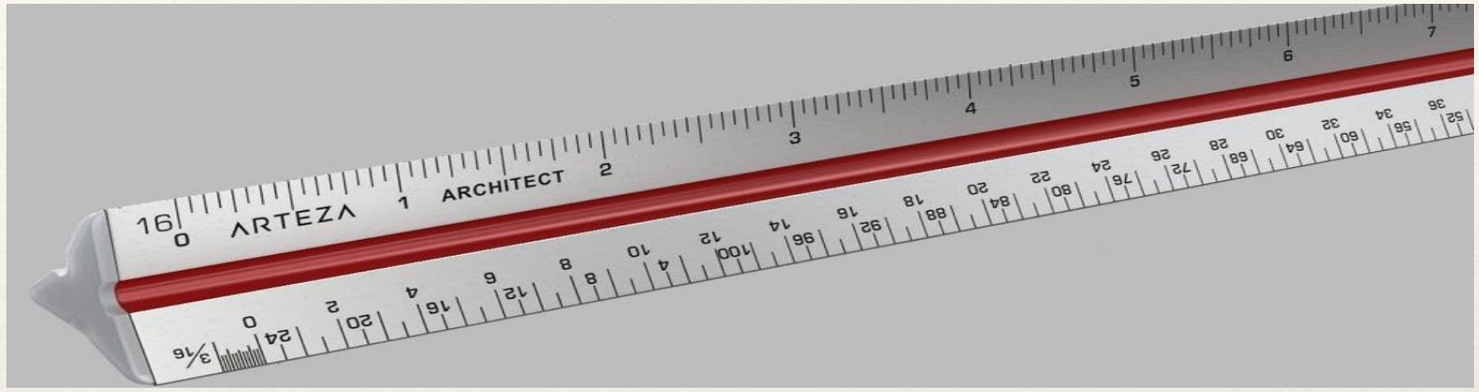
השלמה - KNN

פיתוח:
ד"ר יהונתן שלר
משה פרידמן

KNN – השלמה

סילום (Scaling) של מאפיינים - תזכורת



סילום (Scaling):

סילום מאפיינים - הוא שיטה המשמשת לנורמליזציה של טווח המאפיינים.

המטרה: סילום מחדש, כמו במעבר מאינץ' לס"מ

standardization (t-distribution) – הופכים את הממוצע החדש ל-0, וסטיית התקן, הופכת ל1

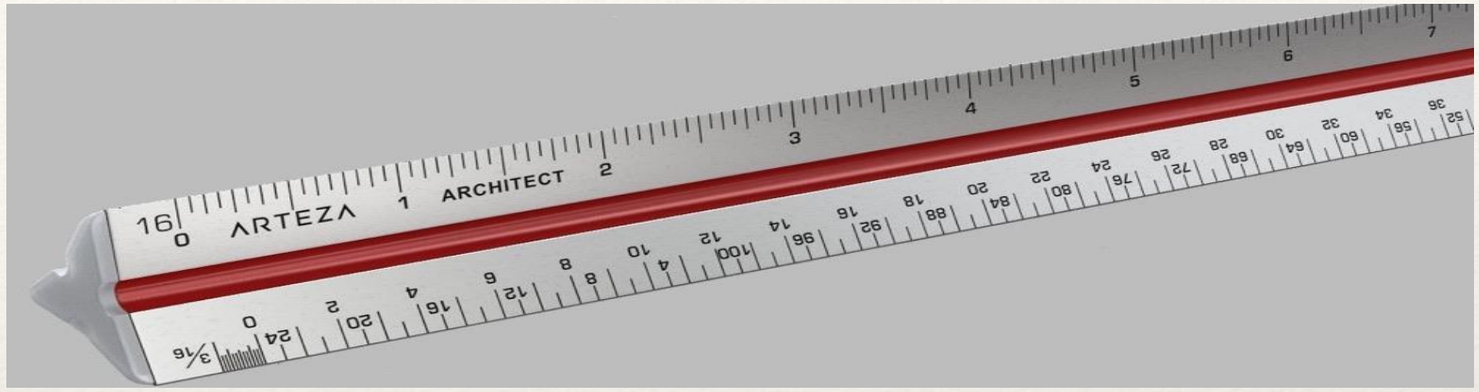
- משתמשים בהתפלגות t

minmax normalization – הסילום מתבצע כך שערך המינימום והמקסימום החדשים, הינם 0 ו-1 בהתאמה.

