HW5 – Theory + SVM

May 28, 2022

Question 1: Let $K(x,y) = (x \cdot y + 1)^3$ be a function over $\mathbb{R}^2 \times \mathbb{R}^2$, namely $x, y \in \mathbb{R}^2$.

1. Let $\psi : \mathbb{R}^2 \to \mathbb{R}^{10}$ be defined as:

   $$\psi(w) = \left(w_1^3, w_2^3, \sqrt{3}w_1^2 w_2, \sqrt{3}w_1 w_2^2, \sqrt{3}w_1^2, \sqrt{3}w_2^2, \sqrt{6}w_1 w_2, \sqrt{3}w_1, \sqrt{3}w_2, 1\right)$$

   We get:

   $$K(x,y) = (x \cdot y + 1)^3 = (x \cdot y + 1)(x \cdot y + 1)(x \cdot y + 1) =$$
   $$\left((x \cdot y)^2 + 2xy + 1\right)(x \cdot y + 1) = \left((x \cdot y)^3 + 3(x \cdot y)^2 + 3xy + 1\right) =$$
   $$\left((x_1 y_1 + x_2 y_2)^3 + 3(x_1 y_1 + x_2 y_2)^2 + 3x_1 y_1 + 3x_2 y_2 + 1\right) =$$
   $$(x_1 y_1)^3 + (x_2 y_2)^3 + 3(x_1 y_1)^2 x_2 y_2 + 3x_1 y_1 (x_2 y_2)^2 +$$
   $$3\left[(x_1 y_1)^2 + (x_2 y_2)^2 + 2x_1 x_2 y_1 y_2\right] + 3x_1 y_1 + 3x_2 y_2 + 1 =$$
   $$x_1^3 y_1^3 + x_2^3 y_2^3 + 3x_1^2 y_1^2 x_2 y_2 + 3x_2^2 y_2^2 x_1 y_1 + 3x_1^2 y_1^2 + 3x_2^2 y_2^2 +$$
   $$6x_1 x_2 y_1 y_2 + 3x_1 y_1 + 3x_2 y_2 + 1 =$$
   $$\left(x_1^3, x_2^3, \sqrt{3}x_1^2 x_2, \sqrt{3}x_1 x_2^2, \sqrt{3}x_1^2, \sqrt{3}x_2^2, \sqrt{6}x_1 x_2, \sqrt{3}x_1, \sqrt{3}x_2, 1\right) \cdot$$
   $$\left(y_1^3, y_2^3, \sqrt{3}y_1^2 y_2, \sqrt{3}y_1 y_2^2, \sqrt{3}y_1^2, \sqrt{3}y_2^2, \sqrt{6}y_1 y_2, \sqrt{3}y_1, \sqrt{3}y_2, 1\right) =$$
   $$\psi(x) \cdot \psi(y)$$

   As required.

2. we called this function the full rational variety of order 3

3. With the Kernel we do 4 multiplications instead of doing 10 without it, saving 6 multiplication operations.

Question 2: Let $f(x,y) = 2x - y$. Find the minimum and the maximum points for f under the constraint $g(x,y) = \frac{x^2}{4} + y^2$.

Let $L(x,y) : \mathbb{R}^2 \to \mathbb{R}$ s.t

$$L(x,y) = 2x - y - \lambda\left(\frac{x^2}{4} + y^2 - 1\right)$$

We derive:

$$\frac{\partial L}{\partial x}(x,y) = 2 - \frac{\lambda x}{2} = 0 \Rightarrow \lambda = \frac{4}{x}$$

$$\frac{\partial L}{\partial y}(x,y) = -1 - 2\lambda y = 0 \Rightarrow \lambda = -\frac{1}{2y}$$

$$\frac{\partial L}{\partial \lambda}(x,y) = -\frac{x^2}{4} - y^2 + 1 = 0 \Rightarrow 1 - \frac{x^2}{4} = y^2 \Rightarrow$$

