Machine learning

Scaling

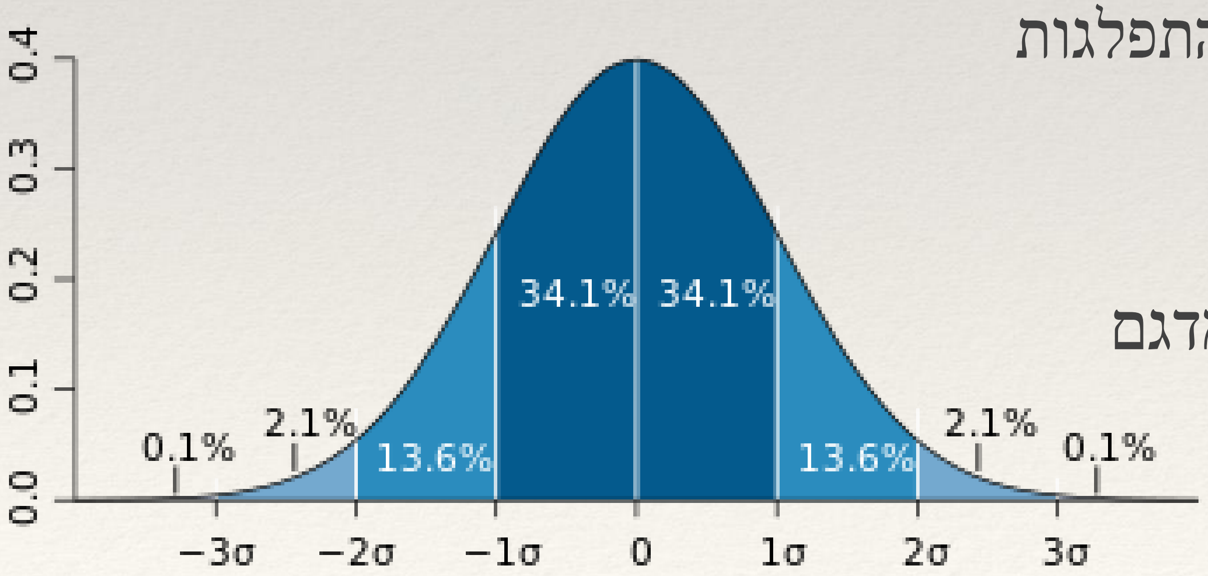
Lecture II

פיתוח:
ד"ר יהונתן שלר
משה פרידמן

הקדמה הסתברותית וסטטיסטית

מושגים:

- משתנה מקרי (בדיד ורציף)
- מרחב המדגם, מאורע
- התפלגות
  - פונקציית צפיפות (PDF), פונקציית התפלגות מצטברת (CDF)
  - תוחלת, שונות וסטיית תקן (באוכלוסייה ובמדגם).
- התפלגות בדידה ורציפה, התפלגות נורמלית, התפלגות z
- מדגם
  - ממוצע וסטיית תקן במדגם
  - התפלגות t



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 בקובייה.
- גובהו של אדם שנבחר באקראי הוא גם כן משתנה מקרי.

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)

משתנה מקרי בדיד ורציף - תזכורת

- משתנה מקרי בדיד – קבוצת הערכים האפשרית (מרחב המדגם) סופית
  - למשל: קובייה

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

מאורע / תצפית על מאורע (observation)

מאורע: תוצאה נצפית מסוימת בניסוי מסוים.

דוגמאות:

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

הסתברות מאורע - תזכורת

הסתברות: מידת הסבירות שמאורע מסוים יתרחש.

- ההסתברות של מאורע יכולה לקבל ערך מספרי שבין 0 ל-1
- למשל, הסתברות 1/6 לקבלת הערך 4 בקובייה הוגנת

התפלגות - distribution - תזכורת

התפלגות היא מרכיב בסיסי בתיאור ההתנהגות של תופעה או תהליך שיש בהם היבטים אקראיים.

