MCQ

1. Which of the following priors is/are expected to yield sparse weight vectors?

   Gaussian prior with spherical covariance

   Gaussian prior with diagonal covariance

   Spike and slab prior

   Laplace prior

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

   Probability that θ is less than some value θ_*

   The MAP estimate of \theta

   Probability that θ is between some range (a,b)

   MLE solution of \theta

   Note: In general, you can't extract the MLE solution from the posterior distribution since the posterior is a combination of the likelihood and the prior and you can't just "remove" the prior from the posterior to extract the likelihood and its maxima, unless in very special cases, such as the posterior having a simple analytical form and us knowing what the prior and its hyperparameters were.

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

   Posterior in closed form

   Posterior predictive in closed form

   MLE has closed form solution

   MLE has a unique solution

   Note: Perhaps the simplest example of a GLM for which the first 3 options aren't true is logistic regression.

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

   It can be used to approximately compute the posterior predictive distribution

   It can be used to approximately compute the marginal likelihood

If used for probabilistic linear regression (assuming fixed hyperparameters), it will give an exact answer for the posterior predictive distribution

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

Note: It is always an approximation (only in theory when you use infinite many samples, you would converge to the true value of the integral). Also, even if the posterior is Gaussian but the likelihood isn't Gaussian (say sigmoid-Bernoulli), we won't be able to compute the PPD without using something like Monte Carlo approximation (or some other approximation technique)

5. 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?

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

Select the model whose posterior distribution has the smallest variance.

Select the model with the largest marginal likelihood

Select the model with the largest posterior probability

Note: Value of the likelihood at the MAP solution isn't a reliable metric of how good the model is since it is still using a very specific value of the parameter to compute the likelihood (probability of the data). Comparing the posterior's variances also doesn't give us any reliable information about how good the model is (comparing variances of difference models may not even be meaningful since different models may have different sets of parameters and the "variance" may not be a single number, unlike marginal probability or the posterior probability of the model)

15 Short-Answer Questions

The answers and grading rubric used for grading these questions are as follows:

==========
Q1) Is the generalized linear model (GLM) a generative model? Briefly justify your answer using not more than 3-4 sentences (or 60 words).

A) No since it does not model the inputs x directly models the conditional distribution of output y given input x.

Grading Rubric: Give 3 marks if the answer mentions that the inputs are not modeled and only the conditional distribution is directly modeled. No partial marking.

==========
Q2) 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).

A) Generative model since it requires estimating the parameters of the class prior p(y|\pi) as well as the parameters of each of the class conditional distributions p(x|y=k). In contrast, a discriminative model only needs to learn the parameters of p(y|x). For example, for binary classification, logistic regression (a discriminative model) only needs a weight vector to model p(y|x), whereas a generative classification model will need a mean and covariance matrix for each of the class conditionals, in addition to the parameters of the class prior.

Grading Rubric: Give 3 marks if the answer correctly mentions that the generative model will require more parameters + mentions the reason + gives an example like the one given in the reference answer. If only the reason is correct but the example is not fully correct but is only somewhat correct then give 2 marks. Otherwise, no marks.
==========
Q3) 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).

A) It does not require us to train the same model multiple times to select the best hyperparameter values but directly optimizes w.r.t. them. Also, it does not waste any training data since we don't need to set aside part of our training data as held-out data in order to do cross-validation.

Grading Rubric: If at least one of the reasons mentioned in the reference answer is given, give 3 marks. Otherwise, no marks.
==========
Q4) 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).

A) Yes, we can use a Gaussian prior with a diagonal precision/covariance matrix. The d-th entry (precision/variance) of the diagonal will control the extent of regularization of the d-th element of the weight vector.

Grading Rubric: 3 marks only if the answer is correct. Otherwise, no marks.

==========

Q5) 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).

A) Posterior predictive is more robust since it does not rely on a single best solution (which could possibly have overfit) but takes into account the uncertainty in the parameter estimates and uses all possible parameter values and reports an averaged prediction.

Grading Rubric: 3 marks only if the answer is correct (must mention either the fact that uncertainty is taken into account which results in robustness, or mention about posterior average). If the explanation is somewhat incorrect/imprecise but the answer contains the basic idea, give 1.5 marks. Otherwise, no marks.

