10-701 Machine Learning, Fall 2011: Homework 3 Solutions

November 2, 2011

1 Hidden Markov Model [25 points, Bin]

1.1 General Questions

- [**4 points**] For each of the following data sets, is it appropriate to use HMM? Provide a one sentence explanation to your answer.

  - Stock market price data

    Answer: True. Stock market price is time sensitive.

  - Collaborative filtering on a database of movie reviews: for example, Netflix challenge: predict about how much someone is going to enjoy a movie based on their and other users' movie preferences

    Answer: False. User's preferences do not change much overtime.

  - Daily precipitation data in Pittsburgh

    Answer: True. Whether it rains or not depends largely on whether it rained yesterday or not (may not be true for Pittsburgh though).

  - Optical character recognition

    Answer: True. Word recognition is sensitive to the character sequence. "Rib" vs "Rob". (Note: if you only consider single character recognition and answer 'No', you will also get full credit.)

- [**2 points**] True or false: (if true, give a 1 sentence justification; if false, give a counter example.) When learning an HMM for a fixed set of observations, assume we do not know the true number of hidden states (which is often the case), we can always increase the training data likelihood by permitting more hidden states.

  Answer: True. For the worst case, we could give one hidden sate for each output value in the training sequence, and achieve perfect fitting. (Note: if you consider the scenario where the number of states is even larger than the number of observed output values, and answered 'False', you will also get full credit.)

- [**4 points**] Show that if any elements of the parameters π (start probability) or A (transition probability) for a hidden Markov model are initially set to zero, then those elements will remain zero in all subsequent updates of the EM algorithm.

  Answer: In the E step, since π and A are initialized to be zero, there wouldn't be any training example associated with the zero probability states, nor transition to any zero probability transitions. Hence, in the M step, the updated probabilities will remain zero.

1.2 HMM for DNA Sequence

In this problem, you will use HMM to decode a simple DNA sequence. It is well known that a DNA sequence is a series of components from $\{A, C, G, T\}$. Now let's assume there is one hidden variable S that controls the generation of DNA sequence. S takes 2 possible states $\{S_1, S_2\}$. Assume the following transition probabilities for HMM M

$$P(S_1|S_1) = 0.8,\ P(S_2|S_1) = 0.2,\ P(S_1|S_2) = 0.2,\ P(S_2|S_2) = 0.8$$

emission probabilities as following

$$P(A|S_1) = 0.4,\ P(C|S_1) = 0.1,\ P(G|S_1) = 0.4,\ P(T|S_1) = 0.1$$
$$P(A|S_2) = 0.1,\ P(C|S_2) = 0.4,\ P(G|S_2) = 0.1,\ P(T|S_2) = 0.4$$

and start probabilities as following

$$P(S_1) = 0.5,\ P(S_2) = 0.5$$

Assume the observed sequence is $x = CGTCAG$, calculate:

- **[5 points]** $P(x|M)$ using the forward algorithm. Show your work to get full credit.

  Answer:

  Initialization:

  $$\forall S_k \in \{S_1, S_2\}, \alpha_1^k = P(x_1|\pi_1 = S_k)P(\pi_1 = S_k) \tag{1}$$

  Iteration:

  $$\forall S_k \in \{S_1, S_2\}, t \in \{2, 3, 4, 5, 6\} \tag{2}$$
  $$\alpha_t^k = P(x_t|\pi_t = S_k) \sum_{i \in \{1,2\}} \alpha_{t-1}^i a_{i,k} \tag{3}$$

  Results:

α	$S_k = S_1$	$S_k = S_2$
α_1^k	0.05	0.2
α_2^k	0.032	0.017
α_3^k	0.0029	0.008
α_4^k	3.9200e-04	0.0028
α_5^k	3.4880e-04	2.3120e-04
α_6^k	1.3011e-04	2.5472e-05

- **[5 points]** The posterior probabilities $P(\pi_i = S_1|x, M)$ for $i = 1, \ldots, 6$. Show your work to get full credit.

  Answer:

  First, use backward algorithm to get β_t^k.

