Announcements

- Additional office hours on
- Final exam
  - Format of the final will be similar to the midterm
  - Expectation is that exam will take between 2-2.5 hrs but you'll have 3 hours
  - Additional 10 minutes before exam begins to study the question paper and plan out strategy of solving
  - Broad topics covered
    - Bayesian Networks
    - Hidden Markov Models
    - Factor Graphs
    - Neural Networks
    - SVM, Random Forests
  - One bonus question
  - There will be questions from what we have done before midterm

Hidden Markov Model – Occasionally cheating casino

In a hypothetical dishonest casino, the casino uses a **fair die** most of the time, but occasionally the casino secretly switches to a **loaded die**, and later switches back to the fair die. A probabilistic process determines the switching back-and-forth from loaded to fair die and *vice versa*, the transition matrix for which is given as follows.

$$\begin{array}{cc} & \begin{array}{cc} F & L \end{array} \\ \begin{array}{c} F \\ L \end{array} & \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix} \end{array} \qquad \pi = \begin{bmatrix} 0.75 & 0.25 \end{bmatrix}$$

Assume that the loaded die will come up “six” with probability 0.5 and the remaining five numbers with probability 0.1 each. Therefore, the observation matrix is:

$$\begin{array}{cc} & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{array}{c} F \\ L \end{array} & \begin{bmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/10 & 1/10 & 1/10 & 1/10 & 1/10 & 1/2 \end{bmatrix} \end{array}$$

1/10

The casino hides the die being rolled and you only observe the sequence of rolls. Find the most likely die for each roll given the observed sequence of numbers is: **2, 3, 5, 6, 6, 1, 5, 6, 4**

HMM Solution

Forward algorithm

1. Input: (A, B, π) and observed sequence $E_1, \ldots, E_n$
2. $[\alpha_1, Z_1]$ = normalize$(b_1 \odot \pi)$
3. **for** $t = 2{:}n$ **do**
   $[\alpha_t, Z_t]$ = normalize$(b_t \odot (A^T \alpha_{t-1}))$
4. return $\alpha_1, \ldots, \alpha_n$ and $\log\big(P(E_1, \ldots, E_n)\big) = \sum_t \log(Z_t)$

5. Subroutine: [v, Z] = normalize(u): $Z = \sum_j u_j$; $v_j = u_j / Z$;

NOTE: $\odot$ represents elementwise product (Hadamard product)

Backward algorithm

1. Input: (A, B, π) and observed sequence $E_1, \ldots, E_n$
2. $\beta_n = 1$; // initialize $\beta_n(j)$ to 1 for all states σ_j
3. **for** $t = n - 1{:}1$ **do**
   $\beta_{t-1} = A(b_t \odot \beta_t)$
4. return $\beta_1, \ldots, \beta_n$

t_0 — t $\quad$ $t+1$ $\longrightarrow$

$\alpha_1 = b_1 \odot [\pi]$

$\begin{bmatrix} 1/6 \\ 1/10 \end{bmatrix} \odot \begin{bmatrix} 0.75 \; 3/4 \\ 0.25 \; 1/4 \end{bmatrix} \Rightarrow \begin{bmatrix} 1/8 \\ 1/40 \end{bmatrix}$