==========

Q6) 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.

A) This model will not be faster than SVR since a simple zero-mean Gaussian prior on the weights will not result in any sparsity, resulting in all weights being nonzero. Note that this model is like RVM without a sparse prior on the weights. In contrast, in SVR, at least have the support vectors having nonzero weights.

Grading Rubric: 3 marks only if the answer is fully correct (the model will be slower and mentions the correct reason). If the answer says the model will be slower but the reason is not fully correct/imprecise, give 1.5 marks. Otherwise, no marks.

==========

Q7) 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).

A) Indeed, since GLM has a global optima (hence unimodal posterior) which can easily be found. Hessian also has a simple algebraic form for GLMs.

Grading Rubric: 3 marks only if the answer is fully correct (GLMs have global MLE and/or unimodal posterior and therefore Laplace approximation is a good idea).

==========

Q8) 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 θ, 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.

A) Due to the Gaussian's property, the marginal posterior for each \theta_d will simply be a univariate Gaussian with mean \mu_d and variance being the d-th diagonal entry of \Sigma. This won't be the case with every model if the posterior p(\theta|X) is not a Gaussian.

Grading Rubric: 3 marks only if the answer is fully correct (right answer for the marginal posterior + says that it won't be possible for those models where the posterior is not a Gaussian). If only the first part is correct (right answer for the marginal posterior), give 2 marks.

==========

Q9) 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).

A) The PPD is of the form p(D'|D) =\int p(D'|\theta)p(\theta|D)d\theta where D' is the test data and D is the training data. To see why the marginal likelihood is a special case of PPD, consider the case when the number of training examples N = 0. Then PPD reduces to P(D') and the posterior is simply the prior and p(D') = \int p(D'|\theta)p(\theta)d\theta. This quantity is simply the marginal likelihood of the test data D'.

Grading Rubric: 3 marks only if the answer is fully correct. Otherwise, no marks.

==========

Q10) 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).

A) With a Bayesian approach, we can compute the posterior probability p(m|X,y) for each model and then use these in two ways: (1) Select the "best" model based on which model has the largest p(m|X,y) and report the PPD for that model, i.e., p(y_*|x_*,X,y,m), and (2) Compute the PPD by doing an averaging at the model-level as well, i.e., the "double averaging" as p(y_*|x_*,X,y) = \sum_{k=1}^K p(y_*|x_*,X,y,m)p(m|X,y) where p(y_*|x_*,X,y,m) is the PPD of model m. Note that approach (1) is not 100% Bayesian since we are still only considering a single best model.

Grading Rubric: 3 marks only if the answer is fully correct. If the answer mentions that we can compute p(m|X) (not that for classification, it will be of the form p(m|X,y) but it is okay if the answer only write a generic form p(m|X)) and use it to select the best model, give 1.5 marks since this is not 100% Bayesian. For getting 3 marks, the answer must mention the averaging over all models as mentioned in the reference answer and also discussed in the lectures.

==========

Q11) 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 λ 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?

A) The algebraic forms of the likelihood and prior are the same and therefore, not surprisingly, the posterior will also have the same form and thus computable analytically. The posterior will be the same distribution as the prior with hyperparameters a' = a + \sum_{n=1}^N x_n and b' = b + N. Note that it is a gamma distribution (the likelihood was a Poisson).

Grading Rubric: 3 marks only if the answer is fully correct. If the answer mentions the form of the posterior correct (doesn't need to say gamma distribution) but the posterior's parameters are not fully correct (forgets the summation over all examples in the update equation of a') then give 2 marks. Otherwise, no marks.

==========

Q12) 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).

A) When the number of training observations becomes very very large, the posterior will reduce to a point estimate (MLE) and all hyperplanes of our Bayesian ensemble will collapse/coincide on the MLE hyperplane. Consequently, the equal-probability contours will just become straight lines parallel to that MLE hyperplane.

