Recitation 05/01

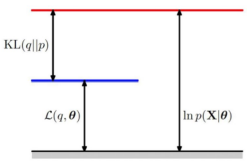
Friday, May 1, 2020 2:26 PM

3) When learning a Gaussian Mixture Model with EM, during the E-step we can choose a distribution over cluster assignments $q(z \mid x_i) = p_\theta(z \mid x_i)$, making the KL-divergence between the two $\mathrm{KL}(q \,||\, p) = 0$.

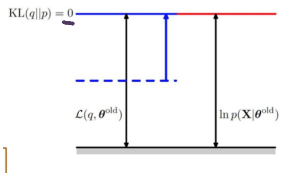
- [x] True
- [ ] False

Explanation if False:

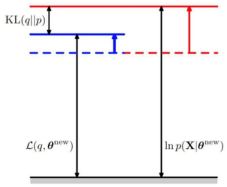
True



E-Step



M-Step



The EM algorithm

① initialize θ^1, $q^1(z)$

② for $t = 1:T$ (or until convergence)

E Step

$$q(z_j) = \sum_{i=1}^{n} p(z_j \mid x_i, \theta^t)$$

M-step

$$\theta^{t+1} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{j=1}^{J} q(z_j) \log\left(\frac{p(x_i, z_j \mid \theta)}{q(z_j)}\right)$$

4) If our data contains two variables x and y both of which are fully observed during training, with no latent variables, then using EM to estimate

$$\widehat{\theta} = \underset{\theta}{\text{argmax}} \ \log p_\theta(x, y)$$

will yield the same parameters as maximizing the data log likelihood directly (excluding numerical errors).

☐ True

☐ False

Explanation if False:

True

like EM if you knew the ground truth z values → you wouldn't need an E-step, just M-step

5) A major advantage of EM is that it is not susceptible to local optimum, unlike other maximum likelihood estimation methods.

☐ True

☐ False

Explanation if False:

False. EM is very susceptible to getting stuck in local optimum.

It's guaranteed to converge to a local optimum but not to global optimum

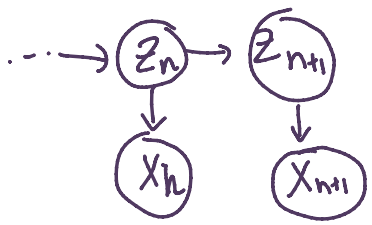
8) Hidden Markov Models assume causal relationships between the latent states defining the model.

☐ True

☐ False

Explanation if False:

True.

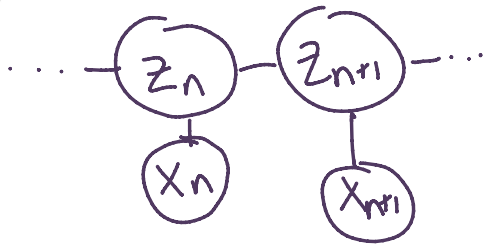


9) Consider a Markov Random Field (MRF) with the same structure as a Hidden Markov Model. The primary difficulty with unsupervised training for the MRF is computing the partition function.

- ☐ True
- ☐ False

Explanation if False:

True



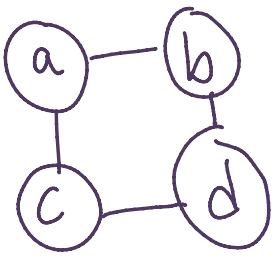
$$p(z,x) = \prod \psi(z_n, z_{n+1}) \prod \psi(x_n, z_n)$$

12) Consider a Markov Random Field with 4 binary random variables (nodes). Each random variable is connected to two neighbors, forming a circle structure. Even with modern, powerful computers, it is impossible to compute exact inference on this graphical model.

- ☐ True
- ☐ False

Explanation if False:

False. There are only 2^4 possible combinations to consider, which is a small number.



$p(a,b,c,d)$

$a \in \{0,1\}$
$b \in \{0,1\}$
$c \in \{0,1\}$
$d \in \{0,1\}$

possible states $= 2 \cdot 2 \cdot 2 \cdot 2$
$= 2^4 = 16$

14) Consider an undirected graphical model with cycles. The Markov Blanket of a node contains all of the node's neighbors.

☐ True

☐ False

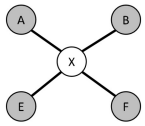
Explanation if False:

True.

Markov Blanket

The absence of "explaining away" makes the Markov blanket simple as well

The Markov blanket of a node contains the neighbors of the node

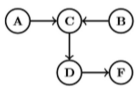


15) Consider the directed graphical model below. A and B are conditionally independent given F.

