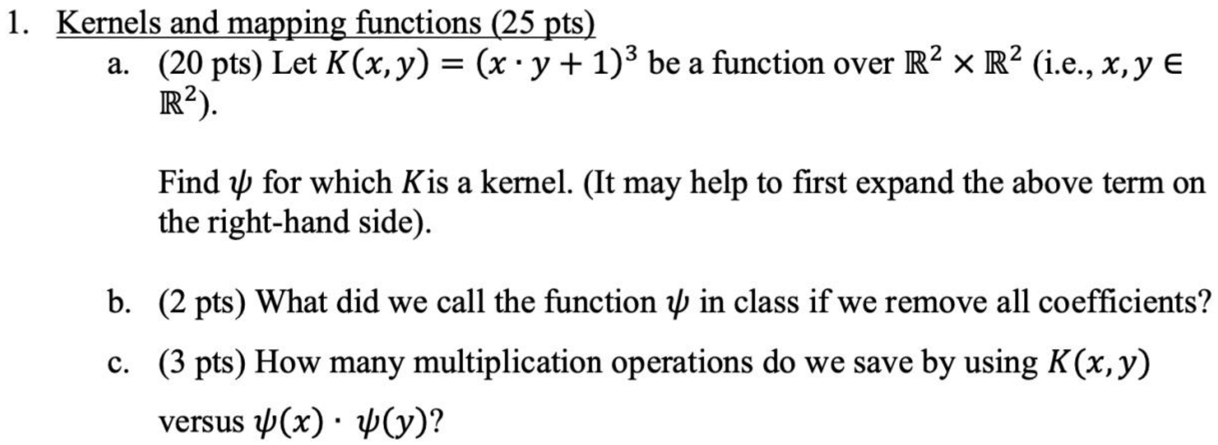
Exercise 5: Theory + SVM

Or Livne – 203972922

Roy Rubin – 201312907



1. Kernels and mapping functions (25 pts)
   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 $\psi(x)\cdot\psi(y)$?

1. Q1 – kernels and mapping functions

a. Finding the mapping function:

- Reminder: a function K is called a kernel if there exists a mapping function ψ so that the following holds: $K(x,y)=\psi(x)\cdot\psi(y)$
- Note: in ex.1.a., K is defined over $\mathbb{R}^2\times\mathbb{R}^2$ which means that x and y are vectors of size 2.
- Using the hint:

$$(x\cdot y+1)^3 \underbrace{=}_{x,y\in\mathbb{R}^2} ((x_1,x_2)^T\cdot(y_1,y_2)+1)^3=(x_1y_1+x_2y_2+1)^3$$

$$=(x_1y_1+x_2y_2+1)\cdot(x_1y_1+x_2y_2+1)\cdot(x_1y_1+x_2y_2+1)$$

$$=(x_1y_1+x_2y_2+1)\cdot((x_1y_1)^2+(x_2y_2)^2+(1)^2+2\cdot x_1y_1\cdot x_2y_2+2\cdot x_1y_1\cdot 1+2\cdot x_2y_2\cdot 1)$$

$$\begin{gathered}=(x_1y_1)^3+x_1y_1\cdot(x_2y_2)^2+x_1y_1+2(x_1y_1)^2\cdot x_2y_2+2(x_1y_1)^2+2x_1y_1\cdot x_2y_2+\\+x_2y_2\cdot(x_1y_1)^2+(x_2y_2)^3+x_2y_2+2x_1y_1\cdot(x_2y_2)^2+2x_1y_1\cdot x_2y_2+2(x_2y_2)^2+\\+(x_1y_1)^2+(x_2y_2)^2+(1)^2+2x_1y_1\cdot x_2y_2+2x_1y_1+2x_2y_2\end{gathered}$$

$$\begin{gathered}=(x_1y_1)^3+(x_2y_2)^3+1+3(x_1y_1)^2\cdot x_2y_2+3x_1y_1\cdot(x_2y_2)^2+3(x_1y_1)^2+3(x_2y_2)^2+3x_1y_1\\+3x_2y_2+6x_1y_1\cdot x_2y_2\end{gathered}$$

