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!

שאלה 1 (4 סעיפים)

א. נתון X משתנה Poisson. מכיר שאז $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ עבור $\lambda > 0$ כלשהי. נניח שעבור סדרת הגרלות בלתי תלויות מהמשתנה X קיבלנו את הערכים $x_1, \ldots, x_n$. כתבי ביטוי המייצג את ה-likelihood של ה-data הנצפה כפונקציה של הפרמטר λ.

ב. הוכיחי שהערך של λ שיתקבל ע"י MLE במקרה זה הוא:

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

ג. נתון ה-data הנצפה הבא הלקוח משתי התפלגויות Bernoulli שונות עם פרמטרים p_1, p_2. כל שורה נוצרה מחמש הגרלות בלתי תלויות מאחת משתי ההתפלגויות לעיל:

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

עבור כל שורה הוטל מטבע שבהסתברות w_1 הוביל לחמש הגרלות עם הפרמטר p_1 ובהסתברות w_2 הוביל לחמש הגרלות עם הפרמטר p_2.
באיזה אלגוריתם היית משתמשת בשביל להעריך את ההתפלגות המשותפת שבבסיס הנתונים המתוארים לעיל. הסבירי את בחירתך.

ד. השתמשי באלגוריתם מסעיף ג' לחשב את הצעד הראשון, כאשר האתחול הינו:

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

<u>שאלה 2 (5 סעיפים)</u>

א. הסבירי מהו ההבדל בין Naïve Bayes לבין Full Bayes.

במשחק הטלת קוביות מטילים 2 קוביות בעלות 6 פאות כל אחת.
בקזינו A, משתמשים בקוביות רגילות, כאשר לכל זוג מספרים בהטלת 2 קוביות הסתברות זהה ושווה ל 1/36.
בקזינו B, מטילים קובייה ראשונה, בה הסתברות זהה לכל אחד מהמספרים. הקוביה השנייה תלויה בתוצאת ההטלה הראשונה, כך שההסתברות הינה 1/3 למס' שיצא בהטלה של הקוביה הראשונה ו-1/3 למספרים שהם ±1 מהמספר שיצא. לדוגמא, אם יצא 3 בקוביה הראשונה, אז בשנייה ייצא 2, 3 או 4 בהסתברות 1/3 כ"א. אם יצא 6 בקוביה הראשונה, אז בשנייה ייצא 5, 6 או 1 בהסתברות 1/3 כ"א.
הסתברות ה-prior לבחור בקזינו A היא 3/5 ובקזינו B היא 2/5.
בהינתן תוצאה של הטלה של זוג הקוביות, אנחנו רוצים לסווג האם ההטלה הגיעה מקזינו A או B.

ב. מה יהיה הקזינו שיבחר ע"י Naïve Bayes כאשר ההטלה שקיבלנו היא 4 בקוביה הראשונה ו-5 בשנייה? הראי את החישובים.

ג. בנתוני השאלה מהסעיף הקודם מה יהיה הקזינו שיבחר ע"י Full Bayes? הראי את החישובים.

ד. מה צריך להיות ה-prior המינימלי של קזינו A כדי שאלגוריתם Full Bayes ייבחר בו ללא תלות בתוצאה שהתקבלה בהטלת 2 הקוביות? ועבור אלגוריתם Naïve Bayes? הראי את החישובים.

ה. בהינתן 2 תוצאות של הטלה של זוג קוביות (2 זוגות מספרים), מה צריך להיות ה-prior המינימלי של קזינו A כדי שאלגוריתם Full Bayes ייבחר בו ללא תלות בתוצאה שהתקבלה בהטלת 2 הקוביות פעמיים (2 זוגות מספרים)? הראי את החישובים.

