Lecture 1: Likelihoods - Results

✕

Attempt 1 of 1

Written 03 February, 2025 10:19 AM - 03 February, 2025 10:21 AM

Attempt Score **80 %**

Overall Grade (Highest Attempt) **80 %**

Question 1 **1 / 1 point**

The likelihood measures the probability of the specific parameter value given the observed data.

- ○ True
- ✓ ◉ False

Question 2 **1 / 1 point**

The likelihood measures the probability of the data given a specific parameter value. True or False?

- ✓ ◉ True
- ○ False

Question 3 **1 / 1 point**

The maximum likelihood estimator is the parameter value that maximizes the likelihood function. True or false?

- ✓ ◉ True
- ○ False

Question 4 **1 / 1 point**

Assuming the likelihood is a binomial distribution, i.e. $p(y|\mu) = \text{Bin}(y|N, \mu)$, what is the maximum likelihood estimate for μ if we observed $N = 10$ and $y = 2$?

- ○ The question cannot be answered without additional information.
- ○ $\mu = 10/2$
- ✓ ◉ $\mu = 2/10$

Question 5 **0 / 1 point**

Assume the following model: $p(y|\mu) = \text{Bin}(y|N, \mu)$ such that the likelihood function is $L(\mu) = \text{Bin}(y|N, \mu)$. If we sum the likelihood function over all possible values of μ we always get 1.

- ✗ ◉ True
- ➡ ○ False

Hide question 5 feedback

Feedback

The correct answer is false, because the likelihood is a distribution over the observed value, but not not a distribution wrt. the parameter μ.

Done

Lecture 2: Prior, likelihood, posterior, posterior predictive - Results

×

Attempt 1 of Unlimited

Written 07 February, 2025 7:48 PM - 07 February, 2025 7:53 PM

Attempt Score **93.14 %**

Overall Grade (Highest Attempt) **93.14 %**

Question 1 **1 / 1 point**

Match the terms: prior, likelihood, posterior, joint to the corresponding distributions.

✓ __3__ posterior distribution	**1.** p(w)
✓ __5__ joint distribution	**2.** p(y|w)
✓ __2__ likelihood	**3.** p(w|y)
✓ __1__ prior distribution	**4.** p(y)
	5. p(y, w)

Question 2 **1 / 1 point**

Suppose we want to use the **product rule** to decompose the joint distribution p(y,w) into a product of likelihood and prior. Identify the correct decomposition.

- ✓ (selected) $p(y, w) = p(y|w)p(w)$
- $p(y, w) = p(y)p(w|y)$
- $p(y, w) = p(w)p(y)$
- $p(y, w) = p(w|y)p(y)$

Question 3 **1 / 1 point**

Let $p(y|w)$ denote the likelihood and let p(w) denote the prior density for w.

Suppose we want to compute the marginal distribution of y using the sum rule. Identify the corresponding equation.

- $$p(y|w) = \frac{p(y, w)}{p(w)}$$

✓ ◉

$$p(y) = \int p(y|w)p(w)dw$$

○

$$p(y) = \frac{p(y, w)}{p(w)}$$

○

$$p(w) = \int p(y|w)p(w)dy$$

Question 4 **0.8 / 1 point**

Suppose now we augment the probabilistic model with another random variable y^*.

Assume y and y^* are **conditionally independent** given w.

Identify **all** the correct decompositions of the joint distribution p(*y*, *y**, *w*).

- ✓ ☑ p(*y*, *y**, *w*) = p(*y*|*w*)p(*y**|*w*)p(*w*)
- ✓ ☐ p(*y*, *y**, *w*) = p(*w*|*y*)p(*w*|y*)p(w)
- ✗ ☐ p(*y*, *y**, *w*) = p(*y**|*w*)p(*w*)p(*y*|*w*)
- ✓ ☐ p(*y*, *y**, *w*) = p(*y**|*w*)p(*w*)p(*y*)
- ✓ ☐ p(*y*, *y**, *w*) = p(*y*)p(*y**)p(*w*)

Question 5 **0.857 / 1 point**

Finally, our goal is to compute the **posterior predictive distribution** of y^* given *y*.

