MVA "Kernel methods"
Homework 4

Jean-Philippe Vert

Due February 26, 2014

Exercice 1. Kernel logistic regression

Given a training set of labeled data $(x_i, y_i)_{i=1,\dots,n}$, with $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$, we consider the ridge logistic regression problem:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n \log\left(1 + e^{-y_i \beta^\top x_i}\right) + \lambda \| \beta \|_2^2 \right\} . \tag{1}$$

a. Show that (1) is equivalent to the penalized maximum likelihood model:

$$\max_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n \log P_\beta \left(Y = y_i \,|\, X = x_i\right) - \lambda \| \beta \|_2^2 \right\} ,$$

for the logit model:

$$\log \frac{P_\beta(Y = 1 \,|\, X = x)}{P_\beta(Y = -1 \,|\, X = x)} = \beta^\top x \, .$$

b. Show that the solution β^* of (1) satisfies

$$\beta^* = \sum_{i=1}^n \alpha_i^* x_i \, ,$$

where $\alpha^* \in \mathbb{R}^n$ is the solution of an optimization of the form

$$\min_{\alpha \in \mathbb{R}^n} \left\{ R(K\alpha) + \lambda \alpha^\top K \alpha \right\} , \tag{2}$$

where K is the $n \times n$ kernel Gram matrix with entries $K_{ij} = x_i^\top x_j$.
c. Rewriting (2) as

$$\min_{\alpha \in \mathbb{R}^n, u \in \mathbb{R}^n} \left\{ R(u) + \lambda \alpha^\top K \alpha \right\} \quad \text{such that} \quad K\alpha = u \,,$$

write the dual problem corresponding to this constrained minimization problem of the form:

$$\max_{\gamma \in \mathbb{R}^n} Q(\gamma) \,, \tag{3}$$

and explain how α^* is obtained from the solution γ^* of the dual problem.
d. For each of the three optimization problems (1), (2) and (3), compute the gradient and Hessian of the objective function.
e. In the language of your choice, implement a simple Newton-Raphson optimization that solves each of the three problems of the form:

$$\min_v F(v)$$

by iterating the steps:

$$v^{new} \leftarrow v^{old} - \left[\nabla_v^2 F \left(v^{old} \right) \right]^{-1} \nabla_v F(v^{old}) \,.$$

f. For different values of n, p (typically varying between 10 and 10^5, depending on your computer), generate $N_{train} = n$ training data and $N_{test} = 1000$ test data according to two separated Gaussian distributions for the two classes $y = \pm 1$. Use the training set to estimate a kernel logistic regression model by solving each of the three optimization problems. Monitor the value of the objective function (1) and the prediction accuracy on the test set as a function of the number of Newton iterations, for each of the three ways to solve the problem[1]. Do you see any difference between the three methods in terms of (i) how fast they decrease the objective function (in time and in number of iterations), and (ii) how well they predict on the test set after a given number of Newton steps?

[1] For the dual optimization problem (3), at each Newton step you should associate to the current estimate γ the primal variable α that minimizes the Lagrangian for γ fixed, in order to recover a corresponding β to estimate the primal loss function.