

# 050/520/600.666 Information Extraction from Speech and Text

Homework # 7

Due April 21, 2005.

## Viterbi Training

Recall from the discussion of the EM algorithm in class that, for any pair of random variables  $\langle \mathbf{t}, \mathbf{y} \rangle$ , and any model parameters  $\theta'$ , if we can find another parameter set  $\theta$  such that

$$\sum_{\mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{y}) \log P_{\theta}(\mathbf{t}, \mathbf{y}) > \sum_{\mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{y}) \log P_{\theta'}(\mathbf{t}, \mathbf{y}) \quad (1)$$

then it will hold that

$$\sum_{\mathbf{t}} P_{\theta}(\mathbf{t}, \mathbf{y}) = P_{\theta}(\mathbf{y}) > P_{\theta'}(\mathbf{y}) = \sum_{\mathbf{t}} P_{\theta'}(\mathbf{t}, \mathbf{y}).$$

If  $P_{\theta}(\mathbf{t}, \mathbf{y})$  corresponds to an HMM, then this leads to the iterative *Baum-Welch* procedure for finding parameters  $\theta$  that maximize the likelihood  $P_{\theta}(\mathbf{y})$  of the observed data. This problem addresses *Viterbi training*, a variation on the Baum-Welch procedure for estimating parameters  $\theta$  of an HMM. In particular, the objective is changed to maximizing  $\max_{\mathbf{t}} P_{\theta}(\mathbf{t}, \mathbf{y})$  instead of  $\sum_{\mathbf{t}} P_{\theta}(\mathbf{t}, \mathbf{y})$ .

Consider an HMM with discrete outputs  $y \in \mathcal{Y}$ , and let the outputs be associated with the transitions (arcs) of the HMM. Let

$\mathcal{T}$	=	the set of all transitions in the HMM
$t_0$	=	an obligatory null transition leading into the initial state; cannot be taken again
$q(y t)$	=	probability of output $y$ given that transition $t$ was taken
$p(t' t)$	=	probability that the next transition is $t'$ given that the previous transition was $t$
$\theta$	=	$\{p(t' t), q(y t), t, t' \in \mathcal{T}, y \in \mathcal{Y}\}$ , the collection of HMM parameter values
$\mathbf{y}$	=	$y_1 y_2 \dots y_n$ , the observed output sequence
$\mathbf{t} = \mathbf{t}_{\theta}(\mathbf{y})$	=	$t_0 t_1 t_2 \dots t_n$ , the most probable transition sequence given $\mathbf{y}$ under $\theta$

Note that the random variable  $\mathbf{t}$  is defined as a deterministic  $\theta$ -dependent function of  $\mathbf{y}$ . Therefore, the value of  $P_{\theta}(\mathbf{t}|\mathbf{y})$  is either 1 or 0! For this special definition of the random variables  $\langle \mathbf{t}, \mathbf{y} \rangle$ , finding a parameter set  $\theta$  that guarantees (1) is therefore equivalent to finding  $\theta$  such that

$$P_{\theta}(\mathbf{t}_{\theta}, \mathbf{y}) > P_{\theta'}(\mathbf{t}_{\theta'}, \mathbf{y}). \quad (2)$$

We will now derive the iterative update formulae for  $\theta$  given the Viterbi path  $\mathbf{t} = \mathbf{t}_{\theta'}$  for the observed  $\mathbf{y}$  under the preceding parameter setting  $\theta'$ .

1. Express the joint likelihood  $P_\theta(\mathbf{t}, \mathbf{y})$  as a function of the parameters  $q(y|t)$  and  $p(t'|t)$  and of the counter contents  $C(t', t)$  and  $C(y, t)$  where

$$\begin{aligned} \forall t', t \in \mathcal{T} \quad C(t', t) &= \text{the number of times } t' \text{ followed } t \text{ in } \mathbf{t}. \\ \forall y \in \mathcal{Y}, \forall t \in \mathcal{T} \quad C(y, t) &= \text{the number of times } y \text{ was emitted from } t \text{ in } \langle \mathbf{t}, \mathbf{y} \rangle. \end{aligned}$$

2. Use the method of Lagrange multipliers to find the values of the parameters  $p(t'|t)$  and  $q(y|t)$  that maximize the joint likelihood  $P_\theta(\mathbf{t}, \mathbf{y})$ . Note that the constraints  $\sum_{t'} p(t'|t) = 1$  and  $\sum_y q(y|t) = 1$  must be satisfied.
3. How do the old parameter values  $\theta'$  enter the update equations of Step 2?
4. The solution  $\theta$  to Step 2, by construction, only guarantees that

$$P_\theta(\mathbf{t}, \mathbf{y}) > P_{\theta'}(\mathbf{t}, \mathbf{y}) \quad \text{for the old Viterbi path} \quad \mathbf{t} = \mathbf{t}_{\theta'}.$$

This is not the same as the condition (2) of the E-M procedure. Argue why the solution  $\theta$  to Step 2 will also guarantee (2), completing the proof of “correctness” of solution.