ההתפלגות קובעת מהו הסיכוי של כל מאורע

פונקציית צפיפות
(Probability density function)

פונקציית צפיפות - של משתנה מקרי היא פונקציה המתארת את צפיפות המשתנה בכל נקודה במרחב המדגם.

$$\text{pdf} = f_{x\in\Omega}(x) = pr_{x\in\Omega}(X = x)$$

- סך כל הערכים שבפונקציית הצפיפות = 1

$$\sum_{X\in\Omega} f(x) = 1$$

פונקציית ההתפלגות המצטברת
(Cumulative distribution function)

פונקציית ההתפלגות המצטברת - של משתנה מקרי היא פונקציה של משתנה מקרי X, שערכיה קובעים את ההסתברות למאורעות מהצורה X<=a, לכל a ממשי.

$$F(x) = \sum_{z\in A, z\leq x} P(X=z)$$

❖ לדוגמה - קובייה הוגנת:

$$F(x) = \begin{cases} 0 & : x < 1 \\ 1/6 & : 1 \leq x < 2 \\ 2/6 & : 2 \leq x < 3 \\ 3/6 & : 3 \leq x < 4 \\ 4/6 & : 4 \leq x < 5 \\ 5/6 & : 5 \leq x < 6 \\ 1 & : x \geq 6. \end{cases}$$

❖ נשים לב שתוצאת הפונקציה מצטברת ל-1

תוחלת (Expected Value) וממוצע

התוחלת מייצגת תוצאה "צפויה" (Expected) של ניסוי זהה החוזר על עצמו פעמים רבות.

- עבור משתנה מקרי בדיד:

$$\mu = E[x] = \sum_{x \in \Omega} pr(X = x) \cdot x$$

- עבור משתנה מקרי רציף: $\mu = \int x\, f(x)\, dx$

דוגמה – קובייה הוגנת – שאלת סקר

שאלה – מהי התוחלת (Expected Value) של קובייה הוגנת?

תזכורת, עבור משתנה מקרי בדיד:

$$\mu = E[x] = \sum_{x \in \Omega} pr(X = x) \cdot x$$

$$\sum_{X \in \Omega} f(x) = 1$$

$$\text{pdf} = f_{x \in \Omega}(x) = pr_{x \in \Omega}(X = x)$$

תשובות אפשריות:

א – 1

ב – 0.5

ג – 3.5

ד – 21

תוחלת (Expected Value)

התוחלת מייצגת תוצאה "צפויה" (Expected) של ניסוי זהה החוזר על עצמו פעמים רבות.

- עבור משתנה מקרי בדיד:

$$\mu = E\left[x\right] = \sum_{x \in \Omega} pr(X = x) \cdot x$$

- דוגמה – קובייה הוגנת

$$\mathrm{E}(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

- עבור משתנה מקרי רציף:

$$\mu = \int x\, f(x)\, dx$$

תוחלת (Expected Value)

התוחלת מייצגת תוצאה "צפויה" (Expected) של ניסוי זהה החוזר על עצמו פעמים רבות.

- עבור משתנה מקרי בדיד:

$$\mu = E[x] = \sum_{x \in \Omega} pr(X = x) \cdot x$$

- אם האוכלוסייה בגודל N, התוחלת שווה לממוצע באוכלוסייה:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

שונות (Variance)

שונות - מדד לפיזור ערכים באוכלוסייה נתונה ביחס לתוחלת שלה.

- עבור משתנה מקרי בדיד:

$$\mathrm{Var}(X) = \mathbb{E}((X-\mu)^2) = \mathbb{E}(X^2) - \mu^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

- אם האוכלוסייה בגודל N:

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \mu^2$$

- עבור משתנה מקרי רציף:

$$\mathrm{Var}(X) = \sigma^2 = \int (x-\mu)^2 f(x)\,dx = \int x^2 f(x)\,dx - \mu^2$$

סטיית תקן (standard deviation)

סטיית תקן – שורש השונות.

$$\sigma = \sqrt{\mathrm{E}[(X-\mu)^2]}$$

- אם האוכלוסייה בגודל N:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

התפלגות בדידה אחידה

התפלגות אחידה בדידה:

למשל עבור משתנה מקרי X שמייצג "פסים", נניח שיש הסתברות זהה לערכים הבאים:

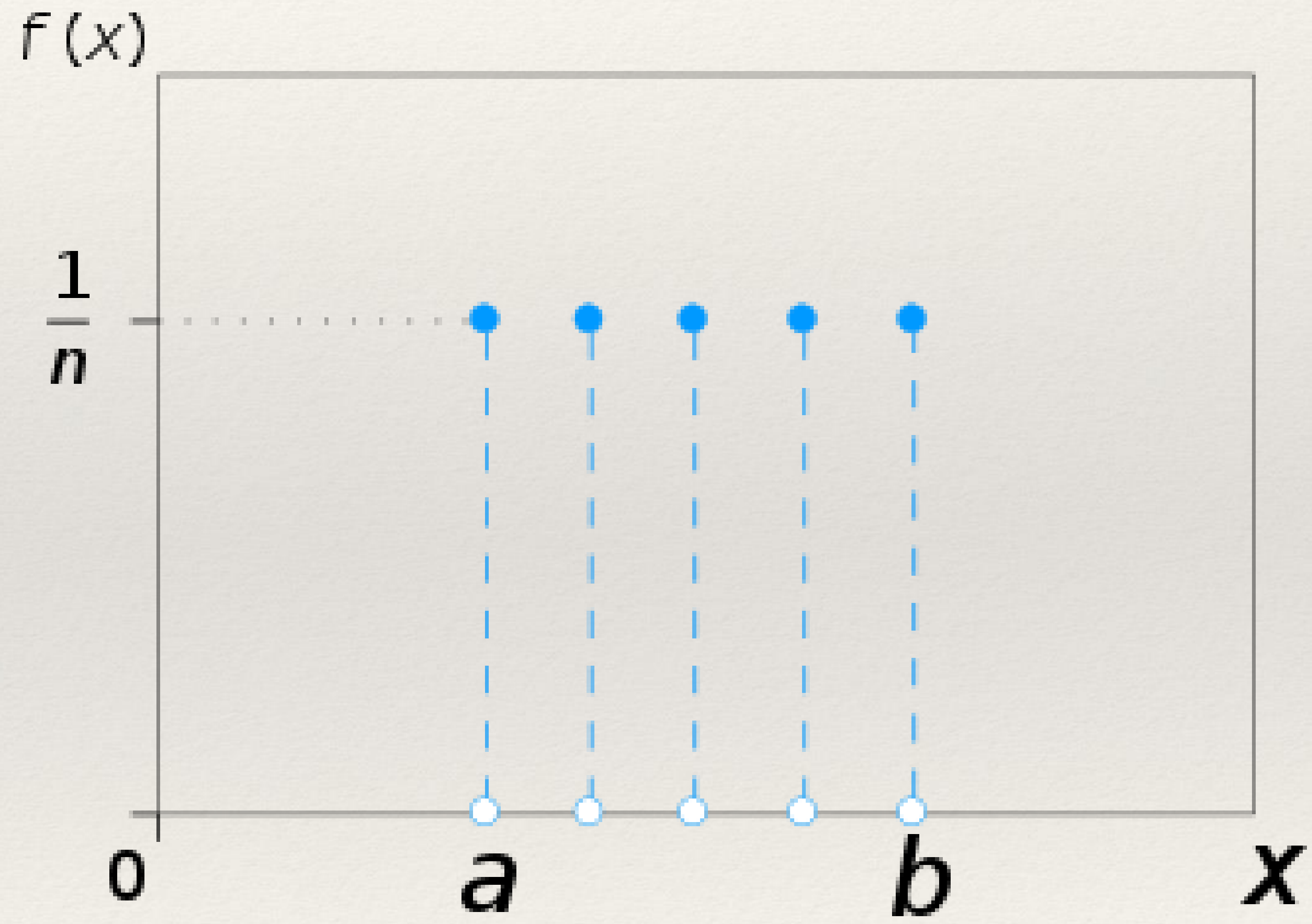
horizontal (אופקי), vertical (אנכי), none (ללא פסים) ונניח שאין ערכים נוספים.

