CS698X: Topics In Probabilistic Modeling And Inference
Users Online : 2

End-sem Exam

Q.1 Consider Bayesian linear regression and logistic regression, trained using datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively (assume hyperparameters to be known). For each of the two models, suppose we have computed the Kullback-Leibler (KL) divergence between the true posterior and its Laplace approximation. Can you say for sure as to which of the two KL divergences will be larger? Justify your answer briefly.

(Must answer within the text-box below)

Max. score: 4; Neg. score: 0; Your score: 2

Your answer:

For the bayesian linear regression the KL divergence between the true posterior and its laplace approximmation is small as true posterior is Normal distritution and even the laplace will also approximate it with normal

But for Logistic regression the KL divergence will be large as the poterior is not the normal distributition which is approximated with normal distribution using the Laplace app.

Feedback:

Q.2 Can the standard Stochastic Gradient Langevin Dynamics (SGLD) algorithm be used to perform inference in a model like gamma-Poisson latent factor model? Briefly justify your answer.

(Must answer within the text-box below)

Max. score: 4; Neg. score: 0; Your score: 4

Your answer:

No we cant use the SGLD algo for performing the inference in the gamma-poission latent factor model. As in this model we are having constraints on the parameter that it should not be negative. which makes of non aplicabilty of the SGLD

Feedback:

Q.3 Which of these is/are true about Gaussian Processes?

model

□ Posterior distribution over the GP function is always available in closed form

✓ ■ The prediction-cost scales in the number of training examples

✓ ■ They can be used for supervised as well as unsupervised learning

---

Q.4 Which of these models can be learned using variational inference or variational EM?

Max. score: 2; Neg. score: 0; Your score: 2

✓ ■ Latent Dirichlet Allocation

✓ ■ VAE

□ GAN

✓ ■ Probabilistic PCA

Q.5 EM can only be used in probabilistic models that have local as well as global unknowns.

Max. score: 1; Neg. score: 0; Your score: 0

✓ ○ false

● true

Q.6 Which of the following is/are true regarding MCMC vs VI?

Max. score: 2; Neg. score: 0; Your score: 0

✓ □ The output of MCMC can be used to find the optima (approximately) of the target distribution whereas the same cannot be done using the output of VI

✓ ■ MCMC requires more storage at test time as compared to VI

■ Both MCMC and VI can, at best, only converge to a local optima of the target distribution

✓ □ Assuming local-conjugacy, per-iteration cost of Gibbs sampling will be slower than mean-field VI updates

Typesetting math: 100%

○ false

✓ ● true

Q.8 Consider a binary classification problem given training data $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$ with $\boldsymbol{x}_n \in \mathbb{R}^D$ and $y_n \in \{0,1\}$. Assume each label $y_n \in \{0,1\}$ to be generated as follows

$$z_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$$
$$\epsilon_n \sim \mathcal{N}(0,1)$$
$$y_n = \mathbb{I}[z_n \geq 0]$$

where $\mathbb{I}[.]$ is the indicator function which returns 1 if the condition is true, and returns 0 otherwise.

Derive the expression for the conditional $p(y_n|\boldsymbol{x}_n, \boldsymbol{w})$ (which will require you to integrate out z_n). Your answer must be in closed form. Hint: You will need to use the definition/properties of cumulative density function (CDF).

For the likelihood model $p(y_n|\boldsymbol{x}_n, \boldsymbol{w})$, is closed form MLE possible for $\boldsymbol{w}$? Irrespective of your answer to this, briefly state a possible way to do MLE for this model, with necessary expressions.

Max. score: 20; Neg. score: 0; Your score: 15

Your answer:

Uploaded file: WhatsApp Image 2021-05-09 at 11.49.25.jpeg (Click to view the uploaded file.)

Feedback:

MLE expressions missing

Q.9 A Bayesian linear regression model with Gaussian likelihood, an isotropic Gaussian prior on weight vector, and gamma priors on the precision hyperparmeters of likelihood and prior will have a closed form expression for the joint posterior distribution over its unknowns.

Max. score: 1; Neg. score: 0; Your score: 1

✓ ● false

○ true

Q.10 Which of these methods can be used to obtain the MLE solution for the hyperparameters of a

Typesetting math: 100%

odel?

