COMP652/ECSE608 - Assignment 3

March 30, 2017

1. [20 points] **PCA**

   Consider the data set available in the file `hw3pca.txt`; each row represents an instance and the columns represent features. You should split the data into 80% representing the training set and 20% to test the representation. Perform PCA on the data and plot the reconstruction error as a function of the number of dimensions, both on the training set and on the test set, as well as the fraction of the variance accounted for obtained by looking at the top eigenvalues. Explain what you see and what are the implications for choosing dimensionality of the data.

2. [30 points] **Spectral method for weighted automata**

   Let $A = (\alpha_0, \{A_0, A_1\}, \alpha_\infty)$ be a weighted automaton with n states (i.e. $\alpha_0, \alpha_\infty \in \mathbb{R}^n$ and $A_0, A_1 \in \mathbb{R}^{n \times n}$) on strings built on the alphabet $\Sigma = \{0, 1\}$. Let $f : \Sigma^* \to \mathbb{R}$ be the function computed by A where Σ^* is the set of all strings built on the alphabet Σ. We denote the empty word by λ (note that $\lambda \in \Sigma^*$).

   (a) [5 points] Consider the function g that counts the number of 1's in a word, e.g. $g(0110) = 2$, $g(10000) = 1$, $g(011101) = 4$... Can the function g be computed by a weighted automaton? Justify your answer.

   (b) [5 points] Let $f_{substring} : \Sigma^* \to \mathbb{R}$ be the function defined by

   $$f_{substring}(w) = \sum_{u \in \Sigma^*, v \in \Sigma^*} f(uwv)$$

   for all $w \in \Sigma^*$. When f is a probability distribution over Σ^*, give a probabilistic interpretation of the function $f_{substring}$.

   (c) [10 points] For any word $w = w_1 w_2 \cdots w_k \in \Sigma^*$, where each w_i is a symbol in Σ, let $A_w = A_{w_1} A_{w_2} \cdots A_{w_k}$ (and $A_\lambda = I$ where λ is the empty word). Show that if the sum $\sum_{w \in \Sigma^*} A_w$ converges then we have the identity

   $$\sum_{w \in \Sigma^*} A_w = (I - A_0 - A_1)^{-1} \tag{1}$$

   and use this identity to give a formula to compute the sum of the function f over all words ($\sum_{w \in \Sigma^*} f(w)$) and to show that the function $f_{substring}$ can be computed by a weighted automaton.

(d) [10 points] Suppose you are given a training sample S drawn from some probability distribution f over Σ^*. Briefly explain how the spectral method could be used to learn the function $f_{substring}$ from S, and how you could recover an estimate of f given the weighted automaton $\hat{A} = (\hat{\alpha}_0, \{\hat{A}_0, \hat{A}_1\}, \hat{\alpha}_\infty)$ returned by the spectral method. What would be the benefits and disadvantages of this method relative to directly learning f from the training sample S?

3. [10 points] **Method of moments and multiview model**.

Let $h \in \{1, \cdots, k\}$ be a discrete random variable with $Pr[h = j] = w_j$ for all j. Consider random vectors $x_1 \in \mathbb{R}^{d_1}, x_2 \in \mathbb{R}^{d_2}, x_3 \in \mathbb{R}^{d_3}$ which are conditionally independent given h, and for which the conditional expectations satisfy

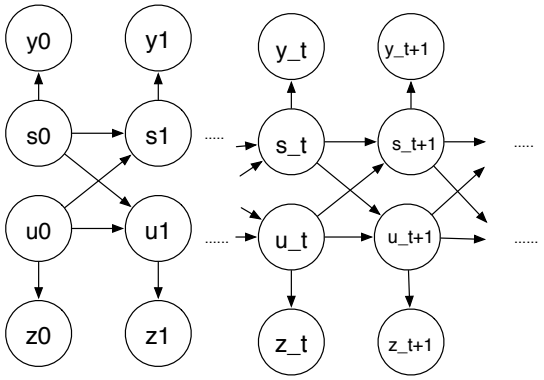
$$E[x_t|h = j] = \mu_{t,j},\ j \in \{1, \cdots, k\},\ t \in \{1, 2, 3\}$$

where the $\mu_{t,j} \in \mathbb{R}^{d_t}$ are the conditional means of the x_t given $h = j$. Using simple properties of expectation, show that the second and third order cross moments $E[x_1 \otimes x_2]$ and $E[x_1 \otimes x_2 \otimes x_3]$ can be expressed as functions of the model parameters w_j, $\mu_{j,t}$. Can we use the tensor method of moments to recover the model parameters from these cross moments? Why/How?

4. [40 points] **Coupled Hidden Markov Models**.

We discussed in class several models for reasoning with sequences of data (trajectories). The HMM is the simplest such example, in which states are hidden, and we see observations that depend on the state. The Coupled Hidden Markov model (CHMM) is a similar kind of graphical model: we have several hidden Markov models running in parallel, and their states interact. This model is quite useful, for example, when you try to parse video, and you consider the observations as being sound and visual data, respectively.

Consider a system with two HMMs, depicted in Figure 1:



Here, s_i and u_i are the states of the two coupled HMMs, y_i and z_i are the observations coming from the two chains, and the two chains interact in the way depicted in the picture.

(a) [5 points] Specify what are the parameters of this model.

(b) [5 points] Derive an algorithm for computing the joint probability of a sequence of observations $(y_0, z_0), (y_1, z_1) \dots (y_T, z_T)$.

(c) [10 points] Derive a forward algorithm that computes the most likely sequence of hidden states given a sequence of observations.

(d) [10 points] Suppose that instead of the chains being coupled at every time step, the coupling only happens every k time steps (on time step 0, k $2k$ etc). For $k = 1$, you get the same model as above. If k is fairly large compared to the length of sequences, the chains are called **loosely coupled**. Describe how your model and the inference algorithms change in this case.

(e) [10 points] Suppose that you observe several sequences of two time series and you know that they come from a loosely coupled HMM; you know the number of possible states for each individual chain, but you do not know k. Describe a learning algorithm for this problem.