COMP 652: MACHINE LEARNING

ASSIGNMENT 3

November 10, 2018

Student ID: 260800927
Samin Yeasar Arnob
McGill University

1

Answer to the Question No. 1 (a)

- At each round, t we select action, K and observe a reward. This can be changed into linear setting. Added noise $\mathcal{N}(0, \sigma^2)$ can be shown to be sub-gaussian.

So we select an action from action space x, where x is 1-of-K hot encoding vectors. We parameterize with θ which is to be learned & is the same dimension-K

So received reward, $r_t = \langle X_t, \theta_t \rangle + \varepsilon_t$

So at each time step, t the reward will be of gaussian with mean θ_t & variance of ε_t.

So, action space is 1 of K hot encoding.

Now $E[r_t] = E\left[\theta_t^T X_{K_t} + \varepsilon\right]$

$\Rightarrow \mu_{K_t} \approx \theta_*^T X_{K_t}$ --- (a)

Answer to the question No 1 (b)

R_T =

In linear case, we are supposed to minimize

$$R_T(\pi, \theta_*) = \sum_{t=1}^{T} \theta_*^T x_* - \sum_{t=1}^{T} \theta_*^T x$$

$$= \sum_{t=1}^{T} \theta_*^T x_* - \sum_{t=1}^{T} \mu_{x_t} \quad \text{[from (a)]}$$

Answer to the question No 1 (c)

UCB implements the optimism in the face of uncertainty principle, according to which one should choose the action as if the environment was as nice a plausible possible. In finite-action stochastic bandit problem the principle dictates to choose the action with largest upper confidence.

(2)

In the case of Linear bandit problems this still holds, but now to calculate the upper confidence bounds one should also ~~better~~ take account the information conveyed by all reward observed because all the data $(X_1, R_1 \ldots X_{t-1}, R_{t-1})$ [R_t = reward] is now connected through the unknown parameter vector.

One idea is to construct confidence set C_t based on $(X_1, R_1 \ldots X_{t-1}, R_{t-1})$ that contain unknown parameter θ_* with high probability. Assuming confidence set indeed contain θ_* for any action K

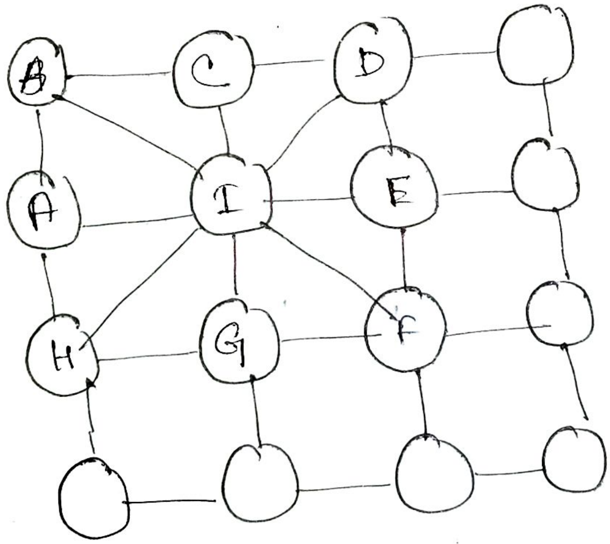
$$UCB_{Kt}(X) = \max_{\theta \in C_t} \langle X, \theta \rangle$$

will be upper bound of mean payoff of K. The UCB algorithm that uses the confidence set C_t at time t then selects

$$K_t = \arg\max_{k} UCB_{Kt}(X)$$

Answer to the Question No 2(a)

Connecting pixels in a 8 neighbourhood instead of 4 will give following diagram:



since pixel I is connected to 8 neighbours, the parameters of this undirected graph will be 8 clique potentials, This is given as

ψ_{ABI}, ψ_{BCI}, ψ_{CDI}, ψ_{DEI}, ψ_{EFI}, ψ_{FGI}, ψ_{GHI}

ψ_{HAI}

Answer to the Question No 2(b)

The possible advantages would be each pixels will have more information since it is connected to eight neighbours instead of 4. Hence each pixels would be dependent on the values of eight pixels. This would allows for information capture and hence lead to better denoising. Also another advantage is the computational efficiency. Since now we have cliques over the undirected graphs, where we can now present the 3 connected nodes with a single clique instead of a single clique potential for a pair of nodes as was the case when it was connected with a 4-neighbour.

2

The disadvantage of 8 neighbours would possiblely be that belief propagation algorithm will be more complex since we are connecting in a 8 neighbourhood model. Also, since the model become large, sampling might not be able to calculate the local information accurately.

