

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

## Homework # 6

Due April 7, 2005.

### Entropy Rate of Markov Chains

We discussed the entropy  $H(X)$  of a random variable, and  $H(X, Y)$  of a pair of random variables, in class. We will now extend the notion to the entropy of a sequence of random variables  $X_1, X_2, \dots$ , taking values in a discrete and finite set  $\mathcal{X}$ . Specifically, it is straightforward to define the *average per-symbol entropy*

$$\frac{1}{n}H(X_1, X_2, \dots, X_n) = \frac{1}{n} \sum_{\langle x_1, \dots, x_n \rangle \in \mathcal{X}^n} p(x_1, \dots, x_n) \log \frac{1}{p(x_1, \dots, x_n)}, \quad (1)$$

of a finite sequence of random variables. The *entropy rate* of the random sequence is the limit

$$H(\mathcal{X}) = \lim_{n \rightarrow \infty} \frac{1}{n}H(X_1, X_2, \dots, X_n), \quad (2)$$

provided it exists. Show that this limit exists when  $X_1, X_2, \dots$ , is a Markov chain, and compute it in terms of its one-step transition probability matrix, as follows.

Consider a time-homogeneous Markov chain  $X_1, X_2, \dots$ , taking values in  $\mathcal{X}$ . i.e.

$$\text{Prob}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \pi(X_1 = x_1) \prod_{t=1}^{n-1} P(X_{t+1} = x_{t+1} | X_t = x_t),$$

where  $\pi$  is the initial distribution of the Markov chain, and the time-homogeneity provides that

$$P(X_{t+1} = j | X_t = i) = P(X_2 = j | X_1 = i) = P_{ij} = [\mathbf{P}]_{ij}, \quad \forall t = 1, 2, \dots,$$

where  $\mathbf{P}$  denotes the  $|\mathcal{X}| \times |\mathcal{X}|$  transition probability matrix.

1. Write down an expression for the average per-symbol entropy (1) in terms of the entropy of the rows of  $\mathbf{P}$  and  $\pi$ .
2. The stationary distribution of  $\mathbf{P}$  is a distribution  $\pi^*$  such that

$$\pi^* \times \mathbf{P} = \pi^*.$$

It is called the stationary distribution because if we set  $\pi = \pi^*$ , then the distribution of all  $X_t$  is also  $\pi^*$ :

$$P(X_2 = j) = \sum_{i \in \mathcal{X}} \pi(X_1 = i)P(X_2 = j|X_1 = i) = \sum_{i \in \mathcal{X}} \pi_i^* P_{ij} = \pi_j^*.$$

Simplify your answer above for the average per-symbol entropy when the Markov chain starts in its stationary distribution:  $\pi = \pi^*$ .

3. Compute the entropy rate (2) for this particular Markov chain.
4. A Markov chain is said to be ergodic if  $\lim_{n \rightarrow \infty} P(X_n = i) = \pi_i^*$  no matter what the original distribution  $\pi$  is. Compute the entropy rate for an ergodic Markov chain. What parameters of the Markov chain does this entropy rate depend on?

## 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

$$\sum_{\mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{y}) \log P_{\theta}(\mathbf{t}, \mathbf{y}) > \sum_{\mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{y}) \log P_{\theta'}(\mathbf{t}, \mathbf{y}) \quad (3)$$

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 <i>given</i> $\mathbf{y}$ 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 (3) is therefore equivalent to finding  $\theta$  such that

$$P_{\theta}(\mathbf{t}_{\theta}, \mathbf{y}) > P_{\theta'}(\mathbf{t}_{\theta'}, \mathbf{y}). \quad (4)$$

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 (4) of the E-M procedure. Argue why the solution  $\theta$  to Step 2 will also guarantee (4), completing the proof of “correctness” of solution.

## Alternate Presentation of the E-M Algorithm

Read the sections titled

- Maximum Likelihood from Incomplete Data (page 1)
- Classical Formulation of the EM Algorithm (page 3)
- An Example of Censored Data (page 5)
- Revisiting the EM Procedure: An Analytical View (page 19)
- Other Conditions for Convergence to Local Extrema (page 21)

from the lecture-notes provided alongside, and summarize your reading in 1 page or less.