Course **Artificial Intelligence Techniques (CS4375)**
Course e-mail cs4375-support-ewi@tudelft.nl
Exercise Sheet **Machine Learning**

Delft University of Technology

Exercise Sheet - Machine learning

First two exercises are obtained from Russell & Norvig

Exercise 1. Consider the noisy-OR model for fever described in Russell & Norvig Section 14.3. Explain how to apply maximum-likelihood learning to fit the parameters of such a model to a set of complete data. (Hint: use the chain rule for partial derivatives.)

Exercise 2. Consider the application of EM to learn the parameters for the network in Figure 20.13(a), given the true parameters in Equation (20.7):

a) Explain why the EM algorithm would not work if there were just two attributes in the model rather than three.

b) Show the calculations for the first iteration of EM starting from Equation (20.8).

Bonus (hard)

Exercise 3. Look at the EM transition parameter update:

$$\theta^{trans(k+1)}_{y\to v} = \frac{\sum_t Q^{(k+1)}(S_{t-1}=y, S_t=v)}{\sum_t Q^{(k+1)}(S_{t-1}=y)}$$

and its derivation (in the lecture slides, see also https://www.worldscientific.com/doi/pdfplus/10.1142/S0218001401000836 (An introduction to hidden Markov models and Bayesian networks of Z Ghahramani), but note the notation is quite dissimilar from R&N).

For the observation probabilities, the update rule is:

$$\theta^{obs}_{y\to w} = \frac{\sum_{t \text{ s.t. } o_t=w} \quad Q^{(k+1)}(S_t=y)}{\sum_t Q^{(k+1)}(S_t=y)}.$$

Derive this update step.

Solution:

Exercise 1

There are a couple of ways to solve this problem. Here, we show the indicator variable method described on page 743. Assume we have a child variable Y with parents $X_1, ..., X_k$ and let the range of each variable be $\{0,1\}$. Let the noisy-OR parameters be $q_i = P(Y=0|X_i=1, X_{-i}=0)$. The noisy-OR model then asserts that

$$P(Y=1 \mid x_1, \ldots, x_k) = 1 - \prod_{i=1}^{k} q_i^{x_i}$$

Assume we have m complete-data samples with values y_j for Y and x_{ij} for each X_i. The conditional log likelihood for $P(Y|X_1, ..., X_k)$ is given by

$$\begin{aligned} L &= \sum_j \log \left(1 - \prod_i q_i^{x_{ij}}\right)^{y_j} \left(\prod_i q_i^{x_{ij}}\right)^{1-y_j} \\ &= \sum_j y_j \log \left(1 - \prod_i q_i^{x_{ij}}\right) + (1-y_j) \sum_i x_{ij} \log q_i \end{aligned}$$

The gradient with respect to each noisy-OR parameter is

$$\begin{aligned}\frac{\partial L}{\partial q_i} &= \sum_j -\frac{y_j x_{ij} \prod_i q_i^{x_{ij}}}{q_i \left(1 - \prod_i q_i^{x_{ij}}\right)} + \frac{(1-y_j)\, x_{ij}}{q_i} \\ &= \sum_j \frac{x_{ij} \left(1 - y_j - \prod_i q_i^{x_{ij}}\right)}{q_i \left(1 - \prod_i q_i^{x_{ij}}\right)}\end{aligned}$$

Exercise 2

a. Consider the ideal case in which the bags were infinitely large so there is no statistical fluctuation in the sample. With two attributes (say, $Flavor$ and $Wrapper$), we have five unknowns: θ gives the relative sizes of the bags, θ_{F_1} and θ_{F_2} give the proportion of cherry candies in each bag, and θ_{W_1} and θ_{W_2} give the proportion of red wrappers in each bag. In the data, we observe just the flavor and wrapper for each candy; there are four combinations, so three independent numbers can be obtained. This is not enough to recover five unknowns. With three attributes, there are wight combinations and seven numbers can be obtained, enough to recover the seven parameters.

b. The combination for $\theta^{(1)}$ has eight nearly identical expressions and calculations, one of which is shown. The symbolic expression for $\theta_{F_1}^{(1)}$ is hown, but not its evaluation; it would be reasonable to ask students to write out the expression in terms of the parameters, as was done for $\theta^{(1)}$, and calculate the value. The final answers are given in the chapter.

