RICE UNIVERSITY, DEPARTMENT OF COMPUTER SCIENCE

Assignment 6, COMP 540

Chen Zeng(cz39), Zhihui Xie(zx18)

April 16, 2018

1 Hidden Markov Models

Answer1 Formulate this problem as a Hidden Markov model as below:
The set of observables: $O = \{l, m, h\}$, low, medium, or high
The set of hidden states: $S = \{H, U\}$, healthy or unhealthy
The parameters $\lambda = [\pi, a, b]$:

- The initial state distribution: $\lambda : [0.5, 0.5]$
- The transition matrix: $\mathbf{a} = \begin{array}{cc} \begin{matrix} \text{H} & \text{U} \end{matrix} & \\ \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix} & \begin{matrix} \text{H} \\ \text{U} \end{matrix} \end{array}$
- The emission matrix: $\mathbf{b} = \begin{array}{cc} \begin{matrix} \text{l} & \text{m} & \text{h} \end{matrix} & \\ \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{bmatrix} & \begin{matrix} \text{H} \\ \text{U} \end{matrix} \end{array}$

The corresponding graphical model: please help to refer to figure 1.1

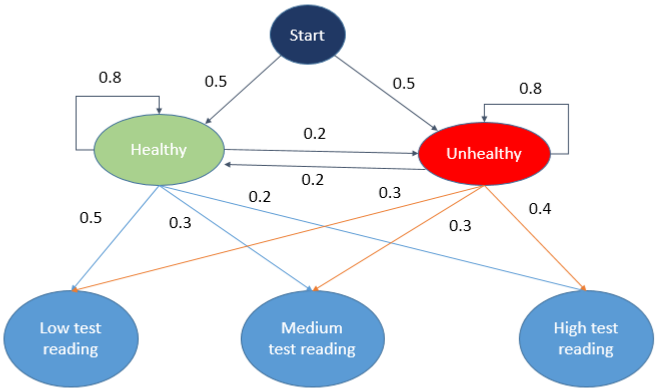

Figure 1.1: Markov Chain

Answer2

$$P(S_1 = H) = P(S_0 = H)\,P(S_1 = H|S_0 = H) + P(S_0 = U)\,P(S_1 = H|S_0 = U) = 0.5 * 0.8 + 0.5 * 0.2 = 0.5$$

$$P(S_1 = U) = P(S_0 = H)\,P(S_1 = U|S_0 = H) + P(S_0 = U)\,P(S_1 = U|S_0 = U) = 0.5 * 0.2 + 0.5 * 0.8 = 0.5$$

$$\begin{aligned} P =& P(S_2 = H|O_1 = l, O_2 = l) \\ =& \frac{P(S_2 = H, O_1 = l, O_2 = l)}{P(O_1 = l, O_2 = l)} \\ =& \frac{P(S_2 = H, O_1 = l, O_2 = l|S_1 = H)\,P(S_1 = H) + P(S_2 = H, O_1 = l, O_2 = l|S_1 = U)\,P(S_1 = U)}{P(O_1 = l, O_2 = l|S_1 = H)\,P(S_1 = H) + P(O_1 = l, O_2 = l|S_1 = U)\,P(S_1 = U)} \\ =& \frac{P(S_2 = H, O_1 = l, O_2 = l|S_1 = H) + P(S_2 = H, O_1 = l, O_2 = l|S_1 = U)}{P(O_1 = l, O_2 = l|S_1 = H) + P(O_1 = l, O_2 = l|S_1 = U)} \end{aligned}$$

