בית ספר ״אפי ארזי״ למדעי המחשב המרכז הבינתחומי

The Efi Arazi school of computer science

The Interdisciplinary Center

סמסטר ב׳ תשע״ז

Spring 2018

מבחן מועד ב בלמידה ממוכנת

Machine Learning Exam B

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: 3 שעות
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. 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 $x_1, x_2, \dots, x_n$.
   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 Bernoulli distribution with parameters p_1, p_2.

   0 0 0 0 1
   1 1 1 1 1
   0 0 1 1 0
   1 1 1 0 1
   0 0 0 0 0

   For each row, a coin was tossed. With probability w_1 it led to 5 independent Bernoulli samples with p_1 and with probability w_2 it led to 5 independent Bernoulli samples with p_2.
   Which algorithm would you use assess the probabilities 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:

$$p_1 = 0.5, \qquad p_2 = 1, \qquad w_1 = 0.25$$

Question 1 (4 parts) – Solution

A.

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

B.

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

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

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

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

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

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 distribution.

D. We provide the full calculations below. Note that all binomial coefficients don't really have to be calculated to compute the responsibilities as they cancel out.

Instance 1:

$$P_A(x_1) = 0.25 \binom{5}{1} 0.5^1 0.5^4 = \frac{5}{2^7}, \qquad P_B(x_1) = 0.75 \binom{5}{1} 1^1 0^4 = 0$$

$$r(x_1, A) = \frac{\frac{5}{2^7}}{\frac{5}{2^7} + 0} = 1, \qquad r(x_1, B) = \frac{0}{\frac{5}{2^7} + 0} = 0$$

Instance 2:

$$P_A(x_2) = 0.25 \binom{5}{5} 0.5^5 0.5^0 = \frac{1}{2^7}, \qquad P_B(x_2) = 0.75 \binom{5}{5} 1^5 0^0 = \frac{3}{4}$$

$$r(x_2, A) = \frac{\frac{1}{2^7}}{\frac{1}{2^7} + \frac{3}{4}} \approx 0.01, \qquad r(x_2, B) = \frac{\frac{3}{4}}{\frac{1}{2^7} + \frac{3}{4}} \approx 0.99$$

Instance 3:

$$P_A(x_3) = 0.25 \binom{5}{2} 0.5^2 0.5^3 = \frac{10}{2^7}, \qquad P_B(x_3) = 0.75 \binom{5}{2} 1^2 0^3 = 0$$

$$r(x_3, A) = \frac{\frac{10}{2^7}}{\frac{10}{2^7} + 0} = 1, \qquad r(x_3, B) = \frac{0}{\frac{10}{2^7} + 0} = 0$$

Instance 4:

$$P_A(x_4) = 0.25 \binom{5}{4} 0.5^4 0.5^1 = \frac{5}{2^7}, \qquad P_B(x_4) = 0.75 \binom{5}{4} 1^4 0^1 = 0$$

$$r(x_4, A) = \frac{\frac{5}{2^7}}{\frac{5}{2^7} + 0} = 1, \qquad r(x_4, B) = \frac{0}{\frac{5}{2^7} + 0} = 0$$

Instance 5:

$$P_A(x_5) = 0.25 \binom{5}{0} 0.5^0 0.5^5 = \frac{1}{2^7}, \qquad P_B(x_5) = 0.75 \binom{5}{0} 1^0 0^5 = 0$$

$$r(x_5, A) = \frac{\frac{1}{2^7}}{\frac{1}{2^7} + 0} = 1, \qquad r(x_5, B) = \frac{0}{\frac{1}{2^7} + 0} = 0$$

$$New\, w_A = \frac{1}{5}\sum_{i=1}^{5} r(x_i, A) = \frac{4}{5}$$

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

$$p_A = \frac{1}{4}\left(1 * \frac{1}{5} + 0.01 * 0 + 1 * \frac{2}{5} + 1 * \frac{4}{5} + 1 * \frac{0}{5}\right) = \frac{7}{20}$$

$$p_B = 1\left(0 * \frac{1}{5} + 0.99 * 1 + 0 * \frac{2}{5} + 0 * \frac{4}{5} + 1 * \frac{0}{5}\right) = 0.99$$

Question 2 (5 parts)

