Scaling exercise

Exercise II

תרגול רקע סטטיסטי בסיסי וסילום

נושאים:

- מושגים ותרגול – תזכורת (מההרצאה בשבוע שעבר)

משתנה מקרי - תזכורת

משתנה מקרי: הוא פונקציה המתאימה כל אירוע אפשרי במרחב הסתברות לערך מספרי.

דוגמאות:

- התאמת צד מטבע לערך 0, וצדו השני לערך 1;
- התאמת ערך של 1,...,6 בהתאם לאחת הפאות בקוביה.
- גובהו של אדם שנבחר באקראי הוא גם כן משתנה מקרי.

משתנה מקרי בדיד ורציף - תזכורת

- משתנה מקרי בדיד – קבוצת הערכים האפשרית (מרחב המדגם) סופית
  - למשל: ערכי המספרים בקוביה
- משתנה מקרי רציף – קבוצת הערכים האפשרית (מרחב המדגם) אין סופית
  - למשל: טמפרטורה של מים

מרחב המדגם Ω - תזכורת

מרחב המדגם Ω: קבוצת כל התוצאות האפשריות בניסוי.

דוגמאות:

- מטבע לערך 0, וצדו השני לערך 1; {0,1}
- התאמת ערך של 1,...,6 בהתאם לאחת הפאות בקוביה. { 1,2,3,4,5,6}
- קבוצת המספרים הממשיים בין 0 ל-100, (היכולה לתאר למשל טמפרטורה אפשרית של מים). [0,100]

הערה: אין חובה שמ"מ יתאר בהכרח משהו מהעולם האמיתי (יכול סתם לתאר מס' ממשי אקראי בין 0 ל-100)

מאורע / תצפית על מאורע (observation) - תזכורת

מאורע: תוצאה נצפת מסויימת בניסוי מסויים.

דוגמאות:

- התוצאה 3 בזריקת קוביה;
- גובה 1.72 של סטודנט.

דוגמה 1 – תמונות מכוניות – הסתברויות בסיסיות - תזכורת

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

שאלה: כמה מכוניות אדומות יש ב-dataset?

- תשובה: 5

שאלה: מהי ההסתברות להמצאות מכונית אדומה?

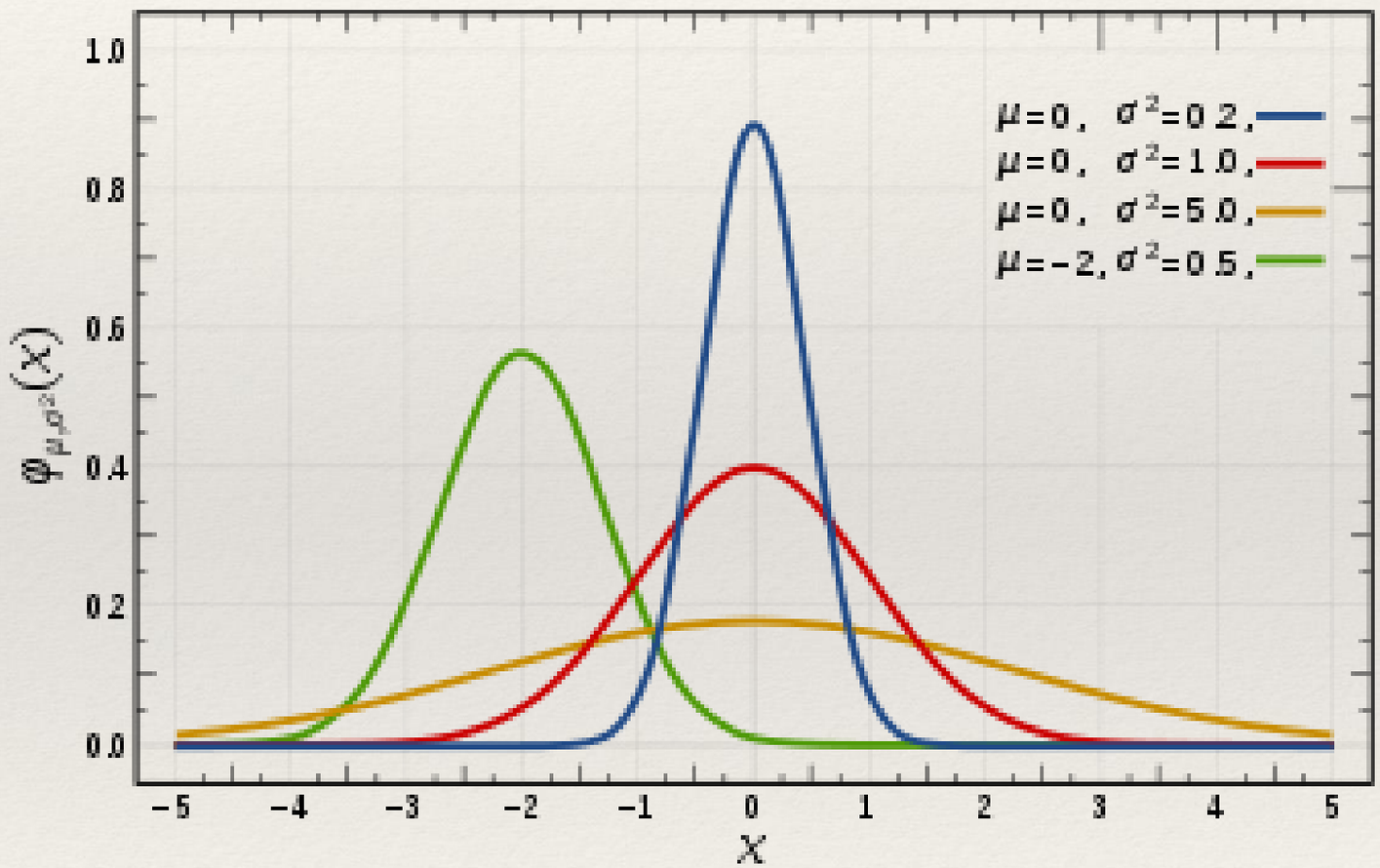
- תשובה: $p(Color = red) = \frac{5}{10} = 0.5$