So, what we are actually looking for is:

$$\begin{aligned}&<\psi(\vec{x}),\psi(\vec{y})>\\&\qquad=(x_1y_1)^3+(x_2y_2)^3+1+3(x_1y_1)^2\cdot x_2y_2+3x_1y_1\cdot(x_2y_2)^2+3(x_1y_1)^2\\&\qquad+3(x_2y_2)^2+3x_1y_1+3x_2y_2+6x_1y_1\cdot x_2y_2\end{aligned}$$

- Definition of ψ :

$$\forall\, \vec{x} \in \mathbb{R}^2 \;\; we\ define\text{:}\;\; \psi(\vec{x}) = \psi(x_1, x_2)$$
$$= (x_1^3, x_2^3, 1, \sqrt{3}x_1^2x_2, \sqrt{3}x_1x_2^2, \sqrt{3}x_1^2, \sqrt{3}x_2^2, \sqrt{3}x_1, \sqrt{3}x_2, \sqrt{6}x_1x_2)$$

- Proof:

$$< \psi(\vec{x}), \psi(\vec{y}) >$$
$$= \left(x_1^3, x_2^3, 1, \sqrt{3}x_1^2x_2, \sqrt{3}x_1x_2^2, \sqrt{3}x_1^2, \sqrt{3}x_2^2, \sqrt{3}x_1, \sqrt{3}x_2, \sqrt{6}x_1x_2\right)^T$$
$$\cdot \left(y_1^3, y_2^3, 1, \sqrt{3}y_1^2y_2, \sqrt{3}y_1y_2^2, \sqrt{3}y_1^2, \sqrt{3}y_2^2, \sqrt{3}y_1, \sqrt{3}y_2, \sqrt{6}y_1y_2\right)$$
$$= x_1^3y_1^3 + x_2^3y_2^3 + 1 + 3x_1^2y_1^2x_2y_2 + 3x_1y_1x_2^2y_2^2 + 3x_1^2y_1^2 + 3x_2^2y_2^2 + 3x_1y_1$$
$$+ 3x_2y_2 + 6x_1y_1x_2y_2 = \underbrace{\cdots}_{\substack{opposite\ way\ was \\ shown\ earlier}} = (x \cdot y + 1)^3$$

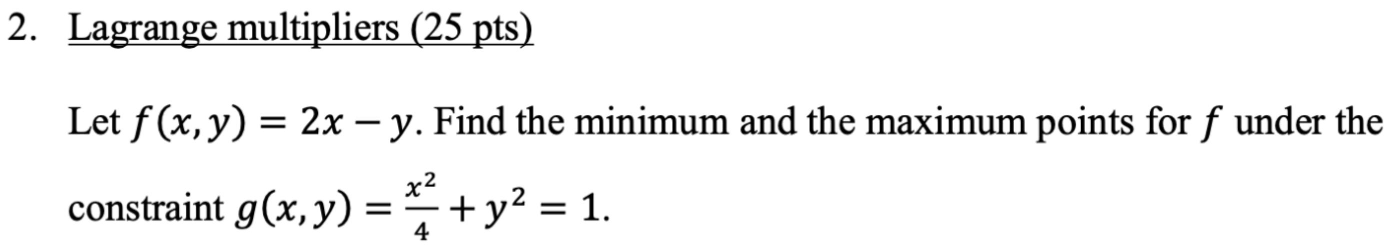
b. If we remove all coefficients, we can call the function ψ a full rational variety of order 3 in an input space of dimension 2.

c. How many multiplications are saved?

- Note that the kernel takes 4 multiplication operations:

$$K(x, y) = (x \cdot y + 1)^3 = \left((x_1, x_2) \cdot \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + 1\right)^3 = \left(\underbrace{x_1y_1}_{mul.1} + \underbrace{x_2y_2}_{mul.2} + 1\right)^{\underbrace{3}_{mul.3, mul.4}}$$

- Note that takes the inner product $< \psi(\vec{x}), \psi(\vec{y}) >$ of takes 10 multiplication operations: inner product of 2 vectors of size 10
- That means we saved 6 multiplications.
- Notes: we did not count the scaler multiplications, we did not count the transformation multiplications, etc.