A. Explain the difference between Naïve Bayes and Full Bayes.

In a dice game, you roll 2 dice. Where each die has 6 faces.
At Casino A they use fair dice, so on each roll, every face has the same probability.
The probability for each pair of numbers in 2 rolls is 1/36.
At the B casino, the first die is fair, where each face has the same probability, but the second die is skewed so that with probability $\frac{1}{3}$ it will land on the same face as the first die and with probability $\frac{1}{3}$ it will land on each of ± 1 from the result of the first die.
For example, if the first die roll is 3 then the second die, in Casino B, will either be 2,3 or 4, each with probability $\frac{1}{3}$. And if the first die roll is 6 then the second die will either be 5,6, or 1, again, each with probability $\frac{1}{3}$.
The prior probability for playing in Casino A is $\frac{3}{5}$ and for playing Casino B it's $\frac{2}{5}$.
Given a game outcome (2 numbers), we want to classify whether the game was played in Casino A or B.

B. Given the following game outcome: 1st die is 4, 2nd die is 5. Which casino will a Naïve Bayes classifier predict? Explain your answer.

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? And for the Naïve Bayes? Show/explain your calculations.

E. Given the results from 2 pairs of rolls (2 pairs of 2 numbers) played in the same casino, what is the minimal prior we need to assign to Casino A in order for the Full Bayes classifier to predict it regardless of the results? Show your calculations.

<u>Question 2 (5 parts) – Solution</u>

A. The difference between Full Bayes and Naïve Bayes is that in Naïve Bayes we assume that the attributes are independent of each other, given the class. Therefore, we can calculate the data likelihood per class as a multiplication of that likelihood for each attribute value and estimate the distribution of each attribute separately. In Full Bayes, on the other hand, we are not assuming the naïve conditional independence assumption and therefore we need to estimate the multi-dimensional distribution for all the attributes together in order to calculate the likelihood.

B. Casino A according to Naïve Bayes (MAP):

$$P\big(A\big|(4,5)\big) = P\big((4,5)\big|A\big)P(A) = \frac{1}{6} * \frac{1}{6} * \frac{3}{5} = \frac{3}{180} = \frac{1}{60}$$

Casino B according to Naïve Bayes (MAP):

$$P\big(B\big|(4,5)\big) = P\big((4,5)\big|B\big)P(B) = \frac{1}{6} * \frac{1}{6} * \frac{2}{5} = \frac{2}{180} = \frac{1}{90}$$

Note that, because of the Naïve assumption, in the probability calculation for Casino B does not use the dependence of the second die.

Also note that the marginal distribution of both dice, in both casinos, is uniform, as we sum over the values in the corresponding row (or column) of the matrix that describes the joint probability distribution.

Conclusion: Naïve Bayes will predict Casino A.

C. Casino A according to Full Bayes:

$$P\big(A\big|(4,5)\big) = P\big((4,5)\big|A\big)P(A) = \frac{1}{6} * \frac{1}{6} * \frac{3}{5} = \frac{3}{180} = \frac{1}{60}$$

Casino B according to Full Bayes:

$$P\big(B\big|(4,5)\big) = P\big((4,5)\big|B\big)P(B) = \frac{1}{6} * \frac{1}{3} * \frac{2}{5} = \frac{2}{90} = \frac{1}{45}$$

Full Bayes will predict Casino B.

D. <u>Full Bayes</u>:

Let p be the prior of Casino A (1-p will be the prior of Casino B).

Note that any possible outcome in Casino A has probability of $\frac{1}{36}$, and in Casino B $\frac{1}{18}$ (there are only 18 possible outcomes in Casino B when considering the full joint distribution).

For any possible outcome we therefore have:

$$P(outcome|A)P(A) = \frac{1}{36}p = \frac{1}{18}(1-p) = P(outcome|B)P(B)$$
$$p = 2(1-p)$$
$$3p = 2$$
$$p = \frac{2}{3}$$

If the prior of A is greater than $\frac{2}{3}$, then a Full Bayes classifier will predict A regardless of the game outcome

<u>Naïve Bayes</u>:

Let p be the prior of casino A (1-p will be the prior of casino B).

Note that each outcome in casino A and B in Naïve Bayes has probability of $\frac{1}{36}$ (=the likelihoods in casino A and B will be the same).

Therefore, only the prior matters and if the prior of A is greater than $\frac{1}{2}$, then the Naive Bayes classifier will predict A regardless of the game outcome.

E. <u>Full Bayes</u>:

Let p be the prior of casino A (1-p will be the prior of casino B).

