CS698X: Topics In Probabilistic Modeling And Inference
Users Online : 1

Mid-sem Exam

Q.1 Consider the following "two-step" model for binary classification:

(1) Draw a count-valued random variable m_n from $Poisson(\theta_n)$ where the Poisson's rate parameter θ_n itself is drawn from $Gamma(r, \exp(w^\top x_n))$ where $r > 0$ is another fixed hyperparameter (note: The Poisson distribution for a count-valued random variable x is of the form $Poisson(x|\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$ where λ is the rate parameter of this Poisson, and here the gamma distribution is assumed to have the shape-scale parametrization with $Gamma(x|a,b)$ = $\frac{1}{\Gamma(a)b^a} x^{a-1} \exp(-x/b)$).

(2) Set the label $y_n = 1$ if $m_n > 0$, otherwise set $y_n = 0$.

For the above model:

(a) Derive the distribution of y_n, conditioned on x_n, w, and r. In particular, since y_n is binary, just derive the expression for $p(y_n = 1|x_n, w, r)$. Hint: To do this, first derive the expression for $p(y_n = 1|\theta_n)$ and then integrate out θ_n.

(b) Briefly explain why the expression $p(y_n = 1|x_n, w, r)$ makes intuitive sense.

(c) Would any specific value of the hyperparameter $p(y_n = 1|x_n, w, r)$ make this model equivalent to a logistic regression model? If yes, for what value?

(d) Assuming a Gaussian prior on w, can posterior over w be obtained in closed form for this model?

Max. score: 13; Neg. score: 0; Your score: 0

Your answer:

Feedback:

Q.2 Which of the following is/are true about Monte Carlo approximation in the context of Bayesian inference?

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

✓ ☑ It can be used to approximately compute the marginal likelihood

□ If the posterior is Gaussian, we do not need to use it.

✓ ☑ It can be used to approximately compute the posterior predictive distribution

Q.3 Suppose you have learned a Bayesian logistic regression model (its posterior and posterior predictive distribution). What will happen to the shapes of the equal-probability contours (i.e., on which all inputs have the same posterior probability of belonging to a given class), as the number of training observations becomes very very large? Briefly justify your answer using not more than 3-4 sentences or 60 words (may use some symbols if needed).

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

Your answer:

As the number of samples increases the variance in w decreases as hessian of likelihood value decreases. so when doing estimation equal probability counter almost will become linear as all the weight vector ideally becomes W_map

Feedback:

Q.4 Is the generalized linear model (GLM) a generative model? Briefly justify your answer using not more than 3-4 sentences (or 60 words).

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

Your answer:

in GLM y is modeled $p(y|\eta)$ in terms of x in the form of which appers in the form $\boldsymbol{w}^T\boldsymbol{x}$ in which we are not modeling the x .

So GLM are discriminative models

Feedback:

Q.5 Will Laplace approximation in general be a good idea to approximate the posterior distribution of a generalized linear model (GLM)? Ignore any computational cost related issues. Briefly justify your answer using not more than 3-4 sentences (or 60 words).

Max. score: 3; Neg. score: 0; Your score: 1

Your answer:

yes it good idea as we get easily get the derivatives ($\boldsymbol{w}_{map}$and $\boldsymbol{H}^{-1}$)

Feedback:

CS698X: Topics In Probabilistic Modeling And Inference
Users Online : 1

□ Gaussian prior with diagonal covariance

✓ ■ Laplace prior

□ Gaussian prior with spherical covariance

✓ ■ Spike and slab prior

Q.7 Which of these is more robust against overfitting: (1) Plug-in predictive distribution, (2) Posterior predictive distribution? Briefly justify your answer using not more than 3-4 sentences (or 60 words).

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

Your answer:

PPD is more robust to overfitting as plug in prediction distribution tires to fix the training data exactly. where as the PPD is integrated over the all the possible weights which specifies the variance in the predicted output(also depends on the input)

Feedback:

Q.8 Consider a generative model for multi-class classification and two ways to learn the model: MLE and MAP. If you have very little amount of training data for some of the classes, would there be any advantage of using the MAP approach over the MLE approach? Answer this using not more than 3-4 sentences or 60 words (may use some symbols if needed).

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

Your answer:

MAP estimate takes into consideration of class prior distribution which acts as regularizer. As there is very less training data MLE tries to overfit the data by minimizing only likelihood

Feedback:

Q.9 Suppose you want to estimate the probability of each face of a six-faced dice and have assumed a Dirichlet prior on the vector of these probabilities. Intuitively, what role does the concentration parameter of this Dirichlet play when the prior is used in this parameter estimation problem? Briefly explain your answer using not more than 3-4 sentences or 60 words (may use some symbols if needed).

