ה א ו נ י ב ר ס י ט ה ה פ ת ו ח ה

20595

כריית מידע

חוברת הקורס אביב ב2012

כתבה: ד"ר מיה הרמן

מרץ 2012 - סמסטר אביב – תשע"ב

פנימי – לא להפצה.

© כל הזכויות שמורות לאוניברסיטה הפתוחה.

תוכן העניינים

אל הסטודנט

אנו מקדמים את פניך בברכה עם הצטרפותך אל הלומדים בקורס "כריית מידע".

בחוברת זו תמצא את "לוח הזמנים ופעילויות", תנאים לקבלת נקודות זכות ומטלות הקורס.

לקורס קיים אתר באינטרנט בו תמצאו חומרי למידה נוספים, אותם מפרסם/מת מרכז/ת ההוראה. בנוסף, האתר מהווה עבורכם ערוץ תקשורת עם צוות ההוראה ועם סטודנטים אחרים בקורס. פרטים על למידה מתוקשבת ואתר הקורס, תמצאו באתר שה"ם בכתובת:
http://telem.openu.ac.il
מידע על שירותי ספרייה ומקורות מידע שהאוניברסיטה מעמידה לרשותכם, תמצאו באתר הספריה באינטרנט www.openu.ac.il/Library.

ייעוץ ינתן ביום ד' בין השעות 9:30-11:30 בטלפון 09-7781260. פגישה יש לתאם מראש.
ניתן לפנות גם בדואר אלקטרוני maya@openu.ac.il

אני מאחלת לך לימוד פורה ומהנה.

בברכה,

ד"ר מיה הרמן
מרכזת הקורס

1. לוח זמנים ופעילויות (20595 / ב2012)

שבוע לימוד	תאריכי שבוע הלימוד	יחידת הלימוד המומלצת	מפגשי ההנחיה*	תאריך אחרון למשלוח ממ"ן (למנחה)
1	16.3.2012-11.3.2012	יחידה 1 מבוא		
2	23.3.2012-18.3.2012	יחידה 2 תורת המידע	מפגש ראשון	
3	30.3.2012-25.3.2012	יחידה 3 הכנת נתונים		
4	6.4.2012-1.4.2012 (ו ערב פסח)	יחידה 4 סיווג וחיזוי	מפגש שני	
5	13.4.2012-8.4.2012 (א-ו פסח)	יחידה 5 עצי החלטה- חלק א		
6	20.4.2012-15.4.2012 (ה יום הזכרון לשואה)	יחידה 6 עצי החלטה- חלק ב	מפגש שלישי	ממ"ן 11 20.4.2012
7	27.4.2012-22.4.2012 (ד יום הזכרון) (ה יום העצמאות)	יחידה 7 למידה בייסיאנית ולמידה מבוססת תצפיות		ממ"ן 21 27.4.2012
8	4.5.2012-29.4.2012	יחידה 8 חוקי הקשר – חלק א		
9	11.5.2012-6.5.2012 (ה ל"ג בעומר)	יחידה 9 חוקי הקשר – חלק ב	מפגש רביעי	

** התאריכים המדויקים של המפגשים הקבוצתיים מופיעים ב"לוח מפגשים ומנחים".*

לוח זמנים ופעילויות - המשך

שבוע הלימוד	תאריכי שבוע הלימוד	יחידת הלימוד המומלצת	מפגשי ההנחיה*	תאריך אחרון למשלוח הממ"ן (למנחה)
10	18.5.2012-13.5.2012	יחידה 10 ניתוח אשכולות- חלק א		
11	25.5.2012-20.5.2012 (א יום ירושלים)	יחידה 11 ניתוח אשכולות- חלק ב	מפגש חמישי	ממ"ן 12 25.5.2012
12	1.6.2012-27.5.2012 (א שבועות)	יחידה 12- *פיזיקת* רשתות אינפו-עמומות	מפגש שישי	ממ"ן 22 1.6.2012
13	8.6.2012-3.6.2012	יחידה13 *פיזיקת* בחירת מאפיינים		
14	15.6.2012-10.6.2012	יחידה 14 *פיזיקת* נושאים מתקדמים בכריית מידע		
15	22.6.2012-17.6.2012	חזרה	מפגש שביעי	

