Exercises

Machine Learning: Foundations and Applications — Paul J. Atzberger
MATH 260J — http://atzberger.org/

1. (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

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

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

   (b) 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.

   (c) 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?)

2. (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 $B = \min_{\mathbf{w}}\{\|\mathbf{w}\| \text{ s.t. } y_i\langle \mathbf{w}, \mathbf{x} \rangle \geq 1\}$ 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.

3. (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 $(-1 \text{ for False}, 1 \text{ for True})$. 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 $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}) \rangle_{\mathcal{H}}$.