אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

תאריך הבחינה: 17/07/2016

שם המרצה: פרופ' מרק לסט

שם הקורס: כריית נתונים ומחסני נתונים

מספר הקורס: 372-1-3105

שנה: תשע"ו (2016) סמסטר: ב' מועד: א'

משך הבחינה: 3 שעות

חומר עזר: דף נוסחאות (מצורף לבחינה) + מחשבון

# חלק 1 [50 נקודות]

- יש לענות על כל השאלות
- משקל של כל שאלה – **5 נקודות**
- יש לבחור **בתשובה הנכונה ביותר**
- יש לנמק בקצרה את התשובה בכתב-יד ברור במקום המיועד לכך על-גבי שאלון הבחינה בלבד. ניתן להשתמש בטיוטה לצורך עריכת התשובה. תשובה לא מנומקת (גם אם הבחירה נכונה) **תקבל ציון של אפס**

א. איזו משימה אינה מהווה משימה של חיזוי / רגרסיה (prediction / regression) ?

1) חיזוי שער החליפין בין דולר לשקל
2) חיזוי המועמד הזוכה בבחירות לנשיאות ארה"ב
3) חיזוי משך אשפוז של חולה בבית-חולים
4) חיזוי ציון ממוצע של סטודנט בלימודים באוניברסיטה

הסבר:

---

ב. מה איננו מוגדר כתפקיד (role) של תכונה בתוכנת RapidMiner ?

1) Attribute
2) Id
3) Label
4) Real

הסבר:

---

ג. חלוקה לאינטרוולים (binning)

1) מגדילה את שונות הנתונים
2) מקטינה את שונות הנתונים
3) מזהה נתונים חריגים
4) מנקה את הנתונים

הסבר:

---

ד. עליכם לחזות את הצלחת המועמד/ת ללימודים אקדמיים. באיזה טווח נמצאת האינפורמציה ההדדית המקסימלית בין ציון פסיכומטרי לבין ציון מאוני ממוצע באוניברסיטה?

1) [0, 1]
2) [6, 7]
3) [9, 10]
4) [90, 100]

הסבר:

---

ה. הערך המקסימלי של מדד ה-Gini של משתנה סיווג בינארי שווה ל-

1) 0.25
2) 0.50
3) 0.75
4) 1.00

הסבר:

---

ו. העלייה בערכו של הפרמטר Minimum significance level ברשת IFN
1) עשויה להגדיל את מספר השכבות ברשת
2) עשויה להגדיל את מספר הקודקודים ברשת
3) שתי התשובות נכונות
4) אף תשובה אינה נכונה

הסבר:

---

ז. מה אנחנו יודעים על משתנה חבוי (hidden variable) ברשת בייסיאנית?
1) יודעים רק שהמשתנה קיים
2) יודעים רק שהמשתנה קיים ואיך הוא משפיע על משתנים אחרים
3) את הערך שלו
4) אף תשובה אינה נכונה

הסבר:

---

ח. עקרון Apriori אומר:
1) תת-קבוצה של קבוצת פריטים שכיחה חייבת להיות שכיחה
2) קבוצת-על של קבוצת פריטים שכיחה חייבת להיות שכיחה
3) שתי התשובות נכונות
4) אף תשובה אינה נכונה

הסבר:

---

ט. משמעות התכונה Non-volatility ("אי-נדיפות") של מחסני נתונים היא
1) לא ניתן לגשת לנתונים במחסן
2) לא ניתן להוסיף נתונים למחסן
3) לא ניתן לעדכן נתונים במחסן
4) לא ניתן למחוק נתונים מהמחסן

הסבר:

---

י. השימוש בטבלאות סיכום (aggregation / summary tables) תורם ל-
1) מהירות הרצת השאילתות
2) איכות הנתונים במחסן
3) צמצום היקף הנתונים במחסן
4) אף תשובה אינה נכונה

הסבר:

---

## חלק 2 [50 נקודות]

- יש להציג את כל התוצאות עם **שלוש ספרות אחרי נקודה עשרונית** אלא אם צוין אחרת!
- יש לרשום את כל התשובות **על-גבי שאלון הבחינה בלבד**
- טיוטות החישוב ייגרסו **ללא בדיקה**