2. Lagrange multipliers (25 pts)

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

2. Q2 - Solving with LaGrange multipliers:

- Finding the maximum and minimum points for the function f under the constraint g: we will extract x, y from the equation: $\nabla f(x,y) = \lambda \nabla g(x,y)$ which is equivalent to:

$$\left(\frac{df}{dx}, \frac{df}{dy}\right) = \left(\lambda \cdot \frac{dg}{dx}, \lambda \cdot \frac{dg}{dy}\right)$$

Meaning we can extract:

$$\frac{df}{dx} = \lambda \cdot \frac{dg}{dx}$$
$$\frac{df}{dy} = \lambda \cdot \frac{dg}{dy}$$

- Finally, we will solve a 3-way equation with the help of the original given constraint:

$$\begin{cases} \frac{df}{dx} = \lambda \cdot \frac{dg}{dx} \\ \frac{df}{dy} = \lambda \cdot \frac{dg}{dy} \\ constraint \end{cases}$$

This will give us 2 λ values. We will extract them, and use them to find the min and max points:

$$\begin{cases} 2 = \lambda \cdot \frac{x}{2} \\ -1 = \lambda \cdot 2y \\ \frac{x^2}{4} + y^2 = 1 \end{cases}$$

$$\begin{cases} x = \frac{4}{\lambda} \\ y = -\frac{1}{2\lambda} \\ x^2 + 4y^2 = 4 \end{cases}$$

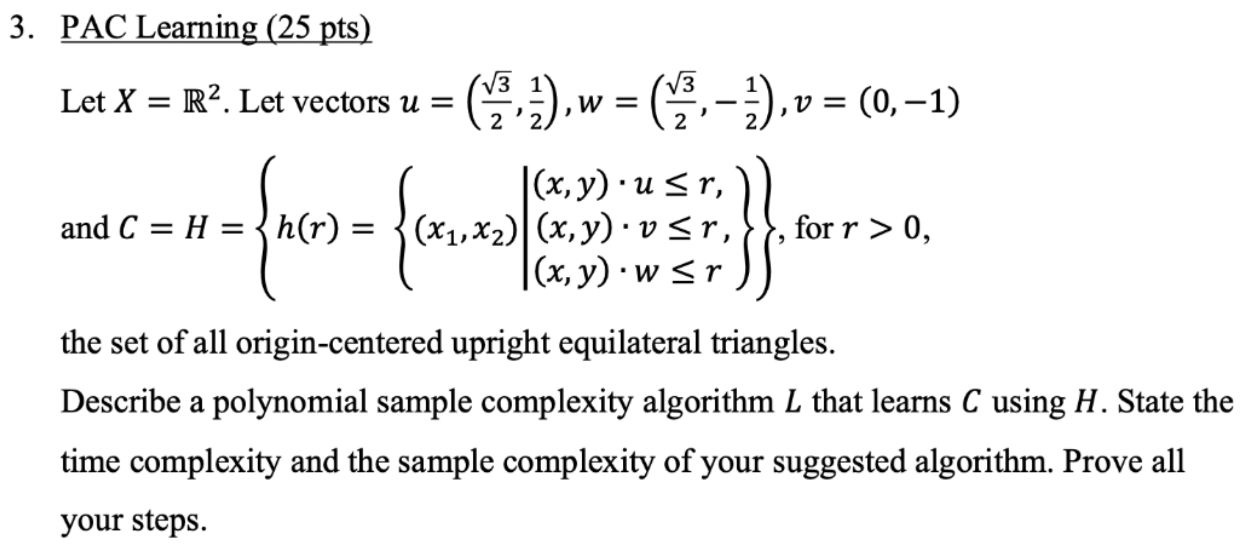
$$\left(\frac{4}{\lambda}\right)^2 + 4 \cdot \left(-\frac{1}{2\lambda}\right)^2 = 4$$
$$\frac{16}{\lambda^2} + 4 \cdot \frac{1}{4\lambda^2} = 4 \quad | \cdot \lambda^2$$
$$16 + 1 = 4\lambda^2$$
$$\lambda^2 = \frac{17}{4}$$
$$\lambda = \pm\frac{\sqrt{17}}{2}$$