p(X=vertical)=

p(X=horizontal)=

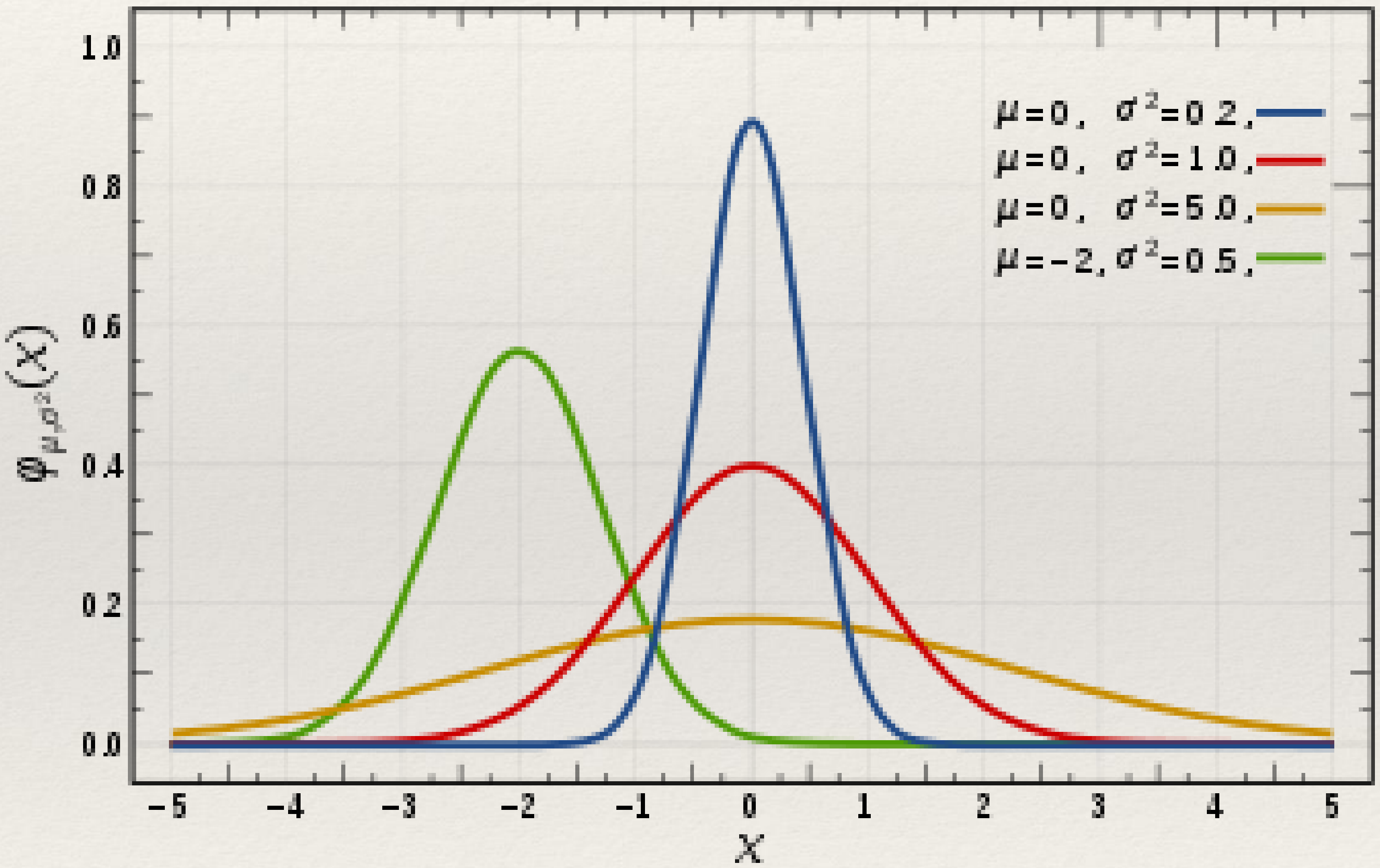
p(X=none)=

1/3



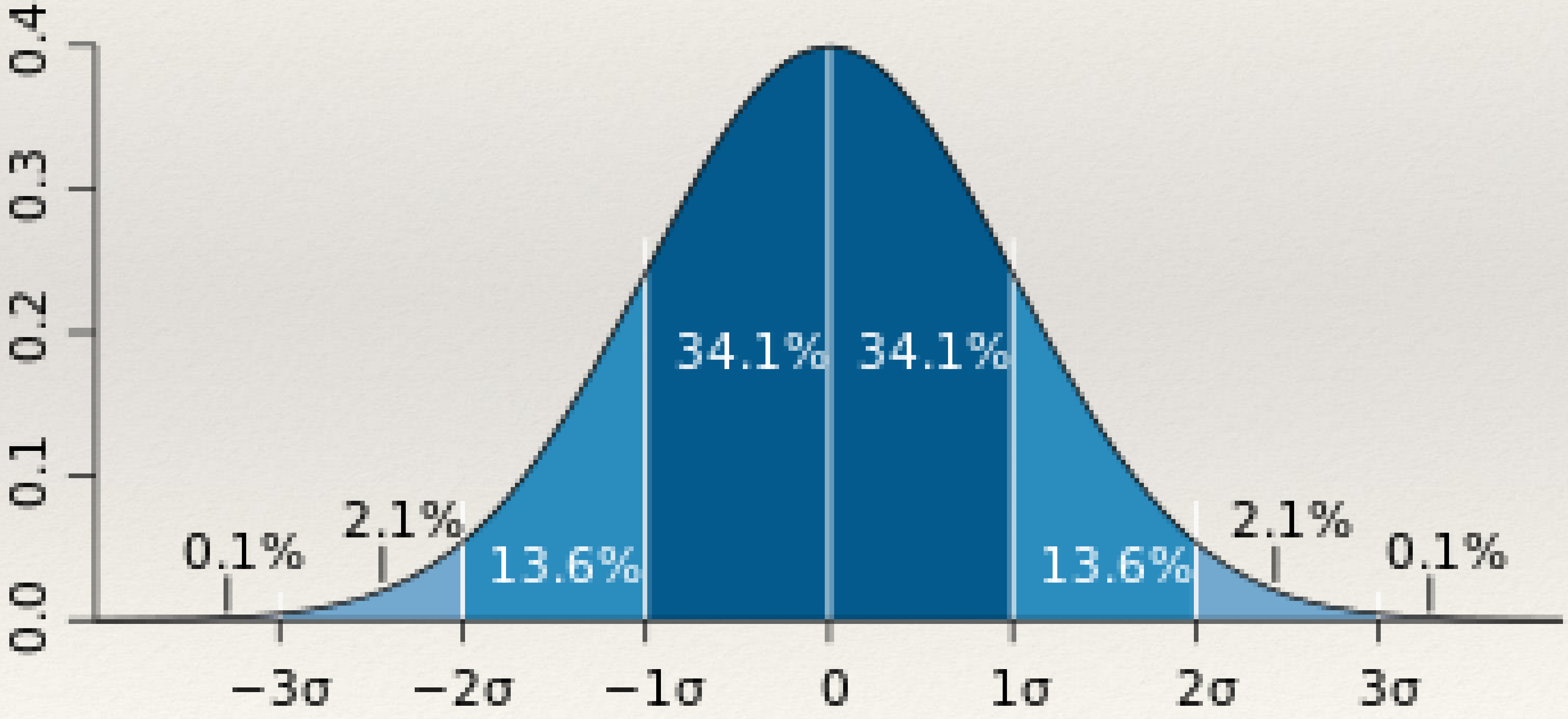
התפלגות נורמלית

התפלגות נורמלית:



התפלגות z – ההתפלגות הנורמלית הסטנדרטית

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



מדגם (sample)

(ויקיפדיה) - מדגם הוא קבוצת פרטים, המהווה מודל לאוכלוסייה, שאליה היא שייכת.

- הפרטים במדגם עשויים להיות בני אדם מאוכלוסייה אנושית כלשהי, בעלי חיים ואפילו עצמים דוממים
- למשל, מדגם של גפרורים נבדק כדי לאמוד את אחוז הגפרורים שלא נדלקים, מכלל אוכלוסיית הגפרורים המיוצרת במפעל מסוים).

- השאיפה – מדגם שמייצג את האוכלוסייה

התפלגות במדגם – ממוצע וסטיית תקן

מדגם הוא קבוצת פרטים, המהווה מודל לאוכלוסייה, שאליה היא שייכת.

- במקרה שלנו – המדגם הוא ה- train set (נרחיב עוד לגבי ה-train-set בהמשך הקורס)
- מאורע במדגם, עבור משתמה מקרי – מקביל לערך מאפיין עבור דוגמה מסוימת, ב-training-set.

מדדים סטטיסטים במדגם:

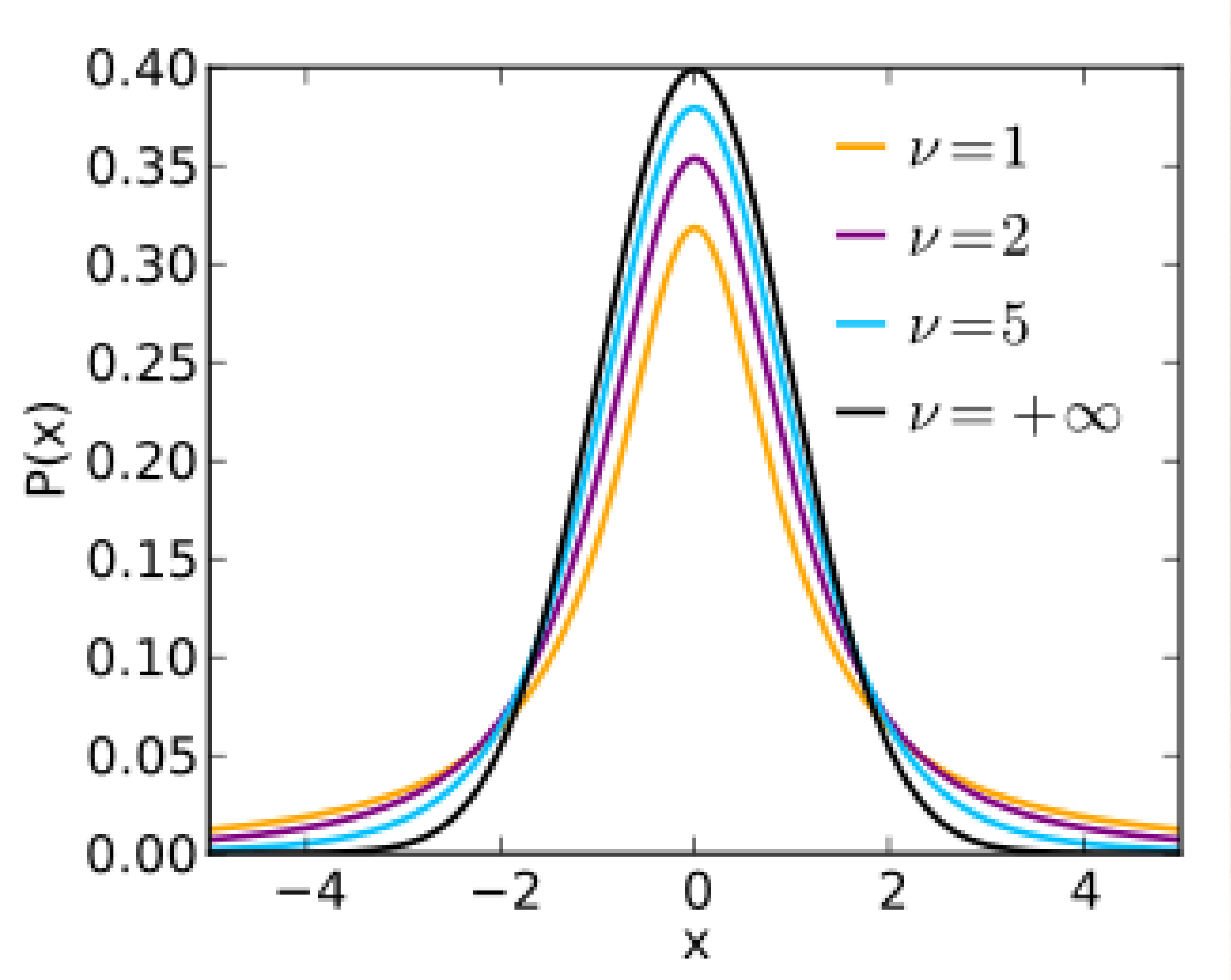
- נהוג לסמן ממוצע במדגם ע"י $\bar{x}$
- סטיית התקן במדגם: $$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
- שימו לב, שבשונות במדגם מחלקים בn-1 (n – מספר הדוגמאות במדגם)

התפלגות t

התפלגות t – התפלגות המבוססת על מידע שנאסף במדגם.

- התפלגות t שואפת להתפלגות z, כאשר גודל המדגם שואף לאינסוף
- בפועל מתייחסים לערכים הרבה יותר קטנים

- בעזרת התפלגות t – נדמה כל התפלגות לנורמלית
- בעזרת התפלגות t – נשווה בעצם את הסולם של מרחב המאפיינים (feature set)



התפלגות t

בסטטיסטיקה – כל התפלגות במדגם (או ב-training-set) ניתן להפוך להתפלגות t, אם ידועות הממוצע וסטיית התקן (במדגם).

- ממוצע במדגם $\bar{x}$ סטיית התקן במדגם: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

שימו לב, שבשונות במדגם מחלקים ב-n-1 (n – מספר הדוגמאות במדגם)

שאלות ביניים

שאלה 1: מהם השונות וסטיית התקן? מה הם באים למדוד?

שאלה 2: מה ההבדל בין סטיית תקן באוכלוסייה וסטיית תקן במדגם?

שאלה 3: מהם התפלגות z והתפלגות t, ומה ההבדל ביניהם?

חזרה לדוגמאות המשמשות ללמידה

מושגים:

- data, dataset
- feature set
- feature vector

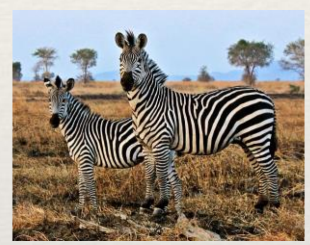
data vs. dataset

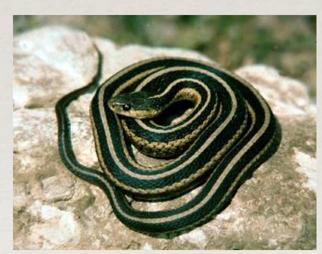
Data – תצפיות או מדידות (גולמיות או מעובדות).

- למשל, כל אחת מהתמונות מתייחסת לתצפית בודדת.