✓ □ Variational EM

■ MCMC

✓ □ EM

Q.11 The posterior distribution for a generalized linear model with Poisson likelihood and Gaussian prior can be approximated using Laplace approximation.

Max. score: 1; Neg. score: 0; Your score: 1

○ false

✓ ● true

Q.12 Mean-field VI can be used only if the model has local conjugacy.

Max. score: 1; Neg. score: 0; Your score: 1

✓ ● false

○ true

Q.13 Which of the following sampling methods use the gradient information of the target distribution?

Max. score: 2; Neg. score: 0; Your score: 2

✓ ■ Hamiltonian Monte Carlo

□ Rejection sampling

✓ ■ Langevin dynamics

□ Metropolis-Hastings sampling

Q.14 For a single hidden layer Bayesian neural network with Gaussian likelihood and Gaussian priors on each weight of the network, we can get closed form Gibbs sampler as well as closed form mean-field VI updates.

Max. score: 1; Neg. score: 0; Your score: 0

Typesetting math: 100%

✓ ○ false

Q.15 The samples generated by MCMC can be used to find (at least approximately close) MAP estimate of the unknowns.

Max. score: 1; Neg. score: 0; Your score: 1

○ false

✓ ● true

Q.16 Suppose you are given a collection of D labeled documents $\{w_d, y_d\}_{d=1}^{D}$ where $w_d = \{w_{d,n}\}_{n=1}^{N_d}$ denotes the set of N_d words in the document d and $y_d \in \{1, \ldots, K\}$ denotes the observed label of this document. Assume the vocabulary size (number of unique words across all the documents) to be V, so each word $w_{d,n}$ is one of these V words from the vocabulary.

Consider the following generative story for each labeled document $\{w_d, y_d\}$

Generate the label y_d from a K-dim multinoulli distribution with parameters $\pi = [\pi_1, \pi_2, \ldots, \pi_K]$

$$y_d \sim \text{multinoulli}(\pi)$$

Generate each word in this document i.i.d. from a V-dim multinoulli with parameters ϕ_{y_d} which is a vector of size V

$$w_{d,n} \sim \text{multinoulli}(\phi_{y_d}), \quad n = 1, 2, \ldots, N_d$$

Assume Dirichlet priors on all the multinoulli parameters: $p(\pi) = \text{Dirichlet}(\alpha, \alpha, \ldots, \alpha)$ and $p(\phi_k) = \text{Dirichlet}(\beta, \beta, \ldots, \beta)$, $k = 1, 2, \ldots, K$, and assume the hyperparameters α and β to be known.

Now answer the following:

Draw the plate notation diagram for this model, clearly showing which parts are unknowns and which parts are observed (usual convention: white nodes = unobserved, shaded nodes = observed/known).

Derive the posterior distribution for π and $\Phi = \{\phi_k\}_{k=1}^{K}$. Also write down the MAP solutions for π and $\{\phi_k\}_{k=1}^{K}$. (you don't need to re-derive the MAP solution from scratch; instead, you should be able to get these easily from the expression of their respective posterior distributions).

How would you use this model for classification, i.e., given a new document $w_* = \{w_{*,n}\}_{n=1}^{N_*}$, where N_* denotes the number of words in it, how would you predict the class label y_* of this document? In particular, give the expression for predictive distribution of y_* given the MAP estimates, i.e., $p(y_* = k|w_*, \pi_{MAP}, \Phi_{MAP})$, as well as the posterior predictive distribution $p(y_* = k|w_*, \{w_d, y_d\}_{d=1}^{D})$. Simplify the expressions for these distributions as much as you can to bring them in their final forms.

Uploaded file: WhatsApp Image 2021-05-09 at 12.04.15.jpeg (Click to view the uploaded file.)

Feedback:

Q.17 What is the proposal distribution used by a Gibbs sampling algorithm and why is Gibbs sampling said to be a special case of Metropolis-Hastings algorithm?

(Must answer within the text-box below)

Max. score: 4; Neg. score: 0; Your score: 3

Your answer:

In Gibbs sampling the proposal distribution is conditional distribution $f(Z|Z_{-i})$

Its the special case of MH as the acceptance prob of MH will become equal to 1 which we use this proposed distribution

Feedback:

Proposal is not the conditional distribution p(z_i|z_-i) but conditional posterior p(z_i|z_-i,X)

Q.18 Which of these is/are true about a generative classification model?