- We received:

$$\begin{cases} x = \dfrac{4}{\lambda} = \dfrac{4}{\pm\dfrac{\sqrt{17}}{2}} = \pm\dfrac{8}{\sqrt{17}} \\ y = -\dfrac{1}{2\lambda} = -\dfrac{1}{2\cdot\left(\pm\dfrac{\sqrt{17}}{2}\right)} = \pm\dfrac{1}{\sqrt{17}} \end{cases}$$

$$possible\ points{:}\left(\pm\frac{8}{\sqrt{17}},\pm\frac{1}{\sqrt{17}}\right)$$

- When analyzing the function f, we note that the possible point that will maximize f will occur when x is positive and y is negative, and the possible point that will minimize f will occur when x is negative, and y is positive.
- So, the maximum point of the function f will occur when: $\left(+\frac{8}{\sqrt{17}},-\frac{1}{\sqrt{17}}\right)$
- The maximum point of the function f will occur when: $\left(-\frac{8}{\sqrt{17}},+\frac{1}{\sqrt{17}}\right)$



3. PAC Learning (25 pts)

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

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

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.

3. Q3 – PAC Learning:

- Note that the points from the original exercise (see above) were not correct. The correct points are: $u = \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right), \mathrm{w} = \left(-\frac{\sqrt{3}}{2}, \frac{1}{2}\right), v = (0, -1)$
- The algorithm will produce a hypothesis which is the smallest relevant origin-centered upright equilateral triangles that contains all the positive points.
- This can be done in $O(m)$ $(m = \#\, of\ instances)$ as follows:
    - Let $\Delta = \Delta^m = (x_i, y_i)_{i=0.}^{m}$ be a set of points in the 2D space, labeled positive and negative. Our algorithm seeks to return a hypothesis $h \in H$.

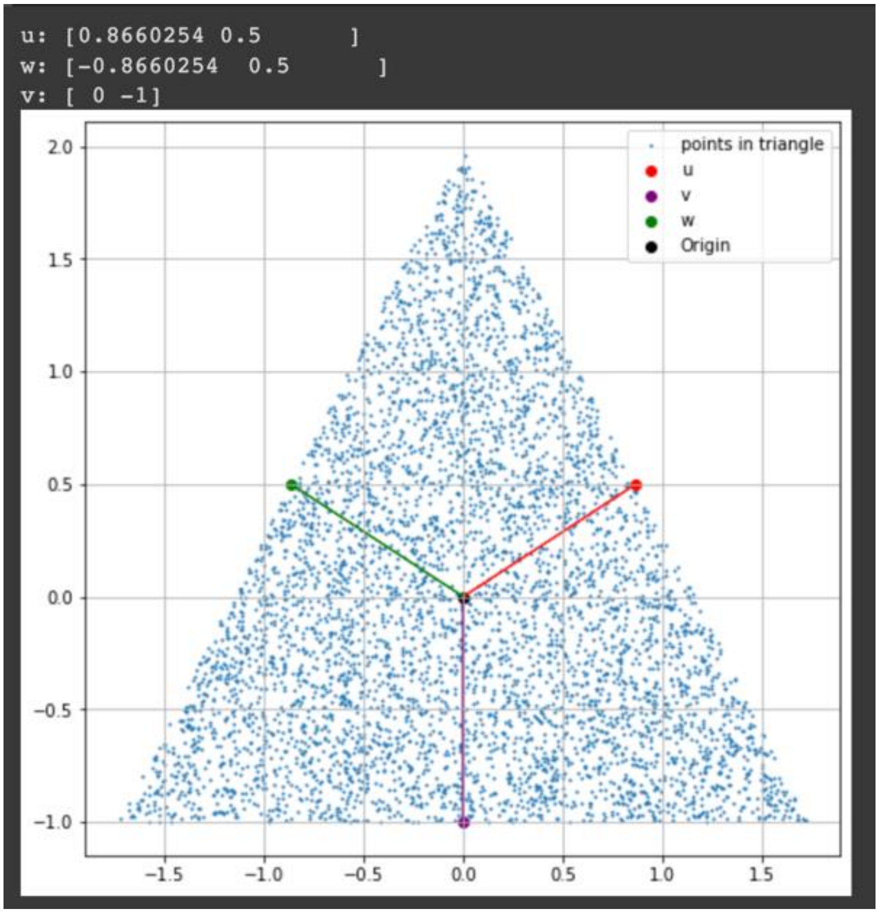
    - Base condition there are 3 equations:

      a) $\mathrm{u}^T(x_i - (0,0)) < r \rightarrow \frac{\sqrt{3}}{2}x + \frac{1}{2}y \le r \rightarrow y \le 2r - \sqrt{3}x$

      b) $\mathrm{v}^T(x_i - (0,0)) < r \rightarrow 0 * x - 1y \le r \rightarrow -y \le r \rightarrow y \ge r$

      c) $\mathrm{w}^T(x_i - (0,0)) < r \rightarrow -\frac{\sqrt{3}}{2}x + \frac{1}{2}y \le r \rightarrow y \ge +2r - \sqrt{3}x$