  Results:

β	$S_k = S_1$	$S_k = S_2$
β_1^k	7.9744e-04	5.7856e-04
β_2^k	0.0022	0.0051
β_3^k	0.0122	0.0150
β_4^k	0.1120	0.0400
β_5^k	0.3400	0.1600
β_6^k	1	1

Then, calculate $P(\pi_i = S_1|x, M)$ by

$$P(\pi_i = S_1|x, M) \quad = \quad \alpha_i^1 \beta_i^1 / \sum_{k \in \{1,2\}} \alpha_1^k \beta_1^k \tag{4}$$

Posterior:

	$S_k = S_1$
$P(\pi_1 = S_1|x, M)$	0.2563
$P(\pi_2 = S_1|x, M)$	0.4476
$P(\pi_3 = S_1|x, M)$	0.4476
$P(\pi_4 = S_1|x, M)$	0.2822
$P(\pi_5 = S_1|x, M)$	0.7622
$P(\pi_6 = S_1|x, M)$	0.8363

- **[5 points]** The most likely path of hidden states using the Viterbi algorithm. Show your work to get full credit.

Answer:
First, calculate V_t^k using Viterbi algorithm, and record the states that maximize the probability for each step $Ptr(k, t)$.
Results:

V_t^k	$S_k = S_1$	$S_k = S_2$
V_1^k	0.0500	0.2000
V_2^k	0.0160	0.0160
V_3^k	0.0013	0.0013
V_4^k	1.0240e-04	0.0016
V_5^k	1.3107e-04	1.3107e-04
V_6^k	4.1943e-05	1.0486e-05

$Ptr(k, t)$	$S_k = S_1$	$S_k = S_2$
$Ptr(k, 1)$	S_1	S_1
$Ptr(k, 2)$	S_1	S_2
$Ptr(k, 3)$	S_1	S_2
$Ptr(k, 4)$	S_1	S_2
$Ptr(k, 5)$	S_2	S_2
$Ptr(k, 6)$	S_1	S_2

The most likely path is $S_2, S_2, S_2, S_2, S_1, S_1$.

2 Bayesian Network

2.1 True or False

- (a) True. Let's say the variables are $X_1, \ldots, X_N$. Construct the full clique $X_1 \rightarrow X_i$ for any $i > 1$, $X_2 \rightarrow X_i$ for any $i > 2$, etc.. This network assumes no conditional independencies, therefore it can encode any probability distribution over $X_1, \ldots, X_N$. (Note: I will not deduct points on this question since I feel the statement is not clear enough.)

- (b) False. Some distributions in $\mathcal{D}$ can have additional independence assumptions not encoded in the network. For example, consider the 3-clique defined by $X_1 \rightarrow X_2 \rightarrow X_3$ and $X_1 \rightarrow X_3$, where all X are boolean. Let the conditional probability table of $P(X_3 \mid X_1, X_2)$ be

	$(X_1, X_2) = (0,0)$	$(X_1, X_2) = (0,1)$	$(X_1, X_2) = (1,0)$	$(X_1, X_2) = (1,1)$
$X_3 = 1$	0.7	0.7	0.4	0.4
$X_3 = 0$	0.3	0.3	0.6	0.6

  Observe that $P(X_3 \mid X_1, X_2 = 0) = P(X_3 \mid X_1, X_2 = 1)$. This proves that X_3 is conditionally independent of X_2 given X_1, even though we cannot derive this independence from the Bayes Net G.

2.2 Joint Probability

$$P(A, B, C, D, E, F, G) \quad = \quad P(A) P(B) P(G) P(C \mid A, B) P(E \mid C, G) P(D \mid C) P(F \mid D)$$

2.3 Number of Parameters

- (a) There are 7 variables, meaning that we need to encode the probabilities for $2^7 = 128$ possible settings of (A, B, C, D, E, F, G). This implies we need $128 - 1 = 127$ parameters.

- (b) We need 1 Bernoulli parameter for each of $P(A), P(B), P(G)$, 2 parameters for each of $P(D \mid C), P(F \mid D)$ (1 Bernoulli parameter for each of the 2 settings of the conditioning variables), and 4 parameters for each of $P(E \mid C, G) P(C \mid A, B)$ (1 Bernoulli parameter for each of the 4 settings of the conditioning variables). In total, we need $3(1) + 2(2) + 2(4) = 3 + 4 + 8 = 15$ parameters.

2.4 Markov Blanket

The Markov Blanket of C includes all immediate ancestors of C (namely A, B), all immediate descendants of C (namely D, E), and all "co-parents" of C (defined as any variable X that shares an immediate descendant with C, which includes just G). Hence the Markov Blanket of C is A, B, D, E, G.

2.5 D-Separation

- $A \perp B \mid C$: False. The only trail from A to B is $A \rightarrow C \leftarrow B$, and it is active because we condition on the inverted fork variable C.

- $A \perp G \mid E$: False. The only trail from A to G is $A \rightarrow C \rightarrow E \leftarrow G$, and it is active since (1) we do not condition on C, and (2) we condition on the inverted fork variable E.

