CS 688: Homework 3 Solutions

April 11, 2024

Note: These are not intended to be comprehensive, just to help you see what the answers should be.

1 The mapping is easy to compute. $[-1,-1]$ maps to $[-1,+1]$; $[-1,+1]$ maps to $[-1,-1]$; $[+1,-1]$ maps to $[+1,-1]$; $[+1,+1]$ maps to $[+1,+1]$. The maximal margin separator in the new space is the line $x_1 x_2 = 0$, with a margin of 1. In the original space, this is equivalent to having either $x_1 = 0$ or $x_2 = 0$, which you can think of as the limit of a hyperbolic separator with two branches.

2 The squared Euclidean distance is the dot product of $\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_j)$ with itself. This can be written as $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_i) + \Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_j) - 2\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$, which is just $K(\mathbf{x}_i, \mathbf{x}_i) + K(\mathbf{x}_j, \mathbf{x}_j) - 2K(\mathbf{x}_i, \mathbf{x}_j)$.

3 This is a relatively simple exercise:

$$\begin{aligned}
\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_j\|^2 &= \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_i) \rangle + \langle \Phi(\mathbf{x}_j), \Phi(\mathbf{x}_j) \rangle - 2\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \\
&= K(\mathbf{x}_i, \mathbf{x}_i) + K(\mathbf{x}_j, \mathbf{x}_j) - 2(\mathbf{x}_i, \mathbf{x}_j) \\
&= 1 + 1 - 2\exp\left(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \\
&< 2
\end{aligned}$$

4 We need to compute the number of different monomials of degree at most. n when the dimensionality is d. One way to think about this is that for each monomial, you have n defrees of power to distribute, and $d+1$ choices for each of those n powers (one of the d input variables or the constant 1). So this is the same as asking how many ways there are of distributing n balls into $d+1$ bins. That is the same as choosing d "separators" out of $n+d$ objects or $\binom{n+d}{d}$. You can subtract 1 if you want to ignore the constant term (where every ball is in the "1" bin). For $n = d = 10$ this latter calculation would give 184755 terms (the former would be one more, which is fine also).

5 This one is tough and intended to get you to think hard. As noted in the problem, start with:

$$\begin{aligned}
L(h) &= \int d\mathbf{x} \int dy P(\mathbf{x}, y)(h(\mathbf{x}) - y)^2 \\
&= \int d\mathbf{x} P(x) \int dy P(y \mid x)(h(\mathbf{x}) - y)^2 \\
&= \int d\mathbf{x} P(x) \int dy P(y \mid x)(h(\mathbf{x})^2 - 2h(\mathbf{x})y + y^2) \\
&= \int d\mathbf{x} P(x) \left(h(\mathbf{x})^2 - 2h(\mathbf{x})\mathbb{E}\left[y \mid x\right] + \mathbb{E}\left[y^2 \mid \mathbf{x}\right]\right) \\
&= \int d\mathbf{x} P(x) \left((h(\mathbf{x}) - \mathbb{E}\left[y \mid \mathbf{x}\right])^2 + \mathbb{E}\left[y^2 \mid \mathbf{x}\right] - \mathbb{E}\left[y \mid \mathbf{x}\right]^2\right)
\end{aligned}$$

The lowest value this can take is:

$$\int d\mathbf{x} P(\mathbf{x}) \left(\mathbb{E}\left[y^2 \mid \mathbf{x}\right] - \mathbb{E}\left[y \mid \mathbf{x}\right]^2\right)$$

which is just $\int d\mathbf{x} P(\mathbf{x}) \text{Var}[y \mid \mathbf{x}]$, and this happens when $h(\mathbf{x}) = \mathbb{E}\left[y \mid \mathbf{x}\right]$.

6 The gradient of the regularized loss function is $\nabla_{\mathbf{w}} L + 2\lambda \mathbf{w}$. Let η be the learning rate. Then the weight update for gradient descent is given by:

$$\begin{aligned}\mathbf{w}(t+1) &= \mathbf{w}(t) - \eta \nabla_{\mathbf{w}} L(\mathbf{w}(t)) - 2\eta\lambda \mathbf{w}(t) \\ &= (1 - 2\eta\gamma)\mathbf{w}(t) - \eta \nabla_{\mathbf{w}} L(\mathbf{w}(t))\end{aligned}$$

The reason for the term "weight decay" is that the weight vector comes down in magnitude at every step before it is updated with the gradient.

7 i. *Axis parallel rectangles*: onsider four points in a "diamond" shape (e.g. $(0,1), (0,-1), (1,0), (-1,0)$). It is easy to see that you can enclose any subset of the 4 points in an axis parallel rectangle. Thus the VC-dimension is at least 4.

If you have any 5 points, if two of points share the same x or y coordinate then an axis parallel rectangle cannot classify those two differently. If they do not, find the ones at the extrema of the x and y axes; that will give you 4 points. Any axis parallel rectangle containing those 4 must also contain the 5th. Therefore, no set of 5 points can be shattered, and the VC-dimension is 4.

ii. *Rays and intervals that can be both positive and negative*: Rays that can be positive or negative– We can shatter any set of two points quite easily, but no set of three because the middle point cannot be different from the outer two. So the VC-dimension is 2.

Intervals that can be positive or negative– Now we can also shatter three points because the middle point can go in its own interval. However, with four points we cannot achieve the dichotomy with alternating labels. So the VC-dimension is 3.

iii. *Concentric spheres:* This is actually equivalent to positive intervals on $\mathbb{R}^+$ because the only thing that matters in terms of implementing a dichotomy is the distance of a point from the origin. For any set of three points on the plane, we can't achieve the dichotomy where the one that is in the middle in terms of distance from the origin is negative and the other two are positive. Two points are easy to shatter. Thus the VC-dimension is 2.