מועדי בחינות הגמר יפורסמו בנפרד

** התאריכים המדויקים של המפגשים הקבוצתיים מופיעים ב"לוח מפגשים ומנחים".*

2. תיאור המטלות בקורס

קרא היטב עמודים אלו לפני שתתחיל לענות על השאלות

פתרון המטלות הוא חלק בלתי נפרד מלימוד הקורס – הבנה מעמיקה של חומר הלימוד דורשת תרגול רב. המטלות תבדקנה ותוחזרנה לך בצירוף הערות המתייחסות לתשובות.

2.1 מבנה המטלות

כל מטלה מורכבת מכמה שאלות.
את הפתרונות למטלה עליך להדפיס. רצוי להשאיר שוליים רחבים להערות המנחה.

אם השאלה בממ"ן אינה ברורה לך, אל תהסס להתקשר אל המנחה בשעות הייעוץ הטלפוני בלבד לצורך קבלת הסבר.

המטלות מלוות את יחידות הלימוד בקורס. להלן פירוט המטלות והיחידות שאליהן מתייחסת כל מטלה.

2.2 חומר הלימוד הדרוש לפתרון המטלות

ממ"ן 11 – יחידות לימוד 1-6 – רשות 2 נקודות.

ממ"ן 21 – יחידות לימוד 1-6 – **חובה** – 13 נקודות
(פרויקט – שלב א).

ממ"ן 12 – יחידות לימוד 7-8 – רשות 2 נקודות.

ממ"ן 22 – יחידות לימוד 7-11 – **חובה** – 13 נקודות
(פרויקט – שלב ב).

ממ"נים 21 ו-22 (פרויקט):
מכיוון ואלו **מטלות חובה** ומהווה משקל רב בציון הסופי, אין להגישן באיחור ללא קבלת אישור מראש. על – כן הקפד לשלוח את המטלות במועד.

2.3 ניקוד המטלות

סה"כ ניתן לצבור 26 - 30 נקודות במטלות.
מטלות החובה בקורס כוללות פרויקט המוגש בשני שלבים ומהוות יחד 26 נקודות.
מומלץ להגיש את כל המטלות.

3. התנאים לקבלת נקודות זכות בקורס

א) הגשת מטלות מנחה 21 ו- 22 (פרויקט חובה).
ב) ציון של לפחות 60 נקודות בפרויקט.
ג) ציון של לפחות 60 נקודות בבחינת הגמר.
ד) ציון סופי בקורס של 60 נקודות לפחות.

לבחינת הגמר רשאי לגשת רק סטודנט שצבר 26 נקודות לפחות.

לתשומת לבכם!

כדי לעודדכם להגיש לבדיקה מספר רב של מטלות הנהגנו את ההקלה שלהלן:

אם הגשתם מטלות מעל למשקל המינימלי הנדרש בקורס, **המטלות** בציון הנמוך ביותר, שציוניהן נמוכים מציון הבחינה (**עד שתי מטלות**), לא יילקחו בחשבון בעת שקלול הציון הסופי.

זאת בתנאי שמטלות אלה **אינן חלק מדרישות החובה בקורס** ושהמשקל הצבור של המטלות האחרות שהוגשו, מגיע למינימום הנדרש.

זכרו! ציון סופי מחושב רק לסטודנטים שעברו את בחינת הגמר בציון 60 ומעלה והגישו מטלות כנדרש באותו קורס.

הכנת המטלות חייבת להעשות על-ידי כל סטודנט בנפרד.

מטלות שלא יבוצעו באופן עצמאי – יפסלו!!!