Grading Rubric: 3 marks only if the answer is fully correct (correctly mentions that the posterior will degenerate to the MLE solution and the equal-probability contours will be straight lines parallel to the MLE based hyperplane). If the answer is not fully correct but shows some basic understanding (e.g., posterior degenerating to the MLE solution) then give 1.5 marks. Otherwise, no marks.

==========

Q13) 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).

A) The entries of the concentration parameter vector can be seen as the number of pseudo-observations, which in this case would correspond to the number of times each of the six faces were shown in our imaginary experiment of rolling the dice. These pseudo-observations help in smoothening our parameter estimates of the Dirichlet probability especially when the number of observations for some of the outcomes are very low.

Grading Rubric: 3 marks only if the answer is correct (mostly along the lines of the reference answer, i.e., mentions that the role is to perform smoothing/regularization especially when the number of observations is very small). If the answer is mostly okay but lacks precision, give 2 marks. Otherwise, no marks.

==========

Q14) 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).

A) Yes, MAP will be advantageous since it will help regularize the probability parameter \pi of the class prior distribution p(y|\pi) if we choose a Dirichlet prior on \pi. Note that this is related to Question 13 as well. The MAP approach will also help in regularizing the parameters of the class-conditional distributions by choosing suitable priors on those (e.g., if we have Gaussian class-conditionals, then we can choose a Gaussian prior on the mean and inverse Wishart on the covariance matrix).

Grading Rubric: 3 marks if the answer correctly mentions that the MAP approach will be better since it will help regularize the model parameters (class prior's parameter as well as class-conditionals' parameters). It is okay if the exact distributions to use as prior are not mentioned. If the answer is mostly okay but lacks precision, give 2 marks. Otherwise, no marks.

==========

Q15) 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).

A) No, they will be the same. When using the full posterior, the predictive mean is \mu_N^\top x_* where \mu_N is the mean of the Gaussian posterior. When using the MAP, the predictive mean will be \mu_MAP^\top x_* where \mu_MAP is the MAP solution of the weight vector, which is the same as the mode of the posterior. Since the mean and mode of a Gaussian are the same, the predictive means computed by both approaches will also be the same.

Grading Rubric: 3 marks for the correct answer with correct explanation. The correct explanation should in some way (either in words or via equations) mention that the mean and mode of a Gaussian are the same and therefore the predictive mean of the distribution of y_* will also be the same. Otherwise, no marks.

5 Medium/Long-Answer Questions

The answers for these questions are on the following pages, after which the grading rubrics are also attached.

Q1

The prior is proportional to $\theta^{a-1}(1-\theta)^{b-1}$

The likelihood corresponding to seeing anything between 0 to K-1 heads will be a sum of Binomial likelihoods

$$P(X|\theta) = \binom{N}{0}(1-\theta)^{N} + \binom{N}{1}\theta(1-\theta)^{N-1} + \dots + \binom{N}{K-1}\theta^{K-1}(1-\theta)^{N-K+1}$$

The posterior will be proportional to

$$\theta^{a-1}(1-\theta)^{b-1} \times P(X|\theta)$$

where $P(X|\theta)$ is given by the above expression

Q2

The MAP estimate is given by (skipping derivation)

(a)

$$\mu_{MAP} = \frac{\sigma^2/N}{\sigma_0^2 + \sigma^2/N}\mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/N}\bar{x}$$

($\bar{x}$ = $\frac{1}{N}\sum_{n=1}^{N} x_n$ = MLE)

The above is indeed a hybrid (a convex combination actually) of the MLE solution $\bar{x}$ and the prior's mean μ_0.

(b) As $\sigma_0^2 \to 0$

$$\mu_{MAP} \to \mu_0$$

This makes intuitive sense because $\sigma_0^2 \to 0$ means that the prior's variance is close to zero and thus we are very certain about about prior belief (thus $\mu_{MAP} \to \mu_0$)

(c) As σ_0^2 becomes very very large, the first term is the expression of μ_{MAP} will tend to zero and

$$\mu_{MAP} \to \bar{x}$$

This also makes intuitive sense since σ_0^2 being very large means a flat/uninformative prior, thereby making the prior virtually uneffective for MAP estimation. Thus μ_{MAP} will approach the MLE solution.