- Figure of triangle:



```
u: [0.8660254 0.5       ]
w: [-0.8660254  0.5       ]
v: [ 0 -1]
```

- Base the figure above, let $(x_i, y_i)_{i=1}^{m^+}$ be all positive labeled data points, which means
  - $a \coloneqq max_{1 \leq i \leq m^+}(x_i)$
  - $b \coloneqq max_{1 \leq i \leq m^+}(y_i)$
  - $s \coloneqq max_{1 \leq i \leq m^+}\,(x_i + y_i)$
- The vertices of the hypothesis triangle $h = L(\Delta)$ will be $\lambda = (a, b), \alpha = (a, s - a), \beta = (s - b, b)$
- Consider $c \in C, and\ let\ \Delta^m(c) = \left(x_i(c), y_i(c)\right)_{i=1}^{m}$ be training data generated from c without errors and by drawing m independent points according to a probability distribution π in $\mathbb{R}^2$. We will denote the probability distribution thus induced on ($\mathbb{R}^2$) by π^m.
- Given $\epsilon > 0\ and\ \delta > 0$, we will now compute a number m (ϵ, δ) so that **Eq1**:

  $m \geq\ m(\epsilon, \delta)$ → e($\Delta^m(c)$) = $\pi^m\left(err_\pi\left(L\left(\Delta^m(C)\right)\right), C\right) > \epsilon\ ) \leq \delta$
- Note that $L\left(\Delta^m(C)\right)$ is the hypothesis h, or the triangle, produced by L when considering data $\Delta^m(c)$ as above.
- e ($\Delta^m(c)$) is a random variable that depends on the stochastic behavior of $\Delta^m(c)$
- it is exactly this behavior that we will want to characterize.
- Consider the strips parallel to the edges of the triangle c as in figure above, these are defined to satisfy:

$$\pi\big(S1(\epsilon)\big) = \pi\big(S2(\epsilon)\big) = \pi\big(S3(\epsilon)\big) = \frac{\epsilon}{3}$$

- Now note that:

$$\{\mathit{\Delta}^m(c)\colon err_\pi\big(L\big(\mathit{\Delta}^m(C)\big),c\big) > \epsilon\} \subseteq D$$

- Where D: = $\{\mathit{\Delta}^m(C)\colon \mathit{\Delta}^m(C) \cap S1(\epsilon) = \Phi\} \cup \{\mathit{\Delta}^m(C)\colon \mathit{\Delta}^m(C) \cap S2(\epsilon) = \Phi\} \cup \{\mathit{\Delta}^m(C)\colon \mathit{\Delta}^m(C) \cap S3(\epsilon) = \Phi\}$
- This because if $\mathit{\Delta}^m(c)$ visits all three strips (note that negative points cannot visit the strips as there are no errors) then, according to our construction, the difference between c and $L\big(\mathit{\Delta}^m(C)\big)\ will\ have\ \pi \le \pi(S1(\epsilon) \cup S2(\epsilon) \cup S3(\epsilon))$.
- In term of probability, we therefore get:

$$\pi^m\left(err_\pi\left(L\big(\mathit{\Delta}^m(C)\big)\right), C\right) > \epsilon\ )$$
$$\le\ \pi^m(\mathit{\Delta}^m(C) \cap S1(\epsilon) = \Phi) +\ \pi^m(\mathit{\Delta}^m(C) \cap S2(\epsilon) = \Phi) +\ \pi^m(\mathit{\Delta}^m(C) \cap S3(\epsilon) = \Phi)$$
$$\le 3\left(1 - \frac{\epsilon}{3}\right)^m$$

- Now we will select $m(\epsilon,\delta)$ = $\frac{3}{\epsilon}\left(\ln(3) + \ln\left(\frac{1}{\delta}\right)\right)$ to get **Eq 1** to hold.

