Bayesian Optimal Design - Weekly Report

Rasmus Hag Løvstad, pgq596

May 1, 2023

1 Week 1

1.1 Objective

The objective of this week is to explore the basics of probabilistic machine learning by implementing a Bayesian linear regression model and explore how it is affected by its different hyperparameters like choice of prior as well as the variance and quantity of data points.

1.2 Theory

The linear model I've chosen to implement is based on the one described in the lecture notes for the PML course.

Given a dataset $\mathcal{D}$ of ℓ datapoints, we let each datapoint i consist of a vector of controlled variables $\mathbf{x}^{(i)}$ and some target variable $y^{(i)}$. The $\mathbf{x}$'s are encoded in a $\ell \times d$ matrix $\mathcal{X}$ and the y's in a ℓ length vector. We are interested in finding a parameter vector θ such that for any i, $f_\theta(x^{(i)}) = \theta^T x^{(i)} \approx y^{(i)}$.

TODO: add citations. Rough draft - terminology will be revised later

To do this, we will start by assuming that the y's are generated by some latent linear model g with some noise ϵ:

$$y^{(i)} = g(\mathbf{x}^{(i)}) + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma_y^2)$ for some σ_y^2.

As is standard practice in linear regression, adding an extra dimension to $\mathcal{X}$ consisting of 1's will be done such that the linear model can account for offset in the data.

We are interested in finding measuring how likely it is for each possible parameter vector to have generated the dataset, which we will measure using the posterior distribution $p(\theta|\mathcal{D}) = p(\theta|\mathcal{X}, \mathcal{Y})$. Using Bayes' rule, we can write this as:

$$p(\theta|\mathcal{X}, \mathcal{Y}) = p(\theta|\mathcal{Y}, \mathcal{X}) = \frac{p(\theta|\mathcal{X})p(\mathcal{Y}|\theta, \mathcal{X})}{p(\mathcal{Y}|\mathcal{X})} = p(\theta|\mathcal{Y}, \mathcal{X}) = \frac{p(\theta)p(\mathcal{Y}|\theta, \mathcal{X})}{p(\mathcal{Y}|\mathcal{X})}$$

where $p(\mathcal{Y}|\mathcal{X})$ is only a normalization constant and can be ignored, since we're interested in finding which θ optimizes the posterior, not the actual value of the posterior. We also assume that $p(\theta|\mathcal{X}) = p(\theta)$ since we assume that the weights of the underlying distribution are independent of the observations.

The prior $p(\theta|\mathcal{X})$ will be assumed to be normal such that $p(\theta|\mathcal{X}) = \mathcal{N}(\theta, \mu_\theta, \Sigma_\theta)$.

By our assumption of the noise ϵ being normally distributed, we can thus also assume that the likelihood $p(\mathcal{Y}|\theta, \mathcal{X})$ is normally distributed. Thus the prior and likelihood are conjugate and we can assume the posterior to be normal as well. This means we can calculate the posterior exactly.

TODO: show why likelihood is normal

insert formula here

TODO: litterature says need to compute $p(\mathcal{D}, \theta)$ - find out why

1.3 Design

There exists a multitude of hyperparameters that can affect the posterior distribution of the model parameters.

TODO: Insert formula, insert calculations for μ and Σ, text about MAP vs Mean vs Posterior Predictive etc.

These can be controlled by the model designer and include

- Choice of prior mean
- Choice of prior variance
- Choice of predictive function or distribution (MAP vs Mean vs Predictive Posterior)

TODO: linear regression is invariant to translation

The model can also be subjected to different kinds of data sets, which can vary by

- Underlying distribution generating the data points
- Number of data points
- Variance of data points (i.e. magnitude of noise)
- Dimensionality

To study the model's behavior, I will first train the same model on multiple data sets generated by the same underlying distribution, but with different amounts of data points. Then, I will train the model on the same amount of data generated by the same distributino, but with different amounts of noise. Finally, I will train the model on the same data set, but with different choices of parameters for the prior.

