Machine Learning from Data – IDC – 2022

HW5 – Theory + SVM

1. <u>Kernels and mapping functions (25 pts)</u>
   a. (20 pts) Let $K(x, y) = (x \cdot y + 1)^3$ be a function over $\mathbb{R}^2 \times \mathbb{R}^2$ (i.e., $x, y \in \mathbb{R}^2$).

      Find ψ for which $K$is a kernel. (It may help to first expand the above term on the right-hand side).

   b. (2 pts) What did we call the function ψ in class if we remove all coefficients?
   c. (3 pts) How many multiplication operations do we save by using $K(x, y)$ versus $\psi(x) \cdot \psi(y)$?

2. <u>Lagrange multipliers (25 pts)</u>

   Let $f(x, y) = 2x - y$. Find the minimum and the maximum points for f under the constraint $g(x, y) = \frac{x^2}{4} + y^2 = 1$.

3. <u>PAC Learning (25 pts)</u>

   Let $X = \mathbb{R}^2$. Let vectors $u = \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right), w = \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right), v = (0, -1)$

   and $C = H = \left\{ h(r) = \left\{ (x_1, x_2) \middle| \begin{array}{l} (x, y) \cdot u \le r, \\ (x, y) \cdot v \le r, \\ (x, y) \cdot w \le r \end{array} \right\} \right\}$, for $r > 0$,

   the set of all origin-centered upright equilateral triangles.
   Describe a polynomial sample complexity algorithm L that learns C using H. State the time complexity and the sample complexity of your suggested algorithm. Prove all your steps.

4. (15 pts) A business manager at your ecommerce company asked you to make a model to predict whether a user is going to proceed to checkout or abandon their cart. You created the model using, and reported 20% error on your test set of size 1000 samples. In the business manager's presentation to upper management, he presented your

model and stated that the company can expect 20% error when deploying the model live on the website.
Luckily, you realize that this is a mistaken assumption, and you correct the statement to say that with 95% confidence, the true error they can expect is up to what percentage? (Just state the error percentage).

5. <u>SVM (10 pts)</u>
   See the notebook in the homework files and follow the instructions there.
   Take a **screenshot** of your resulting graph near the bottom of the notebook (titled "My Graph") and paste into your submission PDF along with your answers to the theoretical questions. Do **NOT** submit your code.

Omer Wachman – 312332372
Adam Vinestock - 209795624
HW No.5

Question 1:

Let $K(x, y)$ = $(x \cdot y + 1)^3$ be a function over $R^2 \times R^2$ (i.e., x, $y \in R^2$).

a. Find ψ for which K is a kernel:

$(\mathrm{x} \cdot \mathrm{y} + 1)^3 =$

$= (\mathrm{x} \cdot \mathrm{y})^3 + (\mathrm{x} \cdot \mathrm{y})^2 + (\mathrm{x} \cdot \mathrm{y}) + 1 = (x_1 y_1 + x_2 y_2)^3 + (x_1 y_1 + x_2 y_2)^2 + (x_1 y_1 + x_2 y_2) + 1$

$= {x_1}^3 {y_1}^3 + 3{x_1}^2 {y_1}^2 x_2 y_2 + 3x_1 y_1 {x_2}^2 {y_2}^2 + {x_2}^3 {y_2}^3 + 3{x_1}^2 {y_1}^2 + 6x_1 y_1 x_2 y_2 + 3{x_2}^2 {y_2}^2 + 3x_1 y_1 + 3x_2 y_2 + 1$

$\Longrightarrow \psi\,(x_1, x_2) = \langle {x_1}^3, \sqrt{3}{x_1}^2 x_2, \sqrt{3} x_1 {x_2}^2, {x_2}^3, \sqrt{3}{x_1}^2, \sqrt{6} x_1 x_2, \sqrt{3}{x_2}^2, \sqrt{3} x_1, \sqrt{3} x_2, 1 \rangle$

b. What did we call the function ψ in class if we remove all coefficients?

We have called this function – cubic mapping.

c. How many multiplication operations do we save by using $K(x, y)$ versus $\psi(x) \cdot \psi(y)$?