שאלה: מהי ההסתברות להמצאות מכונית ספורט?

- תשובה: $p(Type = Sports) = \frac{6}{10} = 0.6$

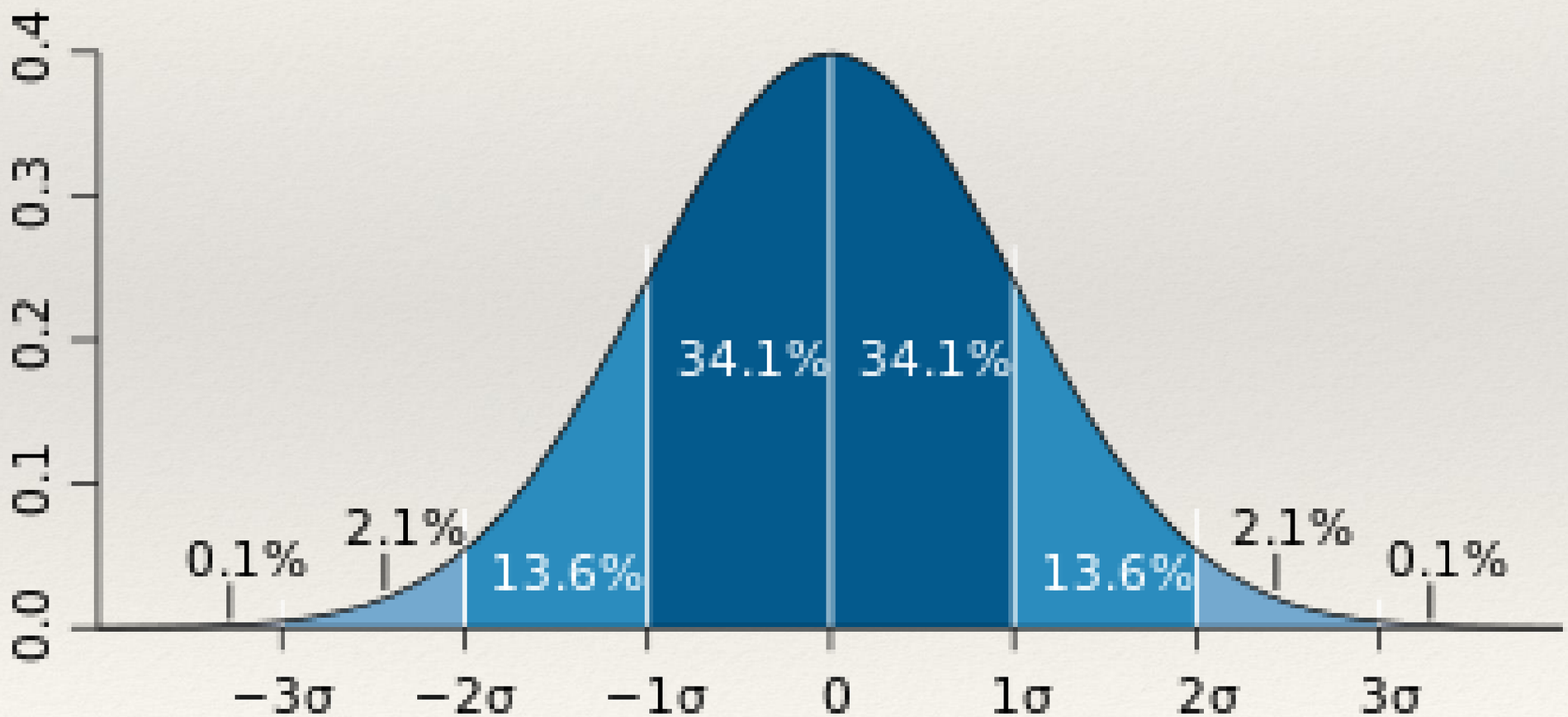
התפלגות נורמלית - תזכורת

התפלגות נורמלית :



התפלגות z – ההתפלגות הנורמלית הסטנדרטית - תזכורת

התפלגות z – סוג מיוחד של התפלגות נורמלית עם תוחלת 0, וסטיית תקן 1



התפלגות במדגם (ב-training set) - תזכורת

בסטטיסטיקה – כל התפלגות ניתן להפוך להתפלגות t, אם ידועות הממוצע וסטיית התקן <u>במדגם</u>

- נהוג לסמן ממוצע במדגם ע"י $\overline{x}$
- סטיית התקן במדגם:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

