REICHMAN UNIVERSITY

MACHINE LEARNING FROM DATA
CS 3141

HW5 - Theory + SVM

Authors:
Gal Almog (806709)
Omri Newman (806646)

May 28, 2022

Question 1

(a) $K(x, y) = (x \cdot y + 1)^3$
Transpose x to get valid matrix multiplication:

$$\begin{aligned}
K(x, y) &= (x^T y + 1)^3 \\
&= (x_1 y_1 + x_2 y_2 + 1)^3 \\
&= (x_1 y_1 + x_2 y_2 + 1)(x_1 y_1 + x_2 y_2 + 1)(x_1 y_1 + x_2 y_2 + 1) \\
&= x_1 y_1^3 + x_1 y_1^2 + x_2 y_2 + x_1 y_1^2 + x_1 y_1^2 x_2 y_2 + x_1 y_1 x_2 y_2^2 + x_1 y_1 x_2 y_2 \\
&+ x_1 y_1^2 + x_1 y_1 x_2 y_2 + x_1 y_1 + x_1 y_1 x_2 y_2^2 + x_2 y_2^3 + x_2 y_2^2 + x_1 y_1^2 x_2 y_2 \\
&+ x_1 y_1 x_2 y_2^2 + x_1 y_1 x_2 y_2 + x_1 y_1 x_2 y_2 + x_2 y_2^2 + x_2 Y - 2 + x_1 y_1^2 + x_1 y_1 x_2 y_2 \\
&+ x_1 y_1 + x_1 y_1 x_2 y_2 + x_2 y_2^2 + x_2 y_2 + x_1 y_1 + x_2 y_2 + 1 \\
&= x_1 y_1^3 + x_2 y_2^3 + 3 x_1 y_1^2 + 3 x_2 y_2^2 + 3 x_1 y_1 + 3 x_2 y_2 + 3 x_1 y_1^2 x_2 y_2 + 3 x_1 y_1 x_2 y_2^2 + 6 x_1 y_1 x_2 y_2 + 1
\end{aligned}$$

Therefore,

$$\psi = \{x_1^3, \sqrt{3} x_1^2, \sqrt{3} x_1, \sqrt{3} x_2^2, \sqrt{3} x_2, \sqrt{3} x_1^2 x_2, \sqrt{3} x_1 x_2^2, \sqrt{6} x_1 x_2, x_2^3, 1\}$$

(b) Full rational variety.

(c) By using $K(x, y)$ instead of $\psi(x) \cdot \psi(y)$, we save 6 multiplication operations.

$\psi(x) \cdot \psi(y) = 10$ operations.
$K(x, y) = 4$ operations.
$10 - 4 = 6$ operations saved.

Question 2

$$
\begin{aligned}
L(x,y,\lambda) &= (2x-y)+\lambda(\frac{x^2}{4}+y^2-1)\\
L_x &= 2+\frac{1}{2}\lambda x = 0\\
x &= -\frac{4}{\lambda} \qquad (i)\\
L_y &= -1+2\lambda y = 0\\
y &= \frac{1}{2\lambda} \qquad (ii)\\
L_\lambda &= \frac{x^2}{4}+y^2-1 = 0
\end{aligned}
$$

Plug (i) and (ii) into L_λ

$$
\begin{aligned}
\frac{(\frac{-4}{\lambda})^2}{4}+(\frac{1}{2\lambda})^2-1 &= 0\\
\frac{4}{\lambda^2}+\frac{1}{4\lambda^2}-1 &= 0\\
\frac{1}{\lambda^2}\cdot\frac{17}{4}-1 &= 0\\
\lambda^2 &= \frac{17}{4}\\
\lambda &= \pm\frac{\sqrt{17}}{2}
\end{aligned}
$$

$$
x = \frac{-4}{\lambda} = \frac{-4}{\frac{\sqrt{17}}{2}} = \frac{-8}{\pm\sqrt{17}} \tag{1}
$$

$$
y = \frac{1}{2\lambda} = \frac{1}{\pm\sqrt{17}} \tag{2}
$$

To get the minimum and maximum values, we plug both pairs of points into f

Min:

$$
\begin{aligned}
f(\frac{-8}{\sqrt{17}},\frac{1}{\sqrt{17}}) &= 2(\frac{-8}{\sqrt{17}})-\frac{1}{\sqrt{17}}\\
&= \frac{-16}{\sqrt{17}}-\frac{1}{\sqrt{17}}\\
&= \frac{-17}{\sqrt{17}}\\
&\approx -4.12
\end{aligned}
$$

Max:

$$
\begin{aligned}
f(\frac{8}{\sqrt{17}},\frac{-1}{\sqrt{17}}) &= 2(\frac{8}{\sqrt{17}})+\frac{1}{\sqrt{17}}\\
&= \frac{17}{\sqrt{17}}\\
&\approx 4.12
\end{aligned}
$$

Therefore the maximum point under the constraint is 4.12 and the minimum point under the constraint is -4.12.