α_1

Calculating β_8

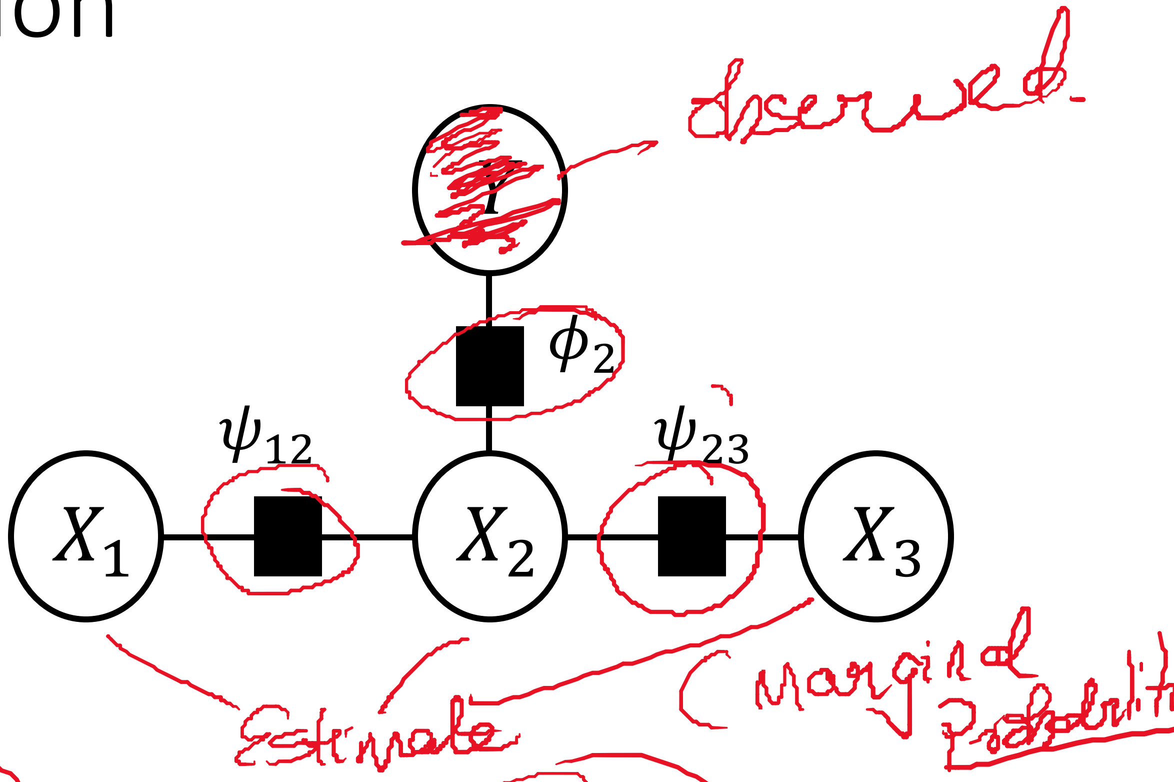
$$A\left(b_9 \odot \beta_9\right)$$

$$\begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}\left[\begin{pmatrix} 1/6 \\ 1/10 \end{pmatrix} \odot \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right] = \begin{bmatrix} 0.17 \\ ? \end{bmatrix}$$

Factor Graphs Belief Propagation

Y is observed to be 0.

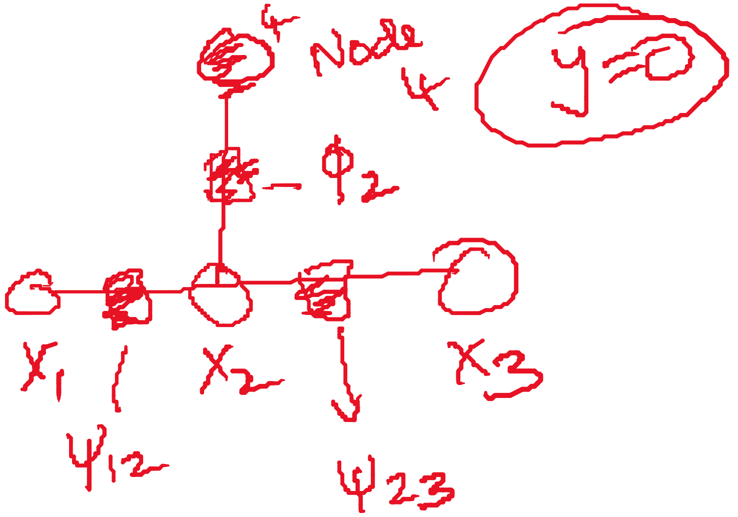
Calculate
$P(X_2|Y=0), P(X_1|Y=0),$
$P(X_3|Y=0)$



