INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

End-Autumn Semester Examination 2023-24

Date of Examination:________________Session: (FN/AN)____Duration: 3 Hrs Full Marks:_100_

Subject No. : AI61004 Subject : Statistical Foundations of AI/ML

Department/Center/School: Centre of Excellence in Artificial Intelligence

Specific charts, graph paper, log book etc., required ________________________________

Special Instructions (if any) : ____

Q1.

A) You are given a 6-faced dice which you initially believe to be fully fair. You carry out 2 rounds of 10 throws each, where the outcomes (frequency of faces) are as below:

	#1	#2	#3	#4	#5	#6
Round 1	3	2	0	0	2	3
Round 2	2	1	1	0	4	2

Using suitable probability distributions, show how your belief about the fairness of the dice gets updated after each round. **[10 marks]**

Solution sketch: This is actually a problem about Bayesian parameter estimate. The observation in each round follows Multinomial distribution. The conjugate prior of Multinomial is Dirichlet distribution, as we saw in Module 2 (Naïve Bayes). As our initial belief is that the dice is fair, the Dirichlet prior has all parameters equal to α. After first round, the Dirichlet parameters change to (n_{11}+ α, n_{12}+ α, …. n_{16}+ α) where n_{11} = 3, n_{12} = 2 etc. After second round, they change to (n_{21}+n_{11}+ α, n_{22}+n_{12}+ α, …. n_{26}+n_{16}+ α). These derivations of Dirichlet posterior need to be shown.

Many students have considered this as a hypothesis testing problem. That does not show how belief gets updated, it only shows if belief (null hypothesis) is rejected or not. I have given upto 5 marks for this approach (depending on how much calculation has been done). Also, some students have used only the Multinomial without Dirichlet prior. In such cases, 5-6 marks have been given.

B) Given the dataset below, check which fits better – a Gaussian distribution, or a Gamma distribution. Use the Method of Moments to estimate the necessary parameters. **[10 marks]**

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
23	14	7	3	9	39	10	1	1	4
D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
5	5	14	7	4	3	9	3	15	3

Solution Sketch: This is straightforward. For both cases, the first two moments have to be calculated empirically from data and compared with their theoretical expressions. By solving equations we can get the values of both parameters for Gaussian and Gamma. Using these values, the joint density of the observations has to be calculated using both Gamma and Gaussian distributions, and the results are compared. There is no need to derive the theoretical moments. Complete calculations will give 10 marks, else marks will be according to the steps done.

Q2.

A) Prove that the samples obtained by rejection sampling actually follow the target distribution **[4 marks]**

Solution Sketch: this has already been discussed in class. However, for full marks it is necessary to generalize to the case where the variable's support is unbounded, i.e. we should use wrapper distribution h(x) instead of U(a,b). 1 mark has been deducted for considering x bounded in (a, b).

B) Explain with a small example why samples drawn using Metropolis Algorithm (not Metropolis-Hastings) may not reflect the target distribution. **[4 marks]**

Solution Sketch: The answer lies in the fact that in Metropolis algorithm, x' is always accepted if f(x')>f(x) where x' is the new candidate sample and x the previous accepted sample. So, there is a greater density of accepted samples as we move towards a region where f(x) is high. You can use examples like multimodal distribution. A theoretical answer based on pdf values is also fine.

C) Consider the following function: f(x) = a*N(x,-5,25) + b*Gamma(x,10,3) + c*Beta(x,10,10) where x is defined over (-ꝏ, ꝏ). Consider that Gamma PDF is 0 in (-ꝏ, 0) and Beta PDF is 0 outside of (0,1). For what values of (a,b,c) is this a valid PDF, and what is the mode of this PDF? **[2 marks]**

Solution Sketch: a+b+c=1. Mode can be calculated either analytically (difficult) or found intuitively by sketching the PDFs for any choice of a, b, c. But writing a specific config like (a=1, b=0, c=0) as the only answer will result in loss of 1 mark.

D) I want to draw samples from the above PDF (choose suitable values of a, b, c) using Metropolis-Hastings Algorithm. The proposal distribution is N(X(t), b), where X(t) is the last accepted sample. Illustrate two situations where i) having small value of b is useful, and ii) having large value of b is useful. **[6 marks]**

Solution Sketch: small 's' means more exploit less explore, which is useful if the domain is very narrow, like c=1, a=b=0, i.e. domain is (0,1) only and our initial point is also in that range. Low 's' ensures that few samples are rejected. But if pdf is multimodal (eg. a=b=c=1/3) with narrow peaks and/or the initial point is in a low-density region, then large 's' is useful to break into how-density regions. Full marks only for full illustration. 3 marks for brief explanation of the idea.

E) Provide an algorithm to draw samples from Gamma(a, 0.5) distribution (a is an integer) using Box-Muller transform [Hint: remember chi-square distribution]**[4 marks]**

Solution Sketch: Gamma(a, 0.5) = chi-square(2a) = distribution of the sum of squares of '2a' variables, each of which follows N(0,1). So generate '2a' such observations from N(0,1) using Box-Muller, square them and sum them.

Q3.

