Exercises

Machine Learning: Foundations and Applications — Paul J. Atzberger
MATH 260J — 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 $y_i = f(\mathbf{x}_i) + \xi_i$, where ξ_i is Gaussian. As discussed in lecture when using Bayesian methods with a Gaussian prior this leads to the optimization problem

$$\min_{\mathbf{w}} J(\mathbf{w}), \quad \text{where} \quad J(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{m}\left(\mathbf{w}^T\phi(\mathbf{x}_i) - y_i\right)^2 + \frac{1}{2}\gamma\mathbf{w}^T\mathbf{w}.$$

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

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

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

   (d) Explain briefly the importance of the term γ 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

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

   Show the solution to this optimization problem yields the same result as in the formulation above using $\boldsymbol{\alpha}$. 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 $k(\mathbf{x}, \mathbf{z}) = \exp\left(-c\|\mathbf{x} - \mathbf{z}\|^2\right)$. 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) $\epsilon_{\text{err}} = \int_0^{2\pi}\left(\mathbf{w}^T\phi(z) - f(z)\right)^2 dz$ when estimated by $\tilde{\epsilon}_{\text{err}} = \frac{2\pi}{N}\sum_{\ell=1}^{N}\left(\mathbf{w}^T\phi(z_i) - f(z_i)\right)^2$. 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 $\tilde{\epsilon}_{\text{err}}$ 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 $\tilde{\epsilon}_{\text{err}}$ 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. (L_1 vs L_2 Regularization) Consider the optimization problem

$$\min_{\mathbf{w}} J(\mathbf{w}), \quad \text{with} \quad J(\mathbf{w}) = \frac{1}{2}(\mathbf{w}-\mathbf{q})^T(\mathbf{w}-\mathbf{q}) + R(\mathbf{w}).$$

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

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

(c) For which solution are most of the components of $\mathbf{w}$ 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.