Identify **all** of the correct distribution given below.

✓ ☑

$$p(y^*|y)$$

✗ ☐

$$\int p(y^*, w|y)dw$$

$$\int p(y^*, w, y) dw$$

$$\int p(y^*|w) p(w|y) dw$$

$$\int p(y^*|w) p(y|w) dy$$

$$\mathbb{E}_{p(w|y)} \left[p(y^*|y) \right]$$

$$\mathbb{E}_{p(w|y)} \left[p(y^*|w) \right]$$

Done

Lecture 3: Bayesian inference - Results

✕

Attempt 1 of Unlimited

Written 17 February, 2025 11:31 AM - 17 February, 2025 11:33 AM

Attempt Score **100 %**

Overall Grade (Highest Attempt) **100 %**

Question 1 **1 / 1 point**

Changing the prior distribution influences the posterior distribution.

- ✓ (●) True
- ( ) False

Question 2 **1 / 1 point**

Changing the prior distribution influences the likelihood.

- ( ) True
- ✓ (●) False

Question 3 **1 / 1 point**

Changing the prior distribution influences the marginal likelihood.

- ✓ (●) True
- ( ) False

Question 4 **1 / 1 point**

Changing the prior distribution influences the posterior predictive distribution.

- ✓ (●) True
- ( ) False

Question 5 **1 / 1 point**

For Bayesian linear regression with a Gaussian likelihood and conjugate Gaussian prior, which of the following statements are true?

- ✓ [x] The predictive posterior distribution is also Gaussian
- ✓ [x] For this model, the posterior mean of the parameters is always identical to the MAP solution
- ✓ [ ] For this model, the posterior mean of the parameters is always identical to the maximum likelihood solution
- ✓ [x] The posterior distribution of parameter is also Gaussian

Done

Lecture 4: Logistic regression - Results

×

Attempt 3 of Unlimited

Written 12 May, 2025 3:56 PM - 12 May, 2025 3:57 PM

Attempt Score **100 %**

Overall Grade (Highest Attempt) **100 %**

Question 1 **1 / 1 point**

This likelihood is equivalent modelling each observation with a Bernoulli distribution as follows

$$p(y_n|w_1) = \mathrm{Ber}(y_n|\sigma(w_1 x_n))$$

- ✓ (●) True
- ( ) False

Question 2 **1 / 1 point**

In standard logistic regression, we use the sigmoid function, i.e. σ(x), to prevent the model from overfitting to the data

- ( ) True
- ✓ (●) False

Question 3 **1 / 1 point**

In standard logistic regression, we use the sigmoid function, i.e. σ(x), to force the output of the model to be in the unit interval [0, 1].

- ✓ (●) True
- ( ) False

Question 4 **1 / 1 point**

The figure above shows the predicted probabilities for three different fits (i.e. three different values of w_1) . Which parameter value has the highest likelihood value? (you don't need to calculate the specific likelihood value)

- ( ) w_1 = 1
- ( ) w_1 = 2
- ✓ (●) w_1 = 3

Question 5 **1 / 1 point**

Increasing the value of w_1 will increase the likelihood for all observations.

- [x] True
- [ ] False

Question 6 **1 / 1 point**

Which value of w_1 maximizes the likelihood for this dataset?

- [ ] $w_1 = 3$
- [ ] We need more details to answer the question
- [ ] $w_1 = 0$
- [x]

$$w_1 = \infty$$

- [ ] $w_1 = -3$

Done

Lecture 5: Key equations for GP regression - Results

×

Attempt 4 of Unlimited

Written 12 May, 2025 4:00 PM - 12 May, 2025 4:00 PM

Attempt Score **100 %**

Overall Grade (Highest Attempt) **100 %**

Question 1 **1 / 1 point**

Gaussian processes can fit non-linear functions, but the posterior mean of a Gaussian process is a linear combination of the training targets

$$\mathbf{y}$$

.

- ✓ (●) True
- ( ) False

Question 2 **1 / 1 point**

Gaussian processes can easily fit non-linear trends in data, and therefore, the posterior predictive distribution is non-Gaussian.