מטלת מנחה (ממ״ן) 11

הקורס: 20595 - כריית מידע

חומר הלימוד למטלה: יחידות 1-6

מספר השאלות: 2 **משקל המטלה:** 2 נקודות

סמסטר: ב2012 **מועד אחרון להגשה:** 20.4.2012

אנא שים לב:
מלא בדייקנות את הטופס המלווה לממ״ן בהתאם לדוגמה שלפני המטלות.
העתק את מספר הקורס ומספר המטלה הרשומים לעיל.

בממ״ן זה **שתי** שאלות המתייחסות לטבלת נתוני האימון הרצ״ב.

ניתן להשתמש בתוכנת WEKA וכן בגיליון אלקטרוני EXCEL .

ענו במפורט על **שתי** השאלות.

בתשובתכם,

- ציינו כל הנחה שבצעתם
- צרפו את התוכנות והקבצים שערכתם בהם שימוש בחישובים.

להלן טבלת נתוני אימון:

מס׳ סטודנט	גיל	מגדר	השלים מעל 85 נ״ז	ממן11	ממן12	ממן21	ממן 22	מבחן
001	24	נ	כ	95	60	95	87	80
002	34	?	ל	85	?	66	69	54
003	119	ז	ל	73	65	82	78	?
004	45	ז	?	100	86	95	83	98
005	?	ז	כ	?	63	68	54	44
006	21	נ	?	72	64	66	58	60
007	33	ג	כ	89	76	88	64	85
008	44	נ	ל	98	100	?	97	100
009	?	נ	ל	72	67	75	?	72

שאלה 1 (25 נקודות)

ציינו והדגימו את שלבי הכנת הנתונים לביצוע כריית מידע כדוגמת טיפול בערכים חסרים, ערכים שגויים ועוד.

שאלה 2 (75 נקודות)

א. בנו עץ החלטה עבור נתוני האימון שבטבלה לחיזוי הצלחה/כישלון במבחן הסופי בקורס כריית מידע. בתשובתכם הדגימו את שלבי בחירת התכונה המפצלת בעץ.

שימו לב: יש לכלול חישובים כדוגמת אנטרופיה, Gain ratio, מדד גיני

ב. איזו מבין התכונה/תכונות ניתן להסיר ומדוע? אם אין תכונה הניתנת להסרה, יש לציין זאת מפורשות.

מטלת מנחה (ממ"ן) 21 - פרויקט גמר

הקורס: 20595 - כריית מידע

חומר הלימוד למטלה: יחידות 1-6

מספר השאלות: 2 **משקל המטלה:** 13 נקודות

סמסטר: ב2012 **מועד אחרון להגשה:** 27.4.2012

אנא שים לב:
מלא בדייקנות את הטופס המלווה לממ"ן בהתאם לדוגמה שלפני המטלות.
העתק את מספר הקורס ומספר המטלה הרשומים לעיל.

הנחיות

נתון בסיס נתונים מתחום תעשיית הרכב המצוי בכתובת:

http://archive.ics.uci.edu/ml/datasets/Automobile

הסברים אודות נתוני הקובץ תוכלו למצוא בכתובת:

http://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.names

מטרת הפרויקט:

לחזות את מחיר הרכב מתוך 26 התכונות הנתונות.

הפרויקט כולל שימוש בחבילת תוכנה WEKA לכריית מידע.

הפרויקט יבוצע בשני שלבים:

א. ממ"ן 21 – בשלב הראשון תידרשו להגדיר את הבעיה, להכין את הנתונים ולפתור את הבעיה בעזרת שיטות סיווג וחיזוי.

ב. ממ"ן 22 – בשלב השני תידרשו לפתור את הבעיה בעזרת חוקי הקשר וניתוח אשכולות. כמו כן, יהיה עליכם לסכם את עיקרי התוצאות שקיבלתם בממ"ן 21 וכן בממ"ן 22 והמסקנות .