## Initializations and Local Maxima in the E-M Algorithm

This problem brings out the fact that the E-M algorithm, at best, converges to a local maximum of the likelihood, and that the point of convergence depends on the starting point. You will have to write a computer program to perform the numerical computations necessary to solve this problem.

Consider a 1-state HMM (!) that produces real-valued outputs  $y_1, y_2, \dots, y_n$ , and denote the lone state by  $s$ . Let the HMM have two output-producing arcs  $t_1$  and  $t_2$ , both from  $s$  to  $s$ , which may be taken with probabilities  $p(t_1) = \alpha_1$  and  $p(t_2) = \alpha_2 = (1 - \alpha_1)$ . Let the output densities be Gaussian.

$$\begin{aligned} \mathcal{N}(y|t_1) &= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y-\mu_1)^2}{2\sigma_1^2}}, \\ \mathcal{N}(y|t_2) &= \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}. \end{aligned}$$

Let  $\theta = [\alpha_1 \quad \mu_1 \quad \sigma_1^2 \quad \mu_2 \quad \sigma_2^2]$  denote the parameters of the HMM.

1. Write down the formula for  $P_\theta(y_1, \dots, y_n)$ . (Hint: this HMM is just a mixture density.)
2. Plot  $P_\theta(y_1)$  for  $\theta = [\frac{1}{3} \quad -2 \quad 1 \quad 2 \quad 1]$ .
3. Implement the E-M algorithm for updating  $\theta$  for this HMM. Make your implementation flexible, so that some parameters may be tied. E.g. we will experiment with arbitrarily fixing the variances and estimating only the means and  $\alpha_1$ , or with setting a minimum (floor) on the estimate of the variance, etc.

4. Let the observed data comprise the following  $n = 25$  values.

+0.608	-1.590	+0.235	+3.949	-2.249
+2.704	-2.473	+0.672	+0.262	+1.072
-1.773	+0.537	+3.240	+2.400	-2.499
+2.608	-3.458	+0.257	+2.569	+1.415
+1.410	-2.653	+1.396	+3.286	-0.712

Set  $\sigma_1^2 = \sigma_2^2 = 1$  and  $\alpha_1 = \frac{1}{3}$ , and obtain the sequence of iterates  $[\mu_1^{(k)}, \mu_2^{(k)}]$  starting with different values  $[\mu_1^{(0)}, \mu_2^{(0)}]$ . Try several different starting points in each of the 4 quadrants in the region  $[-4, +4] \times [-4, +4] \subset \mathbb{R}^2$ .

Plot the “trajectory” of the parameters in the  $x$ - $y$  plane for a few illustrative starting points.

5. How many different points of convergence did you encounter in Step 4? Comment on the value of the (local) maximum at these points, and speculate upon their *regions of attraction*. A region of attraction of a point of convergence is a set of starting points from which the E-M algorithm converges to that particular point.

6. Use a program such as MATLAB to plot the surface corresponding to the likelihood of the 25 data-points of Step 4 as a function of  $[\mu_1, \mu_2] \in [-4, +4] \times [-4, +4]$ , with  $\alpha_1, \sigma_1^2$  and  $\sigma_2^2$  fixed as above.

Comment on the trajectories you plotted in Step 5 in light of this surface plot, and revise your speculation about the regions of attraction if necessary.

7. Repeat Step 4 but this time, make  $\alpha_1$  also a free parameter. In particular, choose the same initial values of  $[\mu_1^{(0)}, \mu_2^{(0)}]$  that you used to plot the illustrative trajectories of  $[\mu_1^{(k)}, \mu_2^{(k)}]$  in Step 4, and plot the (new) trajectories when  $\alpha_1^{(0)} = \frac{1}{3}$ , but it is updated at every iteration instead of being held fixed. Comment on the point of convergence, as well as the (locally) maximum value of likelihood attained, when the HMM has more free parameters.

What happens when you start at  $[\mu_1^{(0)}, \mu_2^{(0)}]$  but an  $\alpha_1^{(0)}$  that is different from  $\frac{1}{3}$ ?

8. Let all five parameters be free, and compute the maximum likelihood that is attained for one or two different starting points.