- $B \perp G \mid C, E$: True. The only trail from B to G is $B \to C \to E \leftarrow G$, and it is d-separated since we condition on C.
- $F \perp G$: True. The only trail from F to G is $F \leftarrow D \leftarrow C \to E \leftarrow G$, and it is d-separated since we do not condition on the inverted fork variable E.

3 Conditional Random Fields [Suyash, 25 points]

3.1 CRFs and HMMs [6 points]

1. Type of model - CRF is a discriminative model, while HMM is generative.
2. Objective function optimized - CRF maximizes conditional probability $P(Y|X)$, while HMM maximizes likelihood.
3. Require a normalization constant - CRF requires a normalization constant, while HMM does not.

3.2 Features in CRFs [8 points]

Consider the standard CRF discussed in class (Lecture 12, slide 22). For each of the following feature functions, explain whether they can be represented by the CRF probability distribution? Briefly explain your answer.

1. $m_k = \mathbb{I}[y_i = y_{i+1}]$ - Yes, it could be written in the form of $f(y_{i+1}, y_i, \mathbf{x})$
2. $n_k = \mathbb{I}[\text{tag}(y_i) = \text{“Proper noun” AND } X_i \text{ is uppercase}]$ - Yes, it could be written in the form of $g(y_i, \mathbf{x})$
3. $o_k = \mathbb{I}[y_i = y_{i+2}]$ - No, standard CRF only permits one step correlation between states, i.e., (y_{i+1}, y_i).
4. $n_k = \mathbb{I}[\text{tag}(y_i) = \text{“Proper noun” AND } X_{i-1} \text{ is an article}]$ - Yes, it could be written in the form of $g(y_i, \mathbf{x})$

3.3 Complex CRFs [7 points]

1.

$$\begin{aligned} P(y|x) &= \frac{1}{z(x)} \exp\{\lambda\left[f(y_2, y_1, x) + f(y_3, y_2, x) + f(y_4, y_2, x) + f(y_5, y_3, x) + f(y_5, y_4, x)\right] \\ &\quad +\mu\left[g(y_1, x) + g(y_2, x) + g(y_3, x) + g(y_4, x) + g(y_5, x)\right]\} \end{aligned}$$

2.

$$\begin{aligned} z(x) &= \sum_{y_1, y_2, y_3, y_4, y_5} \exp\{\lambda\left[f(y_2, y_1, x) + f(y_3, y_2, x) + f(y_4, y_2, x) + f(y_5, y_3, x) + f(y_5, y_4, x)\right] \\ &\quad +\mu\left[g(y_1, x) + g(y_2, x) + g(y_3, x) + g(y_4, x) + g(y_5, x)\right]\} \end{aligned}$$

3.4 Computing the normalization constant [4 points]

1. 2^5

2. 2^n

3. The computational time is exponential with the length of x

4 Gibbs sampling for an infinite gaussian mixture model

4.1 Uniform discrete prior for z

1.

$$\begin{aligned}
p\left(z_i = k \mid \mathbf{x}, \boldsymbol{\mu}, \mathbf{z} \setminus \{z_i\}\right) &\propto p\left(x_i \mid z_i = k, \boldsymbol{\mu}\right) p\left(z_i = k\right) \\
&\propto (2\pi)^{-D/2} \exp\left(-\frac{1}{2}\left\|x_i - \mu_k\right\|_2^2\right) \frac{1}{K} \\
&\propto \exp\left(-\frac{1}{2}\left\|x_i - \mu_k\right\|_2^2\right)
\end{aligned}$$

2.

$$\begin{aligned}
p\left(\mu_k = u \mid \mathbf{x}, \mathbf{z}, \boldsymbol{\mu} \setminus \{\mu_k\}\right) &\propto \left[\prod_{i=1}^{N} p\left(x_i \mid z_i, \mu_k = u, \boldsymbol{\mu} \setminus \{\mu_k\}\right)^{\delta(z_i = k)}\right] p\left(\mu_k = u\right) \\
&\propto \left[\prod_{i=1}^{N} (2\pi)^{-D/2} \exp\left(-\frac{1}{2}\left\|x_i - u\right\|_2^2\right)^{\delta(z_i = k)}\right] (2\pi)^{-D/2} \exp\left\{-\frac{1}{2}\left\|u\right\|_2^2\right\}, \\
&\propto \left[\prod_{i=1}^{N} \exp\left(-\frac{1}{2}\left\|x_i - u\right\|_2^2\right)^{\delta(z_i = k)}\right] \exp\left\{-\frac{1}{2}\left\|u\right\|_2^2\right\}, \\
&\propto \exp\left(-0.5\left[\left\|u\right\|_2^2 + \sum_{i=1}^{N} \delta\left(z_i = k\right)\left\|x_i - u\right\|_2^2\right]\right)
\end{aligned}$$

