Assignment Four

Matthew C. Scicluna
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
Montréal, QC H3T 1J4
matthew.scicluna@umontreal.ca

November 8, 2017

1 Entropy and Mutual Information

Let X be a discrete random variable on a finite space $\mathcal{X}$ with $|\mathcal{X}| = k$.

1. (a) Prove that the entropy $H(X) \geq 0$, with equality only when X is a constant.

   PROOF: WLOG we can assume that $p(x) > 0 \; \forall x \in \mathcal{X}$, (using that $0 \cdot log 0 = 0$). We have that $H(X) = -\sum_x p(x) log p(x) = \sum_x p(x) log p(x)^{-1} \geq 0$, since $p(x) > 0$ and $p(x)^{-1} \geq 1$. If $H(X) = 0$ then $\exists \alpha$ such that $p(\alpha)^{-1} = 1 \Rightarrow p(\alpha) = 1$. Hence X must be a constant, as needed.

   (b) Let $X \sim p$ and q be the Uniform distribution on $\mathcal{X}$. What is the relation between $D(p||q)$ and $H(X)$.

   CLAIM: $D(p||q) = -H(X) + logk$

   PROOF:

$$\begin{aligned} D(p||q) &= \sum_x p(x) log \frac{p(x)}{q(x)} \\ &= \sum_x p(x) log p(x) - \sum_x p(x) log q(x) \\ &= -H(X) + \sum_x p(x) log k \\ &= -H(x) + logk \end{aligned}$$

(c) An upper bound for $H(X)$ is $logk$ since $H(X) = logk - D(p||q) \Rightarrow H(X) \leq logk$.

We consider a pair of discrete random variables (X_1, X_2) defined over the finite set $\mathcal{X}_1 \times \mathcal{X}_2$. Let $p_{1,2}$, p_1 and p_2 denote respectively the joint distribution, the marginal distribution of X_1 and the marginal distribution of X_2. Define the mutual information as:

$$I(X_1, X_2) = \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1, x_2) log \frac{p_{1,2}(x_1, x_2)}{p_1(x_1)p_2(x_2)}$$

We again assume WLOG that $p(x_1, x_2) > 0 \ \forall (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$

2. (a) CLAIM: $I(X_1, X_2) \geq 0$

   PROOF: Notice that $I(X_1, X_2) = D(p_{1,2}||p_1p_2) \geq 0$ by the positiveness of $D(\cdot||\cdot)$.

   (b) We want to express $I(X_1, X_2)$ as a function of $H(X_1), H(X_2)$ and $H(X_1, X_2)$.

$$\begin{aligned}
I(X_1, X_2) &= \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1, x_2) log \frac{p_{1,2}(x_1, x_2)}{p_1(x_1)p_2(x_2)} \\
&= \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1, x_2) log p_{1,2}(x_1, x_2) - p_{1,2}(x_1, x_2) log p_1(x_1)p_2(x_2) \\
&= -H(X_1, X_2) - \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} \Big(p_1(x_1)p_2(x_2) log p_1(x_1) - p_1(x_1)p_2(x_2) log p_2(x_2) \Big) \\
&= -H(X_1, X_2) - \sum_{j=1}^{2} \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_1(x_1)p_2(x_2) log p_j(x_j) \\
&= -H(X_1, X_2) - \sum_{j=1}^{2} \sum_{x_j\in\mathcal{X}_j} p_j(x_j) log p_j(x_j) \\
&= -H(X_1, X_2) + H(X_1) + H(X_2)
\end{aligned}$$

   And so we can represent $I(X_1, X_2)$ using $H(X_1), H(X_2)$ and $H(X_1, X_2)$, as needed.