We know that O_1 is independent to both S_1 and O_2, so we have:

$$P(S_2 = H, O_1 = l, O_2 = l|S_1 = H) = P(S_2 = H, O_2 = l|S_1 = H)\,P(O_1 = l|S_1 = H)$$

$$P(S_2 = H, O_1 = l, O_2 = l|S_1 = U) = P(S_2 = H, O_2 = l|S_1 = U)\,P(O_1 = l|S_1 = U)$$

Then we have:

$$P = \frac{P(S_2 = H, O_2 = l|S_1 = H)\,P(O_1 = l|S_1 = H) + P(S_2 = H, O_2 = l|S_1 = U)\,P(O_1 = l|S_1 = U)}{P(O_1 = l|S_1 = H)\,P(O_2 = l|S_1 = H) + P(O_1 = l|S_1 = U)\,P(O_2 = l|S_1 = U)}$$

We have:

$$P(S_2 = H, O_2 = l|S_1 = H)\,P(O_1 = l|S_1 = H) = (0.8 * 0.5) * 0.5 = 0.2$$

$$P(S_2 = H, O_2 = l|S_1 = U)\,P(O_1 = l|S_1 = U) = (0.2 * 0.5) * 0.3 = 0.03$$

$$
\begin{aligned}
&P(O_2 = l|S_1 = H)\\
&= P(O_2 = l; S_2 = H|S_1 = H) + P(O_2 = l; S_2 = U|S_1 = H)\\
&= \frac{P(O_2 = l; S_2 = H|S_1 = H)}{P(S_1 = H; S_2 = H)} \cdot \frac{P(S_1 = H; S_2 = H)}{P(S_1 = H)} + \frac{P(O_2 = l; S_2 = U|S_1 = H)}{P(S_1 = H; S_2 = U)} \cdot \frac{P(S_1 = H; S_2 = U)}{P(S_1 = H)}\\
&= P(O_2 = l|S_2 = H; S_1 = H) \cdot P(S_2 = H|S_1 = H) + P(O_2 = l|S_2 = U; S_1 = H) \cdot P(S_2 = U|S_1 = H)\\
&= P(O_2 = l|S_2 = H) \cdot P(S_2 = H|S_1 = H) + P(O_2 = l|S_2 = U) \cdot P(S_2 = U|S_1 = H)
\end{aligned}
$$

$$
\begin{aligned}
&P(O_2 = l|S_1 = U)\\
&= P(O_2 = l; S_2 = H|S_1 = U) + P(O_2 = l; S_2 = U|S_1 = U)\\
&= \frac{P(O_2 = l; S_2 = H|S_1 = U)}{P(S_1 = U; S_2 = H)} \cdot \frac{P(S_1 = U; S_2 = H)}{P(S_1 = U)} + \frac{P(O_2 = l; S_2 = U|S_1 = U)}{P(S_1 = U; S_2 = U)} \cdot \frac{P(S_1 = U; S_2 = U)}{P(S_1 = U)}\\
&= P(O_2 = l|S_2 = H; S_1 = U) \cdot P(S_2 = H|S_1 = U) + P(O_2 = l|S_2 = U; S_1 = U) \cdot P(S_2 = U|S_1 = U)\\
&= P(O_2 = l|S_2 = H) \cdot P(S_2 = H|S_1 = U) + P(O_2 = l|S_2 = U) \cdot P(S_2 = U|S_1 = U)
\end{aligned}
$$

So we also have:

$$
\begin{aligned}
&P(O_1 = l|S_1 = H)\, P(O_2 = l|S_1 = H)\\
&= P(O_1 = l|S_1 = H)\,(P(O_2 = l|S_2 = H)\, P(S_2 = H|S_1 = H) + P(O_2 = l|S_2 = U)\, P(S_2 = U|S_1 = H))\\
&= 0.5 * (0.5 * 0.8 + 0.3 * 0.2)\\
&= 0.23
\end{aligned}
$$

$$
\begin{aligned}
&P(O_1 = l|S_1 = U)\, P(O_2 = l|S_1 = U)\\
&= P(O_1 = l|S_1 = U)\,(P(O_2 = l|S_2 = H)\, P(S_2 = H|S_1 = U) + P(O_2 = l|S_2 = U)\, P(S_2 = U|S_1 = U))\\
&= 0.3 * (0.5 * 0.2 + 0.3 * 0.8)\\
&= 0.102
\end{aligned}
$$

