Bayesian Machine Learning

Instructor: Andrew Gordon Wilson

Homework 1
Due: Tuesday September 14 (EOD) via NYU Brightspace

Show all steps, and any code used to answer the questions.

1. Suppose we have data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, and n is the total number of training points. Assume we want to learn the regression model

$$y = ax + \epsilon_x \,, \tag{1}$$

where ϵ_x is independent zero mean Gaussian noise with variance σ^2: $\epsilon_x \sim \mathcal{N}(0, \sigma^2)$.

(a) (2 marks): Let $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and $X = \{x_i\}_{i=1}^n$. Derive the log likelihood for the whole training set, $\log p(\mathbf{y}|X, a, \sigma^2)$.

(b) (2 marks): Given data $\mathcal{D} = \{(4, 21), (9, 59), (7, 25), (15, 127)\}$, find the maximum likelihood solutions for a and σ^2.

(c) (2 marks): Suppose we instead consider the regression model

$$x = by + \epsilon \,. \tag{2}$$

Is the maximum likelihood solution for $b = \frac{1}{a}$? Explain why or why not – with derivations if necessary.

(d) (2 marks): Suppose we place a prior distribution on a such that $p(a|\gamma^2) = \mathcal{N}(0, \gamma^2)$. Use the sum and product rules of probability to write down the *marginal likelihood* of the data, $p(\mathbf{y}|X, \sigma^2, \gamma^2)$, conditioned only on X, σ^2, γ^2.

(e) (2 marks): Without explicitly using the sum and product rules, derive $p(\mathbf{y}|X, \sigma^2, \gamma^2)$, by considering the properties of Gaussian distributions and finding expectations and covariances. This expression should look different than your answer to the previous question. Comment on the differences in computational complexity. **Bonus (1 mark)**: show that both representations in (d) and (e) are mathematically equivalent.

(f) (2 marks): What are the maximum marginal likelihood solutions $\hat{\sigma}^2 = \text{argmax}_{\sigma^2} p(\mathbf{y}|X, \sigma^2, \gamma^2)$ and $\hat{\gamma}^2 = \text{argmax}_{\gamma^2} p(\mathbf{y}|X, \sigma^2, \gamma^2)$?

(g) (2 marks): Derive the predictive distribution for $p(y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, \mathcal{D})$ for any arbitrary test point x_*, where $y_* = y(x_*)$.

(h) (2 marks): For the dataset $\mathcal{D}$ in (b), give the predictive mean $\mathbb{E}[y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, \mathcal{D}]$ and predictive variance $\text{var}(y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, \mathcal{D})$ for $x_* = 14$.

(i) (2 marks): Suppose we replace x in Eq. (1) with $g(x, w)$, where g is a non-linear function parametrized by w, and $w \sim \mathcal{N}(0, \lambda^2)$: e.g., $g(x, w) = \cos(wx)$. Can you write down an analytic expression for $p(\mathbf{y}|w, X, \sigma^2, \gamma^2)$? How about $p(\mathbf{y}|X, \sigma^2, \gamma^2, \lambda^2)$? Justify your answers.