4.2 An infinite prior over z

1.

$$\begin{aligned}
p\left(z_i = k \mid \mathbf{x}, \boldsymbol{\mu}, \mathbf{z} \setminus \{z_i\}\right) &\propto p\left(x_i \mid z_i = k, \boldsymbol{\mu}\right) p\left(z_i = k \mid \mathbf{z} \setminus \{z_i\}\right) \quad \text{if } \#\left[\mathbf{z} \setminus \{z_i\} = k\right] > 0 \\
&\propto p\left(x_i \mid z_i = k, \boldsymbol{\mu}\right) p\left(z_i = k \mid \mathbf{z} \setminus \{z_i\}\right) \\
&\propto (2\pi)^{-D/2} \exp\left\{-\frac{1}{2}\left\|x_i - \mu_k\right\|_2^2\right\} \frac{\#\left[\mathbf{z} \setminus \{z_i\} = k\right]}{N + \alpha} \\
&\propto \exp\left\{-\frac{1}{2}\left\|x_i - \mu_k\right\|_2^2\right\} \left(\#\left[\mathbf{z} \setminus \{z_i\} = k\right]\right)
\end{aligned}$$

2.

$$\begin{aligned}
p\left(z_i = k \mid \mathbf{x}, \boldsymbol{\mu}, \mathbf{z} \setminus \{z_i\}\right) &\propto p\left(x_i \mid z_i = k, \boldsymbol{\mu}\right) p\left(z_i = k \mid \mathbf{z} \setminus \{z_i\}\right) \\
&\qquad k \text{ is the smallest positive integer such that } \#\left[\mathbf{z} \setminus \{z_i\} = k\right] = 0 \\
&\propto p\left(x_i \mid z_i = k, \boldsymbol{\mu}\right) p\left(z_i = k \mid \mathbf{z} \setminus \{z_i\}\right) \\
&\propto (2\pi)^{-D/2} \exp\left\{-\frac{1}{2}\left\|x_i - \mu_k\right\|_2^2\right\} \frac{\alpha}{N+\alpha} \\
&\propto \exp\left\{-\frac{1}{2}\left\|x_i - \mu_k\right\|_2^2\right\}
\end{aligned}$$

4.3 A few subtleties

1.

$$\begin{aligned}
p\left(z_i = k \mid \mathbf{x}, \boldsymbol{\mu}, \mathbf{z} \setminus \{z_i\}\right) &\propto \left[\int_u p\left(x_i \mid z_i = k, \mu_k = u, \boldsymbol{\mu} \setminus \{\mu_k\}\right) p\left(\mu_k = u\right)\, du\right] p\left(z_i = k \mid \mathbf{z} \setminus \{z_i\}\right) \\
&\qquad k \text{ is the smallest positive integer such that } \#\left[\mathbf{z} \setminus \{z_i\} = k\right] = 0. \\
&\propto \left[\int_u p\left(x_i \mid z_i = k, \mu_k = u, \boldsymbol{\mu} \setminus \{\mu_k\}\right) p\left(\mu_k = u\right)\, du\right] p\left(z_i = k \mid \mathbf{z} \setminus \{z_i\}\right) \\
&\propto \left[\int_u (2\pi)^{-D/2} \exp\left\{-\frac{1}{2}\left\|x_i - u\right\|_2^2\right\} (2\pi)^{-D/2} \exp\left\{-\frac{1}{2}\left\|u\right\|_2^2\right\} du\right] \frac{\alpha}{N+\alpha} \\
&\propto \left[\int_u \exp\left\{-\frac{1}{2}\left\|x_i - u\right\|_2^2 - \frac{1}{2}\left\|u\right\|_2^2\right\} du\right]
\end{aligned}$$

2. The likelihood terms in both cases (except when we have to add a new Gaussian in the infinte prior) are the same. In the K-Gaussians case, the prior over z gives each gaussian uniform weight but in the infinite prior, each Gaussian is weighted by how many data points are assigned to it.

3. With an infinite uniform prior, we would have an infinite number of potential z values with identical non-zero probability mass assigned to them. This would be impossible to do under the restriction that the total probability mass for all z values should be equal to 1.