שאלה 3 (4 סעיפים)

נתונה קבוצת מופעים S שאנו רוצים לחלק ל-k קבוצות (כלומר לבצע clustering). להלן אלגוריתם אשר נקרא k-medoids ודומה ל k-means:

Initialize $c_1, \ldots, c_k$ by randomly selecting k different elements from S
Loop:
  Assign all n samples to their closest c_i and create k clusters $S_1, \ldots, S_k$
  For each cluster S_i ($1 \le i \le k$) 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 $c_1, \ldots, c_k$
Return $c_1, \ldots, c_k$

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

א. נניח כי הקבוצה S מונה 7 מופעים בעלי 2 תכונות כנתון בטבלה משמאל. הריצי את אלגוריתם k-medoids לחלוקה לשתי קבוצות (כלומר k=2) כאשר <u>מאתחלים את הריצה עם p_1 ו-p_5</u> בתור המרכזים הראשונים כלומר בשלב הראשון $c_1=p_1$ ו-$c_2=p_5$. (המלצה: ראשית ציירו את המופעים על מישור דו ממדי). בכל שלב צייני מי המרכזים ומה חלוקת המופעים לכל קבוצה - אין צורך להראות את כל החישובים בכל שלב.

ב. הסבירי בנוסחה איזה פונקציה, J_S, מביא אלגוריתם k-means למינימום?

ג. מה ההבדל העיקרי בין אלגוריתם k-means ו- k-medoids? כיצד היית משנה את הפונקציה מסעיף ב להתאים אותה לאלגוריתם k-medoids? קראי לפונקציה החדשה J_{med}.

ד. נניח כי אנו מריצים את שני האלגוריתמים על אותה הקבוצה ושני האלגוריתמים התכנסו. האם אנו מצפים שערך הפונקציה J_{med}, שאותה הביא k-medoids למינימום יהיה קטן, גדול או שווה לערך הפונקציה J_S, שאותה הביא k-means למינימום? הסבירי מדוע!

<u>שאלה 4 (4 סעיפים)</u>

א. כתבי את הנוסחה לחישוב הפיצול הכי טוב בעץ החלטה. הנחי שמדד ה-impurity הוא φ.

ב. הוכיחי כי בהינתן n תכונות בינאריות ו-2 מחלקות (classes), GiniGain הוא תמיד אי-שלילי (≥ 0).

ג. נתון עץ החלטה שעושה פיצולים בינאריים בלבד. בהינתן תכונה בדידה הפיצול יהיה ע"י חלוקה לשתי תתי קבוצות. לדוגמא: חתולים, כלבים ופילים לצד אחד ונחשים וציפורים לצד השני.
כמה חישובי Goodness of Split יבצעו באלגוריתם בניית עץ בשורש, כאשר נלקחות בחשבון k תכונות בדידות ולכל תכונה i יש v(i) ערכים? הסבירי את תשובתך.

ד. בהינתן קבוצת האימון הבאה, בצומת S, החליטי על התכונה שתבחר ע"פ ההנחיות בסעיף ג' ובשימוש ב- GiniGain – הראי את החישובים. אין צורך להגיע לתשובה סופית.

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

<u>שאלה 5 (5 סעיפים)</u>

א. מצאי את המינימום והמקסימום של הפונקציה $3x+2y$ עם האילוץ $4x^2+y^2=1$.

ב. בהינתן ה dataset הבא:

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

השתמשי ב-Lemma הבאה להראות שה-dataset הנ"ל אינו ניתן להפרדה לינארית.

<u>Lemma:</u>

נניח שמפריד לינארי h חוזה פרדיקציה $y \in \{-1,+1\}$ על קבוצת נקודת $z, z' \in \mathbb{R}^2$ כך ש: $h(z)=h(z')=y$.

אזי, המפריד h ייתן את אותה פרדיקציה על כל נקודות ביניים, כלומר:

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

ג. מצאי מיפוי φ למרחב עם ממד לבחירתך, אשר ממפה את ה-dataset מהסעיף הקודם כך שבממד החדש הוא יהיה ניתן להפרדה לינארית ומצאי את המפריד הלינארי בממד החדש.

ד. בהינתן $X=\mathbb{R}^2$ מצאי את פונקציית הקרנל המתאימה למיפוי הבא:

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

ה. בהינתן $X=\mathbb{R}$ הראי ש- $K(x,y)=e^{xy}$ היא פונקציית קרנל עם מיפוי מתאים הממפה למרחב אינסופי. (רמז: $e^z=\sum_{n=0}^{\infty}\frac{z^n}{n!}$ לכל $z \in \mathbb{R}$).

