Homework 3 for #70240413
"Statistical Machine Learning"

Instructor: Prof. Jun Zhu

May 15, 2015

1 Probabilistic Graphical Models

1.1 Marginal Inference for HMM

Given the following Hidden Markov Model (Fig. 1) which indicates a factorized full probability distribution as follows,

$$\begin{aligned} p(x, y) &= p(x_1, x_2, \cdots, x_T, y_1, y_2, \cdots, y_T) & (1) \\ &= p(y_1)p(x_1|y_1)p(y_2|y_1)p(x_2|y_2)\cdots p(y_T|y_{T-1})p(x_T|y_T), & (2) \end{aligned}$$

please show how to compute the following conditional queries for $t = 1, \cdots, T$:

1. $p(y_t|x_1, \cdots, x_t)$ (this is called a "filtering"; Note that each of these is conditioned only on observations up to time step t.)

2. $p(y_t|x_1, \cdots, x_T)$.

Hint: you may want to use recursions and to use the results from 1 to answer 2.

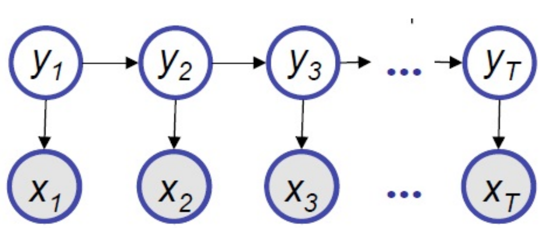

Figure 1: Hidden Markov Model

1.2 Message Passing on a Tree

Consider the DGM in Fig. 2 which represents the following fictitious biological model. Each G_i represents the genotype of a person: $G_i = 1$ if they have a healthy gene and $G_i = 2$ if they have an unhealthy gene. G_2 and G_3 represent the descendants of G_1 and therefore may inherit this specific gene from G_1. $X_i \in \mathbb{R}$ is a continuous measure of blood pressure, which is low if the person is healthy or high if unhealthy.

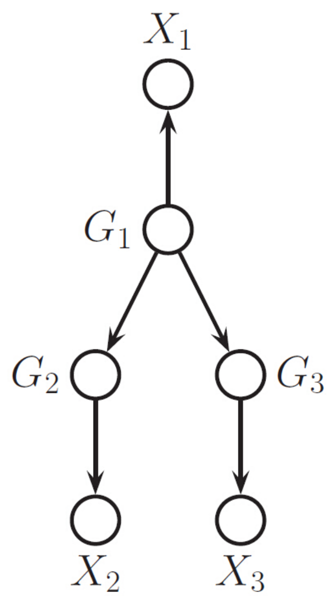

Figure 2: A simple DAG representing inherited diseases

We define the CPDs as follows

$$P(G_1) = (0.5, 0.5) \tag{3}$$

$$P(G_i|G_1) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \quad (i = 2, 3) \tag{4}$$

$$p(X_i|G_i = 1) = \mathcal{N}(X_i|\mu = 55, \sigma^2 = 10) \quad (i = 1, 2, 3) \tag{5}$$

$$p(X_i|G_i = 2) = \mathcal{N}(X_i|\mu = 65, \sigma^2 = 10) \quad (i = 1, 2, 3) \tag{6}$$

1. Suppose you only observe $X_2 = 50$. What is the posterior belief on G_1, i.e., $P(G_1|X_2 = 50)$?
2. Now suppose you observe both $X_2 = 50$ and $X_3 = 50$. What is $P(G_1|X_2, X_3)$? Explain your answer intuitively.

2 Learning Theory

2.1 VC Dimension

Consider the instance space X to be $\mathbb{R}^2$. Please derive the VC dimension of the following hypothesis space:

$H = \{$All the axes-parallel rectangles in $\mathbb{R}^2$, where points inside the rectangle are classified as positive.$\}$.

2.2 Generalization Bound

Consider a learning problem in which instances $X = \mathbb{R}$ are all the real numbers, and the hypothesis space $H = \{(a < x < b) | a, b \in \mathbb{R}\}$ is composed of all the intervals in $\mathbb{R}$. What is the probability that a hypothesis $h \in H$ consistent with m instances $x_1, \cdots, x_m$ will have an error of at least ϵ?

Note: you can use the theoretical results from the lecture notes directly.

3 Topic Modeling

For the LDA model illustrated in the lecture notes P. 55, derive the collapsed Gibbs sampling algorithm for posterior inference. By "collapsed" we mean to first integrate out Θ (topic mixing proportions) and Φ (topics) to perform Gibbs sampling only with $p(Z|W, \alpha, \beta)$ and, after obtaining a good estimate of Z (topic assignments), to then compute the posterior of Θ and Φ through $p(\Theta|Z, \alpha)$ and $p(\Phi|W, Z, \beta)$. Hint: be sure to leverage the conjugacy between the Dirichlet and the Multinomial.

Please implement the sampling algorithm and test it on the "20newsgroup" dataset[1].

Set the number of topics K to be $5, 10, 20, 30$ respectively and show the most-frequent words in each topic for each case. Compare your results with the mixture-of-multinomials model in Homework 1, report the differences and try to explain why.

Bonus: Think about how to specify the prior distributions in LDA (lecture notes P. 17). How do you choose the hyperparameters α and β? Try subjective priors or empirical priors as well and observe the difference.

[1] `http://ml.cs.tsinghua.edu.cn/~wenbo/data/20newsgroup_train.zip`