☐ True

☐ False

Explanation if False:

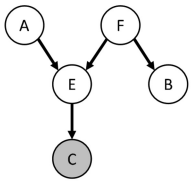


False. The only path from A to B goes through C, which is a head to head node, and

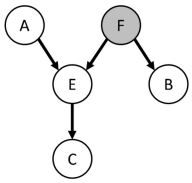
D-Separation Examples

D-Separation Examples

Are A and B d-separated?



No: C is a descendant of head-to-head E



Yes: F is a tail-to-tail node in the path, E is a collider

16) RNNs, HMMs and CRFs all use parameter tying to allow the model to handle variable-length sequences.

☐ True

☐ False

Explanation if False:

True

19) You are considering training a machine learning classifier on a new binary classification task, for which you have a labeled dataset. You ask several colleagues to complete the task, and they average only 50% accuracy on the task. It will be impossible to train a classifier to do much better than 50%.

☐ True

☐ False

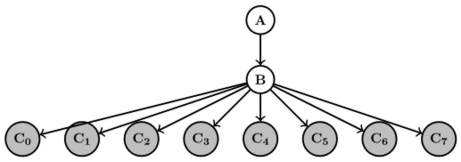
Explanation if False:

False. The task may be difficult for humans but easy for computers.

Short Answer

Unless otherwise indicated, each question is worth 20 points.

21) (24 points) Consider the following graphical model:



This model resembles a Naive Bayes model, with the addition of a second latent variable that generates the first. Let all parameters in the model be denoted by θ. In this setup, B is drawn from the distribution $p(B \mid A, \theta)$ and A is drawn from a prior distribution $p(A \mid \theta)$. Our observed variables are all drawn from independent distributions $p(C_i \mid B, \theta)$ for $0 \leq i \leq 7$. All variables in this graph are discrete.

Answer the following questions in terms of $p(A \mid \theta)$, $p(B \mid A, \theta)$, and $p(C_i \mid B, \theta)$?

1. What is $p(A, B, \vec{C} \mid \theta)$, where $\vec{C} = \{C_0, C_1, ..., C_7\}$?

$$p(A, B, \vec{C} \mid \theta) = p(A \mid \theta)p(B \mid A, \theta) \prod_{i=0}^{7} p(C_i \mid B, \theta)$$

2. What is $p(\vec{C} \mid \theta)$?

$$p(C|\theta) = \sum_{A} p(A|\theta) \sum_{B} p(B|A, \theta) \prod_{i=0}^{7} p(C_i|B, \theta)$$

3. Suppose we have a dataset $\mathcal{X}$ of M points x_j, where $x_j = \vec{C}^{(j)}$ (i.e., each x_j only contains an observation of $\vec{C} = \{C_0, ..., C_7\}$, and A and B are unobserved). We want to learn parameters θ that maximize $p(\mathcal{X}|\theta)$ using EM.

   During the E-step, we want to find, for each example $x_j = \vec{C}^{(j)}$, a $q(A, B \mid \vec{C}^{(j)}, \theta)$ that minimizes the KL divergence between q and $p(A, B \mid \vec{C}^{(j)}, \theta)$, the posterior distribution over A and B defined by our current model parameters, θ. What should we set $q(A, B \mid \vec{C}^{(j)}, \theta)$ equal to, to make $\mathrm{KL}(q \,||\, p) = 0$?

$$q(A, B|\vec{C}^{(j)}, \theta) = \frac{p(A)p(B|A)\prod_{i=0}^{7} p(\vec{C}_i^{(j)}|B)}{\sum_A p(A)\sum_B p(B|A)\prod_{i=0}^{7} p(\vec{C}_i^{(j)}|B)}$$

4. What is $Q(\theta, \theta^{old})$ for the M-step? (You may use $q(A, B|\vec{C}^{(j)}, \theta)$ in your answer.)

$$Q(\theta, \theta^{old}) = \sum_{j=i}^{M} \sum_A \sum_B q(A, B|\vec{C}^{(j)}, \theta^{old}) \log p(A, B, \vec{C}^{(j)}|\theta)$$

The EM algorithm

① initialize θ^1, $q^1(z)$

② for $t = 1 : T$ (or until convergence)

E Step

$$q(z_j) = \sum_{i=1}^{n} p(z_j \mid x_i, \theta^t)$$

M-step

$$\theta^{t+1} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{j=1}^{J} q(z_j) \log \left(\frac{p(x_i, z_j \mid \theta)}{q(z_j)} \right)$$