   (c) From the previous result we have that $I(X_1, X_2) \geq 0 \Rightarrow H(X_1) + H(X_2) \geq H(X_1, X_2)$, and so the maximal entropy of (X_1, X_2) is $H(X_1) + H(X_2)$. By definition this only occurs when $I(X_1, X_2) = 0$, which only occurs if $p_{1,2}(x_1, x_2) = p_1(x_1)p_2(x_2) \ \forall x_1, x_2 \in \mathcal{X}_1 \times \mathcal{X}_2$. This can be seen directly from the definition of I and using the strict positivity of $p(x_1, x_2)$.

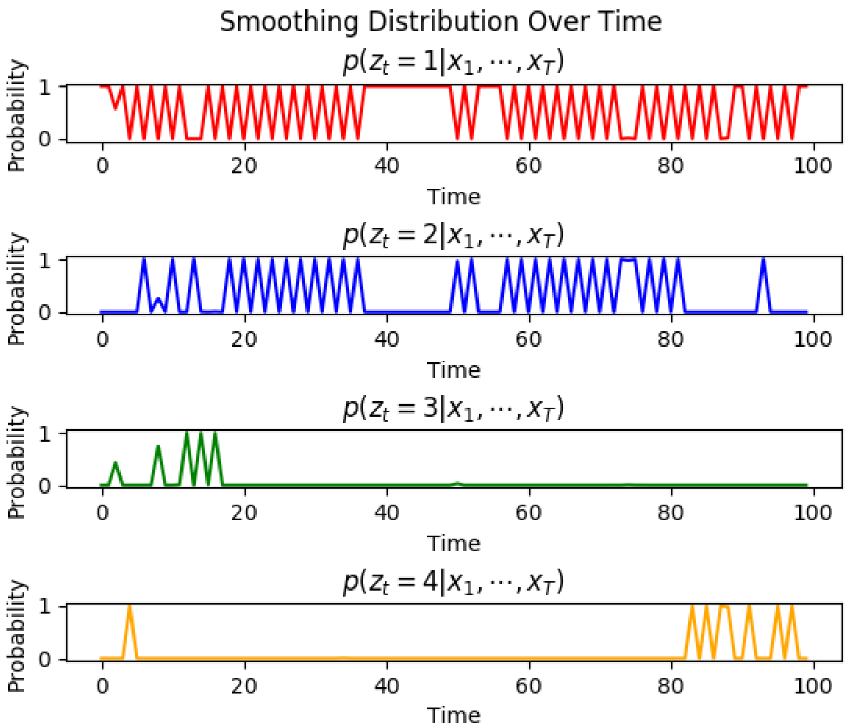

Figure 2.1: The smoothing distribution for the first 100 time points.

2 HMM – Implementation

We use an HMM model to account for the possible temporal structure of some data. We consider the following HMM model: the chain $(z_t)_{t=1}^T$ has $K = 4$ possible states, with an initial probability distribution $\pi \in \Delta_4$ and a probability transition matrix $A \in \mathbb{R}^{4\times4}$ where $A_{ij} = p(z_t = i|z_{t-1} = j)$ and conditionally on the current state z_t, we have observations obtained from Gaussian emission probabilities $x_t|(z_t = k) \sim N(x_t|\mu_k, \Sigma_k)$.

2.1 Fake Parameters Inference

We computed the vectors $\alpha(z_t) = p(z_t, x_{1:t})$ and $\beta(z_t) = p(x_{(t+1):T}|z_t)$ on the test set from Assignment 3 using the following parameters: $\pi_k = \frac{1}{4}$, $A_{ii} = \frac{1}{2}$ and $A_{ij} = \frac{1}{6}, i \neq j$, and μ_k, Σ_k as defined in the homework. We used these to compute the posterior of the latent variable over time $p(z_t \mid x_1, \cdots, x_T)$. We plotted the first 100 datapoints in figure 2.1.

