Machine learning 2
Exercise sheet 9

FLEISCHMANN Kay, Matrnr: 352247
ROHRMANN Till, Matrnr: 343756

June 16, 2013

1 Hidden Markov Models
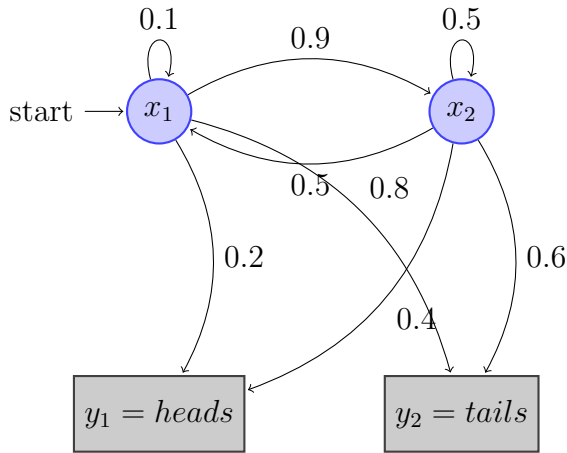
Let $A_{i,j}$ the transition matrix between hidden states x_i and x_j. $B_{i,j}$ is the the probability, beeing in state x_i to observe y_j. The following matrices A and B describe two hidden states and two possible observations.

$$A = \begin{pmatrix} 0.1 & 0.9 \\ 0.5 & 0.5 \end{pmatrix}$$

$$B = \begin{pmatrix} 0.2 & 0.8 \\ 0.4 & 0.6 \end{pmatrix}$$

a. Draw the graph of the model

The Hidden-Markov-chain for A and B looks like the following4

Because of $\pi = (1|0)$ all state-sequences start with the initial state x_1.

b.

We interpret the above as a model for an experiment with two hidden (unfair) coins and two visible coins. Describe such an experiment which can be modelled by the Markov model given above. Here, heads should correspond to the first indices in A, B, π, heads to the second.

A guy is standing behind an curtain and is throwing two unfair coins, just one each time but starts always with the first one. A second guy in front of the curtain just get the information about the observed heads or tails as a sequence. The task of the second guy is to find the most likely sequence of the thrown coins.

c.

Given the bayes rule:

$$P(X|Y(1) = tails, Y(2) = tails) = \frac{P(Y(1) = tails, Y(2) = tails|X)P(X)}{P(Y(1) = tails, Y(2) = tails)} \tag{1}$$

Because of the statistical independence, we can write

$$P(Y(1), Y(2)|X(1), X(2)) = P(Y(1)|X(1)) \cdot P(Y(2)|X(2)) \tag{2}$$

Furthermore we have

$$
\begin{aligned}
P(X(1) = x_i, X(2) = x_j) &= P(X(2) = x_j \mid X(1) = x_i) \cdot P(X(1) = x_i) && (3) \\
&= B_{i,j} \cdot \pi_i && (4)
\end{aligned}
$$

All possible sequences with length 2 are

	State sequence	$P(X)$	$P(Y(1) = tails, Y(2) = tails \mid X)$
1	$x_1 \rightarrow x_1$	$P(X(1) = x_1, X(2) = x_1) = 0.1$	$0.8 \cdot 0.8 = 0.64$
2	$x_1 \rightarrow x_2$	$P(X(1) = x_1, X(2) = x_2) = 0.9$	$0.8 \cdot 0.6 = 0.48$
3	$x_2 \rightarrow x_1$	$P(X(1) = x_2, X(2) = x_1) = 0$	$0.6 \cdot 0.8 = 0.48$
4	$x_2 \rightarrow x_2$	$P(X(1) = x_2, X(2) = x_2) = 0$	$0.6 \cdot 0.6 = 0.36$

$$
\begin{aligned}
P(Y(1) = y_2, Y(2) = y_2) &= \sum_{X \in S^2} P(Y(1) = y_2, Y(2) = y_2 \mid X) P(X) && (5) \\
&= \sum_{X \in S^2} P(Y(1) = y_2, Y(2) = y_2 \mid X) P(X) && (6) \\
&= 0.64 \cdot 0.1 + 0.48 \cdot 0.9 + 0.48 \cdot 0 + 0.36 \cdot 0 && (7) \\
&= 0.496 && (8)
\end{aligned}
$$