Note that each outcome in casino A has probability of $\left(\frac{1}{36}\right)^2$, and in casino B $\left(\frac{1}{18}\right)^2$.

$$P(outcome|A)P(A) = \left(\frac{1}{36}\right)^2 p = \left(\frac{1}{18}\right)^2 (1-p) = P(outcome|B)P(B)$$

$$\left(\frac{1}{36}\right)^2 p = \left(\frac{1}{2*18}\right)^2 p = \left(\frac{1}{18}\right)^2 \frac{p}{4} = \left(\frac{1}{18}\right)^2 (1-p)$$

$$p = 4(1-p)$$

$$5p = 4,\ p = \frac{4}{5}$$

If the prior of A is greater than $\frac{4}{5}$, then the Full Bayes classifier will predict A regardless of the game outcome

Naïve Bayes:

Same as above, really.

Let p be the prior of casino A (1-p will be the prior of casino B).

Note that each outcome in casino A and B in Naïve Bayes has probability of $\left(\frac{1}{36}\right)^2$ (=the likelihoods in casino A and B are equal).

Therefore, if the prior of A will be greater than $\frac{1}{2}$, the Naive Bayes classifier will predict A regardless of the game outcome.

Question 3 (4 parts)

We want to cluster to k groups a set S of instances. Below is a pseudo code of an algorithm that is called k-medoids and is similar to k-means:

```
Initialize $c_1, …, c_k$ by randomly selecting k elements from S
Loop:
    Assign all n samples to their closest $c_i$ and create k clusters $S_1, …, S_k$
    For each cluster $S_i$ ($1 \leq i \leq k$) define a new $c_i$:
            choose $c_i \in S_i$ whose distance to all other members in $S_i$ is the smallest
Until no change in $c_1, …, c_k$
Return $c_1, …, c_k$
```

A. Assume that the set S has 7 instances with 2 features as given in the table. Execute the k-medoids algorithm to cluster the set into two groups (i.e. k=2) when you initialize the execution with p_1 and p_5 as the initial centers. This means that you start at c_1=p_1 and c_2=p_5.
(it is recommended to first plot the instances on a 2D Euclidean plane).
In every step indicate what the centers are and how instances are assigned. You don't have to show all intermediate calculations.

instance	x	y
p_1	**2**	**6**
p_2	4	7
p_3	5	8
p_4	6	1
p_5	**6**	**4**
p_6	7	3
p_7	5	6

B. Explain, using a formula, the function, J_S , that k-means seeks to minimize.

C. What is the main difference between k-means and k-medoids? How would you change the function from B so that it would fit the k-medoids algorithm? Call this new target function J_{med}.

D. Assume that we execute both algorithms on the same set and with the same k. Do we expect the value of J_{med} obtained by k-medoid to be smaller, greater or the same as the value of J_S that k-means has obtained? Explain why!

Question 3 (4 parts) – Solution

A. Below are the stages of the process. Centroids are in boldface and clusters are colored red and green.

1

X_1	**2**	**6**
X_2	4	7
X_3	5	8
X_4	6	1
X_5	**6**	**4**
X_6	7	3
X_7	5	6

Initialization

2

X_1	**2**	**6**
X_2	4	7
X_3	5	8
X_4	6	1
X_5	**6**	**4**
X_6	7	3
X_7	5	6

Assignment 1

3

X_1	2	6
X_2	**4**	**7**
X_3	5	8
X_4	6	1
X_5	**6**	**4**
X_6	7	3
X_7	5	6

Medoids 2

4

X_1	2	6
X_2	**4**	**7**
X_3	5	8
X_4	6	1
X_5	**6**	**4**
X_6	7	3
X_7	5	6

Assignment 2

5

X_1	2	6
X_2	**4**	**7**
X_3	5	8
X_4	6	1
X_5	6	4
X_6	**7**	**3**
X_7	5	6

Medoids 3

6

X_1	2	6
X_2	**4**	**7**
X_3	5	8
X_4	6	1
X_5	6	4
X_6	**7**	**3**
X_7	5	6

Final Assignment

B. k-means minimizes, over all possible selections of centroids μ_i , $i = 1 \dots k$, the total distances of all points to their cluster centroids – the geographic mean of the cluster member points. Formally:

$$J_S = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|_2^2$$

C. The main difference is that in k-means the cluster representative is the geometric center of the cluster member points and doesn't have to be a data point. In k-medioids the algorithm insists on centroids that are actual data points. As a result, the optimization criterion is also different:

$$J_{med} = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - c_i\|_2^2$$

Where the center c_i is an actual data point.