Therefore, we can get the result:

$$
P = \frac{0.2 + 0.03}{0.23 + 0.102} = 0.693
$$

Answer3 The Viterbi's algorithm to calculate the most likely state sequence can refer to figure 1.2

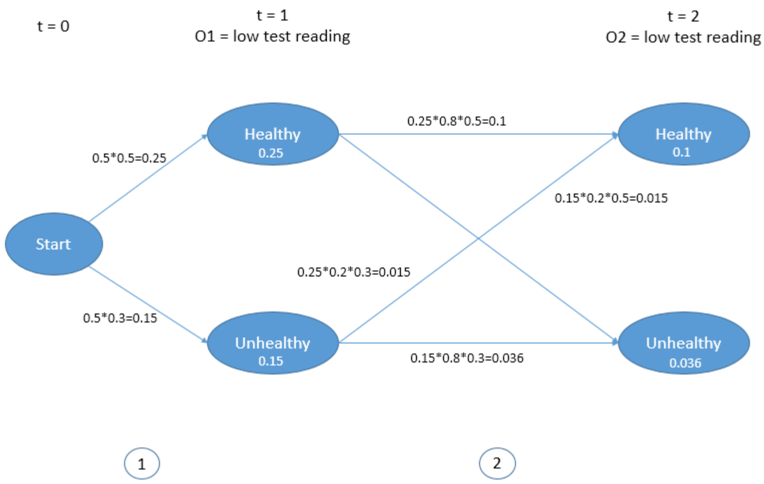

Figure 1.2: Viterbi's Algorithm

The algorithm can be divided into two stages:
1. For the stage 1 ($t_0->t_1$), we have $P(newState) = P_{start}(state)*P_{observation}("lowtestreading")$. Thus, the most likely state in t_1 is Healthy (P=0.5).
2. For the stage 2 ($t_1->t_2$), we have $P_{newState} = P_{oldState}*P_{trans}(oldState->newState)*P_{oberservation}("lowtestreading"|newState)$. Thus, the most likely state in t_2 is Healthy (P=0.1).
As a result, the most likely state sequence for t = 0,1,2 given the evidence from the previous subpart is Healthy, Healthy.

2 EM for mixtures of Bernoullis

Answer1 For a Bernoullis distribution models, we have P(X=x) = μ , where $x \in \{0,1\}, \mu \in [0,1]$.

For the mixture of K Bernoullis distrutions, we have $P(X = x|\mu, \pi_k) = \sum_{k=1}^{K} \pi_k P(X = x|\mu_k)$, where $\mu = \{\mu_1, ..., \mu_k\}, \pi = \{\pi_1, ..., \pi_k\}\, and P(X = x|\mu_k) = \prod_{i=1}^{D} \mu_{k_i}^{x^{(i)}} (1-\mu_{k_i})^{(1-x^{(i)})}$. π_k are the mixture proportions.

$$\begin{cases} z^{(i)} \sim Multibernoullis(\pi); \pi_k > 0, \sum_k \pi_k = 1 \\ x^{(i)}|_{z^{(i)}=k} \sim B(\mu_k, \sum_k) \end{cases}$$

Given data set $X = \{x_1, ..., x_m\}$, we have the loss function and log likelihood:

$$\begin{aligned} L &= \prod_{i=1}^{m} P(x^{(i)}|\mu,\pi) \\ &= \prod_{i=1}^{m} \sum_{k=1}^{K} P(z^{(i)};\pi) P(x^{(i)}|z^{(i)}=k;\mu) \end{aligned}$$

$$\begin{aligned} l &= \sum_{i=1}^{m} \sum_{k=1}^{K} r_k^{(i)} log \frac{\left[P(z^{(i)};\pi)P(x^{(i)}|z^{(i)}=k;\mu)\right]}{r_k^{(i)}} \\ &= \sum_{i=1}^{m} \sum_{k=1}^{K} r_k^{(i)} \left[\sum_{j=1}^{D} \left(x_j^{(i)} log \mu_{kj} + \left(1 - x_j^{(i)} log(1-\mu_{kj})\right)\right) + log\pi_k - log r_k^{(i))} \right] \end{aligned}$$