  - ❖ $r^h\colon= max_{1\le i\le m}((\mathrm{v}^T x_i - (0,0))$ = $(0,-1) * (x,r) = -r$
    a. The maximal product between point in X to the v vector from the origin

  - ❖ $r^h\colon= max_{1\le i\le m}(\big(\mathrm{w}^T x_i - (0,0)\big) = \left(-\frac{\sqrt{3}}{2}, +\frac{1}{2}\right) * (x,r) = -r$
    a. The maximal product between point in X to the w vector from the origin
  - ❖
  - ❖ $a\colon= max_{1\le i\le m}\big(\mathrm{u}^T x_i - (0,0)\big) = \left(\frac{\sqrt{3}}{2},\frac{1}{2}\right) * (0,2r) = r$
    a. The maximal product between point in X to the u vector from the origin

  - ❖ $r^h\colon= max_{1\le i\le m}((\mathrm{v}^T x_i - (0,0))$ = $(0,-1) * (x,r) = -r$
    a. The maximal product between point in X to the v vector from the origin

  - ❖ $r^h\colon= max_{1\le i\le m}(\big(\mathrm{w}^T x_i - (0,0)\big) = \left(-\frac{\sqrt{3}}{2}, +\frac{1}{2}\right) * (x,r) = -r$
    a. The maximal product between point in X to the w vector from the origin

- Simply, this is an origin-centered equilateral triangles where the sizes r,r are the maximal distances we've seen in the training set.
- It spans from (r^h, r^h), this box is contained in the ground-truth box (concept c).
- Time complexity is: O(m) for each vector → Total O(2m) = O(m)
- Now we consider the area between our h to c (remember that $h \subseteq c$).
  There are ___ such areas:
  - ▪ for each of the coordinates $(r^h, r^c), (r^h, r^c)$
    - ○ $B_1 = (r^c - r^h) \cdot 2r$
    - ○ $B_2 = (r^c - r^h) \cdot 2r$

- Consider training data, $D \in X^m$.
- Assume that D visits each one of the 2 sets B, defined above.
- What can we say about Err(h, c)? $P(B_i) \leq \frac{\varepsilon}{2}$
- So, the probability of a point in $D \in X^m$ to be in **either** of those areas B_i is $\geq 1 - \frac{\varepsilon}{2}$
- For a given ε and δ, the number of samples needed is:

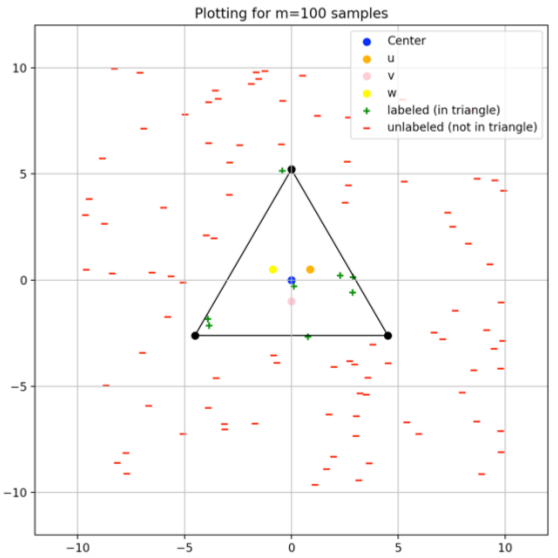
$$P(\{D \in X^m \colon Err(h = L(D), c) > \varepsilon\}) \leq \delta$$