1.4 Results

1.4.1 Varying number of data points

TODO: Add section for change of noise

TODO: Make plots nicer, add "true" line for reference, add legend

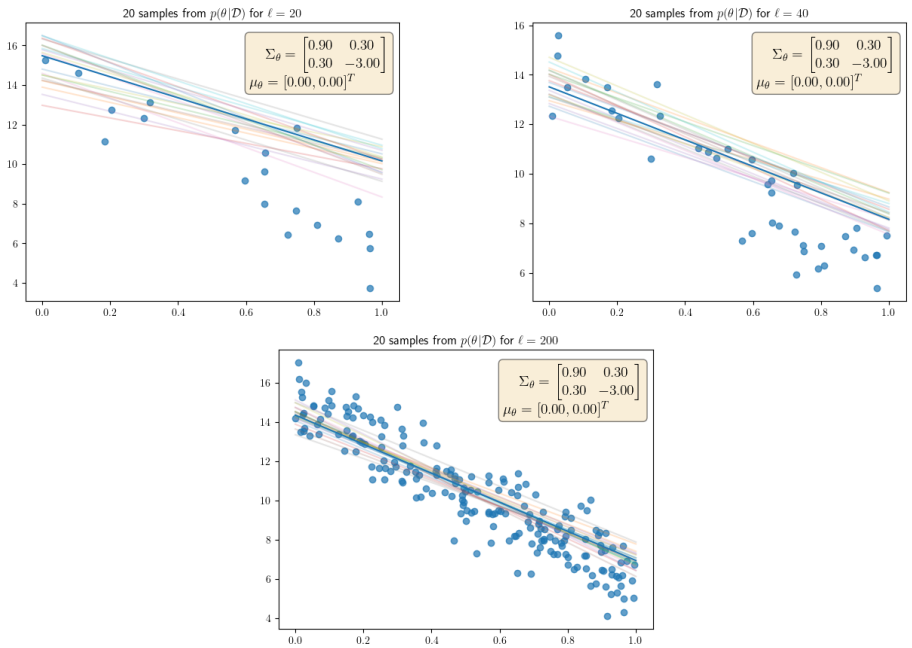

Figure 1: The model trained on data sets with 20, 40 and 200 data points respectively with a fixed prior. The blue line is $\mu_{\theta|\mathcal{D}}$.

1.4.2 Variable prior

TODO: make explicit that it is Σ_θ that has been sampled

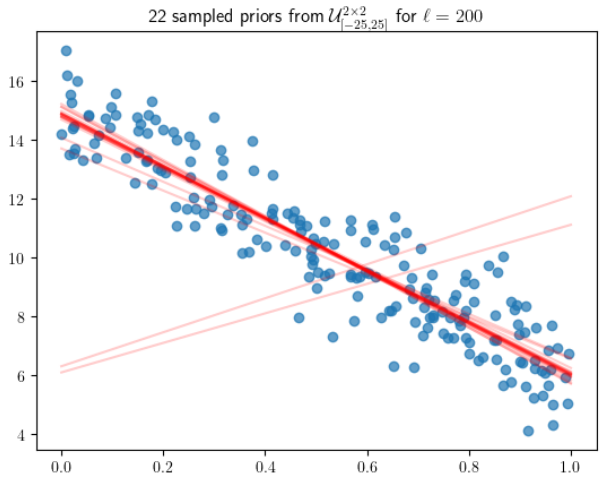

Figure 2: $\mu_{\theta|\mathcal{D}}$ for 22 sampled priors for $\ell = 200$

1.4.3 Posterior predictive function

TODO: fix plot, maybe include Σ_θ? include $\mu_{\theta|\mathcal{D}}$

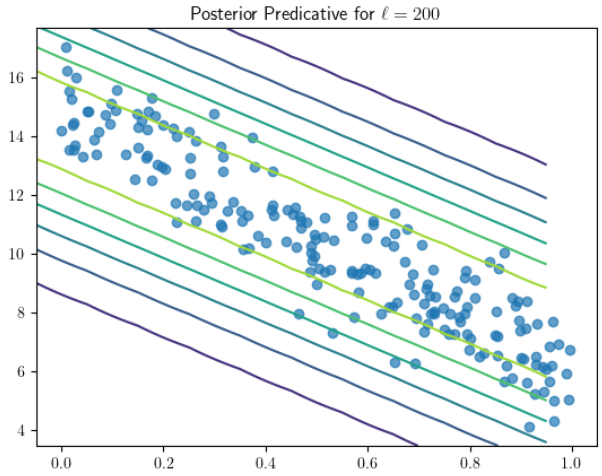

Figure 3: Contour plot of the posterior predicative distribution for each value along the x axis.

1.5 Evaluation

From figure 1, we see that the model is very sensitive to the amount of datapoints in the dataset. The last plot with $\ell = 200$, we can see that larger amounts of data is not sufficient enough when the prior is bad - a good model should have both sufficient data and a good prior.

TODO: add sublabels

From figure 2, we can see that when data is sufficient, sampling random covariance matrices for the prior is likely to yield a good result. From figure 3, we can see that the posterior predictive distribution also looks to be quite representative of the data points, since the points become much more scarse as the become more distant from the yellow region.