על מנת לסייע לכם בפתרון הפרויקט מורכב הממ"ן ממספר שאלות הנחיה. **<u>היכם נדרשים לענות על כל השאלות לפי סדר הופעתן.</u>** לצורך פתרון השאלות הינכם רשאים להניח כל הנחה ו/או הפשטה סבירה שתידרשו לה. יש לציין במפורש בתחילת הפרויקט את ההנחות ו/או ההפשטות בהן הנכם משתמשים. כמו כן, יש לציין כל הנחה ו/או הפשטה עליה אתם מתבססים במהלך הפרויקט.

במסגרת הפרויקט יש להשתמש בחבילת תוכנה חופשית WEKA המצויה בכתובת:

WEKA: http://www.cs.waikato.ac.nz/ml/weka/index.html

צורת הגשה:

יש להגיש את הפרויקט מודפס. הקפידו על כתיבה בהירה ומאורגנת וכן על תרשימים ברורים וקריאים. יש להקפיד על תיעוד מפורט של כל שלבי הפרויקט. **אין צורך** לצרף לפרויקט נתונים טכניים של חבילת התוכנה WEKA.

1. הגדרת הבעיה והכנת הנתונים (50%)

(6%) א. הגדירו את הנתונים בהם השתמשתם בפרויקט כדוגמת: תכונות, סוג הנתונים, נתונים חסרים, תחומי ערכים ועוד.

(7%) ב. הגדירו את מטרות כריית המידע. ציינו את ההנחות וההפשטות בהן השתמשתם.

(7%) ג. בהמשך לסעיפים א ו-ב , הגדירו ותארו את שלבי ה- KDD עבור הבעיה הנתונה.

(15%) ד. בהמשך לסעיפים א ו-ב ערכו סקירה השוואתית לכלל החלופות האפשריות (לפחות 4 חלופות) לביצוע כריית מידע. בתשובתכם יש להתייחס ליתרונות/חסרונות כל אחת מהחלופות בהקשר לבעיה הנתונה.

(15%) ה. תארו את שלבי הכנת הנתונים. בתשובתכם יש להתייחס לבעיות באיכות הנתונים כולל טיפול בערכים חסרים, תצוגה גרפית של הנתונים, ניקוי הנתונים, שילוב והמרה של נתונים ועוד.

הערה:

בשלב זה ניתן להשתמש בתוכנת Excel.

2. סיווג וחיזוי (50%)

(5%) א. בחרו שתי שיטות לחיזוי הנתונים.
הסבירו את השיטות ונמקו את בחירתכם.

(15%) ב. תארו את שלבי השיטות שבחרתם בסעיף ב.

(8%) ג. עבור כל שיטה דווחו את תוצאות הניתוחים.

(7%) ד. העריכו את מידת הדיוק של כל שיטה.

(15%) ה. נתחו השוואתית את התוצאות והסיקו מסקנות כולל הצעות לשיפורים.

מטלת מנחה (ממ"ן) 12

הקורס: 20595 - כריית מידע

חומר הלימוד למטלה: יחידות 7-11

מספר השאלות: 3 **משקל המטלה:** 2 נקודות

סמסטר: ב2012 **מועד אחרון להגשה:** 25.5.2012

אנא שים לב:
מלא בדייקנות את הטופס המלווה לממ"ן בהתאם לדוגמה שלפני המטלות.
העתק את מספר הקורס ומספר המטלה הרשומים לעיל.

שאלה 1 (40%):

הפעילו את האלגוריתם Naïve-Bayes על קובץ נתוני ניסוי בפסיכולוגיה קוגניטיבית המצוי בכתובת:

http://archive.ics.uci.edu/ml/datasets/Balloons

הסברים אודות נתוני הקובץ תוכלו למצוא בכתובת:

http://archive.ics.uci.edu/ml/machine-learning-databases/balloons/balloons.names

וענו במפורט על הסעיפים הבאים:

א. ממשו את האלגוריתם בקוד או בגליון אלקטרוני Excel.

ב. חלקו את הקובץ באופן אקראי לנתוני אימון ונתוני מבחן.

ג. בפלט יש לכלול את דיוק האלגוריתם על נתוני האימון ונתוני המבחן תוך חישוב רווח בר-סמך ברמת בטחון של 95% .

