COMP 652: Machine Learning - Assignment 2

Posted Thursday, March 9, 2017
Due Thursday, March 23, 2017

1. [40 points] **Properties of entropy and mutual information, and Bayes net construction**

   In this exercise, you have to prove some properties of KL divergence nd mutual information. You can assume that all variables are discrete.

   (a) [5 points] $H(X) \geq H(X|Y)$, with equality achieved when X and Y are independent.

   (b) [5 points] The KL divergence between two probability distributions, P and Q, is defined as:

   $$D_{KL}(P,Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

   Show that $D_{KL} \geq 0 \forall P, Q$, and give an example of P and Q for which $D_{KL}(P,Q) \neq D_{KL}(Q,P)$

   (c) [5 points] The mutual information of two random variables X and Y is defined as:

   $$I(X;Y) = D_{KL}(P(X,Y)||P(X)P(Y))$$

   Show that $I(X;Y) = H(X) + H(Y) - H(X,Y)$

   (d) [5 points] Show that $MI_P(Y,Z) \geq 0$, with equality if and only if $Y \perp\!\!\!\perp Z$.

   (e) [10 points] Show that the likelihood score of the graph G underlying a Bayes net can be expressed as:

   $$\log L(G|D) = \sum_{j=1}^{m} \log p(\mathbf{x}_j|G) = m \sum_{i=1}^{n} MI_{\hat{P}}(X_i, X_{\pi_i}) - m \sum_{i=1}^{n} H_{\hat{P}}(X_i)$$

   where m is the number of instances, n is the number of random variables in the network, $\mathbf{x}_j$ denotes the jth instance and $\hat{P}$ is the empirical distribution of the data:

   $$\hat{P}(x_i|x_{\pi_i}) = \frac{N(x_i, x_{\pi_i})}{N(x_{\pi_i})}$$

   (f) [10 points] Consider two structures G_1 and G_2 which are identical except for the fact that G_2 has one extra arc. Using the maximum likelihood score formula you proved above, show that G_1 has a lower score than G_2.

2. [15 points] **Sigmoid Bayes nets**

Suppose we have a Bayes net over a set of binary random variables. Each node is parameterized as:

$$P(X_i|X_{\pi_i}) = \sigma\left(\sum_{j\in\pi_i} w_{ij}X_j\right)$$

where W_{ij} are real-valued weights and σ is the sigmoid function. Suppose we have such a network with two layers: a layer H of hidden nodes, and a layer V of visible nodes. Give a gradient-based learning rule for W_{ij}, in such a way as to maximize the likelihood of a given set of data, in which only the visible variables V are recorded.

3. [20 points] **Markov Random Fields**

   Consider the 2D spin glass model we discussed in the lecture.

   (a) [5 points] Suppose that instead of connecting pixels in a 4-neighborhood, we want to connect them in an 8-neighborhood. Describe what the parameters of the undirected graphical model will be.

   (b) [5 points] Suppose that we want to use such a model to capture natural scenes in images. Describe the advantages and disadvantages of this model compared to connecting a pixel only to 4 neighbours.

   (c) [10 points] For the 2D Ising model connected as in class, write a Gibbs sampling algorithm, assuming that potentials are represented using linear energy functions and that evidence can be injected along the leftmost edge of the model. Assume the model is an $n \times n$ lattice.

4. [25 points] **EM algorithm**

   In this question we will explore a mixture model for modelling text. Suppose you have a vocabulary of M words. We consider each word in a document as a random variable W whose value is a vector of M components, such that $W(i) = 1$ if the value of W is the ith word in the vocabulary, and 0 otherwise. Hence, $\sum_{i=1}^{M} W(i) = 1$ (this is also known as a one-hot encoding). Suppose the words are generated from a discrete mixture of K latent topics:

$$P(W) = \sum_{k=1}^{K} \pi_k P(W|\mu_k)$$

   where π_k is the prior for the latent topic k and $P(W|\mu_k)$ is modelled as:

$$P(W|\mu_k) = \prod_{i=1}^{M} (\mu_k(i))^{W(i)}$$

   Hence, we generate a word by drawing a topic k from π and then drawing the word from the topic's distribution, according to μ_k.

   (a) [5 points] Suppose we have documents consisting of N words, which have been drawn i.i.d. according to this process. Suppose that for each document we have a given topic, which is known. Compute the maximum likelihood values for the π and μ parameters.

(b) [15 points] Suppose now that the topics are not known, and in fact, one document may cover multiple topics. Derive an expectation maximization algorithm for learning the parameters π and μ. In this case, for the expectation step, you need to compute the probability of the topic associated with each word W_j, in order to complete the data, and in the maximization step, you need to re-compute the parameters that maximize the likelihood of the data.

(c) [5 points] The assumption that words are drawn iid from a topic is quite strong. It would make more sense to assume a word's probability is conditioned on the topic as well as the previous work in the document. Explain how many parameters would the model have in this case, and what is the bias-variance trade-off compared to the previous model.