NICTA - Statistical Machine Learning advanced course

Third assignment: Prior knowledge, model assessment and selection

The goal of this assignment is to compare different framework used to introduce prior knowledge in a learning problem to find out that they can lead to the same kind of optimization problem.

In this assignment we will consider the following regression (or prediction) problem: let $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ and $\mathbb{P}(x, y)$ be an unknown joint density $\mathbb{P}(x, y)$. Given

- a training sample $(x_i, y_i), i = 1, m$ i.i.d. from $\mathbb{P}(x, y)$,
- a kernel $k(x, y)$ and a set $H = \{f : \mathbb{R}^d \to \mathbb{R}; f(x) = \sum_{i=1}^m \alpha_i k(x, x_i), \alpha_i \in \mathbb{R}\}$,
- a quadratic cost function $J(x, y, f) = (f(x) - y)^2$

find a function $f \in \mathcal{H}$ that for a given x gives the best prediction for y.

1. The target function of the learning procedure is the one minimizing the expected cost. Formulate it as a function of the conditional density of y given x.

SRM
(a) Define the empirical error and the generalization error,
(b) find out a Vapnik bound that controls the variation of the empirical risk around its expectation in the regression case[1],
(c) express the learning problem in terms of minimization of the empirical risk plus the bound.

Regularization
(a) Formulate the learning problem as the minimization of the empirical error using Tikhonov regularization.
(b) show that the solution of this problem is given by $\alpha = (K + \lambda I)^{-1} y$
(c) what is the effect of the Tikhonov regularization on the eigen values of the gram matrix.

Bayes
(a) Formulate the learning problem within the Bayesian framework using a Gaussian distribution for the error $f(x_i) - y_i$
(b) recall the Bayes theorem and show how to use it with a Gaussian prior (with isotropic equal variance) by doing Maximum a posteriori (MAP). Formulate the function to be minimized as the sum of two terms.

MDL
(a) Express the regression problem in term of minimum description length principle using the length of the data and length of the model.
(b) what is the length of a code needed to encode a probability distribution and its entropy?
(c) put it in relation with the Bayesian and the frequentist point of view.

Prior Sumarize previous results in a table showing the different priors used, the criterion to be minimized and some comments. Why do we need prior knowledge to solve this problem?

Occam Who was Occam and why is it known to machine learning researchers. What is the relationship between Occam's principle, prior knowledge, the "*no free lunch*" theorem and the "*Ugly Duckling*" theorem. Give a simple example illustrating these statements.

[1] see for instance http://www.ece.umn.edu/users/cherkass/comparison_paper_Mar05.pdf