D. The value attained by the k-medioids process is likely to be larger than that attained by k-means. The inner sum, for k-means, is optimized by the selection of the centroid. In k medioids this selection is further constrained and therefore the value for using a medioid as a centroid is larger.

However, initiate the k-means and k-medioid in different places can result a smaller J_{med} value – it can converge to a different minimum that can be smaller than the minimum that k-means converged to.

Question 4 (4 parts)

A. State the gain formula for computing the best split of a node in a decision tree. Assume that the impurity function is a generic φ.

B. Prove that when the attributes as well as the class labeling are all binary then GiniGain is always non-negative (0≥).

C. Consider a decision tree that only performs binary splits. That is: considering an attribute with discrete (or categorical) values, the split will be into two subsets – cats, dogs and elephants to one side and birds and snakes in the other side.
How many Goodness of Split calculations will a tree construction algorithm perform in the root when there are k discrete attributes and the i-th attribute has V(i) possible values? Explain your answer.

D. Consider the training data described below, at a node S. Decide on which attribute to use according to the rules defined in C and using GiniGain. State your calculations. There is no need to get to a final answer.

X1	X2	Y
Green	Ronaldo	-
Green	Ronaldo	+
Green	Messi	-
Blue	Messi	+
Blue	Messi	+
Blue	Neymar	-
Blue	Neymar	-
Blue	Neymar	-

<u>Question 4 (4 parts) – Solution</u>

A.

$$\Delta\varphi(S,A) \equiv \varphi(S) - \sum_{v\in Values(A)} \frac{|S_v|}{|S|}\varphi(S_v)$$

B. We will prove the binary case. The general case is very similar, just involves higher dimensional calculus.

Recall that the <u>Gini index</u> of a dichotomy ($\oplus$ vs $\ominus$, a binary partition) of a set S, partitioned into $p|S|$ and $(1-p)|S|$ elements (labeled $\oplus$ and $\ominus$) is:

$$G(p) = 1 - p^2 - (1-p)^2 = 2p(1-p)$$

Assume that the attribute A partitions S into two sets, one has $q|S|$ elements and the other has $(1-q)|S|$ elements. Call the subsets S_1 and S_2 respectively.

The <u>gain</u> of this split is:

$$Gain(S,A) = G(p) - \big(qG(p_1) + (1-q)G(p_2)\big)$$

Where p_1 and p_2 are the fractions of $\oplus$ elements in S_1 and S_2 respectively. That is: in S_1 we have $p_1|S_1|$ elements labeled $\oplus$ and $(1-p_1)|S_1|$ elements labeled $\ominus$.

We want to prove that

$$G(p) \geq qG(p_1) + (1-q)G(p_2)$$

We note that the function G is concave down (aka convex):

$$G'^{(x)} = 2 - 4x$$

$$G''^{(x)} = -4 \leq 0$$

Recall that for a concave down function $\ell\colon \mathbb{R} \to \mathbb{R}$, for any $0 \leq \lambda \leq 1$ and for <u>any</u> 2 points t and s we have

$$* \ \ \ell(\lambda t + (1-\lambda)s) \geq \lambda\ell(t) + (1-\lambda)\ell(s)$$

Note that

$$p|S| = p_1|S_1| + p_2|S_2|$$

And therefore

$$p = p_1\frac{|S_1|}{|S|} + p_2\frac{|S_2|}{|S|} = qp_1 + (1-q)p_2$$

Now use equation $*$ above with $\lambda = q$ and with $s,t = p_1,p_2$ to get:

$$G(p) = G(qp_1 + (1-q)p_2) \geq qG(p_1) + (1-q)G(p_2)$$

QED.

C. First, we compute how many possible splits an attribute with V(i) values can have. You can think of this problem has the number of strings of size V(i) with alphabet {1,0}. In each position, 0/1 indicates to which subset a value j belongs.

This will give us $2^{V(i)}$. But this will also contain strings with all zeros and all ones which are irrelevant splits. So, we get $2^{V(i)} - 2$. And, in our case the string 01 and 10 are the same so we need to divide by 2;

$$\frac{2^{V(i)} - 2}{2}$$

If we sum up all the attributes we get:

$$\sum_{i=1}^{k} \frac{2^{V(i)} - 2}{2}$$

D. First, we compute the Gini impurity with respect to a split on X_1:

$$\frac{3}{8}\left(1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right)\right) + \frac{5}{8}\left(1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right)\right)$$

Secondly, we compute the Gini impurity with respect to a split on X_2:

