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

EECS 281A / STAT 241A Statistical Learning Theory

Problem Set 5
Fall 2012

Issued: Tues. Oct. 16, 2012 **Due:** Tues. Oct. 30, 2012

Reading: Chapters 9, 10, 11, 12

Problem 5.1
EM algorithm for hidden Markov models

(a) Implement the EM algorithm for HMMs with Gaussian emission probabilities $p(y_t \mid x_t)$, where $y_t \in \mathbb{R}^2$ and $x_t \in \{0, 1, \ldots, m-1\}$. Restrict the covariance matrices to be isotropic (i.e., $\Sigma = \sigma^2 I$).

(b) Fit an HMM with $m = 4$ states to the two-dimensional data in *hmm-gauss.dat* and evaluate the log likelihood on the training and test data in *hmm-test.dat*. Plot the data together with the means of the component densities.

(c) Fit a Gaussian mixture model with 4 states to the same data (again with isotropic covariance matrices $\sigma^2 I$). Compare the performance with that of the HMM.

Problem 5.2
EM and missing values: Suppose you have a random sample of twins and are interested in studying *identical* twins. However, you observe only:

- $m \equiv$ the total number of male twins (both identical and fraternal)
- $f \equiv$ the total number of female twins, and
- $b \equiv$ the number of twins of opposite gender.

Let θ be the probability that a pair of twins are identical. Assume that, given identical twins, the probability the twins are male is p. Given fraternal twins, assume the number of males is $\text{Binomial}(2, q)$.

Give an algorithm for calculating the MLEs for θ, p, and q. (*Hint*: If you knew exactly how many identical male and female twins there are, then the MLEs would be easy to calculate.)

Problem 5.3

(Social network analysis and IPF:) Frank is studying dependencies in voting patterns of a collection of d US senators. For any given bill, he collects a vector $x \in \{0,1\}^d$, where $x_i = 1$ means that senator i voted yes on that bill. He models the random vector $(X_1, X_2, \ldots, X_d)$ as a pairwise Markov random field

$$\mathbb{P}_\theta(x_1, x_2, \ldots, x_d) \quad \propto \prod_{(s,t) \in E} \exp(\theta_{st}(x_s, x_t)).$$

(a) For a set $d = 4$ senators, the data file `Pairwise.dat` contains a 4×30 matrix, summarizing the data from $n = 30$ bills that were voted on in the senate. Implement and apply the IPF updates to estimate the model parameters for each of the following graphs: (i) the graph with edge set $E = \{(12), (23), (34), (14)\}$; and (ii) the graph with edge set $E = \{(12), (23), (13), (14)\}$; and (iii) the fully connected graph with all $\binom{4}{2}$ edges.

(b) Of models (i) and (ii), which model has a higher likelihood?

(c) Of all three models (i), (ii) and (iii), which has the highest likelihood? Do you think that it is the "best" model?

Problem 5.4

Model selection for trees: Recall that for a given tree T with edge set $E(T)$, the MLE for the exponential parameters takes the form

$$\widehat{\theta}_s(x_s) = \log \widehat{\mu}_s(x_s) \quad \text{for all } s \in V, \text{ and}$$

$$\widehat{\theta}_{st}(x_s, x_t) = \log \frac{\widehat{\mu}_{st}(x_s, x_t)}{\widehat{\mu}_s(x_s)\widehat{\mu}_t(x_t)} \quad \text{for all } (s,t) \in E(T).$$

Here $\widehat{\mu}$ are the empirical marginals computed from the data (e.g., $\widehat{\mu}_{st}(j,k) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{st;jk}(x_{is}, x_{it})$), and we assume that they take strictly positive values.

(a) Define the rescaled log likelihood $\ell(\theta) = \frac{1}{n}\sum_{i=1}^n \log \mathbb{P}_\theta(x_{i\cdot})$. Letting $\widehat{\theta}(T)$ denote the MLE for trees, show that $\ell(\widehat{\theta}(T))$ depends on the empirical marginals only via

$$\text{Singleton entropy:} \quad H(\widehat{\mu}_s) = -\sum_{x_s} \widehat{\mu}_s(x_s) \log \widehat{\mu}_s(x_s) \quad \forall\, s \in V, \text{ and}$$

$$\text{Joint edge entropy:} \quad H(\widehat{\mu}_{st}) = -\sum_{x_s, x_t} \widehat{\mu}_{st}(x_s, x_t) \log \widehat{\mu}_{st}(x_s, x_t) \quad \forall\, (s,t) \in E(T).$$

(b) In the model selection problem for trees, the goal is to choose, from *all trees on d nodes*, the highest likelihood tree i.e., $\widehat{T} \in \arg\max_T \ell(\widehat{\theta}(T))$. Show how this problem can be cast as a maximum weight spanning tree calculation. This is important, because it allows us to select the best tree by simple algorithms. (*Hint:* The mutual information $I(\widehat{\mu}_{st}; \widehat{\mu}_s, \widehat{\mu}_t) = H(\widehat{\mu}_s) + H(\widehat{\mu}_t) - H(\widehat{\mu}_{st})$ should be relevant in your calculations.)

(c) Use the technique from (b) to select the best fitting tree for the data in `Pairwise.dat`.