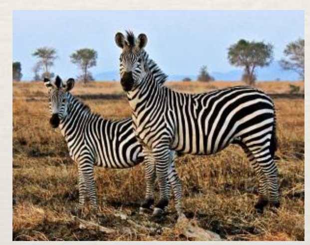



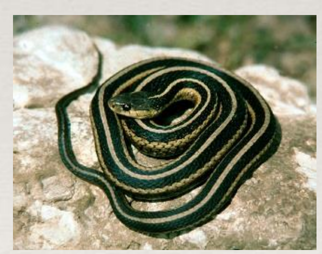



The feature set





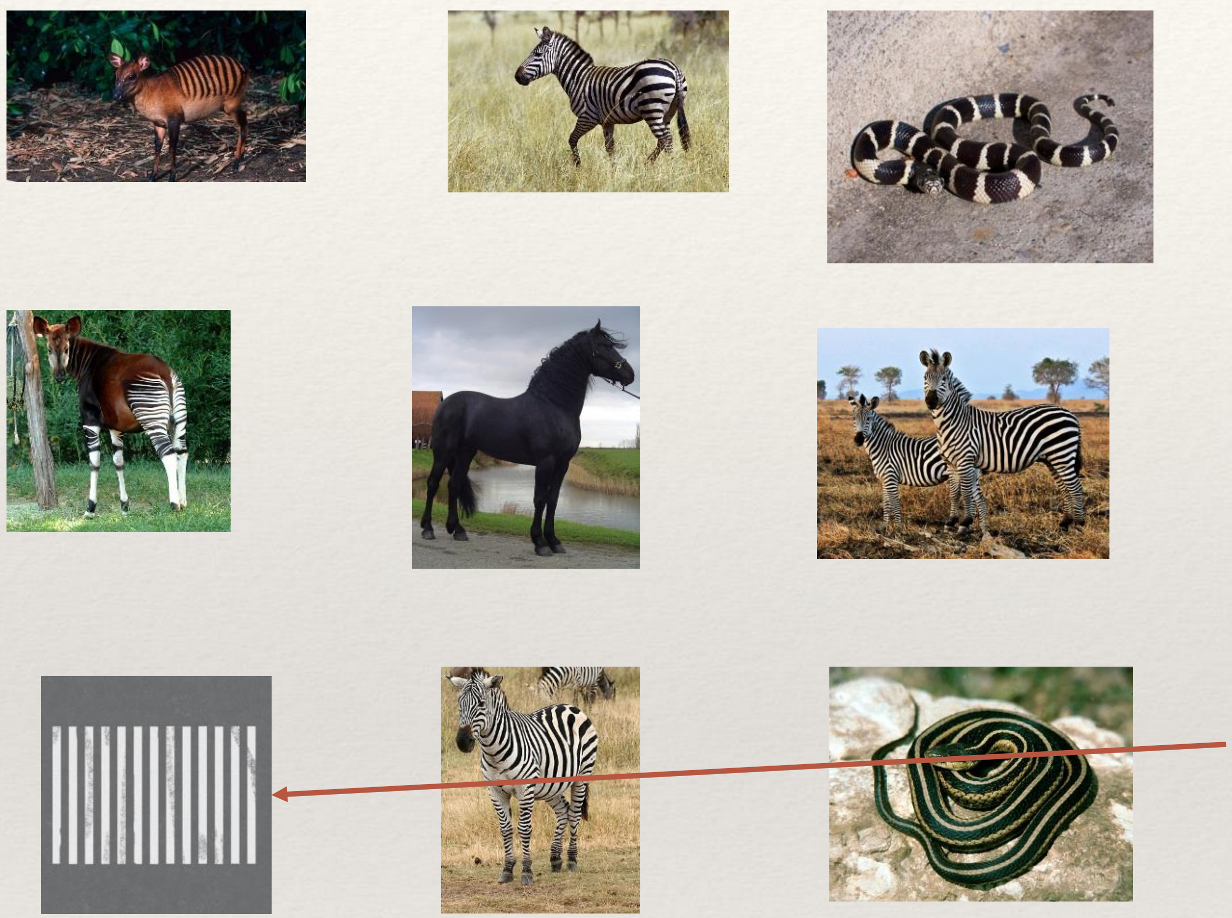



Feature set (אוספת המאפיינים) – את הדוגמאות נייצג על ידי קבוצה סופית של מאפיינים.

- ניתן להניח שישנם עוד תכונות רבות עבור כל דוגמה
- בדוגמה שלנו ה- feature set: "האובייקט הוא חיה?", "פסים אנכיים?", "צבעי שחור-לבן?", "בעלת 4 רגליים?", "חיה גדולה?"

data vs. dataset



Dataset– אסופת תצפיות או מדידות מעובדות (processed).