When using the mapping function of ψ we map from R^2 to R^{10}, therefore when applying the operation ψ (x1,x2)·ψ (y1,y2) we use 10 multiplications.
When applying the kernel $K(x, y)$ = (x1·y1+x2·y2+1)3 we use 4 multiplications.
Therefore, in total we save 6 multiplication operations.

Question 2:

Let $f(x, y)$ = $2x - y$. Find the minimum and the maximum points for f under the constraint $g(x,y)$ = x^2 + y^2 = 1.

Solution:

Lagrange equation:

$\exists\lambda \text{ s.t.}, \mathrm{L}(\mathrm{x}, \mathrm{y}) = \mathrm{f}(x, y) + \lambda g(x, y)$

Find the partial derivates and equal to zero

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

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

$$\frac{\partial}{\partial \lambda}L(x,y) = \frac{x^2}{4} + y^2 - 1 = 0 \Rightarrow \frac{\left(\frac{-4}{\lambda}\right)^2}{4} + \left(\frac{1}{2\lambda}\right)^2 - 1 = \frac{16}{4\lambda^2} + \frac{1}{4\lambda^2} - 1 = \frac{17}{4\lambda^2} - 1 \Rightarrow$$

$$\Rightarrow \frac{17}{4\lambda^2} - 1 = 0 \Rightarrow (\lambda_1, \lambda_2) = (+\frac{\sqrt{17}}{4}, -\frac{\sqrt{17}}{4})$$

Because we squared λ after the substitution, we must consider that the resulted value of λ can be also negative. Hence, we got two different λ.

Substitute(λ_1, λ_2) into X and Y values:

$$x = -\frac{4}{\lambda} \Rightarrow (x_1, x_2) = (-\frac{8}{\sqrt{17}}, +\frac{8}{\sqrt{17}})$$

$$y = \frac{1}{2\lambda} \Rightarrow (y_1, y_2) = (+\frac{1}{\sqrt{17}}, -\frac{1}{\sqrt{17}})$$

Now we have to different point to substitute in f(x,y):

1. f$(x_1, y_1) = f\left(-\frac{8}{\sqrt{17}}, +\frac{8}{\sqrt{17}}\right) = -\frac{16}{\sqrt{17}} - \frac{1}{\sqrt{17}} = -\frac{17}{\sqrt{17}} = -\sqrt{17} \Rightarrow \boldsymbol{Minimum\ point}$

2. $f(x_2, y_2) = f\left(+\frac{1}{\sqrt{17}}, -\frac{1}{\sqrt{17}}\right) = \frac{16}{\sqrt{17}} - (-\frac{1}{\sqrt{17}}) = \frac{17}{\sqrt{17}}$ = $\sqrt{17} \Rightarrow \boldsymbol{Maximum\ point}$

Question 3:

PAC Learning
Let X = R2. Let vectors $\boldsymbol{u}$ **= (√3 , 1) ,** $\boldsymbol{w}$ **= (√3 , − 1) ,** $\boldsymbol{v}$ **= (0, −1)**

and $C = H$ = {h(r) = {(x1, x2)| $(x, y) \cdot v \leq r$, $(x, y) \cdot w \leq r$,$(x, y) \cdot u \leq r$}}, for $r > 0$}

the set of all origin-centered upright equilateral triangles.
Describe a polynomial sample complexity algorithm L that learns C using H. State the time complexity and the sample complexity of your suggested algorithm. Prove all your steps.

Answer:
We will define an algorithm that learns the r parameter according to given data.

