CPSC-340 **Machine Learning and Data Mining** 2012

Homework # 3 - Due Monday 15th of October

NAME:______________________________________

Signature:______________________________________

STD. NUM: ______________________________________

1. **Hidden Markov Models for Assistive Technologies**

Assistive technologies are automated devices to help the elderly and disabled, and are an extremely active area of robotics (and one where UBC researchers and alums are involved). An example is the Misubishi Wakamaru domestic robot, which assists the elderly.

HMMs are an important tool for these robots, as they need to infer and predict a human being's internal states and motivations from their actions. For this exercise, you will be implementing an HMM for a robot pet, which is designed to play with or entertain its owner when it thinks the owner is sad, and remain unobtrusive when the owner is happy. Of course, the robot cannot directly access the human's internal state, but it can observe his or her actions.

For the sake of this exercise, humans have two hidden states: happy (`H`) and sad (`S`), which the robot will need to deduce from five actions: cooking (`CO`), crying (`CR`), sleeping (`SL`), socializing (`SO`), and watching TV (`W`).

From a series of experiments, the robot designers have determined that as people move from one task to another, sad people remain sad 80% of the time and happy people remain happy 90% of the time (the remainder of the time they switch from sad to happy or happy to sad). Initially, 60% of people are happy. That is, $P(x_0 = H) = 0.6$.

The designers have also determined that for happy and sad people, the activity breakdown is as follows:

	S	H
CO	10%	30%
CR	20%	0%
SL	40%	30%
SO	0%	30%
W	30%	10%

Let $\Theta = p(x_t|x_{t-1})$ and $\Phi = p(y_t|x_t)$ denote the transition and observation probability tables respectively. Implement a Python function that allows you to do inference (filtering) and prediction for a given Φ, Θ, $p(x_0)$ and $\mathbf{y}$ for T time steps.

Your function should look something like this:

```
def HMM(px0, theta, phi, y, T):

    ??? (setup) ???

    for t in xrange(T):
        if t < len(y):
            ???
        else:
            ???

    return px
```

(i) Filtering: Write down the probability tables for $p(x_t|x_{t-1})$ and $p(y_t|x_t)$.

(ii) Filtering: If at time $t-1$ we have $p(x_{t-1} = H|y_{1:t-1}) = 1$, what are the values of $p(x_t = H|y_{1:t-1})$ and $p(x_t = H|y_{1:t-1}, y_t = W)$? You can do this derivation by hand.

(i) Filtering:

For the sequence of observations $y_{1:5} = \{SO, SO, CO, W, SL\}$, what are the predictions of the user's internal state at each time step? In other words, what is the filtering distribution $p(x_t|y_{1:t})$ for $t = 1, 2, 3, 4$ and 5? You can generate these results using your program.

(ii) Prediction:

Given the above sequence of observations, what is the state prediction $p(x_8|y_{1:5})$?

(iii) Smoothing: Show that the smoothing distribution $p(x_t|y_{1:T})$, where $t \leq T$, can be written in terms of the filtering distribution $p(x_t|y_{1:t})$, state predictive distribution $p(x_t|y_{1:t})$ and the transition model $\theta = p(x_t|x_{t-1})$ as follows:

$$p(x_t|y_{1:T}) = p(x_t|y_{1:t}) \sum_{x_{t+1}} \frac{p(x_{t+1}|y_{1:T})p(x_{t+1}|x_t)}{p(x_{t+1}|y_{1:t})}$$

Hint: First marginalize x_{t+1} from $p(x_t, x_{t+1}|y_{1:T})$, apply conditional probability rules and then use Bayes rule.

This completes the well known **Forward-Backward algorithm** for HMMs. In the forward pass we compute all the filtering distributions. In the backward pass, starting with $p(x_T|y_{1:T})$, we compute all the smoothing distributions. The smoothing distributions are also known as the beliefs. The Forward-Backward algorithm is a special instance of the junction tree algorithm; a very popular one! It has played a central role in speech recognition.

2. **Entropy of a Bernoulli variable:**

The Bernoulli for $x \in \{0,1\}$ can be written as in class or in *exponential family form* as follows:

$$\mathrm{Ber}(x|\mu) = \mu^x(1-\mu)^{1-x} = \exp[x\log(\mu) + (1-x)\log(1-\mu)] = \exp[\boldsymbol{\phi}(x)^T\boldsymbol{\theta}] \tag{1}$$

where $\boldsymbol{\phi}(x) = [\mathbb{I}(x=0), \mathbb{I}(x=1)]$ and $\boldsymbol{\theta} = [\log(\mu), \log(1-\mu)]$. However, this representation is *over-complete* since there is a linear dependendence between the features:

$$\mathbf{1}^T\boldsymbol{\phi}(x) = \mathbb{I}(x=0) + \mathbb{I}(x=1) = 1 \tag{2}$$

Consequently $\boldsymbol{\theta}$ is not uniquely *identifiable*. It is common to require that the representation be *minimal*, which means there is a unique $\boldsymbol{\theta}$ associated with the distribution. In this case, we can just define

$$\mathrm{Ber}(x|\mu) = (1-\mu)\exp\left[x\log\left(\frac{\mu}{1-\mu}\right)\right]$$

Now we have $\phi(x) = x$, $\theta = \log\left(\frac{\mu}{1-\mu}\right)$, which is the *log-odds ratio* or *logit* function, and $Z = 1/(1-\mu)$.

(a) Show that we can recover the mean parameter μ from the canonical parameter using

$$\mu = \mathrm{sigm}(\theta) = \frac{1}{1+e^{-\theta}}$$

(b) Derive an expression for the entropy $H(\mu)$ of a Bernoulli variable. Plot this expression and explain the plot briefly.

(c) Prove that the derivative of the binary entropy can be expressed as the negative of the logit function. That is,

$$dH(\mu)/d\mu = -logit(\mu) = -\log\left(\frac{\mu}{1-\mu}\right).$$

3. **Maximum likelihood for the Poisson distribution**

   The Poisson pmf is defined as $\mathrm{Poi}(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$, for $x \in \{0, 1, 2, \ldots\}$ where $\lambda > 0$ is the rate parameter. Derive the maximum likelihood estimate (MLE).