Max. score: 2; Neg. score: 0; Your score: 2

✓ ■ Predictions at test-time take into account the fraction of each class present in the training data

✓ ■ It can work even if some training examples are unlabeled

□ Posterior distribution over its parameters can always be computed in closed form

□ It requires fewer parameters as compared to a discriminative model

---

Q.19 Can we use VI to perform model selection? If yes, how? If no, why not?

(Must answer within the text-box below)

Typesetting math: 100%

Max. score: 4; Neg. score: 0; Your score: 4

so we can calculate ELBO for each model m and choose the one with largest ELBO.

Thus VI can be used to perform model selection.

Feedback:

Q.20 The total number of tables that will be eventually occupied across all customers in a Chinese Restaurant Process based prior is proportional to the concentration parameter α

Max. score: 1; Neg. score: 0; Your score: 1

○ false

✓ ● true

Q.21 Consider a latent variable model that learns a K-dimensional latent vector z_n for each observation x_n. Which of the following priors would be appropriate if you want z_n to be a non-negative vector?

Max. score: 2; Neg. score: 0; Your score: 0

☑ Laplace distribution

✓ □ Dirichlet distribution

□ Gaussian distribution

✓ □ Product of K univariate gamma distributions

Q.22 When using VI to approximate a posterior p using another distribution q, why directly minimizing $KL(q||p)$ w.r.t. q is hard? How does VI get around this issue?

Max. score: 4; Neg. score: 0; Your score: 2

Your answer:

From the fact that $logp(X|m) = ELBO + KL(q||p_z)$. as LHS is const wrt z min KL equal to maximizing the ELBO

VI tires to maximize the value of ELBO which intern minimized the $KL(q||p)$

Typesetting math: 100%

Feedback:

Max. score: 2; Neg. score: 0; Your score: 2

- □ It learns the size of the latent code automatically
- ✓ ☑ It can learn nonlinear mapping from latent code to the observation
- □ It uses adversarial training
- ✓ ☑ Probabilistic PCA is one of its special cases

Q.24 Suggest a method that can make Gaussian Process (GP) based models faster at test time as compared to the usual O(N) test-time cost incurred by traditional GPs, where N is the number of training examples.

(Must answer within the text-box below)

Max. score: 4; Neg. score: 0; Your score: 4

Your answer:

learn the pseudo data points(M<<N) to learn the GP. Then this would take only O(M) time at test time.

Feedback:

Q.25 A standard generative adversarial network can be used to generate new data as well as compress training data into a latent representation.

Max. score: 1; Neg. score: 0; Your score: 1

- ✓ ● false
- ○ true

Q.26 What are the main challenges in using VI for learning the posterior for Bayesian neural networks. Briefly state some of the techniques that can address such challenges in VI (and how they do it). Please answer using no more 3-4 sentences.

(Must answer within the text-box below)

Typesetting math: 100%

Max. score: 4; Neg. score: 0; Your score: 0

techniques which trains network M times with different seeds with weights \(\theta_1,...\theta_M\)

Now we can approximate the posterior using \({1\over M }\sum_m \delta_{\theta_m}(\theta)\) which is Bayesian Model Averaging using M models

Feedback:

Q.27 Assume we are given a collection of D documents and a vocabulary of total V unique words. Consider a model which assumes that, for document d, first a categorical latent variable \(z_d\) is drawn from a K-dimensional multinoulli, and then depending on the value of \(z_d\), all words in document d are drawn from one of the K V-dimensional multinoulli distributions. Does the latent dirichlet allocation (LDA) allocation model have any advantage over this model? Briefly justify your answer.

(Must answer within the text-box below)

Max. score: 4; Neg. score: 0; Your score: 4

Your answer:

In the model described above every document should belong to only one of topic as Z is sample from multinoulli. which is not the case with LDA model in which each document can belong to multiple topics.