The learning Algorithm(D): # D is the training data

r = 0 # Initializing r
for every instance $\vec{x}$ =(x_1, x_2) ∈ D with "True" label:
  if r < max ($\vec{x}$ · u, $\vec{x}$ · v, $\vec{x}$ · w):
    r = max ($\vec{x}$ · u, $\vec{x}$ · v, $\vec{x}$ · w)
return r

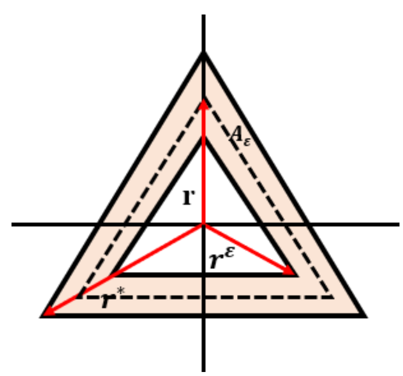
the output represents the "r" for which the data will be classified.
such that:
h = L(D)= {(x_1, x_2)| $(x, y) \cdot v \leq r$, $(x, y) \cdot w \leq r$,$(x, y) \cdot u \leq r$}}

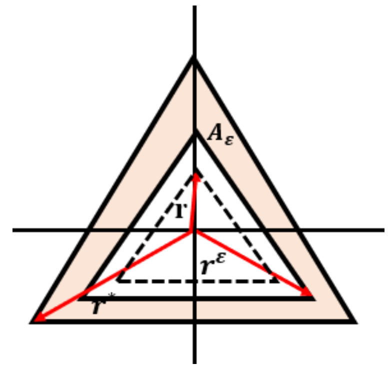
the resulted hypothesis L(D) represents the smallest origin-centered upright equilateral triangle that includes all training data labeled "True".
r represents the distance from the origin to each of the triangle vertices.

Sample complexity:
Given some $\delta \in [0,0.5]$, we can bound our sample size, such that our learning algorithm is not epsilon-bad with confidence 1- δ.
Claim – The concept class of origin-centered upright equilateral triangles is efficiently PAC-learnable.
Note that our learning algorithm is a consistent learner and by taking a large number of examples we get closer to the concept.

Proof:
Let c ∈ C be a concept.
Let r* be the value that defines the concept c: {(x_1, x_2)| $(x, y) \cdot v \leq r^*$, $(x, y) \cdot w \leq r^*$,$(x, y) \cdot u \leq r^*$}}
Let r be the value that defines the hypothesis from the learning algorithm L(D) shown above.
Let $r^{\mathcal{E}}$ be the largest origin-centered upright equilateral triangle such that $P\big((x_1, x_2)\epsilon A_{\varepsilon}\big) \leq \mathcal{E}$
Where A_{ε} is the area between L(D) and the concept.

<u>1st case</u>: $r^{\mathcal{E}} \leq r$
Then $\forall x \in X$, the probability that x ϵ A_{ε} is less than $\mathcal{E}$.





2[nd] case: $r^{\mathcal{E}} \geq r$

Given m samples to the learning algorithm, the probability of an instance $x \in X$ being misclassified is $(1-\mathcal{E})^{\mathrm{m}} \leq e^{-\mathcal{E}m}$

Now, let's bound it with δ:

$$e^{-\mathcal{E}m} \leq \delta \;\Rightarrow \frac{1}{\delta} \leq e^{\mathcal{E}m} \Rightarrow \; ln\left(\frac{1}{\delta}\right) \leq \ln\left(e^{\mathcal{E}m}\right) \Rightarrow m \geq \frac{\ln\left(\frac{1}{\delta}\right)}{\mathcal{E}}$$

So with confidence δ in order for hypothesis L(D) do be epsilon-bad we need at least $\frac{\ln\left(\frac{1}{\delta}\right)}{\varepsilon}$ instances.

Time complexity:

For each sample we perform O(1) calculations.

In total the time complexity is linear to the sample size - O(m)

Question 4:

We can say that with 95% confidence, the true error they can expect is in the interval:

$(\hat{p} - 2se, \hat{p} + 2se)$

$\hat{p} = true\ error$ = 20% = 0.2

$n = number\ of\ samples\ in\ test\ set$ = 1000

$$se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.2(1-0.2)}{1000}} = \sqrt{\frac{0.2*0.8}{1000}} = \sqrt{0.00016} \approx 0.01265$$

$\Longrightarrow we\ can\ say\ with\ confidence\ of\ 95\%\ that\ the\ true\ error\ between:$

$$(0.2 - 1.96 * 0.01265\,, 0.2 + 1.96 * 0.01265\,) = (\,0.1752, 0.2247) \;\approx (17.52\%\,, 22.47\%)$$

Question 5: