

HOMework 2

BAYESIAN LINEAR AND LOGISTIC REGRESSION *

10-424/10-624 BAYESIAN METHODS IN ML
<https://www.cs.cmu.edu/~hchai2/courses/10624>

OUT: 02/06/25

DUE: 02/23/25

Instructions

- **Collaboration Policy:** Please read the collaboration policy in the syllabus at <https://www.cs.cmu.edu/~hchai2/courses/10624/#Syllabus>.
- **Late Submission Policy:** See the late submission policy in the syllabus at <https://www.cs.cmu.edu/~hchai2/courses/10624/#Syllabus>.
- **Submitting your work:** You will use Gradescope to submit answers to all questions and code.
 - **Written:** We will provide you with an Overleaf template for you to complete the written portion of this homework. You may also use the raw \LaTeX source of this assignment (included in the handout .zip) to typeset your answer. **You must use \LaTeX to complete this assignment;** we will not grade any submissions that are not completed using \LaTeX and you will be asked to resubmit (with some penalty). You will submit your completed homework as a PDF to Gradescope.
 - **Programming:** We will provide you with some function headers to help you get started on the programming portion of this assignment. You will submit your completions of these function and any other code you wrote to complete the programming questions to Gradescope. There is no autograder for this assignment; instead we will manually examine your submitted code and award marks for submissions directly.
- **Materials:** Any data that you might need in order to complete this assignment is posted along with the writeup and template on the course website.

*Compiled on Tuesday 18th February, 2025 at 20:02

1 Compositions of Kernels (15 points)

One interesting and useful property of kernel functions is that it is easy to create new kernels from existing ones; this can be especially useful if there are properties or relationships that are not well modeled by any individual kernel but can be described as a function of multiple kernels. Assume that $k_1(\mathbf{x}, \mathbf{y}) = \phi_1(\mathbf{x})^T \Sigma_1 \phi_1(\mathbf{y})$ and $k_2(\mathbf{x}, \mathbf{y}) = \phi_2(\mathbf{x})^T \Sigma_2 \phi_2(\mathbf{y})$ are valid kernel functions and $c_1, c_2 \geq 0$ are non-negative (scalar) constants.

For each of the following functions, $k(\mathbf{x}, \mathbf{y})$, prove that k is a valid kernel function. There are a few different ways to prove a function is a valid kernel, including showing that the Gram matrix $k(X, X)$ is always positive semi-definite for arbitrary sets of inputs X or by expressing $k(\mathbf{x}, \mathbf{y})$ as $\phi(\mathbf{x})^T \Sigma \phi(\mathbf{y})$ for some feature expansion, ϕ , and a positive semi-definite matrix, Σ . You may use any (correct) method to justify your proof.

For full credit, you must show all your work.

- 1.1. (5 points) $k(\mathbf{x}, \mathbf{y}) = c_1 k_1(\mathbf{x}, \mathbf{y}) k_2(\mathbf{x}, \mathbf{y})$

Hint: You may find it useful to call upon the [Schur product theorem](#) in your proof.

- 1.2. (5 points) $k(\mathbf{x}, \mathbf{y}) = c_1 k_1(\mathbf{x}, \mathbf{y}) + c_2 k_2(\mathbf{x}, \mathbf{y})$

- 1.3. (5 points) $k(\mathbf{x}, \mathbf{y}) = e^{k_1(\mathbf{x}, \mathbf{y})}$

Hint: Consider the Taylor series expansion of $e^{k_1(\mathbf{x}, \mathbf{y})}$ around the point 0.

2 LASSO Regression (15 points)

LASSO regression is an alternative to ridge regression that uses ℓ^1 regularization instead of ℓ^2 regularization (this tends to lead to sparse solutions, where many elements of \mathbf{w} are zero). Formally, given data $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^n$ where $\mathbf{x}^{(i)} \in \mathbb{R}^d$, the LASSO estimator is

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^n (\mathbf{x}^{(i)\top} \mathbf{w} - y^{(i)})^2 + \lambda \sum_{j=1}^d |w_j|.$$

- 2.1. (15 points) Prove that LASSO regression is equivalent to finding the *maximum a posterior* (MAP) estimator for \mathbf{w} if the prior on \mathbf{w} is a zero mean, isotropic Laplace distribution i.e.,

$$p(w_j | s^2) = \frac{1}{2s^2} \exp\left(-\frac{|w_j|}{s^2}\right) \forall j \in \{1, \dots, d\}$$

and the residuals are modeled with independent, zero-mean Gaussians with variance σ^2 . **For full credit, you must show all your work.**

3 Laplace Approximation (20 points)

3.1. (10 points) Find a Laplace approximation to the gamma distribution:

$$p(\theta \mid \alpha, \beta) = \frac{1}{Z} \theta^{\alpha-1} \exp(-\beta\theta).$$

Plot the approximation against the true density for $(\alpha, \beta) = (3, 1)$. **For full credit, you must show all your work.**

3.2. (5 points) The true value of the normalizing constant is

$$Z = \frac{\Gamma(\alpha)}{\beta^\alpha}.$$

If we fix $\beta = 1$, then $Z = \Gamma(\alpha)$, so we may use the Laplace approximation to estimate the Gamma function (recall that the Laplace approximation also gives an approximation for the normalizing constant of the distribution being approximated). Plot your approximation of Z based on your answer to the previous part as a function of $\alpha \in [1, 5]$ and analyze the quality of your approximation.

- 3.3. (5 points) Stirling's approximation is a commonly used method for approximating factorials:

$$\log n! \approx n \log n - n + \frac{1}{2} \log(2\pi n)$$

Using the fact that $\Gamma(\alpha) = (\alpha - 1)!$, derive Stirling's approximation starting with your approximation for the Gamma function from the previous part. **For full credit, you must show all your work.**

4 Programming: Bayesian Linear Regression (50 points)

Consider the following dataset:

$$\mathbf{x} = [-1.11, -0.85, -0.76, -0.65, -0.57, -0.56, -0.20, 0.18, 0.59, 1.18]^T$$

$$\mathbf{y} = [1.24, 0.62, 0.14, 0.08, -0.23, -0.22, -1.09, -1.03, -0.35, 5.04]^T.$$

In this question, you will consider three different models that could possibly explain these data, corresponding to three different polynomial basis functions for Bayesian linear regression.

We have provided alongside this handout a file called `bayes_linreg.py`, which contains some function definitions intended to guide and structure your implementation. The use of these functions is optional but *strongly* recommended.

- 4.1. (10 points) Perform Bayesian linear regression on this data set using the polynomial basis functions $\phi_k(x) = [1, x, x^2, \dots, x^k]^\top$ for $k \in \{1, 2, 3\}$; in each case, use the parameter prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{I})$ and fix the noise variance at $\sigma^2 = 1^2$.

Evaluate and plot the posterior means $\mathbb{E}[\mathbf{y}^* | \mathbf{X}^*, \mathcal{D}, \sigma^2]$ on the interval $x^* \in [-4, 4]$ for each model (each value for k). Also plot the posterior mean plus-or-minus two times the posterior standard deviation:

$$\mathbb{E}[\mathbf{y}^* | \mathbf{X}^*, \mathcal{D}, \sigma^2] \pm 2\sqrt{\text{var}[\mathbf{y}^* | \mathbf{X}^*, \mathcal{D}, \sigma^2]}.$$

This (roughly) corresponds to a pointwise 95% credible interval for the regression function.

$k = 1$



$k = 2$ $k = 3$

- 4.2. (5 points) The marginal likelihood of the data is defined as how likely the data is under some specified probabilistic model. For the Bayesian linear regression setting considered here, the marginal likelihood of model k is $p(\mathbf{y} \mid \mathbf{X}, k, \sigma^2) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \phi_k(\mathbf{X})\phi_k(\mathbf{X})^T + \sigma^2\mathbf{I})$.

Compute the *log* marginal likelihood for each model $k \in \{2, 3\}$. As a means of evaluating your implementations, we have provided the log marginal likelihood for $k = 1$ in the table below; you should fill in the two missing values. **Round your answer to four decimal places.** Which model best explains the data i.e. which model has the highest marginal likelihood?

Hint: if you directly compute the marginal likelihood and try to take the log of that value, you may encounter numerical issues. Instead, derive an expression for the log of a Gaussian probability density function and have your code implement that (more stable) computation.

k	$p(\mathbf{y} \mid \mathbf{X}, k, \sigma^2 = 1^2)$
1	-23.1699
2	
3	

- 4.3. (10 points) Next, perform Bayesian linear regression on the same dataset except with the noise variance decreased to $\sigma^2 = 0.1^2$.

Again, evaluate and plot the posterior means as well as the posterior mean plus-or-minus two times the posterior standard deviation for each model $k \in \{1, 2, 3\}$. Comment on the effect of decreasing the noise variance on your posterior credible interval: does this qualitatively seem like a better or worse setting of σ^2 than the previous value you explored?

Comparison with 4.1

$k = 1$

$k = 2$ $k = 3$

- 4.4. (10 points) Finally, apply your Bayesian linear regression implementation to the real-world dataset provided to you in the handout. This (rather old) dataset contains information about (correspondingly old) cars and the goal is to predict the city-cycle fuel consumption in miles per gallon. You can read more about the dataset [here](#).

In the handout, you should find four files `X_test.csv`, `X_train.csv`, `y_train.csv`, `y_test.csv`. We have already begun pre-processed the dataset for you by removing entries with missing values and irrelevant features. Your first task will be to complete pre-processing by normalizing the datasets: subtract the mean of each feature and divide by the standard deviation.

Next, using the partition of the data given in the handout, compute the *log* marginal likelihood, $p(\mathbf{y} \mid \mathbf{X}, k, \sigma^2)$ of each model $k \in \{1, 2, 3\}$ as well as the *log* posterior predictive likelihood on the validation dataset, $p(\mathbf{y}^* \mid \mathbf{X}^*, \mathbf{y}, \mathbf{X}, k, \sigma^2)$. Again, fix the noise variance at $\sigma^2 = 1^2$. We have again provided the log marginal likelihood for $k = 1$ in the table below; you should fill in the five missing values. **Round your answer to four decimal places.**

k	$\log(p(\mathbf{y} \mid \mathbf{X}, k, \sigma^2))$
1	-2524.5633
2	
3	

k	$\log(p(\mathbf{y}^* \mid \mathbf{X}^*, \mathbf{y}, \mathbf{X}, k, \sigma^2))$
1	
2	
3	

- 4.5. (15 points) **Manual Grading of Code:** Submit a single file titled `bayes_linreg.py` that contains all the code you wrote (including code used to generate figures) for this problem to Gradescope. We will manually inspect your code to evaluate the correctness of your implementation and award points accordingly

5 Collaboration Questions (0 points)

After you have completed all other components of this assignment, report your answers to these questions regarding the collaboration policy. Details of the policy can be found in the syllabus.

- 5.1. Did you collaborate with anyone on this assignment? If so, list their name or Andrew ID and which problems you worked together on.

- 5.2. Did you find or come across code that implements any part of this assignment? If so, include full details.