CS 688: Homework 3

Due: Friday March 29 by 10:00 PM

Notes:

- Homework is due **by 10:00 PM on the due date.** Remember that you may not use more than 2 late days on any one homework, and you only have a budget of 5 in total.
- Please keep in mind the collaboration policy as specified in the course syllabus. If you discuss questions with others you **must** write their names on your submission, and if you use any outside resources you **must** reference them. **Do not look at each others' writeups, including code.**
- There are 7 problems on 2 pages in this homework.
- Remember that you will be graded on correctness and clarity. Please write carefully and clearly.

Problems:

1. (20 points) (From Russell & Norvig) Construct a support vector machine that computes the XOR function. Use values of $+1$ and -1 (instead of 1 and 0) for both inputs and outputs, so that an example looks like $([-1, 1], 1)$ or $([-1, -1], -1)$. Map the input $[x_1, x_2]$ into a space consisting of x_1 and $x_1 x_2$. Draw the four input points in this space, and the maximal margin separator. What is the margin? Now draw the separating line back in the original Euclidean input space.

2. (15 points) The key point of the so-called "kernel trick" in SVMs is to learn a classifier that effectively separates the training data in a higher dimensional space without having to explicitly compute the representation $\Phi(\mathbf{x})$ of every point $\mathbf{x}$ in the original input space. Instead, all the work is done through the kernel function that computes dot products $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)\Phi_(\mathbf{x}_j)$.

   Show how to compute the squared Euclidean distance in the projected space between any two points $\mathbf{x}_i$, $\mathbf{x}_j$ in the original space without explicitly computing the Φ mapping, instead using the kernel function K.

3. (15 points) Suppose we use a radial basis kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$. This gives some implicit unknown map $\Phi(\mathbf{x})$. Show that for any $\mathbf{x}_i, \mathbf{x}_j$, $\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_j)\|^2 \leq 2$.

4. (10 points) Consider the transform $\Phi^n_{\text{poly}}(\mathbf{x})$ where $\mathbf{x} \in \mathbb{R}^d$. What is the exact dimensionality of the implied feature space in terms of n (the degree of the polynomial) and d (the dimensionality of the original input data)? Show how you derive this, it is not enough to just write down the answer from the textbook. Evaluate this when $n = d = 10$. (Hint: you will need to use a combinatorial argument – be sure to check your work on some small instances that you can compute by hand)?

5. (15 points) Consider the loss function for linear regression, as a function of the hypothesis h, $L(h) = \mathbb{E}[(h(\mathbf{x}) - y)^2]$. Show that among all hypotheses, the one that minimizes L over the population is given by $h^*(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$. [Hint: You may want to get started by writing out the loss function over the population as $\int d\mathbf{x} \int dy P(\mathbf{x}, y)(h(\mathbf{x}) - y)^2$. The expectation is over $\mathbf{x}$ and y]

6. (10 points) Consider any differentiable loss function $L(\mathbf{w})$ and now consider an L_2-regularized version $L(\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w}$. What is the gradient descent rule for the regularized loss function (you can use $\nabla_{\mathbf{w}}(L)$ in your expression). Could you speculate, based upon this, as to why L_2 regularization is sometimes called weight decay?

7. (15 points) The *VC dimension* of a hypothesis class is defined as the maximum size of a set of data that can be shattered by the hypothesis class (so, the biggest n for which there exists a set of points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ such that $\mathcal{H}$ can achieve all dichotomies on $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$). If there is no such set, the VC-dimension of $\mathcal{H}$ is said to be infinite.

   - (5 points) What is the VC-dimension of axis-parallel rectangles in the plane?
   - (5 points) What are the VC dimensions of (i) rays and (ii) intervals that can be either positive *or* negative (so, for intervals you get to pick both the interval *and* whether points in that interval are all positive or all negative, and for rays you get to pick the start point of the ray *and* whether all points to the right of it are positive or negative)?
   - (5 points) What is the VC-dimension of the hypothesis space of concentric spheres in $\mathbb{R}^d$ – that is, $\mathcal{H}$ contains the functions that are $+1$ for $a \leq \sqrt{x_1^2 + x_2^2 + \ldots + x_d^2} \leq b$?