

Statistical Methods for Artificial Intelligence
(TTIC 103, CMSC 35420)
Autumn 2007
Final Exam

1. This problem is on principal component analysis. Consider a covariance matrix Σ and suppose that we want to compute the k principal eigenvectors and their eigenvalues. Consider a random initial vector Φ_0 and suppose that we compute a series of vector $\Phi_1, \Phi_2, \Phi_3, \dots$ using the following update.

$$\Phi_{t+1} = \frac{\Sigma \Phi_t}{\|\Sigma \Phi_t\|}$$

Let Ψ_1, \dots, Ψ_D be (unknown) orthogonal eigenvectors of Σ with eigenvalues $\sigma_1^2, \dots, \sigma_D^2$. Although we do not know the eigenvectors there exists an unknown decomposition of Φ_t into a linear combination of eigenvectors as follows.

$$\Phi_t = y_{1,t} \Psi_1 + \dots + y_{D,t} \Psi_D$$

- a. Give a closed form expression for $y_{i,t}$ for $t > 1$ as a function of $y_{1,1}, \dots, y_{D,1}$ and the eigenvalues $\sigma_1^2, \dots, \sigma_D^2$.
- b. Describe an algorithm for computing the first principal eigenvector and eigenvalue.
- c. Explain how the answer to part b. might be used to compute the first K eigenvectors and eigenvalues.

2. This problem is on kernel methods. Consider the “quadratic hinge loss” defined as follows. (This is the same as on the midterm).

$$L_{\text{hinge2}}(m) = \begin{cases} (m-1)^2 & \text{for } m \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Define w^* as follows.

$$\begin{aligned} w^* &= \underset{w}{\operatorname{argmin}} \sum_{t=1}^N L_{\text{hinge2}}(m_t(w)) + \frac{1}{2} \lambda \|w\|^2 & (1) \\ m_t(w) &= y_t (w \cdot \Phi(x_t)) \end{aligned}$$

a. Rewrite (1) as an optimization problem on α where we define $w = \sum_{t=1}^N \alpha_t \Phi(x_t)$. The optimization problem must be defined entirely in terms of the kernel function $K(x, y) = \Phi(x) \cdot \Phi(y)$ — you must assume that $\Phi(x)$ cannot be computed but that $K(x, y)$ can be computed. (This part was on the midterm).

b. Let α^* be solution the the kernalized optimization problem in a. Give α_t^* as a function of $m_t(w^*)$. Hint: compute the gradient of (1), set it to zero, and note that you have a form of the representer theorem.

3. This problem is on EM. Suppose we want to model a probability density on pairs $\langle i, j \rangle$ with i and j integers in $\{0, \dots, K-1\}$. We are given a data set x^1, \dots, x^N with $x^t = \langle i_t, j_t \rangle$. We define a model as follows where π and γ are both K dimensional vectors giving distributions on the numbers from 0 to K , i.e., $\pi_i > 0$, $\sum_{i=0}^{K-1} \pi_i = 1$, $\gamma_i \geq 0$ and $\sum_{i=0}^{K-1} \gamma_i = 1$. We assume that the pairs $\langle i_t, j_t \rangle$ are drawn IID where for each such pair we first draw i and then add a random offset to i to get a hidden random variable h , and finally adding another random offset to h to get j . The distribution of the two offsets is assumed to be identicle and given by γ . This is defined mathematically as follows.

$$\Theta = \langle \pi, \gamma \rangle$$

$$P_{\Theta}(\langle i, j \rangle) = \Pi_i \left(\sum_{h=0}^{K-1} \gamma_{(h-i \bmod K)} \gamma_{(j-h \bmod K)} \right)$$

We want to approximately solve the following optimization problem.

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} P_{\Theta}(x^1, \dots, x^N)$$

Assume we are given Θ_{OLD} . Let Θ_{NEW} be the result of a single EM update from Θ_{OLD} . Give a way of computing Θ_{NEW} from the data set and Θ_{OLD} . Your answer can be given in terms of the following quantities.

$$P_u^t = P_{\Theta_{OLD}}(h_t = u)$$

You do not need to give the formula for computing P_u^t , just an expression for Θ_{NEW} as a function of P_u^t and x^1, \dots, x^t .

4. This problem is on junction trees. Suppose that I have a Boolean circuit composed of layers. The inputs of each layer are the outputs of the preceding layer except that the inputs of the first layer are, or course, the inputs to the circuit and the outputs of the last layer are the outputs of the circuit. Each gate (boolean operation) at each layer computes one output of that layer as a function of some number of inputs of that layer. The circuit defines a hypergraph whose nodes are wires (inputs and outputs) and where there is a hyperedge for each Boolean gate containing all its inputs together with its one output.

a. Suppose that the number of wires at each stage is at most n (each stage has at most n inputs and the number of outputs of the circuit is at most n). Define (or draw) a junction tree for this hypergraph with width $2n-1$ (remember the “-1” in the definition of tree width).

b. Suppose that at each layer each input feeds into at most two gates and that n is even. Draw a junction tree with width at most $(3/2)n-1$. Hint: eliminate the inputs one at a time.