בית ספר "אפי ארזי" למדעי המחשב המרכז הבינתחומי

The Efi Arazi school of computer science

The Interdisciplinary Center

סמסטר ב' תשע"ט

Spring 2019

מבחן מועד א בלמידה ממוכנת

Machine Learning Exam A

Lecturer: Prof Zohar Yakhini
Time limit: 3 hours
Additional material or calculators are not allowed in use!

Answer 5 out of 6 from the following question (each one is 20 points)

Good Luck!

מרצה: פרופ זהר יכיני
משך המבחן: 3שעות
אין להשתמש בחומר עזר ואין להשתמש במחשבונים!

יש לענות על 5 מתוך 6 השאלות הבאות
לכל השאלות משקל שווה (20 נקודות)

בהצלחה!

Question 1 (4 parts)

A. In linear regression, why is it common to perform feature normalization prior to the learning phase? Is it common to use feature normalization in the closed-form solution to linear regression ($\vec{\theta} = \text{pinv}\,X \cdot \vec{y}$)? Explain.

B. As part of your job as a machine learning engineer, you are developing a linear regression model using some data your team obtained in laboratory conditions. All of your data (x) comes from a limited range, namely it satisfies $x_i \in [0,10]$. After training the model and getting high accuracy on the training dataset, you start testing your model in the field. You soon realize that the x-values of most of the data points your model encounters in the field are in the range $[100,200]$.

   1. Assume that in reality the function you are modeling is linear. What performance do you expect to see in your field testing? Explain your answer.
   2. One of your colleagues suggests that you use kNN with k=5 for your field predictions. Is this a good suggestion? Explain.

C. A real estate company is interesting in predicting house prices based on 10 measurable features. Furthermore, for some houses it is more important that the prediction be accurate than it is for others. How do you change the linear regression framework to address this task?

D. You are performing linear regression on training data (x_i, y_i). What do you expect the MSE to be under the following conditions:

   1. The Pearson correlation of $\vec{x}$ and $\vec{y}$ is 1.
   2. The Spearman correlation of $\vec{x}$ and $\vec{y}$ is 1.

Question 1 (4 parts) – Solution

A. Feature normalization is common since it may result in faster convergence if the features have different scales. There is no need to use feature normalization when using the closed form solution since there is no convergence involved and the solution will be found regardless of normalization.

B.

1. We will expect to see a good fit since the data came from a linear function. The fit we obtain in the range [0,10] will prevail also in the higher x range since the reality is assumed to be linear.
2. This is not good advice. Using kNN we will return y values that are compatible with the training range. As the x training range is dramatically different from that observed in the fields, these values likely to be wrong. The case when the slope is zero or close to zero is an exception, but note that this means that y doesn't really depend on x.

C. Weighted linear regression assigns a different weight w_i to every instance (x_i, y_i). This allows to adjust the contribution of each instance to the update rule (gradient descent) and control how much every instance will affect the learnt parameters.
In other words, we'll use the following cost function:

$$MSE(\vec{\theta}) = \frac{1}{m}\sum_{i=1}^{m} w_i\left(\vec{\theta}\vec{x}_i - y_i\right)^2$$

D.

1. We expect perfect linear regression since if the Pearson correlation is 1, one vector (y) can be fully described by the other (x) using a linear equation. In this case the MSE will be zero.
2. We can't expect anything in this case. It is possible to have Spearman correlation 1 and Pearson correlation is 1 and thus have perfect linear fit (MSE=0). It is also possible that Spearman correlation is 1 and Pearson's correlation is much lower. No predictions regarding the MSE can be made in this case.

Question 2 (5 parts)

A. Consider classes $(A_1, A_{,2} \dots, A_i)$ consisting of elements with properties $(x_1, x_2 \dots, x_j)$. Write the classification formula for:

1. Naïve Bayes
2. Full Bayes

In rolling two dies with 6 sides each we have the following distributions of results for two casino houses – Casino A and Casino B:

Casino A

Dice2 Dice1	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

Casino B

Dice2 Dice1	1	2	3	4	5	6
1	$\frac{1}{18}$	$\frac{1}{18}$	0	0	0	$\frac{1}{18}$
2	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	0	0	0
3	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	0	0
4	0	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	0
5	0	0	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$
6	$\frac{1}{18}$	0	0	0	$\frac{1}{18}$	$\frac{1}{18}$

The prior probability for playing in Casino A is $\frac{3}{5}$.