- ( ) True
- ✓ (●) False

Question 3 **1 / 1 point**

When the measurement noise goes to infinity (i.e. $\beta \to 0$), the posterior mean approaches zero and the posterior variance approaches the prior variance?

- ✓ (●) True
- ( ) False

Question 4 **1 / 1 point**

Assuming the hyperparameters of the kernel are fixed, then the posterior variances does not depend on the observations

$$\mathbf{y}$$

- [x] True
- [ ] False

Question 5 **1 / 1 point**

The variance of the posterior distribution only depends on the observed targets, i.e.

$$\mathbf{y},$$

indirectly through hyperparameter estimation, e.g. by choosing the hyperparameters that maximized the model evidence

$$p(\mathbf{y}).$$

- [x] True
- [ ] False

Done

Lecture 6: Probabilistic neural networks - Results

×

Attempt 1 of Unlimited

Written 17 March, 2025 12:36 PM - 17 March, 2025 12:37 PM

Attempt Score **100 %**

Overall Grade (Highest Attempt) **100 %**

Given a model with likelihood p(**t**|**W**) and suppose we impose a flat prior on **w** , i.e. p(**w**) $\propto$ 1, then ...

Question 1 **1 / 1 point**

... the maximum a posterior (MAP) solution is the same as the posterior mean.

- ○ True
- ✓ ◉ False

Question 2 **1 / 1 point**

... the maximum likelihood solution and MAP (posterior mode) is the same.

- ✓ ◉ True
- ○ False

Question 3 **1 / 1 point**

... the predictive distribution for MAP is the same as that for Bayesian inference.

- ○ True
- ✓ ◉ False

Question 4 **1 / 1 point**

For models with Gaussian priors, increasing α will cause the MAP estimate of **w** to be numerically larger.

- ○ True
- ✓ ◉ False

Question 5 **1 / 1 point**

For models with Gaussian priors, increasing α will increase the strength of the regularization.

- ✓ ◉ True
- ○ False

Done

Lecture 8: The Monte Carlo Standard Error (MCSE) - Results

×

Attempt 2 of Unlimited

Written 12 May, 2025 4:03 PM - 12 May, 2025 4:06 PM

Attempt Score **100 %**

Overall Grade (Highest Attempt) **100 %**

Question 1 **1 / 1 point**

A Monte Carlo estimator has high bias and low variance.

○ True

✓ ◉ False

Question 2 **1 / 1 point**

A Monte Carlo estimator in an unbiased estimator.

✓ ◉ True

○ False

Question 3 **1 / 1 point**

Suppose

$$z^i \sim p(z)$$

are i.i.d. samples from p(z) for i = 1, ..., S and consider the Monte Carlo estimator:

$$\hat{f} = \frac{1}{S}\sum_{i=1}^{S} f(z^i)$$

then the variance of the estimator, i.e.

$$\mathbb{V}\left[\hat{f}\right] = \frac{1}{S}\mathbb{V}\left[f(z)\right]$$

can be made arbitrarily small if V[f(z)] is finite.

✓ ◉ True

- [ ] False

Question 4 **1 / 1 point**

Suppose we use S samples to estimate some function mean and the resulting Monte Carlo Standard Error (MCSE) is 10. If our goal is to reduce the MCSE by a factor of 10, how many samples S' should we use instead?

- [ ] $S' = S/10$
- [ ] $S' = 10S$
- [x] $S' = 100S$ ✓
- [ ] $S' = S$

Done

Lecture 9: Metropolis-Hastings - Results

✕

Attempt 2 of Unlimited

Written 12 May, 2025 4:16 PM - 12 May, 2025 4:17 PM

Attempt Score **100 %**

Overall Grade (Highest Attempt) **100 %**

Question 1 **1 / 1 point**

It is important to know the normalization constant of the target distribution when using Metropolis-Hastings.

○ True

✓ ◉ False

Question 2 **1 / 1 point**

Once a Metropolis-Hastings sampler reaches its stationary distribution, all future samples will be accepted.

