Coursework 3 : Mathematics for Machine Learning (CO-496)

Vincent Jarasse

15/11/2018

1 Bayesian Linear Regression

1.a

When using the model

$$P(y, w \mid x, \alpha, \beta) = (\prod_{i=1}^{N} \mathcal{N}(y_i \mid w^T \phi_i, \beta))\mathcal{N}(w \mid 0, \alpha I)$$

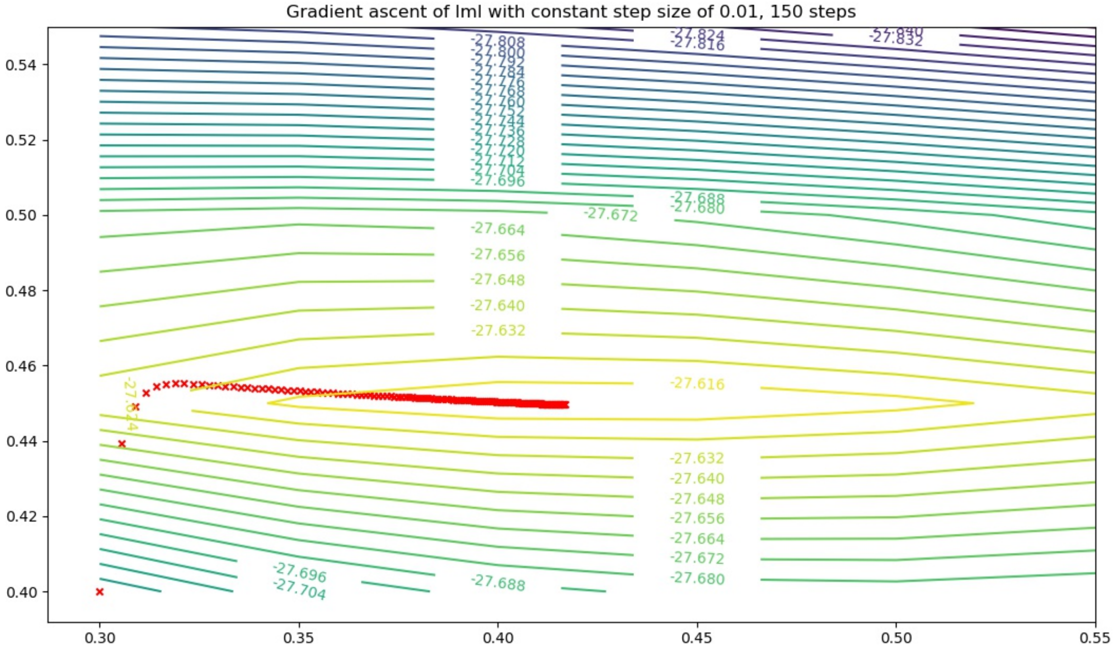
and with the gaussian multiplication formulas, we obtain in the end this log marginal likelihood :

$$log(P(y \mid x)) = -\frac{N}{2} log(2\pi) - \frac{1}{2} log(det(\alpha\phi\phi^T + \beta I)) - \frac{1}{2} y^T(\alpha\phi\phi^T + \beta I)^{-1} y$$

From which we can compute the derivative w.r.t alpha and beta. Please refer to the code submitted via gitlab for this question.

1.b

We maximise the log marginal likelihood by doing a gradient ascent. See below the steps on the contour plot. I used the starting point [0.3,0.4] with 150 steps of size 0.01.
The corresponding values for α and β are $\alpha = 0.42455482$ and $\beta = 0.44923134$.