Given a game outcome (2 numbers), we want to classify whether the game was played in Casino A or B.

B. We observe the following game outcome: 1st die is 6, 2nd die is 1. Which casino will a Naïve Bayes classifier predict? Show your calculations.

C. Given the same game result as in Part B, which casino will a Full Bayes classifier predict? Show your calculations.

D. What is the minimal prior we need to assign to Casino A in order for the Full Bayes classifier to predict A regardless of the game outcome? Show/explain your calculations.

E. You can now change two entries in the joint distribution matrix of Casino B. Given the same results as in Part B (that is: (6,1)), perform a change that will lead the full Bayes classifier to select Casino B under the prior you had found in Part D.

Your newly defined distribution should be an adequate probability distribution.

Question 2 (5 parts) – Solution

A. Assuming n features per instance:

1. Naïve Bayes – $i(\vec{x}) = \underset{i}{\operatorname{argmax}} P(A_i) \cdot \prod_{j=1}^{n} P\left(x_j \middle| A_i\right)$
2. Full Bayes – $i(\vec{x}) = \underset{i}{\operatorname{argmax}} P(A_i) \cdot P(x_1, x_2 \dots, x_n | A_i)$

B. We first notice that $P\left(x_j \middle| B\right) = \frac{1}{6}\ \forall j \in \{1, \dots, 6\}$, and so we get:

$$P\left((6,1) \middle| A\right) \cdot P(A) = \frac{1}{36} \cdot \frac{3}{5} = \frac{1}{60}$$

$$P\left((6,1) \middle| B\right) \cdot P(B) = \frac{1}{36} \cdot \frac{2}{5} = \frac{1}{90}$$

So A will be chosen.

C.

$$P((6,1) \mid A) \cdot P(A) = \frac{1}{36} \cdot \frac{3}{5} = \frac{1}{60}$$

$$P\left((6,1) \middle| B\right) \cdot P(B) = \frac{1}{18} \cdot \frac{2}{5} = \frac{1}{45}$$

So B will be chosen.

D.

We need to solve:

$$\frac{1}{36} \cdot x = \frac{1}{18}(1 - x) \rightarrow x = \frac{2}{3}$$

And so we'll set $P(A) = \frac{2}{3} + \epsilon$

E. We'll change $P((1,6)|B)\ =\ 0$ and $P\left((6,1) \middle| B\right) = \frac{2}{18} = \frac{1}{9}$. This is a valid distribution as the entry in the table sum up to 1. And we get:

$$P((6,1) \mid A) \cdot P(A) = \frac{1}{36} \cdot \frac{2}{3} = \frac{1}{54}$$

$$P\left((6,1) \middle| B\right) \cdot P(B) = \frac{1}{9} \cdot \frac{1}{3} = \frac{1}{27}$$

And now B wins again!

Question 3 (4 parts)

A. Let X be a Poisson random variable.

Recall that then $(\mathrm{X} = k) = e^{-\lambda}\frac{\lambda^k}{k!}$.

Assume that for a series of n independent samples of X we observed the values $x_1, x_2, \ldots, x_n$.

Write down the expression of the probability of the observed data as a function of the parameter λ.

B. Prove that the value of λ given by MLE is:

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$$

C. We are given the following observed data which was sampled from two different Poisson distributions with parameters λ_1, λ_2.

1	1	2
5	11	5
2	0	2
2	1	2
8	4	6

For each row, a coin was tossed. With probability w_1 it led to 3 independent Poisson samples with λ_1 and with probability w_2 it led to 3 independent Poisson samples with λ_2.

Which algorithm would you use to assess the values of the probability parameters for the given scenario? Explain your answer.

D. Use the algorithm from C to compute the first iteration of the algorithm with the following initial values:

$$\lambda_1 = 1, \qquad \lambda_2 = 5, \qquad w_1 = 0.75$$

Question 3 (4 parts) – Solution

A.

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} e^{-\lambda} \cdot \frac{\lambda^{x_i}}{x_i!}$$

B.

$$L(x_1, x_2, \dots, x_n) = \ln \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = \sum_{i=1}^{n} \ln \left(e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \right)$$

$$= \sum_{i=1}^{n} \ln e^{-\lambda} + \sum_{i=1}^{n} \ln \lambda^{x_i} - \sum_{i=1}^{n} \ln x_i!$$

$$= -\lambda n + \ln \lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln x_i!$$

$$\frac{dL}{d\lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i = 0$$

$$\lambda = \frac{1}{n} \sum_{i=1}^{n} x_i$$

C. We will use the EM algorithm. We know that the data came from 2 distributions, but we don’t know which distribution generated each instance. The EM will find the distribution parameters and the probability for each instance to be generated by each of the two distributions.

