Automatic Speech Recognition
Question Set 3

1. We have shown that the maximum-likelihood (ML) estimator for the mean and variance of a multi-variate Gaussian vector is given by the sample mean and variance. Here we look at the case where the random variable X is discrete. In this case, the model parameters are the probabilities in the probability mass function. Suppose there are n elements in the support set of X. Let p_i be the probability of the i-th element, $p_i = Pr(X = x_i)$.

   (a) Given N samples for X, show that the ML estimator for p_i is $\frac{n_i}{N}$, where n_i is the number of samples equal to x_i.

   (b) What is the maximum likelihood?

   solution

   (a) Let k (as n in the problem) be the number of elements in the support set of X. The log data-likelihood is

   $$\log \prod_{i=1}^{k} p_i^{n_i} = \sum_i n_i \log p_i.$$

   We want to maximize the above quantity with the constraint that $\sum_{i=1}^{k} p_i = 1$. Using the Lagrangian

   $$L(p, \lambda) = \sum_i n_i \log p_i + \lambda(\sum_i p_i - 1),$$

   and setting the first derivative to zero, we get

   $$p_i = \frac{n_i}{-\lambda}, \quad i = 1, \ldots, k,$$

   and $\sum_i p_i - 1 = 0$, which is just the constraint. It follows that $\lambda = -N$ and

   $$p_i^* = \frac{n_i}{N}.$$

   (b) The maximum data likelihood is

   $$\prod_{i=1}^{k} (\frac{n_i}{N})^{n_i}.$$

   ■

2. Suppose we have N *labelled* data samples, $(x_j, l_j), j = 1, \ldots, N$, where $x \in R$ and $l \in \{1, \ldots, L\}$. Assume that X_j is Gaussian given l_j.

   (a) Describe how one can obtain the ML parameters for the class probabilities p_l, the class means μ_l, and the class variances Σ_l, for $l = 1, \ldots, L$.

   (b) If the data is *not* labelled, while still assuming that there are L classes and that the class-conditional distribution is Gaussian, state how to use the EM algorithm to obtain ML estimates.

<u>solution</u>

(a) For the class probability,

$$p_l^* = \frac{n_l}{N},$$

where n_l is the number of data points with label l. For the Gaussian parameters,

$$\mu_l^* = \frac{1}{n_l} \sum_{\{j|l_j=l\}} x_j,$$

and

$$\Sigma_l^* = \frac{1}{n_l} \sum_{\{j|l_j=l\}} (x_j - \mu_l)(x_j - \mu_l)'.$$

(b) If the data does not contain labels, one can start with an initial estimates of (p_l, μ_l, Σ_l) for all l, say by assigning random labels to the data and using the result in (a). Then at each iteration, use the current estimates to compute the posterior class probability for each data point to be used in the EM parameter updating equations (9.34), (9.31) and (9.32) in the textbook. Continue until the likelihood converges or a prescribed number of iterations has been reached. ■