Thus, let the derivative equal to 0, we have

$$\sum_{i=1}^{m} r_k^{(1)} \left(\frac{x_j^{(i)}}{\mu_{kj}} - \frac{1-x_j^{(i)}}{1-\mu_{kj}} \right) = 0$$

Finally, we can have:

$$\mu_{kj} = \frac{\sum_{i=1}^{m} r_k^{(i)} x_j^{(ii)}}{\sum_{i=1}^{m} r_k^{(i)}}$$

Answer2 If we have a Beta(α, β) prior, then we have:

$$P(X = x|\mu_k) = \prod_{i=1}^{D} \mu_{k_i}^{\alpha-1} (1-\mu_{k_i})^{\beta-1}$$

According to the last question, we can have:

$$\frac{\alpha-1}{\mu_{kj}} - \frac{\beta-1}{1-\mu_{kj}} + \sum_{i=1}^{m} r_k^{(1)} \left(\frac{x_j^{(i)}}{\mu_{kj}} - \frac{1-x_j^{(i)}}{1-\mu_{kj}} \right) = 0$$

Finally, we can have:

$$\mu_{kj} = \frac{\sum_{i=1}^{m} r_k^{(i)} x_j^{(ii)} + \alpha - 1}{\sum_{i=1}^{m} r_k^{(i)} + \alpha + \beta - 2}$$

3 Principal Components Analysis

Solution:
We know that V = $\{\alpha u : \alpha \epsilon R\}$. Thus, we have:

$$\begin{aligned} f_u(v) &= argmin_{v\epsilon V} ||x - v||^2 \\ &= argmin_{v\epsilon V} (x - \alpha u)^T (x - \alpha u) \\ &= argmin_{v\epsilon V} (x^T x - \alpha^T u^T x - x^T \alpha u + u^T u \alpha^2) \end{aligned}$$

To get the minimum of $f_u(v)$, we can let its derivative equal to 0, so we have:

$$\frac{\partial f_u(v)}{\partial \alpha} = -2u^T x + 2\alpha = 0$$

Then we have:

$$\alpha = u^T x$$

Thus

$$f_u(v) = u^T x u$$

After that, we can have:

$$\begin{aligned}
& argmin_{u:u^t u=1} \sum_{i=1}^{m} ||x^{(i)} - f_u(x^{(i)})||^w \\
&= argmin_{u:u^t u=1} \sum_{i=1}^{m} \left(x^{(i)} - u^T x u\right)^T \left(x^{(i)} - u^T x u\right) \\
&= argmin_{u:u^t u=1} \sum_{i=1}^{m} \left(x^{(i)^T} x^{(i)} - \left(u^T x(i)\right)^2\right) \\
&= argmax_{u:u^t u=1} \sum_{i=1}^{m} \left(\left(u^T x(i)\right)^2\right)
\end{aligned}$$

According to the assumption that the data have zero-mean and unit variance in each dimension, so we have:

$$\begin{aligned}
& argmin_{u:u^t u=1} \sum_{i=1}^{m} ||x^{(i)} - f_u(x^{(i)})||^w \\
&= argmax_{u:u^t u=1} \sum_{i=1}^{m} \left(\left(u^T x(i)\right)^2\right) \\
&= argmax_{u:u^t u=1} \sum_{i=1}^{m} \left(\left(u^T x(i) - u^T \overline{x^{(i)}}\right)^2\right)
\end{aligned}$$

So $argmin_{u:u^t u=1} \sum_{i=1}^{m} ||x^{(i)} - f_u(x^{(i)})||^w$ can be changed to maximize the variance projected data on u, which means that it can find the principal component of the data.