UC Berkeley
Department of Electrical Engineering and Computer Science
Department of Statistics

EECS 281A / STAT 241A Statistical Learning Theory

Problem Set 4
Fall 2011

Issued: Monday, October 24, 2011 **Due:** Monday, November 7, 2011

Reading: For this problem set: Chapters 9, 10, 11, 12

Problem 4.1
Consider data in which 196 individuals are distributed multinomially into four categories $y \in \{120, 16, 22, 38\}$. Let the model for these data be a multinomial probability distribution with probabilities $(\frac{1}{2} + \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi)$, for some π such that $0 < \pi < 1$. We would like to find the maximum likelihood estimate of π. Develop an EM algorithm to find this estimate by letting the complete data consist of five categories, where the original first category is split into two sub-categories. I.e., let the complete data be $x = (x_1, x_2, x_3, x_4, x_5)$, where $y_1 = x_1 + x_2$, and where $y_2 = x_3$, $y_3 = x_4$ and $y_4 = x_5$.

(a) Write out the E step and the M step explicitly for this EM algorithm.

(b) Find a numerical solution by running the algorithm for ten steps.

Problem 4.2
IPF and disease status: Consider an undirected graphical model pairwise factorization of the form

$$p(x_1, \ldots, x_d; \psi) = \frac{1}{Z} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E} \psi_{st}(x_s, x_t).$$

As discussed in class, for binary variables, this model can be used to model disease status in an epidemiological network, where $X_s = 1$ if the individual at node s is infected, and $X_s = 0$ otherwise.

(a) The data file `IPF.dat` contains a 4×30 matrix, summarizing the data from $n = 30$ samples of the disease status of $d = 4$ people. Implement and apply the IPF updates to compute the ML estimates of $\{\psi_s, \psi_{st}\}$ for each of the following graphs:

  (i) the tree-structured graph with edge set $E = \{(1,2), (2,3), (3,4)\}$.
  (ii) the fully connected graph with all $\binom{4}{2}$ edges.

  Which graph gives the higher likelihood to the data? Does higher likelihood mean that it is a "better" model?

(b) For graph (i) in part (a), show the ML estimates $\{\widehat{\psi}_s, \widehat{\psi}_{st}\}$ can be written in closed form as $\widehat{\psi}_s(x_s) = \bar{\mu}_s(x_s)$ and $\widehat{\psi}_{st}(x_s, x_t) = \frac{\bar{\mu}_{st}(x_s, x_t)}{\bar{\mu}_s(x_s)\bar{\mu}_t(x_t)}$, where $\bar{\mu}$ are the empirical marginals computed from the data. Use this fact to verify the correctness of your implementation from part (a). Does the fully connected graph from part (ii) also have this same closed-form solution?

(c) Assume now that we know that the node compatibility functions are constant (i.e., $\psi_s(x_s) = 1$ for all x_s and $s \in V$), and the edge compatibility functions are homogeneous (i.e., $\psi_{st} = \psi_{uv}$ for all edges (s,t) and (u,v)). Describe a modified IPF algorithm for computing the maximum likelihood estimates.

(d) *Model selection for trees:* Assume that we know that the graph is tree-structured, but the edge set $E(T)$ is unknown. For any tree, let $\widehat{\psi}(T)$ denote the ML estimate of the compatibility functions for all vertices, and for all edges in $E(T)$. Letting $\ell(\widehat{\psi}(T))$ denote the maximized log-likelihood for tree T, a natural way to choose the "best" tree is to compute

$$T^* \in \arg\max_T \ell(\widehat{\psi}(T)).$$

Show that any such tree T^* must be a *maximum weight spanning tree*, in the sense that

$$\sum_{(s,t)\in E(T^*)} D(\bar{\mu}_{st} \,\|\, \bar{\mu}_s \, \bar{\mu}_t) \geq \sum_{(s,t)\in E(T)} D(\bar{\mu}_{st} \,\|\, \bar{\mu}_s \, \bar{\mu}_t) \quad \text{for all other trees } T,$$

where $D(\bar{\mu}_{st} \,\|\, \bar{\mu}_s \bar{\mu}_t)$ is the Kullback-Leibler divergence between the joint marginal $\bar{\mu}_{st}$ and the product of marginals $\bar{\mu}_s \bar{\mu}_t$. This relationship is important, because the maximum weight spanning tree problem is easily solved by a greedy algorithm that adds edges at each step.

Problem 4.3

(Conjugate duality) Given a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the dual function is a new function $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, defined as follows:

$$f^*(v) = \sup_{u \in \mathbb{R}^n} \big\{ v^T u - f(u) \big\}. \tag{1}$$

(Note that the supremum can be $+\infty$ for some $v \in \mathbb{R}^n$.)

(a) Compute the conjugate dual of the following functions:

  (i) $f(u) = \log(1 + \exp(u))$.

  (ii) $f(u) = -\log(u)$.

  (iii) $f(u) = \frac{1}{2} u^T Q u$ where $Q \succ 0$ is a symmetric positive definite matrix.

(b) For each of the functions in part (a), specify an example of an exponential family for which f is the cumulant function (denoted A in our description of exponential families). Show how the conjugate dual A^* is related to the entropy function $H(p) = -\int_x p(x;\theta) \log p(x;\theta) dx$.