Max. score: 3; Neg. score: 0; Your score: 2

Feedback:

Q.10 Consider a model with a vector-valued parameter \(\theta \in \mathbb{R}^D\) with posterior distribution \(p(\theta|X) = N(\mu,\Sigma)\). What will be the marginal posterior for each entry of \(\theta\), i.e., \(p(\theta_d|X), d=1,2,\ldots,D\)? Does it have a closed form expression for this model? Also, will \(p(\theta_d|X)\) have a closed form expression for any model in general? Please be brief in answering.

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

Your answer:

yes marginal postirier and \(p(\theta_d|X) \)both have closed form.

Feedback:

Q.11 Consider N i.i.d. observations \(\{x_n\}_{n=1}^N\) from distribution \(p(x|\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}\) where the parameter \(\lambda\) has a prior distribution \(p(\lambda|a,b) \propto \lambda^{a-1} \exp(-b\lambda)\) which is available in closed form (its normalization constant is known). Is the posterior available in closed form? If yes, what is this distribution (up to a proportionality constant) and its parameters (no need to show derivation; just the expression is needed)? If a closed form expression for the posterior can't be found, briefly state why?

Max. score: 3; Neg. score: 0; Your score: 2

Your answer:

yes it has closed form expression as both posion and gamma(\(p(\lambda|a,b) \)) are conjugates to each other. The posterior will be in the form \(gamma(\lambda|a+x,(b+1)^{-1})\).

Feedback:

Q.12 Assume you have K candidate models (assume probabilistic models) that you can possibly try out for a classification problem and don't know which one is the "best". How does a fully Bayesian approach handle this problem? Give a precise answer using not more than 3-4 sentences or 60 words (may use some symbols if needed).

Max. score: 3; Neg. score: 0; Your score: 1.5

Feedback:

Q.13 For any generative model \(p(x|\theta)\) where \(x\) denotes data and \(\theta\) denotes the distribution parameters, it is possible to define a natural notion of ``similarity'' (i.e., a kernel) between any two inputs \(x\) and \(x^\prime\) under \(p(x|\theta)\) as

\(k(x,x^\prime) = g(x,\theta) ^\top \mathbf{F}^{-1} g(x^\prime,\theta)\)

where \(g(x,\theta) = \nabla_\theta \log p(x|\theta)\) and \(\mathbf{F} = \mathbb{E}[g(x,\theta)g(x,\theta)^\top]\) and the expectation is w.r.t. the distribution \(p(x|\theta)\).

Now let's assume \(p(x|\theta)\) to be a Gaussian distribution with mean vector \(\mu\) and a **fixed** covariance matrix \(\Sigma\) (since the covariance is fixed, \(\theta = \mu\)). For this Gaussian distribution, compute the expression \(k(x,x^\prime)\) for similarity between two inputs \(x\) and \(x^\prime\). Does the expression make intuitive sense? What would this be equal to if the covariance matrix \(\Sigma\) is an identity matrix?

Max. score: 10; Neg. score: 0; Your score: 0

Your answer:

Feedback:

Q.14 Consider the probabilistic linear regression model with Gaussian likelihood and Gaussian prior on weight vector \(\mathbf{w}\), and hyperparameters known. If we only care about the predictive mean of test inputs, will the predictive mean be different if computed using the full posterior of \(\mathbf{w}\) as opposed to using the MAP solution of \(\mathbf{w}\)? Give a precise answer using not more than 3-4 sentences or 60 words (may use some symbols if needed).

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

Your answer:

boht are same and will be equal to \(w_{map} = (X^TX + {\lambda \over \beta} I_D)^-1 X^Ty\)

Feedback:

Q.15 Consider three models M1, M2, and M3, learned from the same data X. Which of the following is/are reasonable ways to select the best model?

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

□ Select the model that has the largest likelihood (among all the 3 models) at its MAP solution.

✓ ■ Select the model with the largest marginal likelihood

Q.16 Which of the two approaches to supervised learning - generative model and discriminative model - would usually require a larger number of parameters to be estimated? Justify your answer through a concrete example of each of these two types of models, and not using more than 3-4 sentences (or 60 words).

Max. score: 3; Neg. score: 0; Your score: 2

Your answer:

generative needs large number of parameters to estimmate as we model x,y jointly. But in discriminative we only model y.

Feedback:

Q.17 The gamma distribution (assuming its shape-rate parametrization) is defined as \(Gamma(x|a,b) = \frac{b^a}{\Gamma(a)}x^{a-1}\exp(-bx)\), where \(a\) and \(b\) are the shape and rate parameters, respectively, and \(\Gamma(a)\) denotes the gamma function, and the mean and variance of gamma distribution under the shape and rate parameterization are \(a/b\) and \(a/b^2\), respectively.