מה הקשר בין סילום ל-KNN?

KNN וסילום (Scaling) - מוטיבציה

OR



<u>מוטיבציה</u> –

- למאפיינים שונים פונקציית התפלגות שונה
- KNN לא מניח איזשהם הנחות על התפלגות הנתונים
- סולם (scale) שונה עלול להוביל לעיוות המרחק – מדוע?
- סולם (scale) שונה עלול לתת משקל שונה למאפיינים שונים, רק בגלל הסולם השונה (במקרה של KNN, זה די דומה) – מדוע?

(t-distribution) standardization-ו KNN

מוטיבציה –

- למאפיינים שונים פונקציית התפלגות שונה
- KNN לא מניח איזשהם הנחות על התפלגות הנתונים

פתרון ע"י שיטת הסילום (t-distribution) standardization:

- בסטטיסטיקה – כל התפלגות ניתן להפוך להתפלגות t, אם ידועות הממוצע וסטיית התקן במדגם (התפלגות t, היא קירוב להתפלגות z שהינה סוג של התפלגות נורמלית)
- במקרה שלנו – המדגם הוא ה- train set
- ב- KNN, מהווה תרומה למימדים השונים של פונקציית המרחק

minmax normalization -ו KNN

מוטיבציה –

- סולם (scale) שונה עלול להוביל לעיוות המרחק – מדוע?
- סולם (scale) שונה עלול לתת משקל שונה למאפיינים שונים, רק בגלל הסולם השונה (במקרה של KNN, זה די דומה) – מדוע?

פתרון ע"י שיטת הסילום minmax normalization:

- השוואה פשוטה של הסולם, ע"י קביעת סולם בטווח אחיד.
- מכונה גם נרמול מינימום ומקסימום.

מקובלים למשל הטווחים:

- [0,1] – כפי שהיה לכם בתרגיל – הMin החדש הופך ל-0, והמקסימום ל-1
- [-1,1] – לעיתים מסייע, בדומה למרחק cosine

KNN – איך נקבל החלטה לגבי דוגמה חדשה?

KNN – אלגוריתם הסיווג

Input:

- k – the number nearest neighbors; the set of training examples.

The KNN Algorithm:

- for test instance x_j in the test-set:
  - Calculate $d(x_j,x_i)$
  - Select the k closest training examples, $d(x_j,x_i)$ sorted
  - Use majority voting to classify the test examples

Notations and Terms:

$\mathbf{x_j}$ – example (number j) from the test-set

$\mathbf{x_i}$ – example (number i) from the train-set

$\mathbf{d(x_j,x_i)}$ – distance function – measures distance between x_j and x_i.

KNN – תרגיל

<u>סימונים</u>:

נסמן וקטור (feature vector) עבור ערכי שני מאפיינים X1,X2 כך: (x1,x2)

וקטורים, עבורם ידועה הקטגוריה שלהם c, נסמן כך: (x1,x2|c)

<u>נתונים הווקטורים הבאים</u>:

(0,0|1),(1,0|1),(1,1|-1),(4,2|1),(3,5|-1),(1,4|1),(3,1|-1)

מצאו באמצעות אלגוריתם KNN את הסיווג של הווקטור (3,2)

<u>הערות</u>:

- לפונקצית מרחק, השתמשו בשיטת מרחק אוקלידית
- דלגו כרגע על שלב הסילום
- בחנו את הפתרון עבור k=3,7

KNN – תרגיל

קודם כל, נחשב את המרחקים (לפי הוראות התרגיל משתמשים בפונקצית מרחק אוקלידי)