○ True

✓ ◉ False

Question 3 **1 / 1 point**

Larger proposal variances generally lead to higher acceptance ratios.

○ True

✓ ◉ False

Question 4 **1 / 1 point**

Smaller proposal variances generally lead to higher acceptance ratios.

✓ ◉ True

○ False

Question 5 **1 / 1 point**

A higher acceptance rate is always better.

○ True

✓ ◉ False

Question 6 **1 / 1 point**

The Metropolis-Hastings algorithm is equivalent to the Metropolis algorithm when the proposal distribution is symmetric.

✓ (•) True
( ) False

Question 7 **1 / 1 point**

Stronger correlation in the target distribution generally leads to lower acceptance rates.

✓ (•) True
( ) False

Question 8 **1 / 1 point**

Increasing the number of MCMC samples generally improves the accuracy of the estimated posterior summaries.

✓ (•) True
( ) False

Question 9 **1 / 1 point**

Increasing the number of MCMC samples always improves the predictive accuracy.

( ) True
✓ (•) False

Question 10 **1 / 1 point**

The warm-up samples are discarded to speed up the computations.

( ) True
✓ (•) False

Question 11 **1 / 1 point**

The warm-up samples are discarded because they do not necessarily represent the target distribution.

✓ (•) True
( ) False

Question 12 **1 / 1 point**

MCMC eventually generates perfectly independent and identically distributed samples from the target distribution.

( ) True
✓ (•) False

Done

Lecture 9: Gibbs - Results

×

Attempt 2 of Unlimited

Written 12 May, 2025 4:20 PM - 12 May, 2025 4:20 PM

Attempt Score **100 %**

Overall Grade (Highest Attempt) **100 %**

Question 1 **1 / 1 point**

The Gibbs sampler is a special case of Metropolis-Hastings.

- ✓ (●) True
- ( ) False

Question 2 **1 / 1 point**

It's always possible to derive a Gibbs sampler for a given model.

- ( ) True
- ✓ (●) False

Question 3 **1 / 1 point**

The Gibbs sampler does not have any tuning parameters.

- ✓ (●) True
- ( ) False

Question 4 **1 / 1 point**

If we change a component of the model (e.g. a prior), we may have to re-derive the Gibbs sampler.

- ✓ (●) True
- ( ) False

Question 5 **1 / 1 point**

The acceptance rate for the Gibbs sampler decreases with the dimensionality.

- ( ) True
- ✓ (●) False

Done

Lecture 10: Variational inference - Results

✕

Attempt 1 of Unlimited

Written 28 April, 2025 10:37 AM - 28 April, 2025 10:38 AM

Attempt Score **100 %**

Overall Grade (Highest Attempt) **100 %**

Question 1 **1 / 1 point**

For a given variational family Q and a target distribution p, the optimal variational approximation q is the distribution with smallest KL divergence, i.e.

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}[q||p]$$

- ✓ (◉) True
- ( ) False

Question 2 **1 / 1 point**

The KL-divergence for the best approximation, i.e.

$$\mathrm{KL}[q^*||p]$$

where

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}[q||p]$$

is 0 if the target distribution p belongs to the variational family.

- ✓ (◉) True
- ( ) False

Question 3 **1 / 1 point**

Minimizing KL[q||p] with respect to q leads to the same solution as minimizing KL[p||q] with respect to q.

- ( ) True
- ✓ (◉) False

Question 4 **1 / 1 point**

If we change the approximation q such that the Kullbach-Leibler divergence KL[q||p] increases, the approximation q becomes a better and better approximation of p

- ○ True
- ✓ ◉ False

Question 5 **1 / 1 point**

The contribution to the KL divergence is generally large in regions, where the q is small and p is large.

$$\mathrm{KL}\left[q||p\right] = \int q(\mathbf{z}) \ln \left[\frac{q(\mathbf{z})}{p(\mathbf{z})}\right] \mathrm{d}\mathbf{z}$$

- ○ True
- ✓ ◉ False

Question 6 **1 / 1 point**

The contribution to the KL divergence is generally large in regions, where the q is large and p is small.

$$\mathrm{KL}\left[q||p\right] = \int q(\mathbf{z}) \ln \left[\frac{q(\mathbf{z})}{p(\mathbf{z})}\right] \mathrm{d}\mathbf{z}$$

- ✓ ◉ True
- ○ False

Question 7 **1 / 1 point**

Enlarging the variational family Q generally leads to a more accurate approximation.