D. We provide the full calculations below.

Instance 1:

$$P_A(x_1) = 0.75 \cdot e^{-1} \cdot \frac{1^2}{2!} \cdot e^{-1} \cdot \frac{1^1}{1!} \cdot e^{-1} \cdot \frac{1^1}{1!} = 0.01867$$

$$P_B(x_1) = 0.25 \cdot e^{-5} \cdot \frac{5^2}{2!} \cdot e^{-5} \cdot \frac{5^1}{1!} \cdot e^{-5} \cdot \frac{5^1}{1!} = \ 0.00002$$

$$r(x_1, A) = \frac{0.01867}{0.01867 + 0.00383} \approx 1, \qquad r(x_1, B) = \frac{0.00002}{0.01867 + 0.00002} \approx 0$$

Instance 2:

$$P_A(x_2) = 0.75 \cdot e^{-1} \cdot \frac{1^5}{5!} \cdot e^{-1} \cdot \frac{1^{11}}{11!} \cdot e^{-1} \cdot \frac{1^5}{5!} \approx 0$$

$$P_B(x_2) = 0.25 \cdot e^{-5} \cdot \frac{5^5}{5!} \cdot e^{-5} \cdot \frac{5^{11}}{11!} \cdot e^{-5} \cdot \frac{5^5}{5!} = \ 0.00006$$

$$r(x_2, A) = \frac{0}{0.00006} \approx 0, \qquad r(x_2, B) = \frac{0.00006}{0.00006} \approx 1$$

Instance 3:

$$P_A(x_3) = 0.75 \cdot e^{-1} \cdot \frac{1^2}{2!} \cdot e^{-1} \cdot \frac{1^0}{0!} \cdot e^{-1} \cdot \frac{2^1}{2!} = 0.00934$$

$$P_B(x_3) = 0.35 \cdot e^{-5} \cdot \frac{5^2}{2!} \cdot e^{-5} \cdot \frac{5^0}{0!} \cdot e^{-5} \cdot \frac{5^2}{2!} = \ 0.00001$$

$$r(x_3, A) = \frac{0.00934}{0.00934 + 0.00001} \approx 1, \qquad r(x_3, B) = \frac{0.00001}{0.00934 + 0.00001} \approx 0$$

Instance 4:

$$P_A(x_4) = 0.75 \cdot e^{-1} \cdot \frac{1^2}{2!} \cdot e^{-1} \cdot \frac{1^1}{1!} \cdot e^{-1} \cdot \frac{2^1}{2!} = 0.00934$$

$$P_B(x_4) = 0.25 \cdot e^{-5} \cdot \frac{5^2}{2!} \cdot e^{-5} \cdot \frac{5^1}{1!} \cdot e^{-5} \cdot \frac{5^2}{2!} = \ 0.00006$$

$$r(x_4, A) = \frac{0.00934}{0.00934 + 0.00006} \approx 1, \qquad r(x_4, B) = \frac{0.00024}{0.00934 + 0.00006} \approx 0$$

Instance 5:

$$P_A(x_5) = 0.75 \cdot e^{-1} \cdot \frac{1^6}{6!} \cdot e^{-1} \cdot \frac{1^4}{4!} \cdot e^{-1} \cdot \frac{2^8}{8!} = 0$$

$$P_B(x_5) = 0.25 \cdot e^{-5} \cdot \frac{5^6}{6!} \cdot e^{-5} \cdot \frac{5^4}{4!} \cdot e^{-5} \cdot \frac{5^8}{8!} = \ 0.00042$$

$$r(x_5, A) = \frac{0}{0.00042 + 0} \approx 0, \qquad r(x_5, B) = \frac{0.00042}{0.00042 + 0} \approx 1$$

New W:

$$New\ w_A = \frac{1}{5}\sum_{i=1}^{5} r(x_i, A) = 0.6$$

$$New\ w_B = \frac{1}{5}\sum_{i=1}^{5} r(x_i, B) = 0.4$$

New λ:

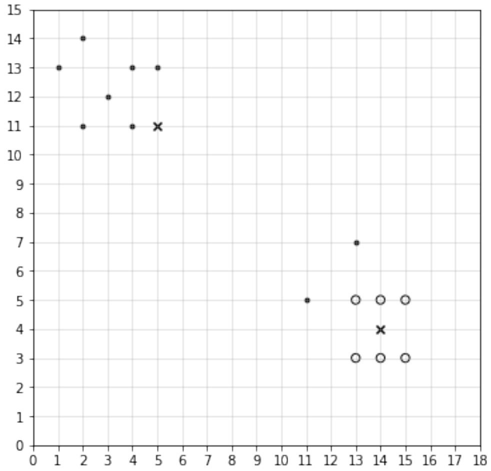
$$\lambda_A = \frac{\left[1 \cdot \frac{(2+1+1)}{3} + 1 \cdot \frac{(2+0+2)}{3} + 1 \cdot \frac{(2+1+2)}{3}\right]}{(1+1+1)} = 1.44$$

$$\lambda_B = \frac{\left[1 \cdot \frac{5+11+5}{3} + 1 \cdot \frac{6+4+8}{3}\right]}{1+1} = 6.5$$

Question 4 (5 parts)

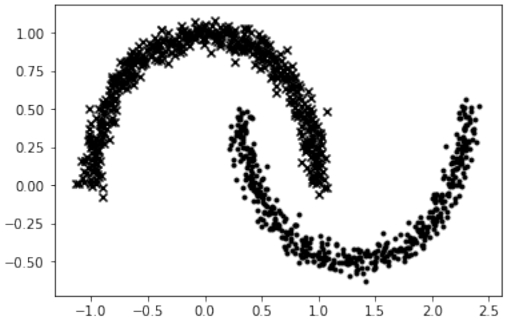
A. What is the function J that the standard k-means algorithm seeks to minimize? Provide a formula.

B. Consider the following status of k-means when applied to data in $\mathbb{R}^2$. The ‘dots’ belong to Cluster 1 and the ‘circles’ belong to Cluster 2. The current cluster centers are given by the ‘x’s.
Compute the exact change in J after updating the assignments (before updating the centers)?



C. Can the function J increase its value during the k-means algorithm? Explain.

D. Consider the following cluster structure. The ‘xs’ (the upper half circle) belong to Cluster 1 and the ‘dots’ (the lower half circle) belong to Cluster 2 (see picture).
Can it be a minimum point for the function J?



E. Suggest a clustering algorithm that can produce the above structure as its output, from the given data.

Question 4 (5 parts) – Solution

A. The function J is the sum of distances of all points from their assigned center.

$$J = \sum_{i=1}^{k} \sum_{x \in D_i} L_p(x, \mu_i)^2 = \frac{1}{m} \sum_{i=1}^{m} L_p(x_i, \mu_{c_i})^2$$

Where:

- k is the number of centers.
- D_i is the cluster of points corresponding to the center i
- x is some instance in the data
- μ_i is the center of the cluster i
- L_p is the p-norm defined by $L_p(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{\frac{1}{p}}$

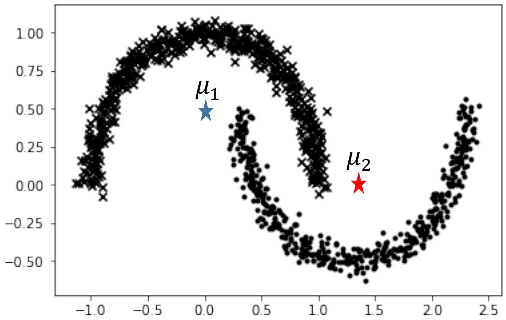
When working with the Euclidean norm we get

$$J = \frac{1}{m} \sum_{i=1}^{m} \left\| x_i - \mu_{c_i} \right\|_2^2$$

B. After updating the assignments, the points at $(11,5), (13,7)$ will belong to the 2nd cluster and all other points will retain the original cluster assignments. These two points contribute $J_1 = 72 + 80$ to the value of J before the update. After the update those two points contribute $J_2 = 10 + 10$. The difference is $\Delta J = (72 + 80) - 20 = 132$.

