable.

C. A bipartite undirected graphical model, where edges exist **only** between visible and hidden nodes, and no edges exist within either group.

D. A Markov random field with local cliques, where visible variables are conditionally independent given their neighbors.

(b) **Marginal Probability of Visible Units**: The joint probability of visible and hidden units is given by the Boltzmann distribution:

$$P(V, H) = \frac{1}{Z} e^{-E(V,H)}$$

where the energy function takes the form:

$$E(V, H) = -\sum_i b_i V_i - \sum_j c_j H_j - \sum_{i,j} W_{ij} V_i H_j.$$

Which of the following correctly expresses the marginal probability $P(V)$ after summing over all hidden variables?

A. $P(V) = \sum_H P(V, H)$, summing out hidden variables explicitly.

B. $P(V) \propto e^{\sum_i b_i V_i}$, ignoring the hidden units.

C. $P(V) \propto \prod_j \left(1 + e^{c_j + \sum_i W_{ij} V_i}\right) e^{\sum_i b_i V_i}$.

D. $P(V) \propto e^{-\sum_{i,j} W_{ij} V_i H_j}$, treating the energy function as directly defining probabilities.

(c) **Role of Hidden Units**: Why do RBMs allow visible units to have statistical dependencies, even though there are no direct connections between them?

A. The hidden units introduce shared dependencies, making visible units **conditionally dependent** even though they are **conditionally independent given** H.

B. The visible units are always independent, since there are no direct edges between them.

C. The bipartite structure forces visible units to be independent **both marginally and conditionally**, meaning RBMs can only model very simple distributions.

D. The energy function forces every visible unit to depend only on itself, meaning hidden units have no real effect.