The result for $P(X \mid Y(1) = tails, Y(2) = tails)$ follows as

	State sequence	$P(X \mid Y(1) = tails, Y(2) = tails)$
1	$X = X(1) = x_1, X(2) = x_1$	$\frac{0.64 \cdot 0.1}{0.496} \approx 0.129$
2	$X = X(1) = x_1, X(2) = x_2$	$\frac{0.48 \cdot 0.9}{0.496} \approx 0.871$
3	$X = X(1) = x_2, X(2) = x_1$	$\frac{0.48 \cdot 0}{0.496} = 0$
4	$X = X(1) = x_2, X(2) = x_2$	$\frac{0.36 \cdot 0}{0.496} = 0$

f.

Prove in the scenario of (e) and the assumptions of the exercise: if $N \rightarrow \infty$, then the sample observation frequency of heads/tails at fixed position n converges, in probability and entry-wise, to the vector $\pi \cdot A^n \cdot B$.

Proof. Given an initial distribution, the probability that we see heads in the sequence of observations Y at position k is $\Pr(Y(k) = y_1) = (\pi \cdot A^k \cdot B)_1$. Thus we see tails at the position k with probability $\Pr(Y(k) = y_2) = (\pi \cdot A^k \cdot B)_2 = 1 - (\pi \cdot A^k \cdot B)_1$.

Running N times the experiment with the same initial distribution, we obtain Y_i observation sequences with $1 \leq i \leq N$. It holds that for all time points t the observations $Y_i(t)$ of the different run of the experiment are conditonally independent. In order to calculate the observations frequency of heads we calculate for each position k with $1 \leq k \leq n$ the empirical frequency:

$$
heads_{emp}(k) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{Y_i(k) = y_1} \tag{9}
$$

The term $\mathbb{1}_{Y_i(k) = y_1}$ can be considered a Bernoulli variable whose probability to be 1 is

$$
\Pr(\mathbb{1}_{Y_i(k) = y_1} = 1) = (\pi \cdot A^k \cdot B)_1
$$

and its probability to be 0 is

$$\Pr(\mathbb{1}_{Y_i(k)=y_1} = 0) = 1 - (\pi \cdot A^k \cdot B)_1$$
$$= (\pi \cdot A^k \cdot B)_2$$

So we can observe that equation (9) is nothing else than the average of i.i.d. Bernoulli variables. This means that

$$\mathbb{E}(heads_{emp}(k)) = (\pi \cdot A^k \cdot B)_1$$

Furthermore by using the Chebyshev inequality we obtain for any $\epsilon > 0$.

$$\Pr(|heads_{emp}(k) - \mathbb{E}(heads_{emp}(k))| \geq \epsilon) \leq \frac{Var(\mathbb{1}_{Y_i(k)=y_1})}{N\epsilon^2}$$
$$\Pr(|heads_{emp}(k) - (\pi \cdot A^k \cdot B)_1| \geq \epsilon) \leq \frac{(\pi \cdot A^k \cdot B)_1 \cdot (\pi \cdot A^k \cdot B)_2}{N\epsilon^2} \qquad (10)$$

Since $(\pi \cdot A^k \cdot B)_1 \cdot (\pi \cdot A^k \cdot B)_2$ is finite equation (10) converges against 0 for $N \to \infty$:

$$\lim_{N\to\infty} \Pr(|heads_{emp}(k) - (\pi \cdot A^k \cdot B)_1| \geq \epsilon) \leq \lim_{N\to\infty} \frac{(\pi \cdot A^k \cdot B)_1 \cdot (\pi \cdot A^k \cdot B)_2}{N\epsilon^2}$$
$$= 0 \qquad (11)$$

The same holds for

$$\lim_{N\to\infty} \Pr(|tails_{emp}(k) - (\pi \cdot A^k \cdot B)_2| \geq \epsilon) \leq 0 \qquad (12)$$

with

$$tails_{emp}(k) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{Y_i(k)=y_2}$$

and

$$\Pr(\mathbb{1}_{Y_i(k)=y_2} = 1) = (\pi \cdot A^k \cdot B)_2$$

if we apply the same reasoning. With equations (11) and (12) we have shown that the observed frequencies $heads_{emp}(k)$ and $tails_{emp}(k)$ converge in probability against $\pi \cdot A^k \cdot B$.

$\square$