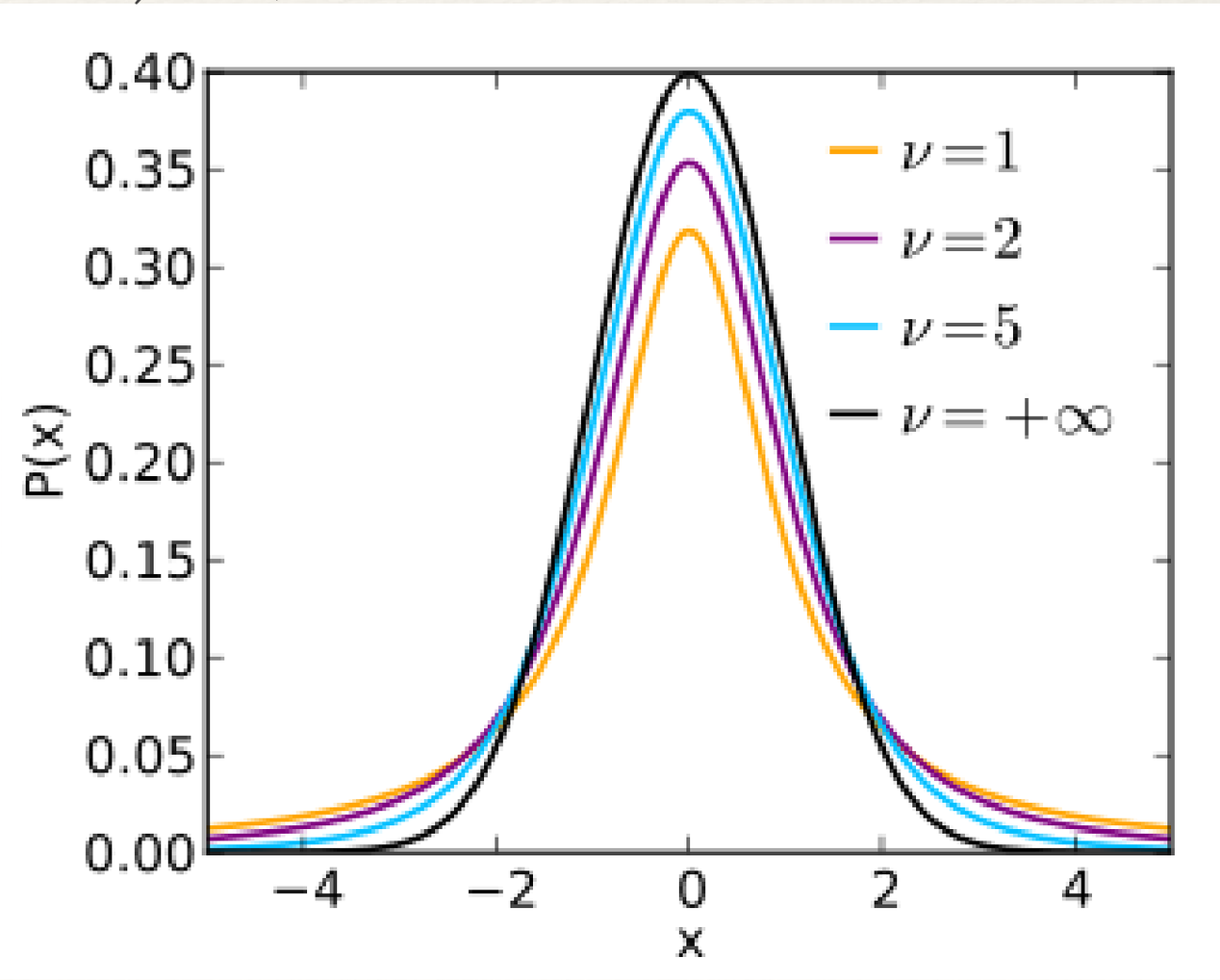
- שימו לב, שבשונות במדגם מחלקים ב-n1

התפלגות t - תזכורת

התפלגות t – התפלגות המבוססת על מידע שנאסף במדגם.

- התפלגות t שואפת להתפלגות z, כאשר גודל המדגם שואף לאינסוף
- בפועל מתייחסים לערכים הרבה יותר קטנים (כדי להחשיב כקירוב להתפלגות z)

- בעזרת התפלגות t – נדמה כל התפלגות לנורמלית
- בעזרת התפלגות t – ניתן בעצם להשוות את הסולם (טווח הערכים) של המאפיינים השונים



דוגמה 2 – The Iris dataset- תזכורת

- אחד ה-datasets המפורסמים
- מכיל feature vectors שמתארים מופעים של אירוסים

- נייצג כל instance ע"י מאפיינים הנוגעים לעלי הכותרת ועלי הגביע

<u>מאפייני ה- Iris Dataset (ה-Feature set):</u>