הערות:

- ציינו כל הנחה שבצעתם
- צרפו את התוכנות והקבצים שערכתם בהם שימוש בחישובים.

שאלה 2 (30%):

נתונה הטבלה:

TID	Items_bought
T100	{M,O,N,K,E,Y}
T200	{D,O,N,K,E,Y}
T300	{M,A,K,E}
T400	{M,U,C,K,Y}
T500	{C,O,O,K,I,E}

בהנחה :

Min_support=60%

Min_confidence=80%

מצאו את כל הקבוצות התדירות תוך שימוש באלגוריתם א-פריורי.

שאלה 3 (30%):

בצעו אשכול לשבע הנקודות הבאות תוך שימוש באלגוריתם **k-means** .

A1(0, 0) , A2(8,0), A3(8,6)

B1(16,0), B2(0,6)

C1(16,6), C2(16,0)

בתשובתכם הניחו:

א. יש לבצע אשכול לשלושה אשכולות בלבד.

ב. מרכזי האשכולות בשלב ההתחלתי הם A1, B1, C1

ג. שימוש במרחק אאוקלידי (Euclidian distance) בריבוע.

שימו לב,

בתשובתכם הסופית יש להדגים את כל השלבים וכן להגדיר את האשכולות הסופיים.

מטלת מנחה (ממ"ן) 22 - פרויקט גמר

הקורס: 20595 - כריית מידע

חומר הלימוד למטלה: יחידות 7-11

מספר השאלות: 3 **משקל המטלה:** 13 נקודות

סמסטר: ב2012 **מועד אחרון להגשה:** 1.6.2012

אנא שים לב:
מלא בדייקנות את הטופס המלווה לממ"ן בהתאם לדוגמה שלפני המטלות.
העתק את מספר הקורס ומספר המטלה הרשומים לעיל.

הנחיות:

בהמשך לממ"ן 21 השלימו במסגרת מטלה זו את הפרויקט בקורס . מומלץ לשוב ולעיין בהנחיות שניתנו בממ"ן 21 .

ענו על השאלות הבאות:

1. חוקי הקשר (45%)

(10%) א. הגדירו חוקים בעלי מינימום confidence & support. הסבירו ונמקו .

(4%) ב. הגדירו את סט הפרמטרים. הסבירו ונמקו.

(10%) ג. בחרו לפחות שני אלגוריתמים של חוקי הקשר.
תארו את האלגוריתמים ונמקו את בחירתכם.

(4%) ד. הריצו ודווחו את התוצאות של כל אחד מהאלגוריתמים.

(7%) ה. לאור סעיף ד הגדירו אילו מהחוקים שהגדרתם בסעיף א יש בהם עניין ועשויים להיות שימושיים.

(10%) ו. נתחו השוואתית את התוצאות של שני האלגוריתמים והסיקו מסקנות.

2. ניתוח אשכולות (45%)

(2%) א. הגדירו מהו ניתוח אשכולות.

(2%) ב. תארו את אופן מדידת איכות האשכולות.

(12%) ג. בחרו שתי גישות לניתוח אשכולות.
בתשובתכם יש לכלול הסבר אודות כל גישה וכן נימוק לבחירתכם.

(12%) ד. תארו את שלבי ניתוח האשכולות עבור 2 הגישות שציינתם בסעיף ג.
בתשובתכם יש להתייחס בין היתר לאופן הכנת הנתונים, מה הם הפרמטרים, ערכי הפרמטרים ועוד.

(7%) ה. עבור כל גישה דווחו את תוצאות הניתוחים

(10%) ו. נתחו השוואתית את התוצאות והסיקו מסקנות.

3. סיכום ומסקנות (10%)

סכמו בקצרה את עיקרי התוצאות שהתקבלו בממ"ן 21 וכן בסעיפים הקודמים בממ"ן הנוכחי ואת המסקנות שניתן לקבל מתוצאות אלה.