LINKÖPING UNIVERSITY
Department of Computer and Information Science
Division of Statistics
Mattias Villani

2014-03-29
Advanced Bayesian Learning
PhD course

Gaussian Processes - Computer Lab

Deadline: April 7 at midnight
Teacher: Mattias Villani
Grades: Pass/Fail
Submission: By email to Mattias Villani, use message header: 'ABL - Gaussian Processes'

The lab is to be reported in a concise report.
You have to write your own code to solve the problems, in whatever language you like.
You are not allowed to use existing GP toolboxes
Existing basic functions for matrix algebra, random number generators etc are allowed.
Attach your code to the email as separate executable files.

---

1. GP Regression. Analyze the Canadian wages data (text file is available in the GitHub repo and on the course page) using a Gaussian process. You probably want to standardarize the Age variable to have zero mean and unit variance. At a minimum you should report the following results:

   (a) Plot of the posterior mean of f and 95% probability intervals.

   (b) Plot of the predictive mean for a new observation y (as a function of x) and 95% predictive intervals. Overlay the training data in the plot.

   (c) Analysis of the effect of different kernel (at least three of them) and inference for the hyperparameters in the kernels (using the marginal likelihood and possibly also cross-validation if you have the time).

2. GP classification. South African heart-disease data. A description of the data set is available here: http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.info.txt, and the actual data can be downloaded here: http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data. See also Section 5.2.2 of the book http://statweb.stanford.edu/~tibs/ElemStatLearn/. Your task is to fit a logistic GP for the binary response *coronary heart disease* (chd) using the following variables as explanatory variables: sdp, tobacco, ldl, obesity and age. Do the following:

   (a) Use standard software (no need to code this part by yourself) to fit a linear logistic regression using only two input variables age and obesity. Plot the posterior probability $Pr(\text{chd} = 1|\mathbf{x})$ in the two-dimensional input space. Overlay the data. Evaluate the accuracy and precision of the classifier using 10-fold cross-validation.

   (b) Implement the Laplace approximation for the logistic case. Use the SE with ARD kernel (separate length scales for all inputs).

(c) Fit a GP classifier using the same two variables used in 2a) as inputs. Plot the posterior probability $Pr(\text{chd} = 1|\mathbf{x})$ in the two-dimensional input space. Overlay the data. Evaluate the accuracy and precision of the classifier using 10-fold cross-validation. Compare with results from 2a).

(d) Fit a GP classifier using all five explanatory variables. Evaluate the accuracy and precision of the classifier.

(e) Bonus question: If you have an ocean of time you may want to compare the GP with other commonly used classifiers (logistic regression with additive splines, naive Bayes, SVM, random forest etc etc). Here you can use standard software packages, no need to code this yourself. This problem is just for fun, no need to do it to pass the course.

Good luck! Hopefully things will go smoothly. If not, just tune some hyperparameter or change the kernel!