Answer to the question No 2 (c)

For a 2D ising model the clique potential between a pair of variables can be written as $\varphi_{i,j}(x_i, x_j) = \begin{bmatrix} e^{w} & e^{-w} \\ e^{-w} & e^{w} \end{bmatrix}$

$w > 0$: ferro magnet

$w < 0$: anti ferro magnet (frustrated system)

$$P(x|\theta) = \frac{1}{Z(\theta)} e^{[-\beta H(x|\theta)]}$$

$$H(x) = -x^T W x = -\sum_{ij} W_{ij} x_i x_j$$

Suppose we have a local evidence in the left most side $P(y|x)$ which is injected from the left of the model

$$P(x,y) = P(x) * P(y|x)$$

$$= \left[\frac{1}{Z} \prod_{i,j} \psi_{ij}(x_i, x_j)\right] \left[\prod_i P(y_i|x_i)\right]$$

Let the local evidence be normally distributed by the following

$$P(y_i|x_i) = N(y_i | x_i, \sigma^2)$$

A way to draw sample given the evidence would be

1) $x_1^{s+1} \sim P(x_1 | x_2^s - x_D^s)$

2) $x_2^{s+1} \sim P(x_2 | x_1^{s+1}, x_3 - - x_D^s)$

3) $x_i^{s+1} \sim p(x_i | x_{1:i-1}^{s+1}, x_{i+1:D}^s)$

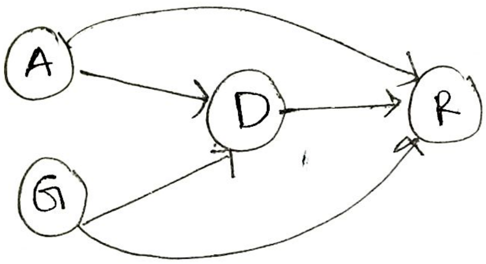
4) $x_D^{s+1} \sim P(x_D | x_1^{s+1} - - x_{D-1}^{s+1})$

1

Answer to the Question 3 (a)

~~P(A,R~~

$$P(A,G,D,R) = P(A)\,P(G)\,P(D \mid A,G)\,P(R \mid D,A,G)$$



$$P(R=1 \mid D=1) = \frac{\sum_{A,G} P(A,G,D=1,R=1)}{\sum_{A,G,R} P(A,G,D=1,R)}$$

$$= \frac{\sum_{A,G} P(A)\,P(G)\,P(D=1 \mid A,G)\,P(R=1 \mid D=1,A,G)}{\sum_{A,G,R} P(A)\,P(G)\,P(D=1 \mid A,G)\,P(R \mid D=1,A,G)}$$

$$P(R=1 \mid D=1, A=0) = \frac{\sum_{G} P(A=0,G,R=1,D=1)}{\sum_{R,G} P(A=0,G,R,D=1)}$$

$$= \frac{\sum_{G} P(A=0)\,P(G)\,P(D=1 \mid A=0,G)\,P(R=1 \mid A=0,G,D=1)}{\sum_{R,G} P(A=0)\,P(G)\,P(D=1 \mid A=0,G)\,P(R \mid A=0,G,D=1)}$$

Answer to the quetion NO 3(b)

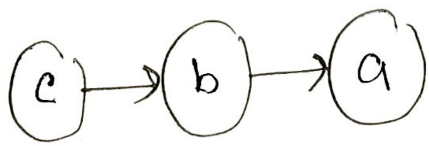
If we make no assumption of joint probability of $P(a,b,c)$, then we have 2 choices for each variable. So we have $2^3=8$ parameters. Although 8^{th} parameter can be known from $(1-\sum 7$ other probability$)$.
So we need 7 parameter to get $P(a,b,c)$.

For joint distribution :

$$P(a,b,c) = P(a|b,c)\,P(b|c)\,P(c)$$
$$= P(a|b)\,P(b|c)\,P(c) \quad \text{(given)}$$

thus a & c are independent.

graphical model would be



$P(c) \sim$ ~~binomial~~ binomial (P)
$P(b|c) \sim$ Beta (α_1, β_1)
$P(a|b) \sim$ Beta (α_2, β_2)

So we need five parameters..

Answer to the question No 4 (a)

we will discretize the given continues feature space.

As nothing mentioned about the complexity of the model, I will at first try simplest model for the task; "Logistic regression" with L2 regularization. And learn parameters using gradient descent.

Answer to the question No 4 (b)

As in this problem by feature space has become too large compare to training data we need to do non-parametric regression & we can use kernel trick: Linear kernel, to do the classification is much more simpler domain. As the in.

Answer to the question No 4(c)

For stock prediction I will ~~fit~~ gaussian process with different kernels (RBF, quadratic) as it will give a confidence bound over next 3 days stock. & I can get a range over which stock may vary.

Answer to the question NO 4(d)

We can use likelihood function to solve this. We have a prior of failure $P(w) = 0.5$ as we considered random failure of machine in the question. After given the variant measurement, D of machine. We will update our posterier $P(w|D) = \frac{P(D|w)\,P(w)}{P(D)}$. Once we learn $P(w|D)$ we can use likelihood $L(w)$ to know how likely it is to require maintanance.