2.2 M-Step Derivation

We now derive the M-Step for the Hidden Markov Model.
Let $\theta^{(s)} = (\pi^{(s)}, A^{(s)}, \mu_1^{(s)}, \cdots, \mu_K^{(s)}, \Sigma_1^{(s)}, \cdots, \Sigma_K^{(s)})$ be the ML parameters learned during

step s of EM. Let $\gamma_{tk} = P(z_t = k|x_{1:T})$ and $\xi_{tlm} = P(z_t = l, z_{t+1} = m|x_{1:T})$ - which are the quantities that were computed in the E-Step using $\theta^{(s)}$. The Expected Complete Data Log-Likelihood (at step $s+1$) is:

$$Q(\theta, \theta^{(s)}) = \sum_{k=1}^{K} \gamma_{tk} log\pi_k + \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_{tk} P(\bar{x}_t|\mu_k, \Sigma_k) + \sum_{t=1}^{T-1}\sum_{l=1}^{K}\sum_{m=1}^{K} \xi_{tlm} logA_{lm} \tag{2.1}$$

To solve for π_k we look for stationary points, subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$

$$\frac{\partial}{\partial \pi_k} Q(\theta, \theta^{(s)}) - \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) = \frac{\gamma_{1k}}{\pi_k} - \lambda = 0 \tag{2.2}$$

$$\Rightarrow \pi_k = \frac{\gamma_{1k}}{\lambda} \tag{2.3}$$

And using that $\sum_{k=1}^{K} \pi_k = 1$ we have that:

$$\pi_k^{(s+1)} = \frac{\gamma_{1k}}{\sum_{l=1}^{K} \gamma_{1l}} \tag{2.4}$$

To solve for A_{lm} we look for stationary points, subject to the constraint $\sum_{l=1}^{K} A_{lm} = 1$

$$\frac{\partial}{\partial A_{lm}} Q(\theta, \theta^{(s)}) - \lambda \left(\sum_{l=1}^{K} A_{lm} - 1 \right) = \sum_{t=1}^{T-1} \frac{\xi_{tlm}}{A_{lm}} - \lambda = 0 \tag{2.5}$$

$$\Rightarrow A_{lm} = \frac{\sum_{t=1}^{T-1} \xi_{tlm}}{\lambda} \tag{2.6}$$

And using that $\sum_{l=1}^{K} A_{lm} = 1$ we have that:

$$A_{lm}^{(s+1)} = \frac{\sum_{t=1}^{T-1} \xi_{tlm}}{\sum_{l=1}^{K} \sum_{t=1}^{T-1} \xi_{tlm}} \tag{2.7}$$

To solve for μ_k we look for stationary points

$$\frac{\partial}{\partial \mu_k} Q(\theta, \theta^{(s)}) = \sum_{t=1}^{T} \frac{\partial}{\partial \mu_k} \frac{-\gamma_{tk}}{2} (\bar{x}_t - \mu_k)^T \Sigma_k^{-1} (\bar{x}_t - \mu_k) \tag{2.8}$$

$$= \sum_{t=1}^{T} \frac{-\gamma_{tk}}{2} (\bar{x}_t - \mu_k) \Sigma_k^{-1} = 0 \tag{2.9}$$

$$\Rightarrow \sum_{t=1}^{T-1} \gamma_{tk} (\bar{x}_t - \mu_k) = 0 \tag{2.10}$$

$$\Rightarrow \mu_k^{(s+1)} = \frac{\sum_{t=1}^{T} \gamma_{tk} \bar{x}_t}{\sum_{t=1}^{T-1} \gamma_{tk}} \tag{2.11}$$

To solve for Σ_k we look for stationary points. As in assignment 1, we take the derivative w.r.t Σ_k^{-1}

$$\frac{\partial}{\partial \Sigma_k^{-1}} Q(\theta, \theta^{(s)}) = \sum_{t=1}^{T} \left(\frac{\partial}{\partial \mu_k} \frac{-\gamma_{tk}}{2} (\bar{x}_t - \mu_k)^T \Sigma_k^{-1} (\bar{x}_t - \mu_k) - \frac{\gamma_{tk}}{2} log|\Sigma_k| \right) \tag{2.12}$$

$$\tag{2.13}$$