Sixth Home Assignment
Data Analysis

Nikolaj Dybdahl Rathcke (rfq695)

June 13, 2015

Question 1

(1)

Given an angle α and a vector w, we can calculate $\|w\|$. When we make a triangle as in the assignment text, we can calculate $\|p\|$ with law of sines, since we know one angle is 90 and we know the angle, β, opposite of p is $180 - 90 - \alpha$. The law of sines, in our example, say that

$$\begin{aligned} \frac{\|w\|}{sin(\pi/2)} &= \frac{\|p\|}{sin(\beta)} \qquad \Leftrightarrow \\ \|p\| &= sin(\beta)\frac{\|w\|}{sin(\pi/2)} \\ &= sin(\beta)\|w\| \end{aligned}$$

This is the length of p, to find the vector p, we need to make it the same ratio as the vector u, this is done by multiplying $\frac{u}{\|u\|}$, thus we get

$$p = sin(\beta)\|w\|\frac{u}{\|u\|} \tag{1}$$

which is the vector p and the projection of w on u.

(2)

The projection vector p is then given by

$$p = sin(\beta)\|w\|u$$

as the denominator in equation (1) is just 1.

(3)

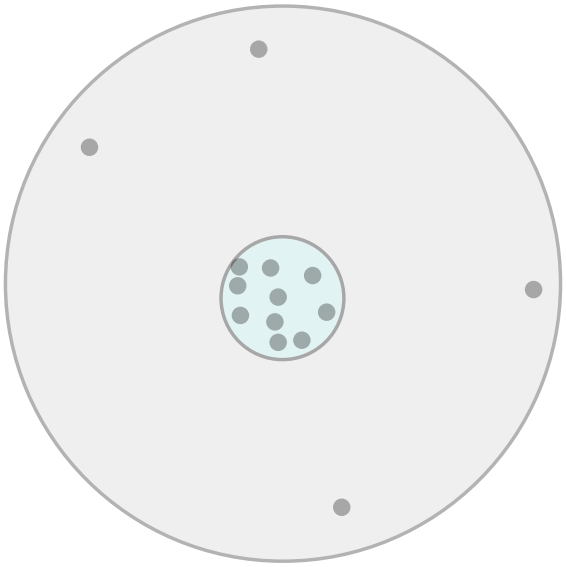
The length, as we found in (1) of p is given by

$$sin(\beta)\|w\|$$

where β is the last unknown angle (which can easily be found) and $\|w\|$ is the length of the given vector w.

Question 3

Consider the following figure where $E_{out}(g)$ is based on the hypothesis g from all datapoints in it. The inner circle is where $E_{out}(g^-)$ is based on hypothesis g^- from the datapoint in that circle.



The out of sample error for on g will be larger as there are outliers, while the error on g^- is less since there are no outliers.

Question 4

(1)

The VC-dimension is 27^d as $\mathcal{H}_d$ is a tree with depth d and branches 27 times at every node. Therefore the VC-dimension is 27^d.

(2)

The VC-dimension is ∞, since the tree has infinite depth.

(3)

We use corralary (1) from the slides on Occam's razor bound, that states

$$\mathbb{P}\left\{\exists h \in \mathcal{H} : E_{out} - E_{in} > \sqrt{\frac{\ln(M/\delta)}{2N}}\right\} \leq \delta$$

We can use this as $\mathcal{H}_d$ is finite, so we replace M by the size and we get the following bound

$$\mathbb{P}\left\{\exists h \in \mathcal{H} : E_{out} - E_{in} > \sqrt{\frac{\ln(27^d/\delta)}{2N}}\right\} \leq \delta$$

This is the bound on $E_{out} - E_{in}$ for learning with $\mathcal{H}_d$.

(4)

From theorem (2) in the slides, we can insert our VC dimension on H_d to get

$$\mathbb{P}\left\{\exists h \in \mathcal{H} : E_{out} - E_{in} > \sqrt{\frac{\ln\left(2^{27^d} 2^{d+1}/\delta\right)}{2N}}\right\} \leq \delta$$

(5)

The idea of Occam's razor bound is to give each hypothesis a probability. This probability is given by $p(h)$ where $\sum_{h\in\mathcal{H}} p(h) \leq 1$ and as such, the "burden" is shared by all hypotheses with a specific weight.

Question 6

This question is exercise 5.4 in [LFD].

(1)

The problem with the bound is that it has been invalidated since the learning model has been chosen after the data is observed. So the choice has been based on observations of the data which *likely* fits some learning method instead of figuring our what learning method *should* be used on the dataset. This means we can not know if the bounds are correct.

(2)

No, as the VC dimension is the outcome of one learning method, that is not nessecarily the right one.