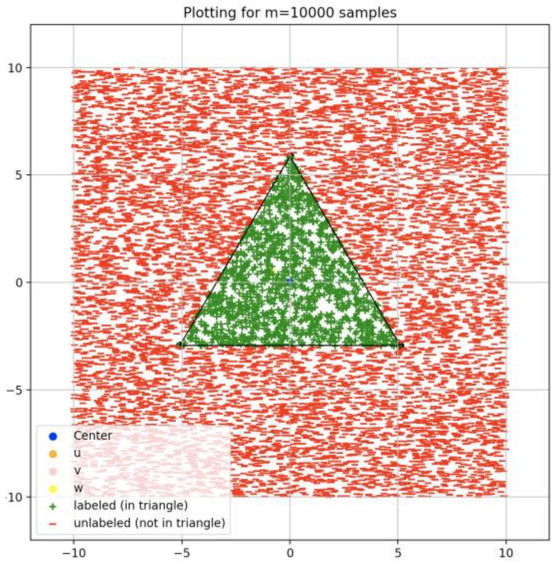
$$P(\{D \in X^m \colon Err(h = L(D), c) > \varepsilon\}) \leq \sum_{i=1}^{2} (P(X - B_i))^m \leq 2(1 - \frac{\varepsilon}{2}) \leq 2e^{-\frac{m\varepsilon}{2}} \leq \delta$$

$$2e^{-\frac{m\varepsilon}{2}} \leq \delta \Rightarrow ln(2e^{-\frac{m\varepsilon}{2}}) \leq ln(\delta) \Rightarrow ln(2) + ln(e^{-\frac{m\varepsilon}{2}}) \leq ln(\delta) \Rightarrow ln(e^{-\frac{m\varepsilon}{2}}) \leq ln(\frac{\delta}{2}) \Rightarrow -\frac{m\varepsilon}{2} \leq ln(\frac{\delta}{2})$$

$$\Rightarrow m \geq \frac{2}{\varepsilon} ln(\frac{2}{\delta})$$

- Meaning, when we want a confidence of $1 - \delta$ to get an error of ε, we will need **at least** $\frac{2}{\varepsilon} ln(\frac{2}{\delta})$ training instances.
- Some plots for intuition:

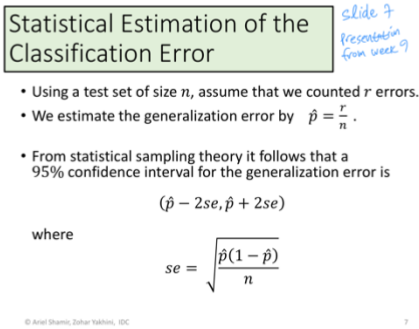




4. (15 pts) A business manager at your ecommerce company asked you to make a model to predict whether a user is going to proceed to checkout or abandon their cart. You created the model using, and reported 20% error on your test set of size 1000 samples. In the business manager's presentation to upper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

4. Q4 – Statistical estimation of the classification error:

- A short reminder from the lecture:



Statistical Estimation of the Classification Error

slide 7
Presentation from week 9

- Using a test set of size n, assume that we counted r errors.
- We estimate the generalization error by $\hat{p} = \frac{r}{n}$.
- From statistical sampling theory it follows that a 95% confidence interval for the generalization error is

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

where

$$se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

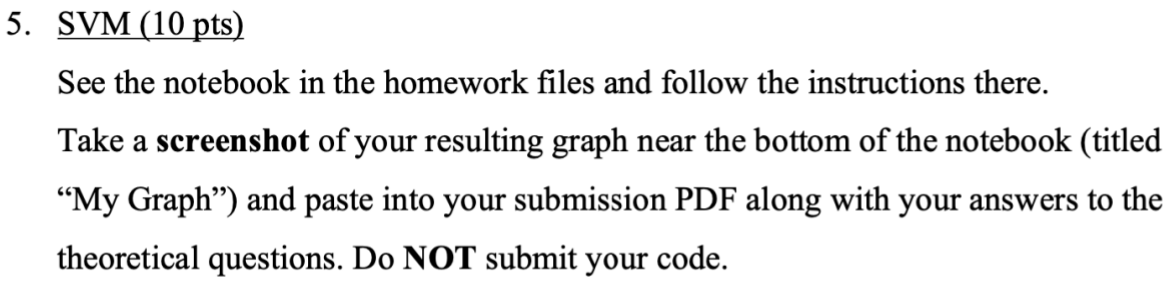
© Ariel Shamir, Zohar Yakhini, IDC

7

- We note the generalization error is: $\hat{p} = \frac{r}{n} = \frac{200}{1000} = 0.2$
- We calculate the "se" value: $se = \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}} = \sqrt{\frac{0.2\cdot(1-0.2)}{1000}} = 0.01264$
- With the calculations above, the error margins are:

$$(\hat{p} - 2se, \hat{p} + 2se) = (0.2 - 2\cdot 0.01264, 0.2 + 2\cdot 0.01264\,) = (0.17472, 0.22528)$$