For the split {Ronaldo, Messi} – {Neymar}:

$$\frac{5}{8}\left(1-\left(\left(\frac{2}{5}\right)^2+\left(\frac{3}{5}\right)^2\right)\right)+\frac{3}{8}\left(1-\left(\left(\frac{3}{3}\right)^2\right)\right)$$

For the split {Ronaldo, Neymar} – {Messi}:

$$\frac{5}{8}\left(1-\left(\left(\frac{1}{5}\right)^2+\left(\frac{4}{5}\right)^2\right)\right)+\frac{3}{8}\left(1-\left(\left(\frac{1}{3}\right)^2+\left(\frac{2}{3}\right)^2\right)\right)$$

For the split {Messi, Neymar} – {Ronaldo}:

$$\frac{6}{8}\left(1-\left(\left(\frac{2}{6}\right)^2+\left(\frac{4}{6}\right)^2\right)\right)+\frac{2}{8}\left(1-\left(\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right)\right)$$

Question 5 (5 parts)

A. Find the minimum and the maximum of $3x + 2y$ under the constraint $4x^2 + y^2 = 1$

B. Given the following dataset:

X1	X2	Y
+1	+1	+1
-1	+1	+1
0	-1	+1
0	0	-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. Find $K\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$ which is a kernel function for

$$\varphi(x) = \big(x_1{}^3, \sqrt{3}x_1{}^2x_2, \sqrt{3}x_1x_2{}^2, x_2{}^3\big)$$

E. Show that the function $K(x, y) = e^{xy}$ for $x, y \in \mathbb{R}$ is a kernel for some mapping into infinite dimensional space. (hint: $e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$ for $z \in \mathbb{R}$)

Question 5 (5 parts) – Solution

A.

$$L(x,y,\lambda) = 3x + 2y \ -\lambda(4x^2 + y^2 - 1)$$

$$\nabla L_x = 3 - 8\lambda x = 0 \Rightarrow x = \frac{3}{8\lambda}$$

$$\nabla L_y = 2 - 2\lambda y = 0 \Rightarrow y = \frac{1}{\lambda}$$

$$\nabla L_\lambda = 4x^2 + y^2 - 1 = 0$$

$$\frac{36}{64\lambda^2} + \frac{1}{\lambda^2} = 1 \Rightarrow \lambda^2 = \frac{100}{64} \Rightarrow \lambda = \pm\frac{5}{4}$$

Then we get:

$$x = \pm\frac{3}{10}$$

$$y = \pm\frac{4}{5}$$

And therefore:

$max = \frac{5}{2}$ is attained at $\left(\frac{3}{10},\frac{4}{5}\right)$ $\qquad$ $min = -\frac{5}{2}$ is attained at $\left(-\frac{3}{10},-\frac{4}{5}\right)$

Note that to determine max/min we only need to plug into $f(x) = 3x + 2y$.

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 point (0,1) as +1. Likewise, use $\alpha = 0.5$ and the third point in the data, together with the point (0,1), to show that the origin must also be classified as +1 to derive a contradiction to the fourth point in the data.

C. $\varphi(x_1, x_2) = (x_1, x_2, x_1^2 + x_2^2)$.

One possible separator has weights $(w_1, w_2, w_3, b) = (0,0,1,-0.5)$, resulting in the predictor

$$h(x_1,x_2) = sgn(< w, \varphi(x_1,x_2) >) =$$
$$= sgn(< (0,0,1,-0.5), (x_1, x_2, x_1^2 + x_2^2, 1) >) = sgn(x_1^2 + x_2^2 - 0.5)$$

D.

$$\begin{aligned}\langle\varphi(x),\varphi(y)\rangle &= \left\langle\left(x_1{}^3, \sqrt{3}x_1{}^2x_2, \sqrt{3}x_1x_2{}^2, x_2{}^3\right), \left(y_1{}^3, \sqrt{3}y_1{}^2y_2, \sqrt{3}y_1y_2{}^2, y_2{}^3\right)\right\rangle \\ &= x_1^3y_1^3 + 3x_1{}^2x_2y_1{}^2y_2 + 3x_1x_2{}^2y_1y_2{}^2 + x_2{}^3y_2{}^3 \\ &= (x_1y_1 + x_2y_2)^3 \\ &= \langle x, y\rangle^3\end{aligned}$$

E. Let

$$\varphi(x) = \left(1, x, \frac{x^2}{\sqrt{2!}}, \frac{x^3}{\sqrt{3!}}, \frac{x^4}{\sqrt{4!}}, \dots\right)$$