(a) Approximate this gamma distribution by a Gaussian using Laplace approximation.

(b) How does the Laplace approximation compare with approximating \(Gamma(x|a,b)\) by another Gaussian whose mean and variance are equal to the mean and variance, respectively, of the actual gamma distribution \(Gamma(x|a,b)\). Under what condition would these two approximations be roughly the same?

Max. score: 10; Neg. score: 0; Your score: 0

Your answer:

Feedback:

Q.18 Briefly state why the marginal likelihood of a model can also be seen as a special case of the posterior predictive distribution, Answer this using not more than 3-4 sentences or 60 words (may use some symbols if needed).

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

No Answer Given

Q.19 Which of the following quantities can be obtained given a posterior distribution \(p(\theta|X)\)?

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

✓ ☑ The MAP estimate of \theta

☐ MLE solution of \theta

✓ ☑ Probability that \(\theta\) is less than some value \(\theta_*\)

✓ ☑ Probability that \(\theta\) is between some range (a,b)

Q.20 A zero-mean Gaussian prior is equivalent to using L2 regularization on the weight vector w. Can such a prior be used to impose different amounts of regularization on different components of the weight vector? If yes, how? If no, why not? Answer this using not more than 3-4 sentences or 60 words (may use some symbols if needed).

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

Your answer:

yes we can use that by changing the different \(\lambda \) pression parameter for different w.

which lead to different amounts of regulization on different components of w.

Feedback:

Q.21 Assume we have \(N\) scalar-valued observations from a Gaussian \(\mathcal{N}(\mu,\sigma^2)\). Assume \(\sigma^2\) to be fixed and \(\mu\) to be unknown with a Gaussian prior \(\mathcal{N}(\mu_0,\sigma_0^2)\) with both \(\mu_0\) and \(\sigma_0^2\) to be known. Assume \(N\) is not too large.

(a) Derive the MAP estimate of \(\mu\) (you may directly write the answer if you know it) and show that it can be expressed as a hybrid of the MLE solution (which is the sample mean of the N observations - you don't need to derive it) and prior's mean.

(b) What happens to the MAP estimate of \(\mu\) when the Gaussian prior's variance \(\sigma_0^2\) becomes very small? Briefly justify your answer.

(c) What happens to the MAP estimate of \(\mu\) when the Gaussian prior's variance \(\sigma_0^2\) becomes very large? Briefly justify your answer.

CS698X: Topics In Probabilistic Modeling And Inference
Users Online : 1

Uploaded file: WhatsApp Image 2021-02-21 at 19.54.22.jpeg (Click to view the uploaded file.)

Feedback:

Q.22 Consider a kernel based regression model which assumes the outputs generated as: \(y_n \sim \mathcal{N}(y_n|\sum_{i=1}^N w_i k(x_i,x_n),\beta^{-1})\) where \(\{x_i\}_{i=1}^N\) denotes the training inputs. Assume a zero-mean Gaussian prior on each of the weights \(\{w_i\}_{i=1}^N\) and the precision/variance hyperparameter of the prior to be fixed or estimated via MLE-II. At test time, will this model be faster than a Support Vector Machine based regression model (SVR)? If yes, why? If no, why not? Please be brief in answering.

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

Your answer:

No its wont be fater that SVR as it we have to taken in consideration of all the training inputs while prediction as w will not sparce. But in SVR we have sparce w which reduce our time for prediction by consdering inputs that have only non zero w attached with them

Feedback:

Q.23 Which of the following is/are true about a generalized linear model (GLM) in general?

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

□ Posterior predictive in closed form

✓ ■ MLE has a unique solution

□ Posterior in closed form

■ MLE has closed form solution

Q.24 What is the advantage of selecting the "best" hyperparameter values using an MLE-II approach as compared to using cross-validation? Your answer should not use more than 3-4 sentences (or 60 words).

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

Your answer:

for MLE-II we do no need to take a part of training data for cross validation. where as in cross validatoin we have to take out a part of traning data for cross validation. In MLE-II we select the h.p

CS698X: Topics In Probabilistic Modeling And Inference
Users Online : 1

Q.25 Suppose you have a coin with probability of heads being \(\theta \in (0,1)\) which is assumed to have a \(Beta(a,b)\) prior. The coin is tossed independently N times and it shows heads less than K (<N) times but you do not know the exact number of heads (but only that this number was less than K, i.e., between 0 to K-1). Write down the expression for the posterior of \(\theta\) up to a proportionality constant (i.e., you don't need to show the calculation of the normalization constant).

Max. score: 6; Neg. score: 0; Your score: 1

Your answer:

Uploaded file: WhatsApp Image 2021-02-21 at 20.02.48.jpeg (Click to view the uploaded file.)

Feedback:

Score: 46.5