	מרחק מ-(3,2)	סיווג	ווקטור
3.6	$\sqrt{(3-0)^2+(2-0)^2}=\sqrt{13}$	1	(0,0)
2.8	$\sqrt{(3-1)^2+(2-0)^2}=\sqrt{8}$	1	(1,0)
2.2	$\sqrt{(3-1)^2+(2-1)^2}=\sqrt{5}$	-1	(1,1)
1	$\sqrt{(3-4)^2+(2-2)^2}=\sqrt{1}$	1	(4,2)
3	$\sqrt{(3-3)^2+(2-5)^2}=\sqrt{9}$	-1	(3,5)
2.8	$\sqrt{(3-1)^2+(2-4)^2}=\sqrt{8}$	1	(1,4)
1	$\sqrt{(3-3)^2+(2-1)^2}=\sqrt{1}$	-1	(3,1)

איזו קטגוריה נבחר עבור K=3? איזו קטגוריה נבחר עבור K=7?

- עבור – k=3 - שני שכנים מצביעים 1-, ושכן אחד מצביע 1, לפי הרוב - נסווג **1-**
- עבור – k=7 - הצבעת הרוב – נסווג **1**

KNN – בחירת הקטגוריה בזמן סיווג
כיצד נבחר את הקטגוריה עבור דוגמה חדשה?

- **<u>1-NN</u>** - Given a new point x, we wish to find it's nearest point and return it's classification.

- **<u>K-NN</u> -** Given a new point x, we wish to find it's k nearest points and return their average classification.

- **<u>Weighted</u> -** Given a new point x , we assign weights to all the sample points according to the distance from x and classify x according to the weighted average.

משמעות של ערכי K שונים

- המשמעות של ערכי K מאוד קטנים, עלולה להיות החלטה המושפעת מאוד מרעש (מדוע?)
- המשמעות של ערכי K מאוד גדולים, משמעה, עליה משמעותית בסיבוכיות של אלגוריתם KNN בו מרכז הכובד הוא בזמן אמת / בזמן הבדיקה.

שאלה: מה יקרה אם K שואף ל-n (כאשר n מסמנת את מספר הדוגמאות באימון)?

- תשובה: התשובה בעצם תשאף להתפלגות הקטגוריות ב-training-set (מדוע?)

משמעות של ערכי K שונים (המשך)

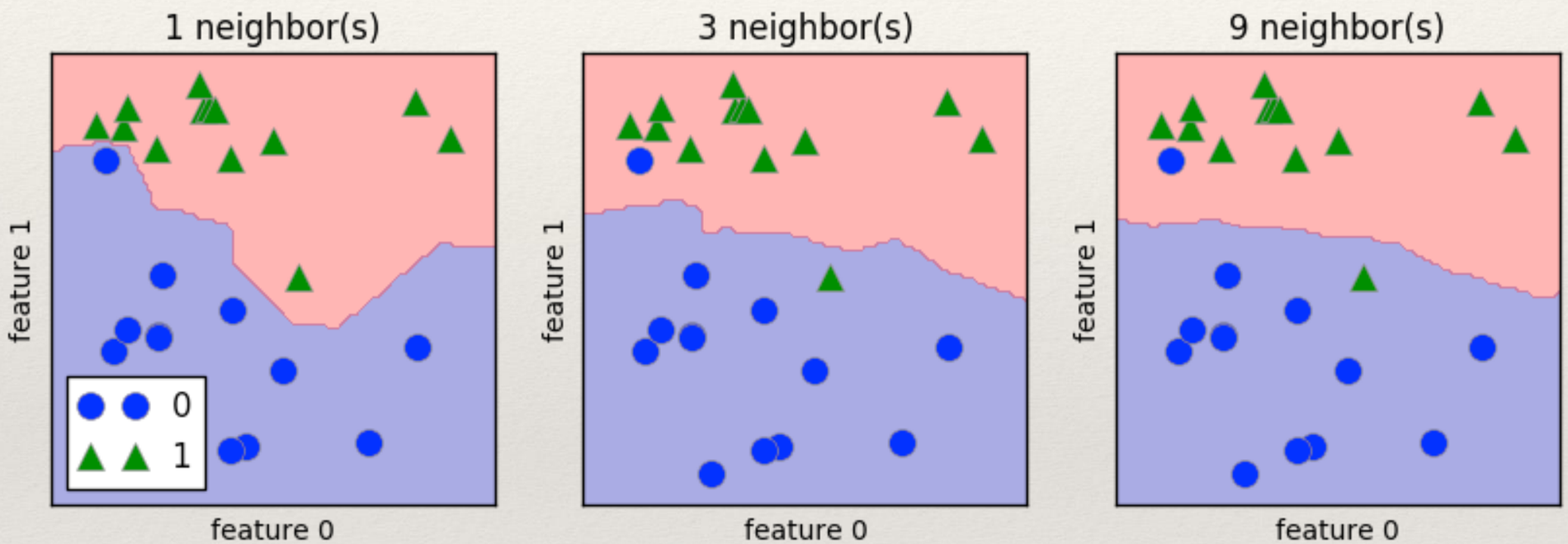
- אם מספר המחלקות הוא 2, נשאף למספר K אי זוגי (מדוע?)
- באופן דומה ננסה לבחור את K באופן שיעזור להכריע ולבחור את המחלקה מבין השכנים הקרובים.
- אחד מכללי האצבע הוא לבחור $K=O(\sqrt{n})$, אבל זה תלוי בבעיה.