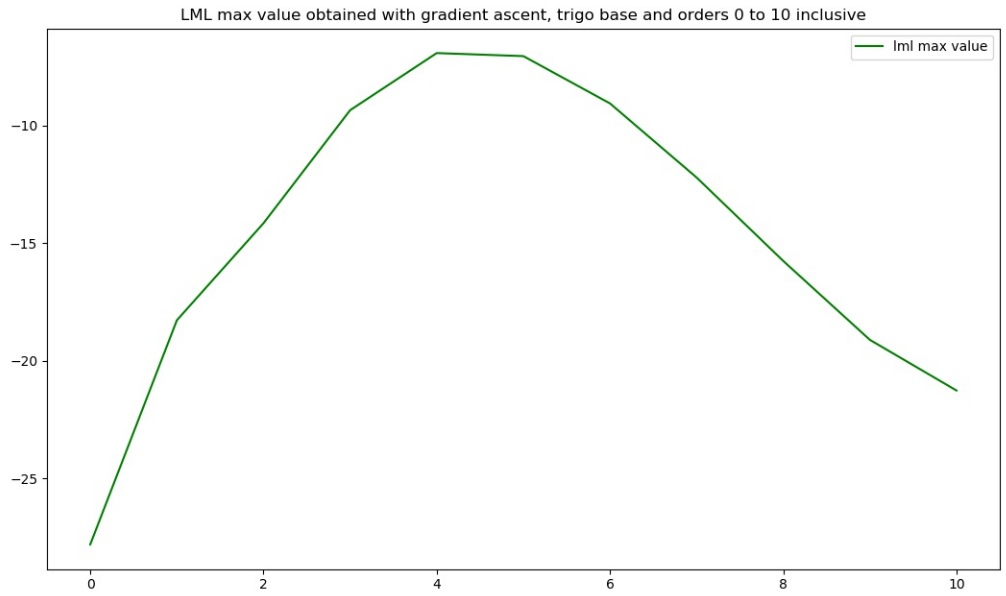


1.c

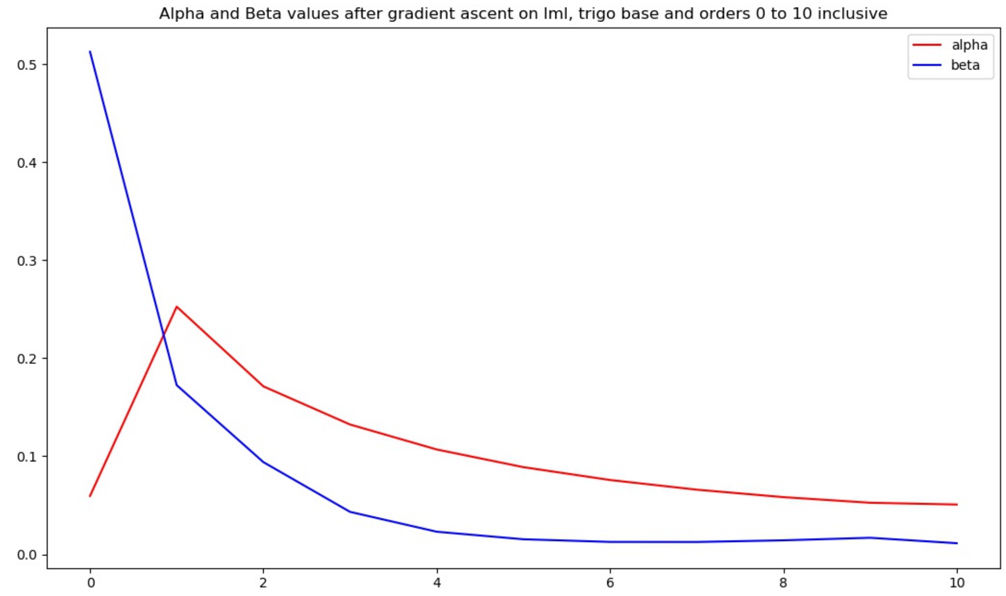
We apply eleven times the gradient ascent on the lml for orders 0 to 10 inclusive (as it was said not to do the 11th order). I used a little optimisation for the gradient ascent algorithm : when the next value is higher than the current one, I increase the step size by 1.5. When the step size is too high and thus the next value is less than the current one, I undo the step and divide the step size by two. This allows me to begin with a very little step size (0.00001) and "only" compute 500 steps, with a convergence at the end.

The graph below is the lml max value against the order of the basis. I also plotted the corresponding values for alpha and beta.

Maximising the lml has the advantage of being less demanding in terms of computations compared to cross validation, and we can deal with many hyperparameters at once. However, cross validation gives a quantitative measure of the generalisability of the model, which is not possible here.



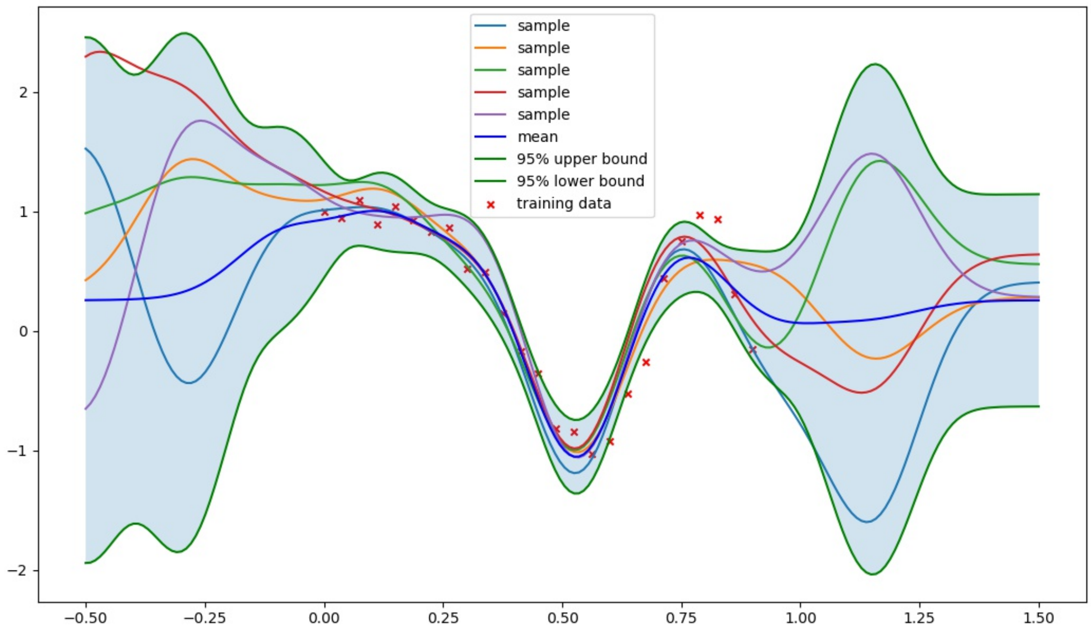
LML max value obtained with gradient ascent, trigo base and orders 0 to 10 inclusive



Alpha and Beta values after gradient ascent on lml, trigo base and orders 0 to 10 inclusive

1.d

Below is the plot of noise-free predicted function using 5 samples for the posterior distribution. The blue curve is the mean function (which corresponds to the MAP estimate). The blue shaded region between the green curves is the 95 % confidence region. We can see the erratic behaviour of the predictions when too far away from the gaussian base means.


sample
sample
sample
sample
sample
mean
95% upper bound
95% lower bound
training data
2
1
0
−1
−2
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50