Small Problem 1: Bayesian Linear Regression

Summary

Given:

A set of training points $\{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$

Generative model for the data in terms of the weight vector w and hyperparameters, as described below.

Find:

Query 1: The posterior probability distribution of w.

Metrics:

Metric 1: Expected squared Euclidean distance between the predicted mean $\hat{w}$ and the true mean w, where the expectation is taken with respect to the posterior distribution.

Metric 2: Total variation distance between the computed posterior and the correct posterior over w.

Details:

The file "problem-1-generator.R" contains R code to generate the true regression coefficients and the input training data. The model is

$$
\begin{aligned}
\Sigma_1 &= 2\mathbf{I}_{5\times 5} \\
\mu &\sim \mathcal{N}(0, \Sigma_1) \\
\Sigma_2 &= \mathbf{I}_{5\times 5} \\
\Sigma_{\text{prior}} &\sim \text{Wishart}(1, \Sigma_2) \\
w &\sim \mathcal{N}\left(\mu, \Sigma_{\text{prior}}^{-1}\right) \\
x_{ij} &\sim \text{Uniform}(-1, 1) \\
\tau &\sim \text{Gamma}(0.5, 2) \\
\epsilon_i &\sim \mathcal{N}\left(0, \frac{1}{\tau}\right) \\
y_i &= \sum_j x_{ij} w_j + \epsilon_i
\end{aligned}
$$

The file contains 500 training examples generated from a single run of the R code. There are four covariates generated uniformly from $[-1, 1]$. The values of the variables that generated the data are

$$\begin{aligned} \mu &= (-1.8195312, 1.2237587, 0.8361809, -2.6017006, -2.3574193) \\ \Sigma_{\text{prior}} &= \text{(see "problem-1-prior.Sigma.csv")} \\ w &= (-1.731855, 2.986017, 2.698284, -3.591651, -3.714157) \\ (x_i, y_i) &= \text{(see "problem-1-data.csv")} \end{aligned}$$

Queries/Metrics:

1. Let $P(\hat{w}|D)$ be the posterior distribution of the estimated weight vector. One metric is the expected squared error $\mathbb{E}\left[\|\hat{w} - w\|^2\right]$ under this distribution.
2. We have provided samples generated from the true posterior distribution $P_{\text{true}}(\hat{w}|D)$. We can estimate the total variation distance between the true distribution and your estimate $P(\hat{w}|D)$ using the samples generated by your estimated distribution:

$$\int_w |P(\hat{w}|D) - P_{\text{true}}(\hat{w}|D)|\, dw$$

Submission:

The metric value should be computed for each elapsed time step (by calling the provided code or by implementing yourself). The metric value should be reported for several elapsed time steps. The number of elapsed time steps should be sufficient to establish an "informative profile".

For further details regarding submission of the metric and your code, please refer to the main CP4 problem description document, e.g. PPAML-Challenge-Problem-4.pdf.

Sample output for this problem has been provided in the "sampleoutput" folder:

problem-1-query-1-metric-1.csv
problem-1-query-1-metric-2.csv