Feedback:

Q.28 Consider the LDA model with K topics, learned using D documents, and and suppose the vocabulary consists of V unique words. In LDA, each document \(d \in \{1,2,\ldots,D\}\) is associated with a vector \(\theta_d\) and there are K topic vectors \(\{\phi_k\}_{k=1}^K\).
What is the meaning of each entry of the vectors \(\theta_d\) and \(\phi_k\) and what are the sizes of these vectors?
Suppose we have performed a collapsed inference by integrating out \(\theta_d\) and \(\phi_k\) and have inferred only word-to-topic assignment variables \(z_{d,n} \in \{1,2,\ldots,K\}\) for the corresponding word \(w_{d,n} \in \{1,2,\ldots,V\}\)which denotes the n-th word in the d-th document. Given these word-to-topic assignment variables for all words in all documents, how would you approximate \(\theta_d\) and \(\phi_k\). Give the expressions (or explain briefly in words) as to how you would compute these using the \(z_{d,n}\)'s.

(You may write the answer in the textbox below or may upload pics/PDF of your solution)

Max. score: 12; Neg. score: 0; Your score: 0

Typesetting math: 100%

Your answer:

Q.29 What is the phenomenon of posterior collapse in VAE? Is a simple model like probabilistic PCA also likely to suffer from this?

(Must answer within the text-box below)

Max. score: 4; Neg. score: 0; Your score: 4

Your answer:

When the decoder is more powerful then the $\log p_\theta(x|Z)$will make it very large which in tern KL in loss function becomes close to zero by collapsing $q_\phi(z|x)$ to $p_\theta(z)$.

No the probabalistic PCA wont suffer.

Feedback:

Q.30 Consider the following model for generating N count-valued observations $y_1,y_2,\ldots,y_N$:

$$y_n \sim \text{Poisson}(\lambda_n), \quad n=1,\ldots,N \\ \lambda_n \sim \text{Gamma}(1,b), \quad n=1,\ldots,N \\ b \sim \text{Gamma}(1,1)$$

Does the model have local conjugacy w.r.t. all the unknowns? Irrespective of the answer, derive the conditional posteriors for all the unknowns, and sketch a Gibbs sampler for sampling from the posterior $p(\lambda_1,\lambda_2,\ldots,\lambda_N,b|y_1,y_2,\ldots,y_N)$. Assume the shape-rate parametrization of the gamma distribution.

Max. score: 14; Neg. score: 0; Your score: 14

Your answer:

Uploaded file: WhatsApp Image 2021-05-09 at 11.10.20-converted.pdf (Click to view the uploaded file.)

Feedback:

Q.31 Rejection sampling will reject lots of samples if the constant M in the condition $Mq(z) \geq \tilde{p}(z)$ is set to a very large value, where q is the proposal and $\tilde{p}$ is the unnormalized target.

Typesetting math: 100%

Max. score: 1; Neg. score: 0; Your score: 1

Q.32 Suppose you have tossed a coin a number of times. Now suppose you want to compute the probability that \(\theta \leq 0.5\) where \(\theta\) is the probability of heads. Briefly describe a Bayesian way to do this.

(Must answer within the text-box below)

Max. score: 4; Neg. score: 0; Your score: 4

Your answer:

each coin toss can be modeled by \(bernoulli(y|\theta)\). and let assume a beta prior on the \(\theta\) then we can get the posteiror \(p(\theta|Y) = Beta(θ | $\alpha + N$_1, $\beta + N$_0)\) from this posteiror distribution we can get the probability that \(p(\theta \le 0.5 | Y) \) = \(\int_0^{0.5} Beta(θ | $\alpha + N$_1, $\beta + N$_0)\)

Feedback:

Q.33 Which of the following is/are true about the Latent Dirichlet Allocation model?

Max. score: 2; Neg. score: 0; Your score: 2

✓ ■ It learns a non-negative vector representation for each document

✓ ■ It can be seen as a model that clusters/groups the documents

□ It learns a discrete/categorical representation for each document

✓ ■ It clusters the words in each document

Q.34 Which of these inference methods can be used to approximate the posterior for the logistic regression model (assume no additional variable introduced in the model)?

Max. score: 2; Neg. score: 0; Your score: 0

✓ ■ Laplace approximation

✓ ■ Metropolis-Hastings

■ Gibbs sampling

✓ ■ Stochastic Gradient Langevin Dynamics

Typesetting math: 100%

Typesetting math: 100%