Take-home Final

Machine Learning: Foundations and Applications
MATH 260J

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

Do any 3 of the following 5 problems.

1. Show that the concept class of hyper-rectangles $[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n] \in \mathbb{R}^n$ in PAC learnable. Hint: Start by considering $n = 2$ and showing this is learnable and then work from there.

2. (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 $J[f] = L(f(x_1), \ldots, f(x_m)) + G(\|f\|_{\mathcal{H}})$.

3. The Support Vector Machine (SVM) is a widely used method that performs classification by finding in some sense the best hyperplane that separates the data. The criteria used by SVM for defining the best hyperplane is to try to obtain good generalization by looking for a hyperplane with largest margin separating the classes of the training data samples $\{x_i, y_i\}_{i=1}^m$. In the case of separable data sets this is captured by the constrained optimization problem

$$\min_{\mathbf{w}, b} \|\mathbf{w}\|^2 \tag{1}$$

$$\text{subject:} \quad \left(\mathbf{w}^T\mathbf{x}_i + b\right) y_i \geq 1. \tag{2}$$

(a) What is the VC-dimension of the set of hyperplane classifiers for $\mathbf{x} \in \mathbb{R}^n$? The hypothesis space is $\mathcal{H} = \{h | h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x}_i + b), \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$.

(b) We discussed in lecture the derivation of the *dual problem* by defining the *dual function* and use of the Karush-Kuhn-Tucker conditions. Derive the dual formulation of the SVM in the separable case.

(c) How does the weight vector $\mathbf{w}$ depend on the training data samples $\{x_i, y_i\}_{i=1}^m$? In particular, which training data samples contribute with non-zero coefficients to $\mathbf{w}$? Hint: Use the KKT conditions to obtain representation formula for $\mathbf{w}$ in terms of the data.

4. (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, where $(F = -1, T = 1)$. However, if we use the feature map $\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \phi_3(\mathbf{x})] = [x_1, x_2, x_1x_2]$ into $\mathbb{R}^3$ there is a linear classifier of the form $h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + b)$.

(a) Find weights $\mathbf{w}$ and b that correctly classifies the points with XOR labels.

(b) Give the kernel function $k(\mathbf{x}, \mathbf{z})$ associated with this feature map into $\mathbb{R}^3$.

(c) 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 $\boldsymbol{\phi}(\mathbf{z}) = k(\cdot, \mathbf{z})$, give the inner-product $\langle f, g \rangle_{\mathcal{H}}$ for two functions $f(\cdot)$ and $g(\cdot)$ from this space.

(d) Show $k(\cdot, \mathbf{z})$ has the reproducing property under this inner-product.

(e) 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 $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}) \rangle_{\mathcal{H}}$.

5. (Neural Networks) Consider a basic Multilayer Perceptron (MLP) with two inputs x_1, x_2, single output y, and a hidden layer with n units h_i. Corresponding to this MLP is the hypothesis space $\mathcal{H} = \{q : \mathbb{R}^2 \to \mathbb{R} | q(x_1, x_2; \mathcal{W}) = \sum_{i=1}^n w_i^{(2)} h_i, \text{where } h_i = \sigma(w_{i1}^{(1)} x_1 + w_{i1}^{(1)} x_2)\}$. The output is $y = q(x_1, x_2; \mathcal{W})$ where the $\mathcal{W}$ denotes the collection of weights.

(a) Consider the case where we set $x_2 = 1$ and $x_1 \in [0, 1]$ with activation the Rectified Linear Unit (ReLU) $\sigma = \max(0, z)$. Show that with at most $n = k + 2$ hidden units we can exactly represent any function $f(x_1)$ that is piece-wise linear with k internal transition points on $[0, 1]$ and $f(x) = 0$ for $x \notin (0, 1)$. For instance, show that $f(x) = 2x$ for $x \leq 1/2$ and $f(x) = 2(1 - x)$ for $x > 1/2$, which has $k = 1$ internal transition points, can be exactly represented on $[0, 1]$ by a MLP with $n = 3$ hidden units.

(b) Consider approximating a general function $f(x)$ on $[0, 1]$ by using a gradient descent $\dot{\mathbf{w}} = -\alpha \nabla_{\mathbf{w}} L$ to minimize the loss $L(q) = \frac{1}{m} \sum_{i=1}^m (f(z_i) - q(z_i; \mathbf{w}))^2$. Consider m data points $z_i \in [0, 1]$ where we take in the MLP $x_1 = z_i$ and $x_2 = 1$. State for the MLP the back-propagation method for computing the gradient in $\mathbf{w}$. Draw the computational graph in the case when $n = 1$ and $m = 1$ for both the "forward pass" and the "backward pass."

(c) Explain techniques for how you might mitigate getting stuck in local minima or overfitting the data?