- עלי גביע (sepal):
  - אורך, רוחב
- עלי כותרת (petal):
  - אורך, רוחב

<u>דוגמאות ל-feature vectors:</u>

sepal		petal	
length	width	length	width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2

דוגמה 2 – The Iris dataset - תזכורת

- נניח שיש לנו נתונים עבור 5 אירוסים ב-dataset (בצורת feature vectors)

sepal		petal	
length	width	length	width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2

שאלה: מהו האורך הממוצע של עלי הגביע (sepal length)?

- תשובה: נסמן את המאפיין כ-X_1; $\overline{X_1} = \frac{5.1+4.9+4.7+4.6+5}{5} = 4.86$

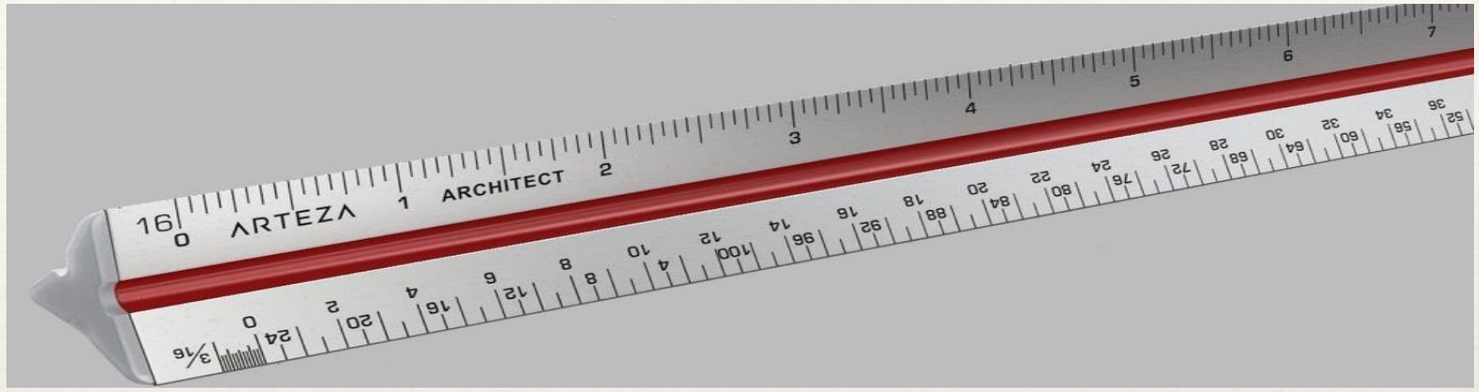
מהי סטיית התקן?