X_1	X_2	ψ_{12}
0	0	1
0	1	0.9
1	0	0.9
1	1	1

X_2	X_3	ψ_{23}
0	0	0.1
0	1	1
1	0	1
1	1	0.1

X_2	Y	ϕ_2
0	0	1
0	1	0.1
1	0	0.1
1	1	1



$P(X_1)$, $P(X_2)$ $(P(X_3)$

$$m_{12}(x_2) = \sum_{x_1} \psi_{12}(x_1, x_2)$$

$$= \begin{bmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{bmatrix} = \begin{bmatrix} 1.9 \\ 1.9 \end{bmatrix} = k \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Marginlize(x_1)

$$P(X_2) = \hat{k} \; \underline{m_{12}(x_2)} \cdot m_{42}(x_2) \; m_{32}(x_2)$$

$$\hat{k} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1.0 \\ 0.1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\frac{1.0}{1.1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$\hat{k}$

$\boxed{\text{Calc } P(X_1) \; ; P(X_3)}$

Gradient Descent

$L(z) = \|y - w^T x\|^2$

Given N training data points $\{(\boldsymbol{x}^{\boldsymbol{k}}, y^k)\}$ for $k = \{1, \ldots, N\}$, $\boldsymbol{x}^k \in R^d$, and labels $y^k \in \{-1,1\}$, we seek a linear discriminant function $f(\boldsymbol{x}) = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}$ optimizing the loss function $L(z) = e^{-z}$ for $z = yf(\boldsymbol{x})$.

1. Find the gradient and gradient descent update equation to find $\boldsymbol{w}$.

2. Suppose you also want to include a penalty term $\lambda||\boldsymbol{w}||^2$ to the overall loss function. Derive the gradient for gradient descent to update $\boldsymbol{w}$.

$L(z)$

$$L(z) = \sum_{i=1}^{N} e^{-z_i} = \sum_{i=1}^{N} e^{-y^i w^T x^i}$$

$$w^T x^i = w_1 x_1^i + \cdots$$

$$w = [w_1, w_2, \ldots, w_d]$$

$$\nabla_w L = \begin{bmatrix} \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_d} \end{bmatrix}$$

$$\frac{\partial L}{\partial w_1} = \sum_{i=1}^{N} \exp(-y^i w^T x^i)(-y^i x_1^i)$$

$$\frac{\partial L}{\partial w_k} = \sum_{i=1}^{N} \exp(-y^i w^T x^i)(-y^i x_k^i)$$

$$\nabla_w L = \begin{bmatrix} \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_d} \end{bmatrix} = \sum_{i=1}^{N} \exp(-y^i w^T x^i) \cdot$$

$$\underbrace{\nabla_w L}_{\text{gradient}} = \sum_{i=1}^{N} \exp(-y^i w^T x^i)(-y^i) \begin{bmatrix} x_1^i \\ x_2^i \\ \vdots \\ x_d^i \end{bmatrix} = -\sum_{i=1}^{N} \exp(-y^i w^T x^i)\, y^i x^i$$

$$w^{t+1} = w^t - \eta \nabla_w L$$

$$= w^t + \eta \sum_{i=1}^{N} \exp(-y^i w^T x^i)\, y^i x^i$$

$$L = \underbrace{\sum_{i=1}^{N} \exp(-y^i w^T x^i)} + \underbrace{\lambda \|w\|^2}$$