Q23

Part 1: Derive the expression for

$$P(y_n = 1 \mid x_n, w, r)$$

Note that $y_n = 0$ only if $m_n = 0$

Since $m_n \sim \text{Pois}(\theta_n)$

$$P(m_n = 0 \mid \theta_n) = \frac{\theta_n^0 \exp(-\theta_n)}{0!} = \exp(-\theta_n)$$

Thus $P(y_n = 1 \mid \theta_n) = 1 - P(m_n = 0 \mid \theta_n)$

$$= 1 - \exp(-\theta_n)$$

To get $P(y_n = 1 \mid x_n, w, r)$ from the above, we need to integrate out θ_n from $P(y_n = 1 \mid \theta_n)$

$$P(y_n = 1 \mid x_n, w, r) = \int P(y_n = 1 \mid \theta_n) P(\theta_n \mid w, r, x_n) d\theta_n$$

Gamma distribution

Substituting the expression for Gamma PDF of θ_n and solving the above integral gives

$$P(y_n = 1 \mid x_n, w, r) = 1 - \frac{1}{\left[1 + \exp(w^T x_n)\right]^r}$$

This model is actually known as "softplus regression"

(See this paper: "Softplus Regression and Convex Polytopes" for further generalizations and properties)

Part 2: The expression makes intuitive sense since as $w^T x_n$ becomes very large +ve, $P(y_n = 1 \mid x_n, w, r)$ approaches 1, and as $w^T x_n$ becomes very large -ve, this probability approaches 0, just like what you would expect from a linear model of classification (probabilistic). Even logistic regression behaves like that.

Part 3: Indeed. $r = 1$ makes the model equivalent to logistic regression.

Part 4: Since the likelihood (a generalization of the likelihood we have in logistic regression) is not conjugate to the Gaussian prior, we can't obtain the posterior easily in closed form.

Q4

$$\log \text{Gamma}(x|a,b) = a\log b - \log\Gamma a + (a-1)\log x - bx$$

x(MAP)

We need the mode of the above and the Hessian.

Mode: $\frac{\partial}{\partial x}\log \text{Gamma}(x|a,b) = 0$

$$\Rightarrow \frac{(a-1)}{x} - b = 0$$

$$\Rightarrow \hat{x} = \frac{(a-1)}{b} \Rightarrow \text{MAP solution}$$

Hessian: $H = -\nabla^2 \log \text{Gamma}(x|a,b)\Big|_{x=\hat{x}}$

$$= +\frac{(a-1)}{\hat{x}^2} = \frac{(a-1)\,b^2}{(a-1)^2} = \frac{b^2}{(a-1)}$$

Thus, the Laplace approx is:

$$N\left(x \,\middle|\, \frac{(a-1)}{b}, \frac{(a-1)}{b^2}\right) \rightarrow N(x|\hat{x}, H^{-1})$$

Also, Gamma has mean $= \frac{a}{b}$ and variance $= \frac{a}{b^2}$ using which the Gaussian approx will be $N(x|\frac{a}{b}, \frac{a}{b^2})$. This will be similar to the Laplace approx if a is very large (i.e., $a-1 \approx a$)

Q5

$$\log p(x|\theta) = \text{Const} - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)$$

$$\nabla \log p(x|\theta) = -\frac{1}{2} \times 2 \times \Sigma^{-1}(x-\mu) \times (-1)$$

$$g(x,\theta) = \Sigma^{-1}(x-\mu)$$

Now, $F = \mathbb{E}[g(x,\theta)\, g(x,\theta)^T]$ → Using def^n of covariance

$$= \text{Cov}[g(x,\theta)] + \mathbb{E}[g(x,\theta)]\,\mathbb{E}[g(x,\theta)]^T$$