<u>שיפורים נוספים</u>:

- בהמשך הקורס, נלמד דרך לבחור את ערכו של K
- בחירת פונקצית המרחק (ברירת המחדל – היא פונקצית מרחק אוקלידית) המיטבית לבעיה
- הכרעה, כאשר יש שיוויון בין המחלקות

משמעות של ערכי K שונים (המשך)

<u>דוגמה להשפעה של ערכי K שונים</u>:



<u>שאלה</u>: האם לא נעדיף את הגרף עבור K=1, הרי נראה שהוא יותר מתאים לנתונים?

<u>תשובה</u>: יש חשש להתאמת יתר ל-training (overfitting), ורגישות גבוהה מידי לרעש (נרחיב על התאמת יתר עוד במהלך הקורס).

שאלות ביניים

שאלה 1:

כיצד סילום ע"י standardization (t-distribution) מסייע ל-KNN?

שאלה 2:

כיצד סילום ע"י minmax normalization מסייע ל-KNN?

שאלה 3:

כיצד נקבל החלטה לגבי הקטגוריה של דוגמה חדשה ע"י KNN? מה פירוש 1-NN בעצם? מה ההבדל בין בחירה לפי הרוב, לבין בחירה ממושקלת?

שאלה 4:

מה המשמעות של ערכי K שונים (קטנים וגדולים)? מדוע נשאף לבחור K אי זוגי אם מספר הקטגוריות הוא 2?

KNN – תכונות

- KNN הינו אלגוריתם עצלן - עיבוד הנתונים מתבצע רק עם קבלת נקודה חדשה לסיווג
  - דוחים את רוב העבודה לזמן הסיווג
- KNN לא מניח איזשהם הנחות על התפלגות הנתונים
- KNN עובד מידית גם על "ריבוי מחלקות" (נדון על ריבוי מחלקות עוד במהלך הקורס)

KNN – סיכום השיטה

1. הגדרת הבעיה כבעיית סיווג ו-modeling
2. איסוף דוגמאות, vectorization
3. חלוקה ל-train-test (ו-validation)
4. cleansing – ניקוי ה-data, וסילום הערכים האפשריים.

- training – אלגוריתם עצלן – לא עושה (כמעט) כלום בשלב זה
- שיערוך (משתמש בסיווג) – בהמשך
- סיווג (דוגמאות לא ידועות)

נתראה בשבוע הבא ☺