C. No. The value of J will monotonically decrease in each iteration of k-means. By *decrease* we mean *decrease or does not change*. First, J decreases in the reassignment step since each instance is assigned to the closest centroid, so the distance it contributes to J decreases. Second, it decreases in the recentering step because the new centroid is the vector μ for which J is minimal given the assignment.

D. No.



The picture shows the centroids of the 2 clusters. Clearly the value of J will decrease if we'll assign 'xs' that are closer to μ_2, to Cluster 2. Similarly, for the ‘dots’. If this were an intermediate configuration reached by k-means then it can’t be a stationary point for the algorithm.

E. Naïve cluster growing (as defined in class) with a small enough threshold T. This algorithm will start with a random, say, ‘dot’. It will then add ‘dots’ until it exhausts the semi-circle of ‘dots’. It will then move to the ‘xs’ and will produce the depicted configuration.

Question 5 (4 parts)

A. Given data in $\mathbb{R}^{10}$ which is not linearly separable, a student suggests to map the data to a full rational variety of degree 10 and then try to find a linear classifier in the higher dimensional space.
What is the problem in this approach and how can you overcome it?

B. Given the following dataset:

X1	X2	Y
+1	0	+1
-1	0	+1
0	+2	+1
0	+1	-1

Use the lemma below to show that it is not linearly separable.

Lemma:

Assume that a linear classifier predicts the same $y \in \{-1,+1\}$ for some two points $z, z' \in \mathbb{R}^2$ (that is $h(z) = h(z')$). Then it will produce the same prediction for any intermediate point. That is:

$$\forall \alpha \in [0,1] \quad h\big((1-\alpha)z + \alpha z'\big) = y$$

C. Find a mapping φ into a space of a dimension of your choosing that maps the dataset from Part B into a linearly separable dataset and define the linear classifier.

D. Let $X = \mathbb{R}^2$. Consider the mapping $\varphi: \mathbb{R}^2 \to \mathbb{R}^3$ given by $\varphi(x_1, x_2) = \left(\sqrt{2}x_1x_2, x_1^2, x_2^2\right)$.

1. Find a kernel for this mapping.
2. Consider the hypotheses space

$$H = \left\{ h: \mathbb{R}^2 \to \{-1,+1\} \;\middle|\; \begin{array}{c} \text{the two sets} \\ \{\varphi(x) \mid h(x) = -1\} \\ \text{and} \\ \{\varphi(x) \mid h(x) = +1\} \\ \text{are linearly separable in } \mathbb{R}^3 \end{array} \right\}$$

What is the VC dimension of H? Justify your answer.

Question 5 (4 parts) – Solution

A. The full rational variety is a mapping to a $\binom{20}{10}$-dimensional space. It is not practical to map and work in this space because of a very large time complexity. In order to overcome this problem, we can use the kernel trick with the kernel $K(\vec{x},\vec{y}) = (\vec{x}\cdot\vec{y}+1)^{10}$.

B. Choose $\alpha = 0.5$ and the first and second points in the data to see that a linear classifier that perfectly predicts the data must classify the origin as +1. Likewise, use $\alpha = 0.5$ and the third and the origin points to show that the fourth point must also be classified as +1 to derive a contradiction.

C. $\varphi(x_1,x_2) = ({x_1}^2, x_2)$.
One possible separator has weights $(w_1,w_2,b) = (2,1,-1.5)$, resulting in the predictor

$$\begin{aligned} h(x_1,x_2) &= sgn(< w, \varphi(x_1,x_2) >) \\ &= sgn(< (2,1,-1.5), ({x_1}^2, x_2, 1) >) \\ &= sgn(2{x_1}^2 + x_2 - 1.5) \end{aligned}$$

D.

1.

$$\varphi(x)\cdot\varphi(y) = \left(\sqrt{2}x_1x_2, x_1^2, x_2^2\right)\cdot\left(\sqrt{2}y_1y_2, y_1^2, y_2^2\right) =$$
$$= 2x_1x_2y_1y_2 + x_1^2y_1^2 + x_2^2y_2^2 = (x_1y_1 + x_2y_2)^2 = (x\cdot y)^2 = K(x,y)$$

2. $VC = 4$.
To see that $VC \geq 4$ consider:

$x \in \mathbb{R}^2$	$\varphi(x) \in \mathbb{R}^3$
(0,0)	(0,0,0)
(0,1)	(0,0,1)
(1,0)	(0,1,0)
(1,1)	$(\sqrt{2},1,1)$

This set of points can be shattered by H.
To see this, consider an assignment $y = (y_1, y_2, y_3, y_4)$ where the y_is represent the labels (+1 or -1) assigned to each of the original points in $\mathbb{R}^2$.
Note that the matrix $A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & \sqrt{2} & 1 & 1 \end{pmatrix}$ is invertible. We now define $w = A^{-1}y$. The hypothesis $h(x) = sgn\left(w\cdot\varphi(x)\right)$ represents the assignment y on our set of points. $h \in H$, since it's defined as a linear separator in $\mathbb{R}^3$.