Also $\mathbb{E}[g(x,\theta)] = \Sigma^{-1}(\mathbb{E}[x] - \mu)$

$$= \Sigma^{-1}(\mu - \mu)$$

$$= 0$$

Thus $F = \text{Cov}[g(x,\theta)]$

$$= \text{Cov}[\Sigma^{-1}x - \Sigma^{-1}\mu]$$

$$= \Sigma^{-1}\,\text{Cov}(x)\,\Sigma^{-1} - 0$$

$$\Rightarrow \Sigma^{-1}\Sigma\Sigma^{-1}$$

$$F = \Sigma^{-1}$$

Thus $K(x,x') = g(x,\theta)^T F^{-1} g(x',\theta)$

$$= (x-\mu)^T \Sigma^{-1}\Sigma\Sigma^{-1}(x'-\mu)$$

$K(x,x') = (x-\mu)^T \Sigma^{-1}(x'-\mu)$ → Akin to a Mahalanobis type similarity b/w x and x'

For $\Sigma = I$, it is like dot product like similarity (Computed after subtracting off the mean)

Grading Rubrics

Question 21

==========
Correct prior specified (Beta distribution which is proportional to \theta^(a-1)(1-\theta)^(b-1): 1 mark

Correct likelihood (sum of K Binomials): 4 marks
Somewhat correct likelihood (sum of K terms, each of which is in form of a product of Bernoullis): 2 marks

Correct posterior (just needs to mention that it is proportional to the product of the prior and likelihood): 1 mark

==========

Question 22

==========
Correct expression for MAP: 2 marks (note: derivation is not required; directly writing the final expression is also acceptable)
Somewhat correct expression for MAP but some minor mistakes: 1 mark

Correct justification for what happens when \sigma_0^2 becomes very very small: 2 marks
Somewhat correct justification for what happens when \sigma_0^2 becomes very very small: 1 mark

Correct justification for what happens when \sigma_0^2 becomes very very large: 2 marks
Somewhat correct justification for what happens when \sigma_0^2 becomes very very large: 1 mark

==========

Question 23

==========
Total marks for this question = 13

Part (a) is worth a total of 8 (3+5) marks for two of the key steps involved, as mentioned below:

Derives the correct expression for p(y_n = 1|\theta_n) and shows it to be 1 - exp(-\theta_n): 3 marks

Shows some basic steps for the derivation of p(y_n = 1|\theta_n) but derivation/result not fully correct: 2 marks

Derives the correct expression for p(y_n = 1|x_n,w,r) by integrating out \theta_n: 5 marks

Shows mostly correct steps for the derivation of p(y_n = 1|x_n,w,r) but derivation/result has minor errors: 4 marks

Shows some basic steps for the derivation of p(y_n = 1|x_n,w,r) but derivation/result has some major errors: 2 marks

==========

Part (b) is worth 2 marks. Give 2 marks if explanation is along the lines of what is mentioned in the solution (basically, similar to logistic regression)

==========

Part (c) is worth 1 mark. Should say that it will happen for r=1

==========

Part (d) is worth 2 marks: Should say that due to lack of conjugacy, we can't get a closed form posterior.
==========

Question 24

==========
Total marks for this question = 10

Correct MAP for gamma distribution: 2 marks
Mostly correct MAP with some minor mistakes: 1 mark

Correct Hessian: 2 marks
Mostly correct Hessian with some minor mistakes: 1 mark

Correct Laplace approximation: 2 mark (says that it is a Gaussian with mean = MAP solution and covariance matrix as the inverse of Hessian derived above)

Correct expression for the Gaussian approximation with mean being Gamma's mean and variance being Gamma's variance: 2 marks

Mostly correct expression (some minor errors) for the Gaussian approximation with mean being Gamma's mean and variance being Gamma's variance: 1 mark

Correct stating when this Gaussian approximation will be the roughly the same as the Laplace approximation: 2 marks (no partial marking for this part)
==========

Question 25

==========

Total marks for this question = 10

Correct gradient g(x,\theta): 2 marks
Mostly correct gradient with some minor mistakes: 1 marks

Correct derivation for the Fisher matrix: 4 marks
Mostly correct derivation for the Fisher matrix: 3 marks
Approach correct for deriving the Fisher matrix but has some major errors: 2 marks

Correct expression for k(x,x'): 2 marks
Mostly correct expression for k(x,x') with some minor mistakes: 1 mark

Correct intuitive meaning of k(x,x') when \Sigma = I: 2 mark
Somewhat correct explanation for the intuitive meaning of k(x,x') when \Sigma = I: 1 mark