We have

$$\begin{aligned}\langle\varphi(x),\varphi(y)\rangle &= \left\langle\left(1, x, \frac{x^2}{\sqrt{2!}}, \frac{x^3}{\sqrt{3!}}, \frac{x^4}{\sqrt{4!}}, \dots\right), \left(1, y, \frac{y^2}{\sqrt{2!}}, \frac{y^3}{\sqrt{3!}}, \frac{y^4}{\sqrt{4!}}, \dots\right)\right\rangle \\ &= 1 + xy + \frac{x^2y^2}{\sqrt{2!}\sqrt{2!}} + \frac{x^3y^3}{\sqrt{3!}\sqrt{3!}} + \frac{x^4y^4}{\sqrt{4!}\sqrt{4!}} + \cdots \\ &= 1 + xy + \frac{x^2y^2}{2!} + \frac{x^3y^3}{3!} + \frac{x^4y^4}{4!} + \cdots = e^{xy}\end{aligned}$$

Question 6 (4 parts)

A. Consider two hypothesis spaces H and H' such that $H \subseteq H'$. Prove that:

$$VC(H) \leq VC(H')$$

B. Recall the 3 sample complexity bounds given in class:

- $m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$
- $m \geq \frac{1}{\varepsilon^2}\left(\ln 2|H| + \ln\frac{1}{\delta}\right)$
- $m \geq \frac{1}{\varepsilon}\left(8 \cdot VC(H)\log_2\frac{13}{\epsilon} + 4\log_2\frac{2}{\delta}\right)$

Consider an instance space $X = [-1,1] \times [-1,1]$.
Let $N \in \mathbb{N}$ and $N \geq 2$, we define the sets $\mathrm{A1} := \{\frac{1}{N}, \frac{2}{N}, \dots, 1\}$, $\mathrm{A}_2 = \{\frac{1}{2N}, \frac{2}{2N}, \dots, 1\}$ and the following hypothesis spaces:

- $H_1 = \{h: h(x_1, x_2) = +1 \Leftrightarrow {x_1}^2 + {x_2}^2 \leq r^2,\ r \in \mathrm{A}_1\}$
- $H_2 = \{h: h(x_1, x_2) = +1 \Leftrightarrow {x_1}^2 + {x_2}^2 \leq r^2,\ r \in \mathrm{A}_2\}$
- $H_3 = \{h: h(x_1, x_2) = +1 \Leftrightarrow {x_1}^2 + {x_2}^2 \leq r^2,\ r \in [0,1]\}$

You can view each hypothesis as a circle centered at the origin. (i.e. an instance is classified as positive iff it's inside the circle with radius r)

For each of the following cases, use one of the above bounds to compute the number of instances needed to guarantee an error of less than 0.1 with probability at least 95%:

1. When trying to learn a concept in H_1 using H_2.
2. When trying to learn a concept in H_3 using H_3.
3. When trying to learn a concept in H_3 using H_2.

Question 6 (4 parts) – Solution

A. Suppose by contradiction that d = VC(H’) < VC(H), then there exists a set of size d + 1 denoted A that is shattered by H (and possibly larger sets as well), i.e., there are hypotheses in H realizing all dichotomies on A. Since H ⊆ H’ these hypotheses are also in H’ and therefore H’ also shatters A, thus d = VC(H’) ≥ d + 1, a contradiction. We conclude that VC(H) ≤ VC(H’) as required.

B. First, observe that $|H_1| = N, |H_2| = 2N, |H_3| = \infty$.

1. $c \in H_1 \subseteq H_2$, and $|H_2|$ is finite, therefore we will use the first bound:
$$m \geq 10(\ln 2N + \ln 20)$$
2. Since $|H_3| = \infty$, we must use the third bound and compute the VC dimension.
$VC \geq 1$:
The set with the single instance (0.5, 0) is shattered by a circle with radius 1, therefore $VC(H_3) \geq 1$.
$VC < 2$:
For any set with two instances of distances r1, r2 from the origin, suppose w.l.o.g. that r1 ≤ r2, then the dichotomy −1, +1 cannot be obtained as any circle containing the second instance will also contain the first, therefore $VC(H_3) < 2$.
We got that $VC = 1$ and then:
$$m \geq 10(8\log_2 130 + 4\log_2 40)$$
3. Possibly $c \notin H_2$ and $|H_2|$ is finite, therefore we will use the second bound:
$$m \geq 100(\ln 4N + \ln 20)$$