Homework 4

Machine Learning: Foundations and Applications   Paul J. Atzberger
MATH CS 120   http://atzberger.org/

1. 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

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

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

   (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.

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

3. (Kernel PCA and Dimension Reduction) Consider the data set of points in $\mathbb{R}^n$ on an embedded circle with random orientation. For concreteness, consider $n = 4$ with ideal data $\mathcal{S} = \{(\cos(\theta_\ell) - \sin(\theta_\ell), \cos(\theta_\ell) + \sin(\theta_\ell), \cos(\theta_\ell) - \sin(\theta_\ell), \cos(\theta_\ell) + \sin(\theta_\ell)) | \theta_\ell = 2\pi(\ell - 1 - \frac{1}{2}(m-1))/(m-1), \ell = 1, 2, \ldots, m, m = 6\}$. You are also welcome to create your own data sets with more points m or add a small amount of noise to explore the methods.

(a) Perform Kernel-PCA to reduce this data set to a 1D description using the feature map $\phi(\mathbf{x}) = [\arccos(\mathbf{w}_1^T\mathbf{x}), \arcsin(\mathbf{w}_2^T\mathbf{x}), \mathbf{x}^T\mathbf{x}] - \phi_0$, where $\mathbf{w}_1 = (1,1,1,1)/4$, $\mathbf{w}_2 = (-1,1,-1,1)/4$, $\phi_0 = \frac{1}{m}\sum_{i=1}^{m}\phi(\mathbf{x}_i)$. This has kernel $k(\mathbf{x},\mathbf{z}) = \phi(\mathbf{x})^T\phi(\mathbf{z})$.

(b) How can this 1D description found (singular vector) be interpreted geometrically back in the original input space $\mathbb{R}^4$?

(c) *Bonus:* If we did not have a good idea for the choice of $\mathbf{w}_1$ and $\mathbf{w}_2$ how might you find them to obtain a data-dependent kernel? Hint: PCA could again be useful here.