- ✓ ◉ True
- ○ False

Question 8 **1 / 1 point**

Mean-field variational families works better because they capture correlation in the posterior

- ○ True
- ✓ ◉ False

Question 9 **1 / 1 point**

Mean-field variational families often lead to faster algorithms because it ignores the posterior correlation

- ✓ ◉ True
- ○ False

Done

Lecture 10: Mixture models - Results

×

Attempt 3 of Unlimited

Written 12 May, 2025 4:27 PM - 12 May, 2025 4:27 PM

Attempt Score **100 %**

Overall Grade (Highest Attempt) **100 %**

Question 1 **1 / 1 point**

If we switch the triplet of values for two components, i.e.

$$(\pi_k, \mu_k, \Lambda_k)$$

with

$$(\pi_j, \mu_j, \Lambda_j)$$

, then likelihood of the model changes.

- ○ True
- ✓ ◉ False

Question 2 **1 / 1 point**

The parameter

$$\Lambda_k$$

describes the number of points in the k'th cluster.

- ○ True
- ✓ ◉ False

Question 3 **1 / 1 point**

The vector

$$\pi$$

represents a probability distribution over clusters such that

$$\pi_j$$

describes the proportion of data points in the j'th cluster.

- [x] True
- [ ] False

Question 4 **1 / 1 point**

Assume one of the mixing is exactly zero, i.e.

$$\pi_j = 0$$

, then the data can equivalently be represented using K -1 components.

- [x] True
- [ ] False

Question 5 **1 / 1 point**

When modelling a data set with N observations, we need N latent variable vectors

$$\mathbf{z}_n$$

, i.e. one for each data point.

- [x] True
- [ ] False

Done

Lecture 12: Variational families - Results

Attempt 3 of Unlimited

Written 12 May, 2025 4:29 PM - 12 May, 2025 4:29 PM

Attempt Score **100 %**

Overall Grade (Highest Attempt) **100 %**

Question 1 **1 / 1 point**

For mean-field Gaussian families, the number of variational parameters grows linearly in the number of model parameters D.

- ✓ (●) True
- ( ) False

Hide question 1 feedback

Feedback

The number of variational parameters is 2D

Question 2 **1 / 1 point**

For full-rank variational families, the number of variational parameters is a cubic function of the number of model parameters D.

- ( ) True
- ✓ (●) False

Hide question 2 feedback

Feedback

The number of variational parameters is D for the mean and O(D^2) for the covariance matrix

Question 3 **1 / 1 point**

The mean-field Gaussian family is a subset of the full-rank Gaussian family. That is, all mean-field Gaussian distributions are contained in the full-rank family.

- ✓ (●) True

○ False

▼ Hide question 3 feedback

Feedback

Yes, a mean-field Gaussian is a special case of a full-rank Gaussian, where the covariance matrix is diagonal.

Question 4 **1 / 1 point**

If we change the model, we need to re-calculate and re-implement the entropy term of the ELBO

○ True

✓ ◉ False

▼ Hide question 4 feedback

Feedback

No, the entropy is independent of the model and only depends on the choice of variational family

Question 5 **1 / 1 point**

If we change the variational family, we need to re-calculate and re-implement the entropy term

✓ ◉ True

○ False

Question 6 **1 / 1 point**

Suppose our model of interest has D binary parameters, i.e.

$$\mathbf{w} \in \{0, 1\}^D$$

, instead of D continuous parameters. What would be an appropriate variational family?