להלן נתונים אמיתיים של רעידות אדמה שנרשמו באחד מאזורי הארץ:

| No | YEAR | Max_Magnitude | Avg_ Magnitude | Num_Events | Class |
|---|---|---|---|---|---|
| 1 | 1996 | 4.5 | 2.952 | 14 | 1 |
| 2 | 1997 | 5.5 | 2.900 | 22 | 0 |
| 3 | 1998 | 3.5 | 2.700 | 12 | 0 |
| 4 | 1999 | 4.2 | 2.771 | 17 | 1 |
| 5 | 2000 | 4.3 | 2.910 | 10 | 1 |
| 6 | 2001 | 4.3 | 3.108 | 12 | 0 |
| 7 | 2002 | 4.2 | 2.667 | 12 | 0 |
| 8 | 2003 | 4.2 | 2.775 | 16 | 1 |
| 9 | 2004 | 4.5 | 3.038 | 8 | 0 |
| 10 | 2005 | 4.2 | 2.714 | 7 | 1 |
| 11 | 2006 | 4.3 | 3.065 | 20 | 0 |
| 12 | 2007 | 4 | 2.713 | 16 | 0 |
| 13 | 2008 | 4.2 | 2.642 | 12 | 0 |
| 14 | 2009 | 4.2 | 2.750 | 12 | 0 |
| 15 | 2010 | 4.1 | 2.783 | 12 | 1 |

- העמודה Max_Magnitude מייצגת את העוצמה המקסימלית של רעידות אדמה במהלך שנה מסוימת
- העמודה Avg_ Magnitude מייצגת את העוצמה הממוצעת של רעידות אדמה במהלך שנה מסוימת
- העמודה Num_Events מייצגת את כמות הרעידות שנרשמו במהלך שנה מסוימת
- העמודה Class מייצגת את העוצמה המקסימלית של רעידות אדמה במהלך השנה העוקבת (0 – מתחת לחציון, 1 –מעל לחציון) ומהווה את **משתנה המטרה**

**א.** האלגוריתם IFN הורץ על טבלת הנתונים הנ"ל עם רמת-המובהקות של 0.10 ושלוש תכונות קלט מועמדות (candidate input attributes): Max_Magnitude , Avg_ Magnitude , Num_Events. עבור השכבה הראשונה של הרשת נבחרה התכונה Avg_ Magnitude. התכונה פוצלה ע"י האלגוריתם לשלושת האינטרוולים הבאים: [2.642, 2.714), [2.714, 3.038), [3.038,∞). נא לחשב את האינפורמציה ההדדית (mutual information) בין השכבה הראשונה של הרשת למשתנה המטרה. **15 נקודות.**

| j' / j | 0 | Cond. | Joint | 1 | Cond. | Joint | Total | Cond. |
|---|---|---|---|---|---|---|---|---|
| [2.642, 2.714) | | | | | | | | |
| [2.714, 3.038) | | | | | | | | |
| > 3.038 | | | | | | | | |
| Total | | | | | | | | |

**Mutual Information**

| j' / j | 0 | 1 | Total |
|---|---|---|---|
| [2.642, 2.714) | | | |
| [2.714, 3.038) | | | |
| > 3.038 | | | |
| **Total** | | | |

ב. יש לחשב את הסטטיסטי Likelihood-Ratio Statistic עבור האינפורמציה ההדדית שחושבה בסעיף הקודם ואת מספר דרגות החופש שלה. **5 נקודות.**

| | | | |
|---|---|---|---|
| $G^2$ = | | DF = | |

ג. יש לנרמל את התכונות Num_Events , Avg_ Magnitude , Max_Magnitude לטווח שבין 0 ל-1 בשיטת min-max. **10 נקודות.**

| | **Max_Magnitude** | **Avg_ Magnitude** | **Num_Events** |
|---|---|---|---|
| Min | | | |
| Max | | | |

**הערכים המנורמלים:**

| **No** | **YEAR** | **Max_Magnitude** | **Avg_ Magnitude** | **Num_Events** |
|---|---|---|---|---|
| 1 | 1996 | | | |
| 2 | 1997 | | | |
| 3 | 1998 | | | |
| 4 | 1999 | | | |
| 5 | 2000 | | | |
| 6 | 2001 | | | |
| 7 | 2002 | | | |
| 8 | 2003 | | | |
| 9 | 2004 | | | |
| 10 | 2005 | | | |
| 11 | 2006 | | | |
| 12 | 2007 | | | |
| 13 | 2008 | | | |
| 14 | 2009 | | | |
| 15 | 2010 | | | |