A) Consider a Hidden Markov Model with 3 possible latent state values, with Gaussian emission distribution with parameters (-3, 9), (0, 4), (3, 9). The state transition parameters follow the relation: p(Z(t)=j|Z(t-1)=i) α 1/(|i-j|+1). All states are equally likely at t=1. You are provided with a short sequence of observations X=(0.8, -2, 1.5). Find the probability distribution of Z(3), using the observations and parameters. **[8 marks]**

Solution sketch: This is based on the principle of forward-backward algorithm. First the transition matrix has to be constructed according to the given rule (2 marks). Emission distribution is given. It needs to be remembered that p(Z(3)|X) depends on Z(2) and Z(1) which have to be marginalized, but the observed emissions X(1), X(2), X(3) need to be accounted for in this marginalization process. Partial marks (upto 5) for failing to show these steps.

B) Now consider an Order-2 HMM where Z(t) depends on both Z(t-1) and Z(t-2), as p(Z(t)=k|Z(t-1)=j,Z(t-2)=i) α 1/(|k-i|+|k-j|1). All states are equally likely at t=1 and t=2. Consider an observation sequence X of length T. You want to estimate the possible values of the corresponding latent states using Gibbs Sampling. Provide the full algorithm, including the sampling steps for all variables, and how you will estimate the marginal distributions of each latent variable from the collected samples. **[8 marks]**

Solution sketch: Gibbs Sampling proceeds by initializing the latent variables, and then updating them one at a time (conditioned on the current values of the rest). We need p(Z(t)|Z(-t),X) which is proportional to p(Z(t), Z(-t), X) which factorizes as p(Z(3)|Z(2),Z(1))*p(Z(4)|Z(3),Z(2))….. *p(Z(T)|Z(T-1),Z(T-2)) * p(X(1)|Z(1))*….p(X(T)|Z(T)). But since we have fixed all variables except Z(t) for now, we need to consider only those terms which involve Z(t), the rest are constant. So p(Z(t)) α p(Z(t)|Z(t-1),Z(t-2))*p(Z(t+1)|Z(t),Z(t-1))*p(Z(t+2)|Z(t+1),Z(t))*p(X(t)|Z(t)). These terms can be calculated for different values of Z(t) (1,2,3,4) using the rules mentioned above. Finally, the samples will be collected at regular intervals of about 10 iterations, after the burn-in phase of 50 iterations. We will finally have many samples of each latent variable (Z(1), Z(2) ….. Z(T)), from which we will estimate the marginal distribution of each of them.

Note that most students have missed the fact that Z(t) should be conditioned not only on Z(t-1) and Z(t-2) but also on Z(t+1) and Z(t+2). We had discussed a similar problem in class regarding Gibbs Sampling in HMM.

C) Now suppose that the mean parameters of the state-specific emission distributions are not known, but they are treated as IID latent variables (m1, m2, m3) following N(0, 50). How will the Gibbs Sampling algorithm now change? **[4 marks]**

Solution sketch: So now you have new latent variables (m1, m2, m3) for Gibbs Sampling. So you bring them into the sampling process. Each of them are first initialized along with the Z variables. Then they are also updated by sampling in each iteration, as p(m1 | m2,m3,m4,Z(1),….Z(T),X). Once again, this is proportional to the product of those terms which involve m1, while the rest are temporarily constant. This includes those observations for which Z(t) is currently equal to 1, will involve m1 as N(X(t)|m1,s1), along with the prior N(m1|0,S). So m1 is sampled in this way, from this product of Gaussians (which is also Gaussian). Another option is to marginalize over m1, m2, m3 but the math is more complicated to derive.

Q4. Consider the following model: $Y_i \sim Ber(p)$, $Zi \sim N(0,1)$, $Xi \sim N(W_kZ_i, s)$ where $k= Y_{i.}$ We have 3 observations of X and Y: [X1=3.6 Y1=1, X2=-1.2 Y2=0, X3=4.2 Y3=1]
Consider the following 2 choices of parameters:
P1: [W0=-5, W1=5, s=9, p=0.5], P2: [W0=0, W1=4, s=25, p=0.8]
Which of these two sets of parameters fits the observations best? Note that we do not observe Z. **[8 marks]**

Solution Sketch: This is similar, but not same as the problem that was present in previous semester's paper. Here Z is unobserved, not Y. Since Y is binary, it can be marginalized over by just adding the two cases (Y=0 and Y=1). But since Z is real (following Gaussian), here it has to be marginalized by integration for each of the three observations separately. For this integration over the product of two Gaussian PDFs $N(Xj\ ;\ W_kZ_i,\ s)*N(Z_i;\ 0,1)$, we can use a formula that was discussed in Module 2 (Bayesian linear regression). Most students have simply ignored the Zs while writing the PDF. This results in loss of 4 marks. Some other students have considered some hypothetical values of Z, or that Z is binary like Y. In such cases too, 3-4 marks have been deducted. Those who have tried to marginalize, I have given at least 7 marks even if the calculation is not complete.