○

$$q(\mathbf{w}) = \prod_{i=1}^{D} \mathcal{N}(w_i | m_i, v_i)$$

○

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \boldsymbol{m}, \boldsymbol{V})$$

$$q(\mathbf{w}) = \prod_{i=1}^{D} \mathrm{Beta}(w_i|a_i, b_i)$$

$$q(\mathbf{w}) = \prod_{i=1}^{D} \mathrm{Ber}(w_i|p_i)$$

Done

Lecture 12: BBVI - Results

×

Attempt 3 of Unlimited

Written 12 May, 2025 4:36 PM - 12 May, 2025 4:37 PM

Attempt Score **100 %**

Overall Grade (Highest Attempt) **100 %**

Question 1 **1 / 1 point**

BBVI allows us to implement and test different models without having to do explicit model-specific calculations.

- ✓ (●) True
- ( ) False

Question 2 **1 / 1 point**

When we use variational inference with multivariate Gaussian (full-rank) variational family, the resulting posterior approximation will be equivalent to the Laplace approximation.

- ( ) True
- ✓ (●) False

Question 3 **1 / 1 point**

If we run classic gradient ascent with constant step size for long enough with stochastic gradients, then it converges to a local optima.

- ( ) True
- ✓ (●) False

Question 4 **1 / 1 point**

When using stochastic gradients, in some iterations we might take a step away (i.e. in the wrong direction) from the local optima

- ✓ (●) True
- ( ) False

Question 5 **1 / 1 point**

If we run stochastic gradient ascent with Robbins-Monro step sizes for long enough with stochastic gradients, then it converges to a local optima.

- ✓ (●) True
- ( ) False

Question 6 **1 / 1 point**

If we increase S, the number of Monte Carlo samples used for the gradient estimators, the computational cost increases while the variance of the gradient decreases. True or False?

- (•) True
- ( ) False

Question 7 **1 / 1 point**

The entropy-term in the ELBO cannot be estimated using Monte Carlo samples.

- ( ) True
- (•) False

Question 8 **1 / 1 point**

The score function gradient estimator generally exhibits lower variance than the re-parametrized gradient.

- ( ) True
- (•) False

Question 9 **1 / 1 point**

Which of the following statements about BBVI is true?

- [ ] The BBVI algorithm approximates the expectations in the ELBO using efficient numerical integration
- [ ] When using the BBVI algorithm, the score function gradient estimator is generally preferred when possible
- [x] When using the BBVI algorithm, the reparametrized gradient estimator is generally preferred when possible
- [ ] Assessing convergence of the BBVI algorithm is generally easy because the ELBO is stochastic
- [x] The BBVI algorithm approximates the expectations in the ELBO using Monte Carlo sampling

Done

Lecture 13: Epistemic and aleatoric uncertainty for classification - Results

×

Attempt 1 of Unlimited

Written 12 May, 2025 1:05 PM - 12 May, 2025 1:07 PM

Attempt Score **100 %**

Overall Grade (Highest Attempt) **100 %**

Question 1 **1 / 1 point**

The uncertainty for a random variable

$$y \sim \mathrm{Ber}(p)$$

is maximized when p = 0.5

- ✓ (selected) True
- False

Question 2 **1 / 1 point**

At x ≈ 10, the GP fit shows low epistemic uncertainty.

- True
- ✓ (selected) False

Hide question 2 feedback

Feedback

Large uncertainty for the latent function values is usually due lack of data (or lack of prior information)

Question 3 **1 / 1 point**

At x ≈ 1, the GP fit shows low aleatoric uncertainty and low epistemic uncertainty.

- ✓ (selected) True
- False

Hide question 3 feedback

Feedback

At x = 1 we have low uncertainty for latent function values and low uncertainty for outcome according to the posterior predictive distribution

Question 4 **1 / 1 point**

At x ≈ 0, the GP fit shows low aleatoric uncertainty and high epistemic uncertainty.

- ○ True
- ✓ ◉ False

Hide question 4 feedback

Feedback

At x = 0, there is low uncertainty for the latent function values (because there is plenty of data nearby), but large uncertainty for outcome according to the posterior predictive distribution (because we are on the boundary between two classes).

Question 5 **1 / 1 point**

Both epistemic and aleatoric uncertainty contribute to the predictive distribution.

- ✓ ◉ True
- ○ False

Done