CS 281A/Stat 241A Homework Assignment 4 (due November 6)

1. **Noisy-OR Model**.

   A "noisy-OR model" is a conditional probability model for a binary output y given m binary input variables $x = (x_1, \ldots, x_m)$. It takes the form:

   $$p(y = 0 \mid x) = \prod_{j=1}^{m} \xi^{x_j},$$

   where the parameters $\xi = (\xi_1, \ldots, \xi_m)$ can be interpreted as the probabilities of an input variable failing to trigger the output variable.

   Derive an EM algorithm for estimating ξ given n data points $\{(x_{i1}, \ldots, x_{im}, y_i) : i = 1, \ldots, n\}$. Each iteration of your algorithm should run in $O(m)$ time. Hint: introduce independent hidden variables $z_{i1}, \ldots, z_{im}$, with $p(z_{ij} = 1) = \xi_j$. Define $s_{ij} = z_{i1} \wedge \cdots \wedge z_{ij}$ and note the relationship between s_{im} and y_i.

2. **Conjugate duality**.

   Let $f(x)$ be a convex function. *Conjugate duality* refers to the fact that to each convex $f(x)$ there corresponds a *conjugate function* $f^*(x)$ that has the following dual relationship to $f(x)$:

   $$\begin{aligned} f(x) &= \sup_{\mu}(\mu x - f^*(\mu)) \\ f^*(\mu) &= \sup_{x}(\mu x - f(x)). \end{aligned}$$

   (a) Derive the conjugate functions for $f(x) = -\log(x)$ and $f(x) = e^x$.

   (b) Show that the negative of the logarithm of the logistic function is convex and derive its conjugate function.

   (c) Prove Jensen's inequality using conjugate duality.

3. **EM and Missing values**

   Suppose you have a random sample of twins and are interested in studying *identical* twins. However, you only observe $m =$ the total number of male twins (both identical and fraternal), $f =$ the total number of female twins, and $b =$ 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 $Binomial(2, p')$.

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

4. **Linear interpolation of language models**.

   In natural language processing, a *language model* is a distribution over sequences of words. It has many applications; for example, in speech recognition, it is used to enforce grammaticality of the sentences produced by the recognizer.

A k-gram language model has the following form:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid x_{i-1}, \ldots, x_1),$$

where each conditional is defined by

$$p(x_i \mid x_{i-1}, \ldots, x_1) = p_k(x_i \mid x_{i-1}, \ldots, x_{i-(k-1)}).$$

Let V denote a finite set of possible words, and let each word $x_i \in V$ for $i = 1, \ldots, n$. Note that in this model, the probability of each word depends only on the previous $k - 1$ words. (Assume that $x_i \stackrel{\text{def}}{=} \text{Boundary-Word}$ for $i < 1$.)

(a) Let θ denote the parameters of the k-gram model. What is θ? How many parameters are there as a function of k? State the maximum likelihood estimate for θ (justification not necessary). Note that if k is too small, then we have a bad model of language, but if k is too big, we need a lot of data to estimate θ.

One way to cope with this tradeoff is to linearly interpolate between k-gram models with different values of k. In particular, we define

$$p(x_i \mid x_{i-1}, \ldots, x_1) = \sum_{k=1}^{K} \lambda_k(x_i) p_k(x_i \mid x_{i-1}, \ldots, x_{i-(k-1)}),$$

where $\lambda = \{(\lambda_1(x), \ldots, \lambda_K(x)) : x \in V\}$, with $\sum_{k=1}^{K} \lambda_k(x) = 1$, are additional parameters which control the weighting between the different k-gram models. Even though there are more parameters, our hope is that lower weight will be given to the higher-order language models, thus guarding against overfitting.

(b) Introduce hidden variables into this model and derive an EM algorithm for maximum likelihood estimation. There is a big flaw with this approach; see if you can spot it before moving on to the next question.

(c) Prove that one maximum likelihood estimate of λ that achieves the optimum is degenerate in the sense that $\lambda_K(x) = 1$ for all $x \in V$. This is bad, because we end up with simply a K-gram model, which is bound to overfit for large K.

What is done in practice to avoid this degeneracy is to first estimate θ using (a) on one part of the data, and then use EM (b) to optimize λ on a different part of the data.

5. **EM algorithm for Hidden Markov Models**.

(a) Implement the EM algorithm for HMM's with Gaussian emission probabilities $p(y_t|q_t)$, where y_t is a two–dimensional real vector. Restrict the covariance matrices to be isotropic: $\Sigma = \sigma^2 I$.

(b) Fit a HMM with 4 states to the two–dimensional data in *hmm-gauss.dat* and evaluate the log likelihood on the training and test data (*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.