A) Explain the E-M algorithm to estimate the parameters of a generic latent variable model p(Z,X), where Z is the set of latent variables and X is the set of observed variables. Explain why it is guaranteed to converge to at least a local minima. **[4+8=12 marks]**

Solution Sketch: This is self-explanatory. But despite mentioning "generic latent variable model", many students have just written the parameter update equations for Gaussian Mixture Model. In that case, I have given 2 marks out of the first 4.

Q5 A) Provide an algorithm based on importance sampling to carry out the integration of $(\sin(x)/(\log(x)+1))*x^5*(1-x)^4$ over the interval (0,1). If your proposal distribution is N(0,1), which samples will have high "importance scores"?. **[6 marks]**

Solution Sketch: Importance Sampling gives us ∫f(x)p(x)dx. Importance sampling is different from plain Monte Carlo integration, where you draw samples over the range from p(x) and evaluate the integrand f(x) at each of them. It is important to observe here that we do importance sampling when p(x) is difficult to draw samples from. We choose another proposal distribution q(x) from which we draw the samples, and then we calculate f(x)*p(x)/q(x) at each sample. In this case, it was mentioned that q(x) = N(0,1). Most students failed to recognize that p(x) is Beta(6,5), whose pdf is proportional to $x^5*(1-x)^4$. f(x) is then the remainder, i.e. (sin(x)/(log(x)+1)). Importance score is given by p(x)/q(x). Since the support of Beta is (0,1) which is already mentioned in the question as a hint, but the proposal distribution N(0,1) has the entire real line as support, the samples which are within (0,1), and are especially close to the Beta mode 5/9, will have high importance scores. 2 marks have been deducted for failing to identify p(x) and solving it as Monte Carlo integration. Note that sin(x)/(log(x)+1) cannot be p(x) as it can be negative in the interval (0,1). For example, for x=10^(-4), log(x)=-4, so the denominator is negative though numerator is positive.

B) You have a dataset X of 10 binary observations. Your Null Hypothesis is that the Bernoulli parameter is 0.5, while the alternate hypothesis is that this parameter is below 0.5. Your test-statistic T(X) is the fraction of '1's, and your rejection region is defined by $(T(X)-0.5<\epsilon)$. What will be possible values of ϵ, if you are to reject the null hypothesis at 5% level of significance? **[4 marks]**

Solution sketch: discussed in class. The important thing to note is that due to H1, the rejection region should be low values of T(X) like 0, 0.1 etc and not high values like 0.9, 1.0 (since they will be even more unlikely under H1). ϵ then will be negative.

C) In the above case, suppose you have obtained 6 heads. What will be the p-value if i) the alternate hypothesis is that parameter is below 0.5, ii) alternate hypothesis is that parameter is different from 0.5? **[4 marks]**

Solution Sketch: Note the def of p-value as discussed in class. It is the probability (under H0) of getting samples that are as 'extreme' as the observations. In case i) extreme means low values (as in part B). For case ii) we must use the 2*min(left end, right end) formula as both low and high values of T(X) define the rejection region.

D) You are provided with observations (x1, x2, …. xN) which are all real numbers. You want to fit a Gaussian Mixture Model of K parameters on it. How will you evaluate the goodness of fit of this model on the data using chi-square test? Write the test statistic, including necessary expressions for how to calculate it. Also write the rejection criteria for the null hypothesis at level of significance α. **[6 marks]**

Solution Sketch: This is a direct question again. The only thing is that we need to specify the 'expected value' f(y) for each interval 'y' based on GMM. For this, the PDF of GMM has to be derived by marginalizing over the latent variables Z. It needs to be mentioned that the GMM parameters will be estimated using E-M (no derivations needed). Also the degree of freedom of chi-square depends on the number of parameters, which is 3K-1 for GMM (pi, mu, sigma for each of the K components).

6 A) Draw 10 samples [x1, …. x10] from N(0, 5) using Box-Muller transformation. Draw 10 more samples [y1, …. y10] from N(x1), N(x2) …. N(x10). **[5+5=10 marks]**

Solution Sketch: Direct question. The scaling using standard deviation and the shifting of mean (for the y samples) need to be shown.

B. Devise a permutation test to check if X and Y follow same distribution. **[6 marks]**

Solution Sketch: Direct question. 3 marks for writing the formulae and approach, 3 more for the numerical computations. Complete computations not needed for full marks.

C. Use Kernel Density Estimate to fit a PDF on Y using i) Parzen Window ii) Gaussian Kernel. Calculate the fitted densities at Y=0 in both cases. **[6 marks]**

Solution Sketch: Direct question. 3 marks for writing the formulae for KDE, 3 more for the numerical computations. Full numerical computation of Parzen Window not needed, only at Y=0.

D. You have a sequence of observations [3.4, -1.5, 7.2, -2.8, -0.4, 6.5]. Would you like to model this with a single high-variance Gaussian, or a mixture of two low-variance Gaussians? Explain your answer using AIC and BIC. **[4 marks]**

Solution Sketch: For mixture of two low-variance Gaussians, obviously the positive values will be one component, negative values another. Mean, variance can be easily calculated in both cases, based on which we can have the likelihood. After that, AIC and BIC can be calculated from formula. At least partial numerical calculations needed for full marks. 1 mark less for intuitive answer.