Mathematics for New Technologies in Finance

Solution sheet 7

Exercise 7.1 (Bayesian optimization)

(a) Recall the definition of prior, likelihood, posterior, and evidence distributions in bayesian statistics.

(b) Consider linear model on $\mathbb{R}$: $Y \sim \theta X + Z$, $\theta \sim \mathcal{N}(0,1)$, $Z \sim \mathcal{N}(0,1)$, and θ independent with X. Compute $p_\theta(y|x)$ and $p(\theta|x,y)$. Prove that maximizing the posterior $p(\theta|x,y)$ is exactly doing Ridge regression (fix λ here).

(c) Consider Lasso regression, what is the prior under Bayesian perspective? Please calculate the posterior under this prior.

(d) Would you expect a sparser weight or denser weight using Lasso regression instead of Ridge regression.

Solution 7.1

(a) Posterior = Likelihood * Prior / Evidence

(b) See the proof here.

(c) Sparser for Lasso.

Exercise 7.2 (Stochastic gradient descent)

(a) Assume that we aim to find the θ^* to maximize the posterior:

$$p(\theta|x_1, \cdots, x_n) = \frac{p(\theta)\prod_{i=1}^n p(x_i|\theta)}{p(x_1, \cdots, x_n)} \tag{1}$$

with stochastic gradient descent method in practice. In each step, do we calculate $\nabla p(\theta|x_1, \cdots, x_n)$? do we calculate $\nabla \log p(\theta|x_1, \cdots, x_n)$? do we calculate $\nabla \log p(\theta)$ or $\nabla \log p(x_i|\theta)$?

(b) If $p(x_1, \cdots, x_n)$ has no closed formula, does it cause a trouble when we do stochastic gradient descent?

(c) Construct a stochastic differential equation with invariant measure to be the posterior distribution $p(\theta|x_1, \cdots, x_n)$.

Solution 7.2

(a) We calculate $\nabla \log p(\theta)$ and $\nabla \log p(x_i|\theta)$.

(b) No, because this is a scaling term

(c) See lecture notebook 3

References

[1] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.