שאלה 6 (4 סעיפים)

א. בהינתן מרחב היפותזות H ומרחב היפותזות H', כך ש $H \subseteq H'$, הוכיחי ש:
$VC(H) \leq VC(H')$

ב. להלן 3 הנוסחאות ל- sample complexity שנלמדו בכיתה:

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

נתון מרחב הדוגמאות $X = [-1,1] \times [-1,1]$, , ומרחבי היפותזות, שמוגדרים במדויק בהמשך, של מעגלים שמרכזם בראשית הצירים, כאשר instance יסווג כחיובי אם ורק אם הוא נופל בתוך המעגל בעל רדיוס r.

יהי N מספר טבעי, נגדיר את הקבוצות $A_1 = \{\frac{1}{N}, \frac{2}{N}, \ldots, 1\}$ ו- $A_2 = \{\frac{1}{2N}, \frac{2}{2N}, \ldots, 1\}$ ואת מרחבי ההיפותזות הבאים:

- $H_1 = \{h: h(x_1, x_2) = +1 \Leftrightarrow {x_1}^2 + {x_2}^2 \leq r^2,\ r \in A_1\}$
- $H_2 = \{h: h(x_1, x_2) = +1 \Leftrightarrow {x_1}^2 + {x_2}^2 \leq r^2,\ r \in A_2\}$
- $H_3 = \{h: h(x_1, x_2) = +1 \Leftrightarrow {x_1}^2 + {x_2}^2 \leq r^2,\ r \in [0,1]\}$

לכל אחד מהמקרים הבאים, השתמשי באחת מהנוסחאות לחישוב כמות ה-instances הנדרשת להבטיח טעות של 0.1 בהסתברות של לפחות 95%:

1. כאשר מנסים ללמוד קונספט, c, שנמצא במרחב H_1 בעזרת מרחב ההיפותזות H_2.
2. כאשר מנסים ללמוד קונספט, c, שנמצא במרחב H_3 בעזרת מרחב ההיפותזות H_3.
3. כאשר מנסים ללמוד קונספט, c, שנמצא במרחב H_3 בעזרת מרחב ההיפותזות H_2.

בהצלחה!

Standard formula sheet – IDC 2018

1. Distributions:

Normal $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Bernoulli trials - $B(n,p)$ $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

Poisson $P(X=k) = e^{-\lambda} \frac{\lambda^k}{k!}$

Geometric $P(X=k) = (1-p)^{k-1} p$

2. Decision Trees:

Gini $Gini(S) = 1 - \sum_{i=1}^{c} \left(\frac{|S_i|}{|S|}\right)^2$

Entropy $Entropy(S) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log \frac{|S_i|}{|S|}$

3. Gradient descent and update steps:

Linear regression $\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{d \in D} \left(h_\theta(x^{(d)}) - y^{(d)}\right) \cdot x_j^{(d)}$

Perceptron $w_j := w_j - \eta \sum_{d \in D} (o^{(d)} - t^{(d)}) x_j^{(d)}$

Dual perceptron If $o^{(d)} \cdot t^{(d)} < 0$ then:

$\alpha_j = \alpha_j + \eta$

4. Logistic regression:

$P(h(x) = 1) = \frac{1}{1 + e^{-w^T x}}$

5. SVM:

Primal objective function $\frac{1}{2}\|w\|^2 + \gamma \sum_d \xi_d - \sum_d \alpha_d (t_d(w^T x_d + w_0) - 1 + \xi_d) - \sum_d \mu_d \xi_d$

s.t $\alpha_d \geq 0$ $\mu_d \geq 0$

Dual objective function $\sum_d \alpha_d - 1/2 \sum_d \sum_e \alpha_d \alpha_e t_d t_e x_d^T x_e$

s.t $\sum_d \alpha_d t_d = 0,\ 0 \leq \alpha_d \leq \gamma$

6. EM (for Bernoulli distributions):

$New\ w_{A_j} = \frac{1}{N} \sum_{i=1}^{N} r(x_i, A_j)$

$p_{A_j} = \frac{1}{(New\ w_{A_j})N} \sum_{i=1}^{N} r(x_i, A_j) v(i)$