$$y = \pm\sqrt{1 - \frac{x^2}{4}}$$

We got that:

$$\frac{4}{x} = -\frac{1}{2y} \Rightarrow$$

$$x = -8y$$

$$y = \pm\sqrt{1 - \frac{64y^2}{4}} = \pm\sqrt{1 - 16y^2}$$

Hence:

$$y^2 = 1 - 16y^2 \Rightarrow$$

$$y = \pm\frac{1}{\sqrt{17}}, x = \pm\frac{8}{\sqrt{17}}$$

We get the following extreme points:

$$\left(\frac{8}{\sqrt{17}}, -\frac{1}{\sqrt{17}}\right), \left(-\frac{8}{\sqrt{17}}, +\frac{1}{\sqrt{17}}\right)$$

$$f\left(-\frac{8}{17}, \frac{1}{\sqrt{17}}\right) = -2\frac{8}{\sqrt{17}} - \frac{1}{\sqrt{17}} = -\sqrt{17} \to \min$$

$$f\left(\frac{8}{17}, -\frac{1}{\sqrt{17}}\right) = 2\frac{8}{\sqrt{17}} + \frac{1}{\sqrt{17}} = \sqrt{17} \to \max$$

Question 3: we can observe that the lines forming the triangles are composed using the unit vectors u, v, w and that the distance of the lines from the origin will be r.

let r^*as the distance of the **target** triangle from the origin so each sample which is drawn independently from a distribution D will have a positive target value if it's inside the triangle and negative value otherwise.

our hypothesis will yield an r by doing the following algorithm:

1. Run over all the positive samples in our training set
2. For each one calculate $r = Max((x_1, x_2)\cdot u, (x_1, x_2)\cdot v, (x_1, x_2)\cdot w)$ - with is basically the maximal scalar projection of the vector formed by the sample (x_1, x_2) on each of the directional vectors.
3. Choose the max r

Then use r and the direction vectors to formulate the hypothesis triangle.

the error region A_r will be the difference area between the target triangle (r^*) and the hypothesis triangle (r).

we define error ε and we want that $Pr[(x1, x2) \in A_r] \leq \varepsilon$

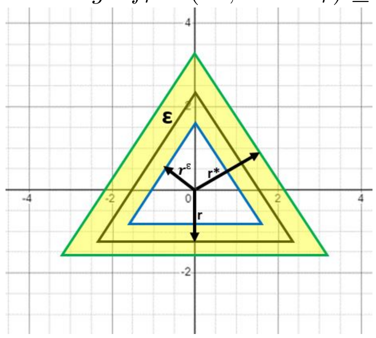
We claim that the concept class of origin-centered upright equilateral triangles is efficiently PAC-learnable

proof:

Case 1:

we define :

$r^\varepsilon = arginf_r Pr(x1, x2 \in A_r) \leq \varepsilon$



If $r^\varepsilon \leq r$ then the probability of the being in A_r is less than ε.

Case 2:

the probability of missing the A_r region with m training samples:

P(m instances missing the error region) $< (1-\varepsilon)^m$

we want this probability to be less then δ, thus the sample complexity will be:

$(1-\varepsilon)^m \leq e^{-\varepsilon m} \leq \delta \Longrightarrow m \geq \frac{1}{\varepsilon} ln \frac{1}{\delta}$

in order to calculate r the algorithm will need to go over all the positive samples which results in $\mathcal{O}(m)$ time complexity, and as m is polynomial in $\frac{1}{\varepsilon}, \frac{1}{\delta}$,so does the running time complexity.

Question 4: Using a test set of size 1000, assume that we counted 200 errors.

We estimate the generalization error by $\hat{p} = \frac{200}{1000} = 0.2$

From statistical sampling theory it follows that a 95% confidence interval for the generalization error is

$(\hat{p} - 1.96se, \hat{p} + 1.96se)$ where $se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

so in our case the confidence interval will be (0.1752, 0.2248)

so our true error could be up to 22.48%

Question 5: