

Exercises

Machine Learning: Foundations and Applications
MATH 260

Paul J. Atzberger
<http://atzberger.org/>

- (Kernel Methods and RKHS) Consider the classification of points $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$ having labels associated with the XOR operation $y = x_1 \oplus x_2$ with $\mathcal{S} = \{(-1, -1, F), (-1, 1, T), (1, -1, T), (1, 1, F)\}$. There is no direct linear classifier $h(\mathbf{x}) = \text{sign}(\mathbf{w}^T \mathbf{x} + b)$ that can correctly label these points. Here, we use $(-1$ for False, 1 for True). However, if we use the feature map $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \phi_3(\mathbf{x})] = [x_1, x_2, x_1 x_2]$ into \mathbb{R}^3 there is a linear classifier of the form $h(\mathbf{x}) = \text{sign}(\mathbf{w}^T \phi(\mathbf{x}) + b)$.
 - Find weights \mathbf{w} and b that correctly classifies the points with XOR labels.
 - Give the kernel function $k(\mathbf{x}, \mathbf{z})$ associated with this feature map into \mathbb{R}^3 .
 - Show the Reproducing Kernel Hilbert Space (RKHS) \mathcal{H} for this feature map consists of all the functions of the form $f(\cdot) = ax_1 + bx_2 + cx_1x_2$. Using that $\phi(\mathbf{z}) = k(\cdot, \mathbf{z})$, give the inner-product $\langle f, g \rangle_{\mathcal{H}}$ for two functions $f(\cdot)$ and $g(\cdot)$ from this space.
 - Show $k(\cdot, \mathbf{z})$ has the reproducing property under this inner-product.
 - Show that we can express $\mathbf{w} = \sum_i \alpha_i k(\cdot, \mathbf{x}_i)$ and that the classifier can be expressed using only kernel evaluations as $h(\mathbf{x}) = \text{sign}(\sum_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b)$.
Hint: Recall that the dot-product expressions are short-hand $\mathbf{w}^T \phi(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle_{\mathcal{H}}$.
- (Support Vector Machine (SVM)). The SVM is a widely used method to perform classification by trying to find hyperplanes that separate the data classes of $\mathcal{S} = \{x_i, y_i\}_{i=1}^m$. SVMs aim to obtain generalization by looking for hyperplanes with the largest margin. In the case with two separable classes, this corresponds to the constrained optimization problem

$$\min_{\mathbf{w}, b} \|\mathbf{w}\|^2 \quad \text{subject to} \quad (\mathbf{w}^T \mathbf{x}_i + b) y_i \geq 1.$$

- What is the VC-dimension of the set of hyperplane classifiers for $\mathbf{x} \in \mathbb{R}^n$? The hypothesis space is $\mathcal{H} = \{h \mid h(\mathbf{x}) = \text{sign}(\mathbf{w}^T \mathbf{x}_i + b), \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$.
 - We derived in lecture the *dual problem* for SVMs in the non-separable case using the Karush-Kuhn-Tucker (KKT) conditions. Derive the dual formulation for the SVM in the separable case.
 - How does the weight vector \mathbf{w} depend on the training data samples $\mathcal{S} = \{x_i, y_i\}_{i=1}^m$? In particular, which training data samples contribute to \mathbf{w} ? Hint: Use the KKT conditions to obtain representation formula for \mathbf{w} in terms of the data. (Which coefficients are non-zero?)
- (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

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

- (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 $\Phi = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_m)]^T$ defined by the data so we can express $\mathbf{w} = \Phi^T \alpha$. 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 Φ 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 $\nabla_{\alpha} J = 0$ to derive equations for the solution of the optimization problem. Express the linear equations for the solution α 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

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

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

4. Consider kernel regression in the case when $k(\mathbf{x}, \mathbf{z}) = \exp(-c\|\mathbf{x} - \mathbf{z}\|^2)$. 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 $i = 1, 2, \dots, m$. Study the L_2 -error (least-squares error) $\epsilon_{\text{err}} = \int_0^{2\pi} (\mathbf{w}^T \phi(z) - f(z))^2 dz$ when estimated by $\tilde{\epsilon}_{\text{err}} = \frac{2\pi}{N} \sum_{\ell=1}^N (\mathbf{w}^T \phi(z_\ell) - f(z_\ell))^2$. 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 $10, 10 \times 2^1, 10 \times 2^2, \dots, 10 \times 2^9$. 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 $[0, 2\pi]$.
5. (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.

6. (Perceptron) Consider the separable case and a dataset $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ represented as $\mathbf{x}_i = (\tilde{\mathbf{x}}_i, 1)$ to handle the bias term. We could try to find a classifying hyperplane $h(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$ using the following procedure: (i) initialize $\mathbf{w}^{(1)} = \mathbf{0}$, (ii) if there is some index i with \mathbf{x}_i misclassified with $y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \leq 0$ then update the weights using $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_i \mathbf{x}_i$.
- (a) Show this method always converges in the separable case to a $\hat{\mathbf{w}}$ so that $y_i \langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle > 0$.
- (b) Show the method converges in at most T iterations with $T \leq (RB)^2$, where $B = \min_{\mathbf{w}} \{\|\mathbf{w}\| \text{ s.t. } y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1\}$ and $R = \max_i \|\mathbf{x}_i\|$.

Hint: Let \mathbf{w}^* be the vector of smallest norm with $y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle \geq 1$, which exists by the separability condition. Show after T iterations $\frac{\langle \mathbf{w}^*, \mathbf{w}^{(T+1)} \rangle}{\|\mathbf{w}^*\| \|\mathbf{w}^{(T+1)}\|} \geq \frac{\sqrt{T}}{RB}$. Cauchy-Schwartz then yields the inequality.