- תשובה: $S_{X_1} = \sqrt{\frac{1}{4} \cdot \left((5.1-4.86)^2+(4.9-4.86)^2+(4.7-4.86)^2+(4.6-4.86)^2+(5-4.86)^2\right)}$

$= \sqrt{(0.0576+0.0016+0.0256+0.0676+0.0196)/4} \approx 0.207$



סילום (Scaling) של מאפיינים



סילום (Scaling):

סילום מאפיינים - הוא שיטה המשמשת לקביעת טווח חדש של ערכי המאפיינים.

המטרה: סילום מחדש, בדומה למעבר מאינץ' לס"מ

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

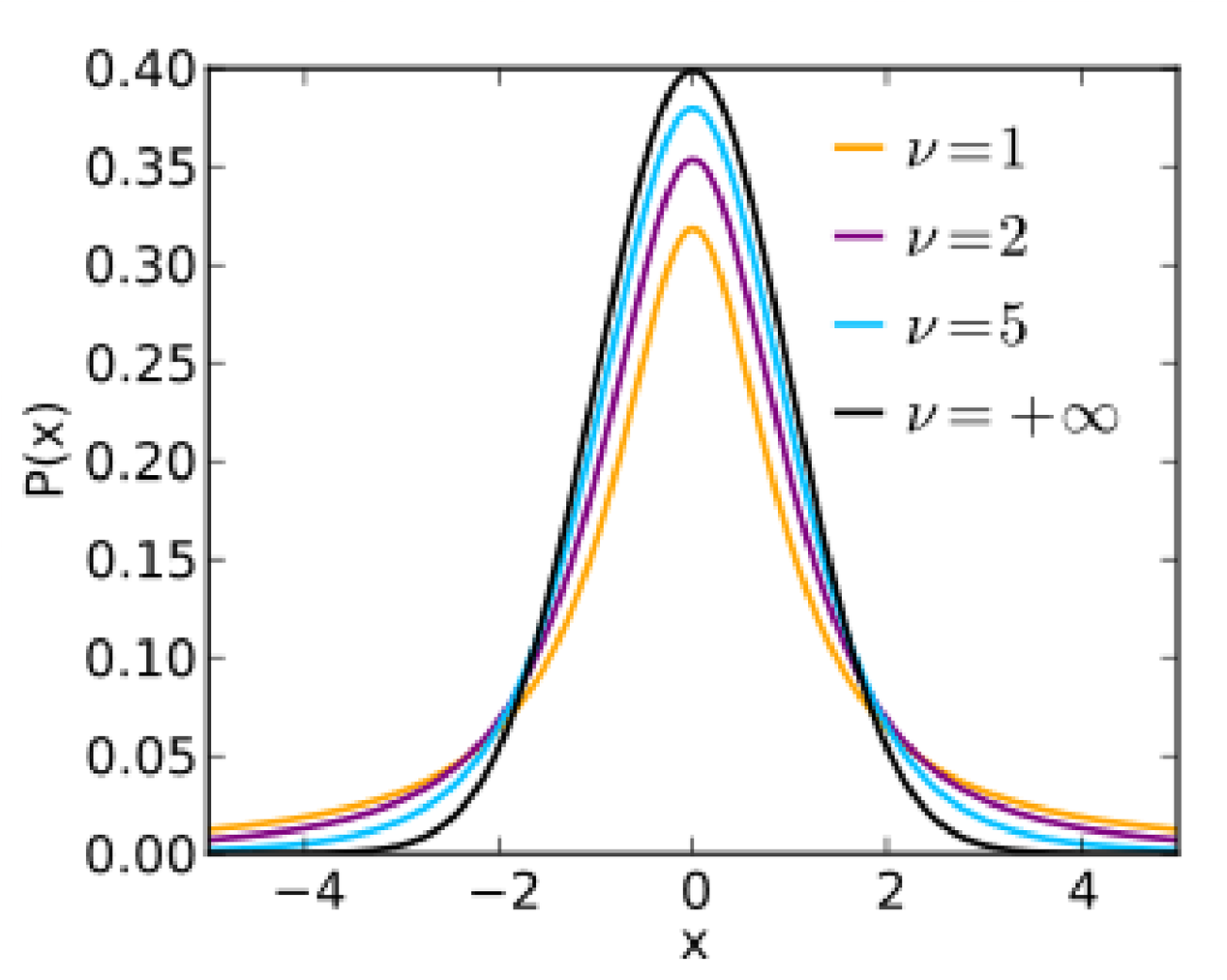
- משתמשים בהתפלגות t

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

t-distribution standardization - סילום

התפלגות t – שואפת להתפלגות z, כאשר גודל המדגם שואף לאינסוף

- בפועל מתייחסים לערכים הרבה יותר קטנים



<u>t-distribution Standardization</u>

– ניקח את ההתפלגות של כל מאפיין, ונעביר אותה להפלגות t

- השיטה: מפחיתים את הממוצע ($\bar{x}$) מהערך של המאפיין ומחלקים בסטיית התקן במדגם (s)