gradient

$$\nabla_w L = -\sum_{i=1}^{N} \exp(-y^i w^T x^i)\, y^i x^i + 2\lambda w$$

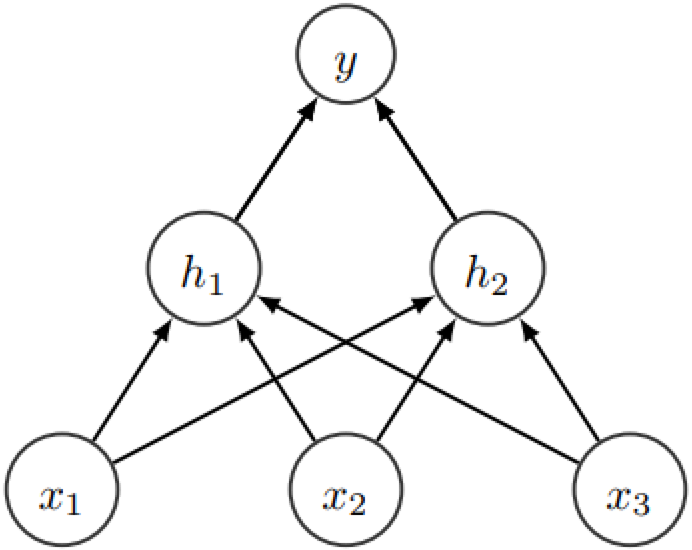
gradient descent update

$$w^{t+1} = w^t - \eta\left(-\sum_{i=1}^{N} \exp(-y^i (w^t)^T x^i)\, y^i x^i + 2\lambda w^t\right)$$

$$= (1 - 2\eta\lambda) w^t + \eta \sum_{i=1}^{N} \exp(-y^i ((w^t)^T x^i))\, y^i x^i$$

Neural Networks Backpropagation

Consider the neural network given alongside. The hidden units and output layer has ReLU activation function. The loss function is given by $L(y, y) = \frac{1}{2}(y - t)^2$ where t is the target value. For simplicity, assume that the bias terms are 0. Weights connecting input to hidden layer and hidden layer to output layer are given by W and V respectively.



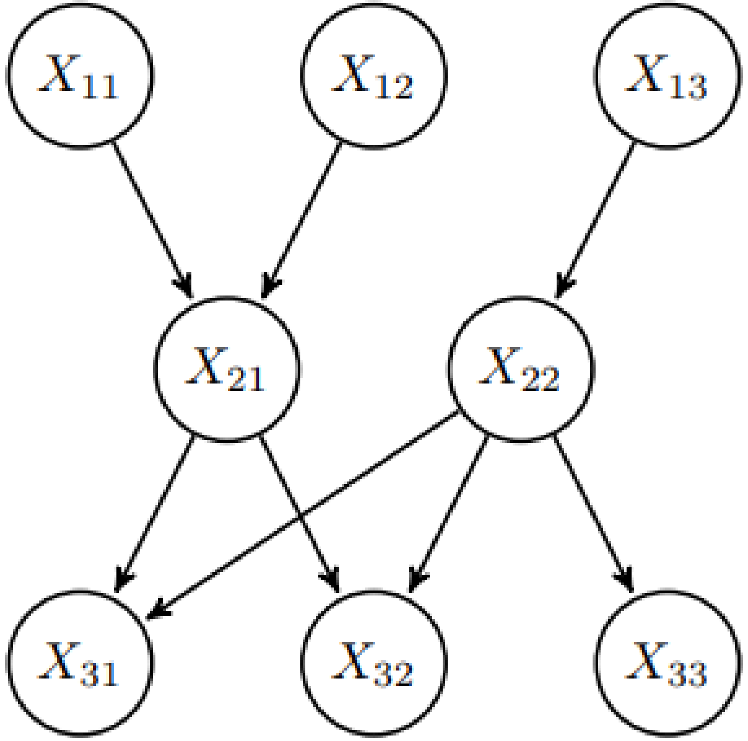
1. Write the forward equation to map input to output.
2. Compute the output and backpropagation for $x = [1,2,1]$ and $t = 1$.

$$W = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}$$

$$V = \begin{bmatrix} 0 & 1 \end{bmatrix}$$

Bayesian Network Question

Consider the Bayesian Network alongside with binary variables. Answer the following questions.



1. Is there any variable(s) conditionally independent of X_{33} given X_{11} and X_{12}? If so, list all.

2. Is there any variable(s) conditionally independent of X_{33} given X_{22}? If so, list all.

3. How many parameters are required to specify the factorized joint distribution?

4. Express $P(X_{13} = 0, X_{22} = 1, X_{33} = 0)$ in terms of the conditional probabilities from the Bayesian Network.