To see that $VC < 5$ note that no 5 points in $\mathbb{R}^3$ can be shattered by linear separators. This was proven in class, as part of showing that the VC-dimension of linear separators in $\mathbb{R}^d$ is d+1.

Question 6 (3 parts)

A. Assume that an RBF kernel was selected in applying SVM to a classification task in $\mathbb{R}^2$. That is – a kernel of the form:

$$K(x,y) = \varphi(x)\cdot\varphi(y) = exp\left(-\frac{1}{2}\|x-y\|^2\right)$$

Prove that for every two points $x, y \in \mathbb{R}^2$ the following holds:

$$\|\varphi(x)-\varphi(y)\|^2 = 2 - 2exp\left(-\frac{1}{2}\|x-y\|^2\right)$$

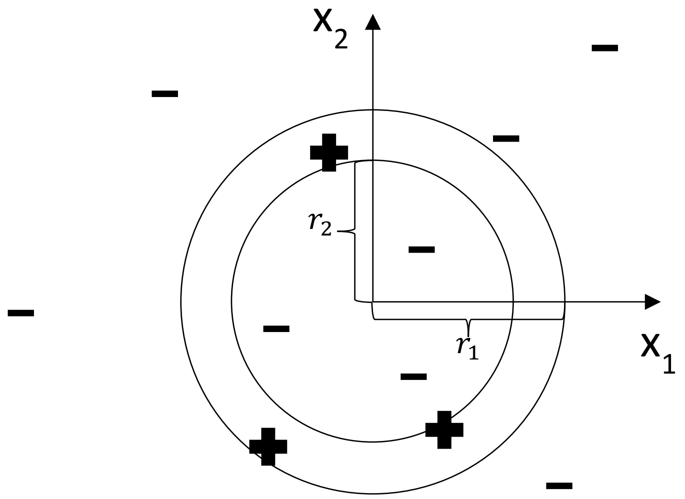
B. TRUE or FALSE

After mapping into higher dimensional space using the RBF it is possible that a kNN classification algorithm, based on unweighted Euclidean distance, achieves better classification results than it achieved in the original space.

Hint: use the identity proven in A.

C. Let $X = \mathbb{R}^2$. Let $C = H$ the set of all concentric rings. For a concept $c \in C$ let r_1 and r_2 be the radii of the concept ring where $r_1 \geq r_2$ (see picture). Each instance in the sampled training data is drawn from an unknown distribution D and consists of 2 features, namely the position of the instance $(x1, x2)$ and a target value $(+1$ if it's inside the ring and -1 otherwise).

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.



Question 6 (3 parts) – Solution

A.

$$\begin{aligned}&\|\varphi(x)-\varphi(y)\|^2\\&=\big(\varphi(x)-\varphi(y)\big)\big(\varphi(x)-\varphi(y)\big)\\&=\varphi(x)\varphi(x)+\varphi(y)\varphi(y)-2\varphi(x)\varphi(y)\\&=2-2exp\left(-\frac{1}{2}\|x-y\|^2\right)\end{aligned}$$

B. FALSE.

Suppose x_i, x_j are two neighbors of the test instant x with the following property:

$$\|x-x_i\| < \|x-x_j\|$$

That is, x_i is closer to x than x_j.

After mapping with RBF, we get:

$$\begin{aligned}\|\varphi(x)-\varphi(x_i)\|^2 &= 2-2exp\left(-\frac{1}{2}\|x-x_i\|^2\right)\\&< 2-2exp\left(-\frac{1}{2}\|x-x_j\|^2\right)=\|\varphi(x)-\varphi(x_j)\|^2\end{aligned}$$

That is, if x_i was the nearest neighbor in the original dimension, it will be the nearest neighbor in the new dimension. This generalizes to k>1, namely, the k-nearest-neighbors in the new space will be exactly the same neighbors as in the original space. Therefore, the classification will not change and the results will be the same.