- And so, the true error could be up to: 22.53%

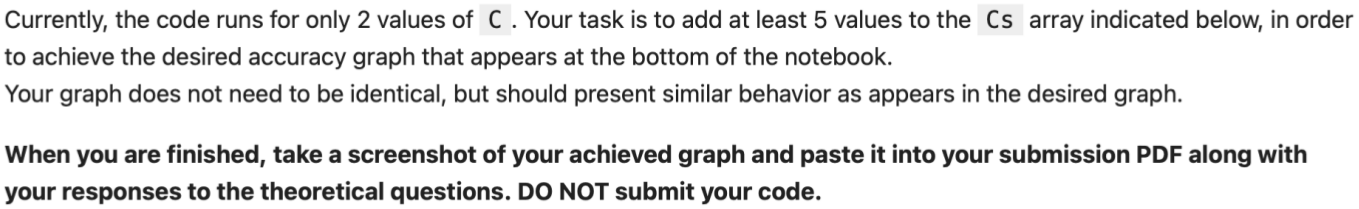


5. SVM (10 pts)

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.

From inside the notebook:

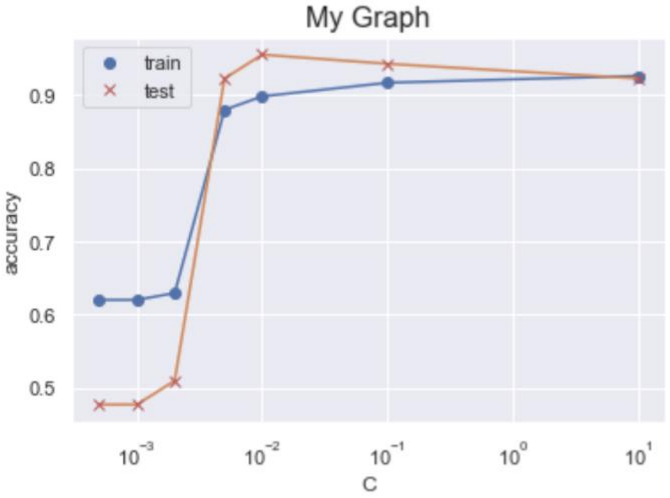


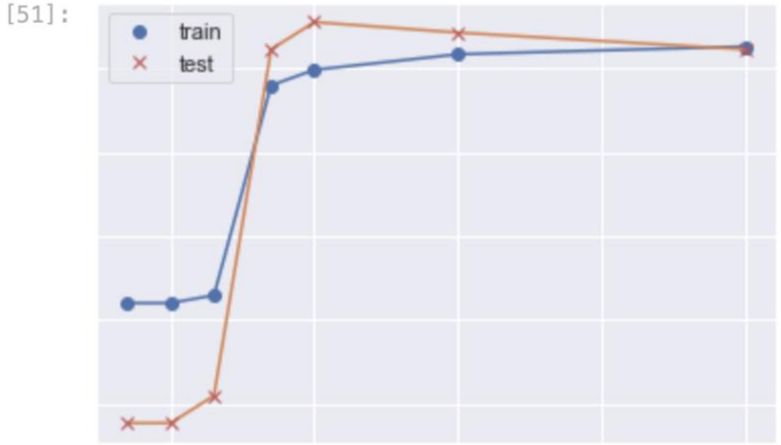
Currently, the code runs for only 2 values of `C`. Your task is to add at least 5 values to the `Cs` array indicated below, in order to achieve the desired accuracy graph that appears at the bottom of the notebook.
Your graph does not need to be identical, but should present similar behavior as appears in the desired graph.

When you are finished, take a screenshot of your achieved graph and paste it into your submission PDF along with your responses to the theoretical questions. DO NOT submit your code.

5. Q5 – SVM with jupyter notebook

- My chosen C values: [0.0005, 0.001, 0.002, 0.005, 0.01, 0.1, 10]
- Final graph in comparison to needed graph:


My Graph
● train
× test
accuracy
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
C


[51]:
● train
× test