The Iris dataset – דוגמה 3
t-distribution standardization

sepal		petal	
length	width	length	width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2

- נניח שיש לנו נתונים עבור 5 אירוסים ב-dataset (בצורת feature vectors)
- ממוצע אורך של עלי הגביע (sepal length):4.86
- סטיית התקן ≈ 0.207

שאלה: בצעו סילום למאפיין 'עלי הגביע' ע"י t-distribution standardization

השיטה: מכל ערך מפחיתים את הממוצע ומחלקים בסטיית התקן

- תשובה:

Sepal length	
before scaling	after scaling
5.1	(5.1-4.86)/0.207= 1.16
4.9	(4.9-4.86)/0.207= 0.19
4.7	(4.7-4.86)/0.207= -0.77
4.6	(4.6-4.86)/0.207= -1.25
5	(5-4.86)/0.207= 0.67

Minmax normalization - סילום

Minmax normalization - השוואה פשוטה של הסולם, ע"י קביעת סולם בטווח אחיד.

- מכונה גם נרמול מינימום ומקסימום.

טווחים מקובלים:

- [0,1] – נשתמש בד"כ בטווח זה ב minmax normalization
- [-1,1]

The Iris dataset – דוגמה 4
Minmax normalization

נניח שיש לנו נתונים עבור 5 אירוסים ב-dataset (בצורת feature vectors)

שאלה: בצעו סילום למאפיין 'עלי הגביע' (sepal length): ע"י

Minmax normalization, כך שהערכים המתקבלים בטווח [0,1]

תשובה: השיטה - מכל ערך מפחיתים את המינימום ומחלקים במקסימום פחות המינימום.

sepal		petal	
length	width	length	width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2

Sepal length	
before scaling	after scaling
5.1	(5.1-4.6)/(5.1-4.6) = 1
4.9	(4.9-4.6)/(5.1-4.6) = 0.6
4.7	(4.7-4.6)/(5.1-4.6) = 0.2
4.6	(4.6-4.6)/(5.1-4.6) = 0
5	(5-4.6)/(5.1-4.6) = 0.8

שאלה: מה יש לעשות, כדי לעשות סילום של Minmax normalization עבור הטווח [-1,1]? בצעו סילום כנ"ל עבור הדוגמה הראשונה והשלישית.

תשובה: נוכל לעשות חישוב דומה, להכפיל ב2 ולהפחית 1:

- עבור הדוגמא הראשונה: 2*(5.1-4.6)/(5.1-4.6)-1 = 1
- עבור הדוגמא השלישית: 2*(4.7-4.6)/(5.1-4.6)-1 = -0.6

The Iris dataset – דוגמה 4
סיכום - Minmax normalization

נניח שיש לנו נתונים עבור 5 אירוסים ב-dataset (בצורת feature vectors)

שאלה: בצעו סילום למאפיין 'עלי הגביע' (sepal length): ע"י

Minmax normalization, סיכום ל-2 הטווחים:

sepal		petal	
length	width	length	width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2

Sepal length - scaling with **minmax normalization** - **2 ranges**		
before scaling	**after** scaling **[0,1] range**	**after** scaling **[-1,1] range**
5.1	(5.1-4.6)/(5.1-4.6) = **1**	2*(5.1-4.6)/(5.1-4.6)-1 = **1**
4.9	(4.9-4.6)/(5.1-4.6) = **0.6**	2*(4.9-4.6)/(5.1-4.6)-1 = **0.2**
4.7	(4.7-4.6)/(5.1-4.6) = **0.2**	2*(4.7-4.6)/(5.1-4.6)-1 = **-0.6**
4.6	(4.6-4.6)/(5.1-4.6) = **0**	2*(4.6-4.6)/(5.1-4.6)-1 = **-1**
5	(5-4.6)/(5.1-4.6) = **0.8**	2*(5-4.6)/(5.1-4.6)-1 = **0.6**