Machine Learning 2
Exercise Sheet 3

Prof. Dr. Dr. Lars Schmidt-Thieme, Brad Baker
Information Systems and Machine Learning Lab
University of Hildesheim

May 3rd, 2016

Submission until May 9th, 8:00 AM by learnweb.
Please put your name in all filenames and somewhere visible in the margins of the pdf.
Non-pdf submissions for non-programming exercises will not be graded.

For further reading on Gaussian Processes specifically, see Rasmussen & Williams' 2006 book "Gaussian Processes for Machine Learning" (`http://www.gaussianprocess.org/gpml/chapters/RW.pdf`). A link to this pdf has been provided on the Moodle.

Exercise 5: Gaussian Processes and Other Models (10 Points)

a) (1 point) What is the kernel function for a linear classification or regression model?

b) (6 points) Consider linear regression for D many features, with a prior on the parameters

$$p(w) = \mathcal{N}(0, \Sigma)$$

. The posterior predictive distribution is given by

$$p(f_*|x_*, X, y) = \mathcal{N}(\mu, \sigma^2) \tag{1}$$

$$\mu = \frac{1}{\sigma_y^2} x_*^T A^{-1} X^T y \tag{2}$$

$$\sigma^2 = x_*^T A^{-1} x_* \tag{3}$$

where $A = \frac{1}{\sigma_y^2} X^T X + \Sigma^{-1}$.

Show that we can "kernelize" linear regression, i.e. that we can define a sensible kernel function, and then show how this kernelized model is equivalent to a gaussian process.

Are there any potential problems with this resulting gaussian process model?

c) (3 points) How might we go about casting a basic support vector machine as a Gaussian Process? i.e. describe how you might kernelize a basic SVM. Are there any problems with converting SVMs into Gaussian Processes?

Exercise 6: Gaussian Processes for Classification (10 Points)

a) (1 points) What kinds of approximative methods are required for Gaussian Process Classification, and why are such methods required?

b) (7 points) Consider the 3-class softmax function

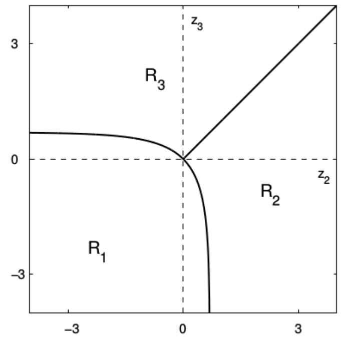
$$p(C_c) = \frac{exp(f_c)}{exp(f_1) + exp(f_2) + exp(f_3)},$$

where $c = 1, 2, 3$ and f_1, f_2, f_3 are the corresponding activation functions.

Let $z_2 = f_2 - f_1$ and $z_3 = f_3 - f_1$, then

$$p(C_1) = \frac{1}{1 + exp(z_2) + exp(z_3)},$$

and similarly for the other classes. The decision boundary relating to $p(C_1) > 1/3$ is the curve $exp(z_2) + exp(z_3) = 2$. The decision regions for the three classes are illustrated in the figure below.



Let $\mho = (f_1, f_2, f_3)^T$ have a gaussian distribution with zero mean, and let $\pi(f) = softmax(f)$. Consider the effect of this distribution on

$$\bar{\pi} = \int \pi(f) p(f) df$$

. For a Gaussian with a chosen covariance, we approximate this integral by drawing samples from $p(f)$. Show that the classification can be made to fall into any of the given decision regions depending on the covariance matrix.

c) (2 points) Read the following paper on approaches to Gaussian Processes for classification: `http://papers.nips.cc/paper/1694-efficient-approaches-to-gaussian-process-classification.pdf`. Briefly summarize one or all of the approaches taken here, and compare with the approximation methods used in class.

Bonus 3: Gaussian Process Classification with R

a) Download the USPS data set from `http://www.gaussianprocess.org/gpml/data/`. read the R documentation for the LaplacesDemon Laplace approximation library `https://artax.karlin.mff.cuni.cz/r-help/library/LaplacesDemon/html/LaplaceApproximation.html`. Preprocess the data as indicated on page 63 of "Gaussian Processes for Machine Learning", in section 3.7.3, and perform Gaussian Process Classification on the USP data set using Laplace Approximation. Output the misclassification rate for your implementation, and perform n-fold cross validation for a sensible value of n.

b) Use your implementation from the previous problem to explore how your model behaves with respect to the hyper-parameters l and σ_f, using a small grid-search. Plot the results and describe the effects of the two hyper-parameters.