Exercise 3

We follow the steps from the slides (starting around slide 77).

$\mathbf{x} = \{(o_{i1}, o_{i2}, ..., o_{iT})\}_{i=1}^{N}$ is the set of N trajectories of observations of the form $x_i = (o_{i1}, o_{i2}, ..., o_{iT})$.

$\mathbf{z} = \{(s_{i1}, s_{i2}, ..., s_{iT})\}_{i=1}^{N}$ is the set of N trajectories of hidden states of the form $z_i = (s_{i1}, s_{i2}, ..., s_{iT})$.

The joint probability for the HMM is defined as

$$P(\mathbf{x}, \mathbf{z}|\theta) = \theta_{s_0}^{init} \prod_{t=1}^{T} \theta_{s_{t-1}\to s_t}^{trans} \theta_{s_t \to o_t}^{obs}$$

where

$$\begin{aligned}\theta_{s_{t-1}\to s_t}^{trans} &\triangleq P(s_t|s_{t-1}) \\ \theta_{s_t\to o_t}^{obs} &\triangleq P(o_t|s_t)\end{aligned}$$

are the parameters for the transition and observation probabilities.

For an EM update we then need

$$\theta^{(k+1)} = argmax_{\theta} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}, \theta^{(k)}) L(\mathbf{x}, \mathbf{z}|\theta)$$

The full (or 'completed') log-likelihood is:

$$L(\mathbf{x}, \mathbf{z}|\theta) = \log \theta_{s_0}^{init} + \sum_{t=1}^{T} \log \theta_{s_{t-1}\to s_t}^{trans} + \sum_{t=1}^{T} \log \theta_{s_t\to o_t}^{obs}$$

Now we are only looking at the update for the observation probabilities of some state y. We should maximize:

$$\begin{aligned}\theta_{y\to .}^{obs(k+1)} =& argmax_{\theta_{y\to .}^{obs}} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}, \theta^{(k)}) L(\mathbf{x}, \mathbf{z}|\theta) \\ =& argmax_{\theta_{y\to .}^{obs}} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}, \theta^{(k)}) \Big[\sum_{t=1}^{T} \log \theta_{s_t\to o_t}^{obs} \Big]\end{aligned}$$

Let's abbreviate $Q^{(k+1)}(\mathbf{z}) = P(\mathbf{z}|\mathbf{x}, \theta^{(k)})$.

Then:

$$\begin{aligned}
&argmax_{\theta^{obs}_{y\rightarrow .}} \sum_{\mathbf{z}} Q^{(k+1)}(\mathbf{z})\Big[\sum_{t=1}^{T} \log \theta^{obs}_{s_t \rightarrow o_t}\Big] \\
&= argmax_{\theta^{obs}_{y\rightarrow .}} \sum_{\mathbf{z}} Q^{(k+1)}(\mathbf{z})\Big[\sum_{t\ s.t.\ \{S_{i,t=y}\}}^{T} \log \theta^{obs}_{y \rightarrow o_t}\Big] \\
&= argmax_{\theta^{obs}_{y\rightarrow .}} \sum_{t} \sum_{\mathbf{z}\ s.t.\ S_{t=y}} Q^{(k+1)}(\mathbf{z}) \log \theta^{obs}_{y \rightarrow o_t} \\
&= argmax_{\theta^{obs}_{y\rightarrow .}} \sum_{t} \sum_{(s_0 \ldots s_{t-1}, s_{t+1} \ldots s_T)} Q^{(k+1)}(s_0 \ldots s_t = y \ldots s_T) \log \theta^{obs}_{y \rightarrow o_t} \\
&= argmax_{\theta^{obs}_{y\rightarrow .}} \sum_{t} Q^{(k+1)}(s_t = y) \log \theta^{obs}_{y \rightarrow o_t} \\
&=argmax_{\theta^{obs}_{y\rightarrow .}} \sum_{w} \sum_{t\ s.t.\ o_t=w} Q^{(k+1)}(s_t = y) \log \theta^{obs}_{y \rightarrow w}
\end{aligned}$$

Then for the maximization we have:

$$\theta^{obs}_{y\rightarrow w} = \frac{\sum_{t \text{ s.t. } o_t=w} \quad Q^{(k+1)}(S_t = y)}{\sum_t Q^{(k+1)}(S_t = y)}.$$

▲