Question 3

$$
\begin{aligned}
(x,y)\cdot u \le r &\iff \frac{\sqrt{3}}{2}x + \frac{1}{2}y \le r \\
(x,y)\cdot v \le r &\iff -y \le r \\
(x,y)\cdot w \le r &\iff \frac{-\sqrt{3}}{2}x + \frac{1}{2}y \le r
\end{aligned}
$$

This region defines an equilateral triangle where $r*$ defines the true target concept and we are learning r.
Learning algorithm: for each sample (x,y), compute r then take the max r from all m samples.
This is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.

$r \le r*$ and we will analyze the error region between the two triangles and want to verify it's bounded by ϵ to satisfy PAC learnability. This can be broken down into two cases.

- If $Pr[(x,y) \in T_r] \le \epsilon$, then we are done.
  Define $r^{\epsilon} = arginf_r Pr[(x_1, x_2) \in T_r] \le \epsilon$
  (the triangle with the largest perimeter with probability less than or equal to ϵ).
- Case 1: if $r^{\epsilon} \le r$, then probability of T_r is less than ϵ.
- Case 2: probability of missing T_r with m training samples.

$(1-\epsilon)^m \le e^{-\epsilon m}$ with sample size $m \le \frac{ln(1/\delta)}{\epsilon}$

$$
\begin{aligned}
\implies e^{-\epsilon m} &\le e^{ln(1/\delta)} \\
&\le e^{ln(\delta)} \\
&\le \delta
\end{aligned}
$$

Therefore, if the probability of the region between the two triangles is small, the error it incurs is also small. The time complexity of this algorithm is $O(m)$, as it has to iterate through each of the m points in the dataset. L operates in time and sample complexity that is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$. Therefore, L is a PAC learnable algorithm.

Question 4

$n = 1000$ (samples)
$\hat{p} = \frac{r}{n} = 0.2$ Therefore $r = 200$ (generalization error)

$$\begin{aligned} se &= \sqrt{\frac{\hat{p}(1-p)}{n}} \\ &= \sqrt{\frac{(0.2)(0.8)}{1000}} \\ &= \sqrt{\frac{0.16}{1000}} \end{aligned}$$

95% confidence interval = $\hat{p} \pm 2se$

$$\begin{aligned} \hat{p} \pm 2se &= 0.2 \pm 2\left(\sqrt{\frac{0.16}{1000}}\right) \\ &= [0.175, 0.225] \end{aligned}$$

Therefore, the true error that they can expect is up to 22.5% .

Question 5

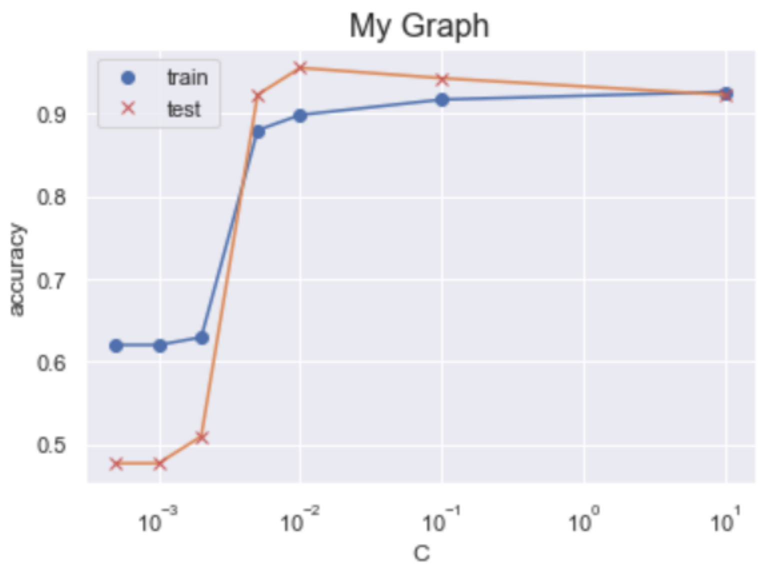

Figure 1: SVM Accuracy Graph