Homework 3

MATH CS 120                                          http://atzberger.org/

1. (Kernel-Ridge Regression) Consider the problem of constructing a model that approximates the relation \( y = f(x) \) from samples obscured by noise \( y_i = f(x_i) + \xi_i \), where \( \xi_i \) is Gaussian. As discussed in lecture when using Bayesian methods with a Gaussian prior this leads to the optimization problem

\[
\min_w J(w), \quad \text{where} \quad J(w) = \frac{1}{2} \sum_{i=1}^{m} (w^T \phi(x_i) - y_i)^2 + \frac{1}{2} \gamma w^T w.
\]

(a) Show that the solution weight vector \( w \) always can be expressed in the form \( w = \sum_{i=1}^{m} \alpha_i \phi(x_i) \). Hint: Compute the gradient \( \nabla_w J = 0 \).

(b) Consider the design matrix \( \Phi = [\phi(x_1), \ldots, \phi(x_m)]^T \) defined by the data so we can express \( w = \Phi^T \alpha \). Substitute this into the optimization problem to obtain the dual formulation in terms of minimizing over a function \( J(\alpha) \). Express this in terms of the design matrix \( \Phi \) and Gram matrix \( K \), where \( K_{ij} = k(x_i, x_j) = \phi(x_i)^T \phi(x_j) \).

(c) Compute the gradient \( \nabla_\alpha J = 0 \) to derive equations for the solution of the optimization problem. Express the linear equations for the solution \( \alpha \) in terms of the Gram matrix \( K \).

(d) Explain briefly the importance of the term \( \gamma \) and role it plays in the solution.

(e) Suppose we consider the regression problem to be over all functions \( f \in \mathcal{H} \) in some Reproducing Kernel Hilbert Space (RKHS) \( \mathcal{H} \) with kernel \( k \) and use regularization \( \|f\|_\mathcal{H}^2 \). This corresponds to the optimization problem

\[
\min_{f \in \mathcal{H}} J[f], \quad \text{with} \quad J[f] = \frac{1}{2} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \frac{1}{2} \|f\|_\mathcal{H}^2.
\]

Show the solution to this optimization problem yields the same result as in the formulation above using \( \alpha \). Hint: Use the representation results we discussed in lecture for objective functions of the form \( J[f] = L(f(x_1), \ldots, f(x_m)) + G(\|f\|_\mathcal{H}) \).

2. Consider kernel regression in the case when \( k(x, z) = \exp(-c\|x - z\|^2) \). Compute the kernel-ridge regression for \( f(x) = \sin(x) \) in the specific case of \( y_i = \sin(x_i) \) with \( x_i = 2\pi(i - 1)/m \) for \( i = 1, 2, \ldots, m \). Study the \( L_2 \)-error (least-squares error) \( \epsilon_{err} = \int_{0}^{2\pi} (w^T \phi(z) - f(z))^2 \, dz \) when estimated by \( \epsilon_{err} = \frac{2\pi}{N} \sum_{i=1}^{N} (w^T \phi(z_i) - f(z_i))^2 \). To try to approximate the integral well take \( z_i = 2\pi(i - 1)/N \) with large \( N \gg m \), say \( N = 10^5 \). Use this to construct a log-log plot of \( \epsilon_{err} \) vs \( m \) when \( m \) varies over the range, say \( 10, 10 \times 2^1, 10 \times 2^2, \ldots, 10 \times 2^9 \). Plot on the same figure the errors \( \epsilon_{err} \) vs \( m \) for a few different choices of the hyperparameter \( c \), say \( c = 100, 10, 1, 0.1, 0.01 \). For \( f(x) = \sin(x) \) for which \( c \) values do you get the best accuracy? Explain briefly for what choice of \( c \) you would expect for the model to generalize the best under a data distribution for \( x_i \) that is uniform on \([0, 2\pi]\).
3. (L₁-Regularization) Consider the optimization problem
\[
\min_w J(w), \quad \text{with} \quad J(w) = \frac{1}{2}(w - q)^T(w - q) + R(w).
\]

(a) Find the solution \(w \in \mathbb{R}^4\) when \(R(w) = \gamma \frac{1}{2} \|w\|_2^2\) with \(q = (1, 1, 1, 4)\) and \(\gamma = 1\). Hint: Consider values \(w\) where \(\nabla_w J = 0\) or the gradient does not exist.

(b) Find the solution \(w \in \mathbb{R}^4\) when \(R(w) = \gamma \|w\|_1\) with \(q = (1, 1, 1, 4)\) and \(\gamma = 1\). Hint: Consider values \(w\) where \(\nabla_w J = 0\) or the gradient does not exist.

(c) For which solution are most of the components of \(w\) zero. Briefly explain why one might expect one of the regularizations to do better in pushing solutions close to the coordinate axes to promote sparsity.