* The equalities derive from section A and the inequality derives from the inequality in the original space

C. The algorithm will produce a hypothesis which is the smallest relevant ring that contains all the positive points. This can be done in O(m) as follows:

Let $\Delta = \Delta^m = (x_i, y_i)_{i=1}^m$ be a set of points in the plane, labeled positive and negative.

Our algorithm seeks to return a hypothesis $h \in H$.

Let $(x_i, y_i)_{i=1}^{m^{(+)}}$ be all positively labeled data points.

Find:

1. $r_1^h := \max_{1\le i\le m^{(+)}}(\|x_i-(0,0)\|)$
2. $r_2^h := \min_{1\le i\le m^{(+)}}(\|x_i-(0,0)\|)$

The radii of the hypothesis ring $h = L(\Delta)$ will be r_1^h, r_2^h

Consider $c \in C$ and let $\Delta^m(c) = (x_i(c), y_i(c))_{i=1}^m$ be training data generated from c without errors and by drawing m independent points according to a probability distribution π on $\mathbb{R}^2$. We will denote the probability distribution thus induced on $(\mathbb{R}^2)^m$ by π^m .

Given $\varepsilon > 0$ and $\delta > 0$ we will now compute a number $m(\varepsilon, \delta)$ so that

(Eq.1) $\quad m \geq m(\varepsilon, \delta) \Longrightarrow e(\Delta^m(c)) = \pi^m \left(err_\pi(L(\Delta^m(c)), c) > \varepsilon\right) \leq \delta$

Note that $L(\Delta^m(c))$ is the hypothesis h, or the ring, 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.

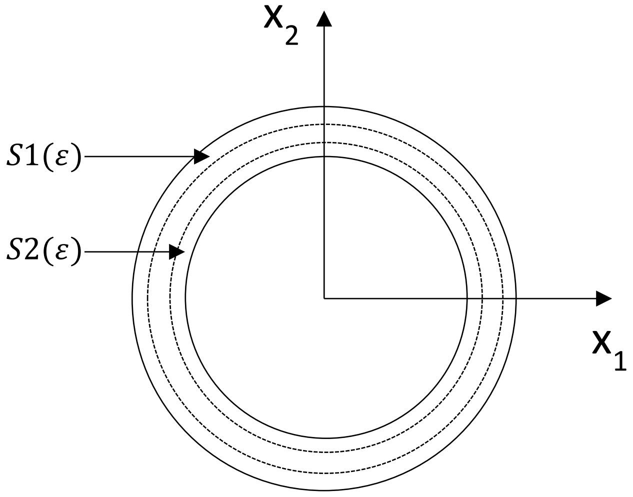


Fig1
The solid circles represent the boundaries of the concept c

Consider the strips $S1, S2$ created from the solid and dashed circles as in Fig1. These are defined to satisfy:

$$\pi(S1(\varepsilon)) = \pi(S2(\varepsilon)) = \frac{\varepsilon}{2}$$

Now, note that

$$\{\Delta^m(c): err_\pi(L(\Delta^m(c)), c) > \varepsilon\} \subseteq$$
$$\{\Delta^m(c): \Delta^m(c) \cap S1(\varepsilon) = \emptyset\} \cup$$
$$\{\Delta^m(c): \Delta^m(c) \cap S2(\varepsilon) = \emptyset\}$$

This is because if $\Delta^m(c)$ visits the two strips (note that negative points cannot visit these strips as there are no errors) then, according to our construction, the difference between c and $L(\Delta^m(c))$ will have $\pi \leq \pi(S1(\varepsilon) \cup S2(\varepsilon)) < \varepsilon$.

In term of probability we therefore get:

$$\pi^m \left(err_\pi(L(\Delta^m(c)), c) > \varepsilon\right) \leq$$
$$\pi^m \left(\Delta^m(c) \cap S1(\varepsilon) = \emptyset\right) +$$
$$\pi^m \left(\Delta^m(c) \cap S2(\varepsilon) = \emptyset\right) \leq$$
$$2\left(1 - \frac{\varepsilon}{2}\right)^m$$

We now select $m(\varepsilon, \delta) = \frac{2}{\varepsilon}\left(\ln(2) + \ln\left(\frac{1}{\delta}\right)\right)$ to get (Eq.1) to hold.

Q.E.D