ד. יש להריץ את האיטרציה הראשונה של האלגוריתם k-means תוך שימוש בערכים המנורמלים של שלוש התכונות שחישבתם לעיל. אין להתייחס ליתר התכונות בטבלה! יש לחשב את מרכזי האשכולות לפני ואחרי ביצוע האיטרציה. **20 נקודות**

| No | YEAR | Old cluster | Distance to 1 | Distance to 2 | New cluster |
|---|---|---|---|---|---|
| 1 | 1996 | **1** | | | |
| 2 | 1997 | **1** | | | |
| 3 | 1998 | **1** | | | |
| 4 | 1999 | **1** | | | |
| 5 | 2000 | **1** | | | |
| 6 | 2001 | **1** | | | |
| 7 | 2002 | **1** | | | |
| 8 | 2003 | **2** | | | |
| 9 | 2004 | **2** | | | |
| 10 | 2005 | **2** | | | |
| 11 | 2006 | **2** | | | |
| 12 | 2007 | **2** | | | |
| 13 | 2008 | **2** | | | |
| 14 | 2009 | **2** | | | |
| 15 | 2010 | **2** | | | |

**לפני האיטרציה:**

| | Max_Magnitude | Avg_ Magnitude | Num_Events |
|---|---|---|---|
| **Centroid 1** | | | |
| **Centroid 2** | | | |

**אחרי האיטרציה:**

| | Max_Magnitude | Avg_ Magnitude | Num_Events |
|---|---|---|---|
| **Centroid 1** | | | |
| **Centroid 2** | | | |

**372-1-3105 : כריית נתונים ומחסני נתונים - תשע"ו, סמסטר ב'**

## Information Theory

- Conditional Entropy $H(Y/X) = -\Sigma\, p(x,y)*\log p\,(y/x)$
- Mutual Information $I(X;Y) =$

$$\sum_{x,y} p(x,y)\bullet\log\frac{p(y/x)}{p(y)}$$

Conditional Mutual Information $I(X;Y/Z) =$

$$\sum_{x,y} p(x,y,z)\bullet\log\frac{p(x,y/z)}{p(x/z)\bullet p(y/z)}$$

- Fano's Inequality: $H\,(Y/X_1 \ldots X_n) \le H\,(P_e) + P_e \log_2 (m-1)$
- The MDL Principle

$$h_{MDL} = \arg\min_{h\in H}\{L_{C_1}(h) + L_{C_2}(D/h)\}$$

## Decision Trees

- Expected information needed to classify a tuple in $D$ (after using $A$):

$$Info_A(D) = \sum_{j=1}^{v}\frac{|D_j|}{|D|}\times I(D_j)$$

- Expected number of records in $C_i$, for class $j$:

$$e'_{ij} = \frac{e_j}{\sum_{j=1}^{c} e_j}\sum_{j=1}^{c} o_{ij}$$

- Chi-Square Statistic:

$$\sum_{j=1}^{c}\sum_{i=1}^{v}\frac{(o_{ij}-e'_{ij})^2}{e'_{ij}}\Big|_{H_0} \sim \chi_\alpha^2((v-1)(c-1))$$

- Apparent (pessimistic) error rate:

$$q = \frac{N - n_C + 0.5}{N}$$

- Entropy induced by threshold $T$:

$$E(A,T;S) = \frac{|S_1|}{|S|}Ent(S_1) + \frac{|S_2|}{|S|}Ent(S_2)$$

- Split Information:

$$SplitInfo_A(D) = -\sum_{j=1}^{v}\frac{|D_j|}{|D|}\times\log_2\left(\frac{|D_j|}{|D|}\right)$$

- Gini index:

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2$$

- Gini split ($T$):

$$gini_{split}(T) = \frac{N_1}{N}gini(T_1) + \frac{N_2}{N}gini(T_2)$$

- Twoing Splitting Rule:

$$\frac{p_L p_R}{4}\left[\sum_j |p(j/t_L) - p(j/t_R)|\right]^2$$

- Cost-complexity function (CART):

$$R_\alpha(T) = R(T) + \alpha\cdot|\tilde{T}|$$

## IFN

- IFN Conditional mutual information at a node z:

MI ($A_{i'}$ ; $A_i$ / z) =

$$\sum_{j=0}^{M_i-1}\sum_{j'=0}^{M_{i'}-1} P(V_{ij};V_{i'j'};z)\bullet\log\frac{P(V_{i'j'}^{ij}/z)}{P(V_{i'j'}/z)\bullet P(V_{ij}/z)}$$

- IFN Likelihood-Ratio Statistic:

$$G^2(A_{i'};A_i/z) = 2\bullet(\ln 2)\bullet E^*\bullet MI(A_{i'};A_i/z)$$

$$G^2|_{H_0} \sim \chi^2((NI_{i'}(z)-1)\cdot(NT_i(z)-1))$$

- Conditional Mutual Information in a Layer $i'$:

$$MI(A_{i'};A_i) = \sum_{\substack{z\in Layer_{i'}\\ Split(z)=true}} MI(A_{i'};A_i/z)$$

- IFN Connection Weight:

$$w_z^{ij} = P(V_{ij};z)\bullet\log\frac{P(V_{ij}/z)}{P(V_{ij})}$$

- Conditional Mutual Information (Split)

$$\sum_{t=0}^{M_i-1}\sum_{y=1}^{2} P(S_y;C_t;z)\bullet\log\frac{P(S_y;C_t/S,z)}{P(S_y/S,z)\bullet P(C_t/S,z)}$$

## Bayesian Learning

- Naïve Bayes Classifier:

$$C_{NB} = \arg\max_{C_i}\; P(C_i)*\prod_{k=1}^{n} P(x_k|C_i)$$

- m-estimate: $\dfrac{n_c + mp}{n+m}$

- Laplacian-estimate: $\frac{n_c+1}{n+K}$
- Joint probability in Bayesian network:

$$P(x_1,\ldots,x_n)=\prod_{i=1}^{n}P(x_i \mid Parents(X_i))$$

**k - Nearest Neighbors**

- Distance-weighted k-NN:

$$\hat{f}(q)=\arg\max_{v\in V}\sum_{i=1}^{k}w_i\delta(v,f(x_i))$$

$$w_i=\frac{1}{d(x_q-x_i)^2}$$

**SVM**

- Linear SVM:

$$\mathbf{w}=\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i$$

$$y_j(w^T x_j+b)\geq 1$$

- Nonlinear SVM:

$$g(x_j)=\sum_{i\in SV}\alpha_i y_i K(x_i,x_j)+b$$

- Polynomial kernel:

$$K(\mathbf{x}_i,\mathbf{x}_j)=(1+\mathbf{x}_i^T\mathbf{x}_j)^p$$

**Clustering**

- Distance measure for <u>symmetric</u> binary variables:

$$d(i,j)=\frac{b+c}{a+b+c+d}$$

- Distance measure for <u>asymmetric</u> binary variables:

$$d(i,j)=\frac{b+c}{a+b+c}$$

- Distance measure for nominal variables:

$$d(i,j)=\frac{p-m}{p}$$

- Distance measure for variables of mixed types:

$$d(i,j)=\frac{\sum_{f=1}^{p}\delta_{ij}^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{p}\delta_{ij}^{(f)}}$$

- Rank for an ordinal variable:

$$z_{if}=\frac{r_{if}-1}{M_f-1}$$

- Cluster centroid: $C_m=\frac{\sum_{i=1}^{N}(t_{ip})}{N}$

**Data Preparation**

- min-max normalization:

$$v'=\frac{v-min_A}{max_A-min_A}(new\_max_A-new\_min_A)+new\_min_A$$

- z-score normalization:

$$v'=\frac{v-mean_A}{stand\_dev_A}$$

- normalization by decimal scaling:

$$v'=\frac{v}{10^j}$$

- Simple Moving Average:

$$\hat{Y}_{t+1}=\frac{Y_t+Y_{t-1}+\ldots+Y_{t-k+1}}{k}$$

- Weighted Moving Average:

$$\hat{Y}_{t+1}=w_tY_t+w_{t-1}Y_{t-1}+\ldots+w_{t-k+1}Y_{t-k+1}$$

$$where:\quad w_t+w_{t-1}+\ldots+w_{t-k+1}=1$$

- Exponential Moving Average:

$$F_t=\alpha Y_{t-1}+(1-\alpha)F_{t-1}$$