- Instance - דוגמה (example) מעובדת, בה נקבע את ערכי המאפיינים (features) מבין המאפיינים ב-feature set.
  - למשל: "פסים אנכיים" = "כן"
- Feature Vector –ייצוג וקטורי של הדוגמה המעובדת, ע"י הערכים של המאפיינים בווקטור
- שימו לב – מאפיין (feature) – הוא בעצם משתנה מקרי
  - ה-dataset – ממנו נוציא את הנתונים הוא בעצם מדגם סטטיסטי.

Feature Vectors

<u>זברות</u>

<u>מאפייני בעיית הזברה (Feature set):</u>

- חיה: (כן, לא)
- פסים: (אנכיים, אופקיים, ללא)
- צבעים: (שחור, לבן, חום, ...)
- רגליים: (2, 4, ללא)
- גודל החיה: (גדולה, בינונית, קטנה)

<u>דוגמה ל- feature vectors (בעלת ערכי 0,1 בלבד):</u>

is_animal	vertical_stripes	black_&_white	4_legs	large
0	1	1	0	0
1	0	1	0	0
1	0	0	1	1
1	1	1	1	1

שאלות

שאלה 1: מהו feature set?

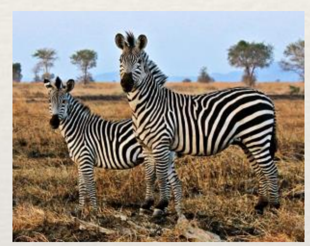
שאלה 2: מהו feature vector?

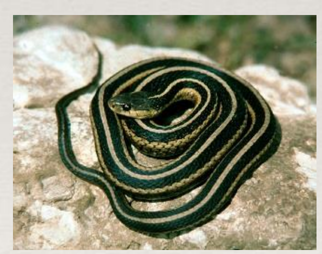
שאלה 3: מה הקשר בין משתנה מקרי ל-feature set?

The feature set





Feature set (אוספת המאפיינים) – את הדוגמאות נייצג על ידי קבוצה סופית של מאפיינים.

- ניתן להניח שישנם עוד תכונות רבות עבור כל דוגמה
- בדוגמה שלנו ה- feature set: "האובייקט הוא חיה?", "פסים אנכיים?", "צבעי שחור-לבן?", "בעלת 4 רגליים?", "חיה גדולה?"

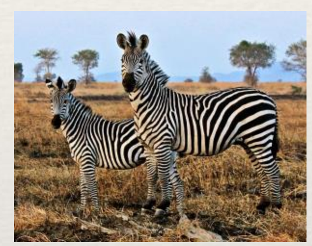
data vs. dataset

Dataset– אסופת תצפיות או מדידות מעובדות (processed).

- Feature Vector - דוגמה (example) מעובדת, בה נקבע את ערכי המאפיינים (features) מבין המאפיינים ב-feature set.
    - למשל: "פסים אנכיים" = "כן"



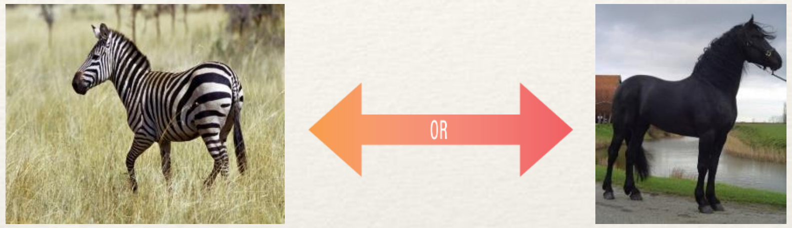



- Feature Vector –יצוג וקטורי של הדוגמה המעובדת, ע"י הערכים של המאפיינים בוקטור
- שימו לב – מאפיין (feature) – הוא בעצם משתנה מקרי
- ה-dataset – ממנו נוציא את הנתונים הוא בעצם מדגם סטטיסטי.

Feature Vectors

<u>זברות</u>



<u>מאפייני בעיית הזברה (Feature set):</u>

- חיה: (כן, לא)
- פסים: (אנכיים, אופקיים, ללא)
- צבעים: (שחור, לבן, חום, ...)
- רגליים: (2, 4, ללא)
- גודל החיה: (גדולה, בינונית, קטנה)

<u>דוגמה ל- feature vectors (בעלת ערכי 0,1 בלבד):</u>

is_animal	vertical_stripes	black_&_white	4_legs	large
0	1	1	0	0
1	0	1	0	0
1	0	0	1	1
1	1	1	1	1

The Iris Dataset

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

- נייצג כל דוגמה 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>

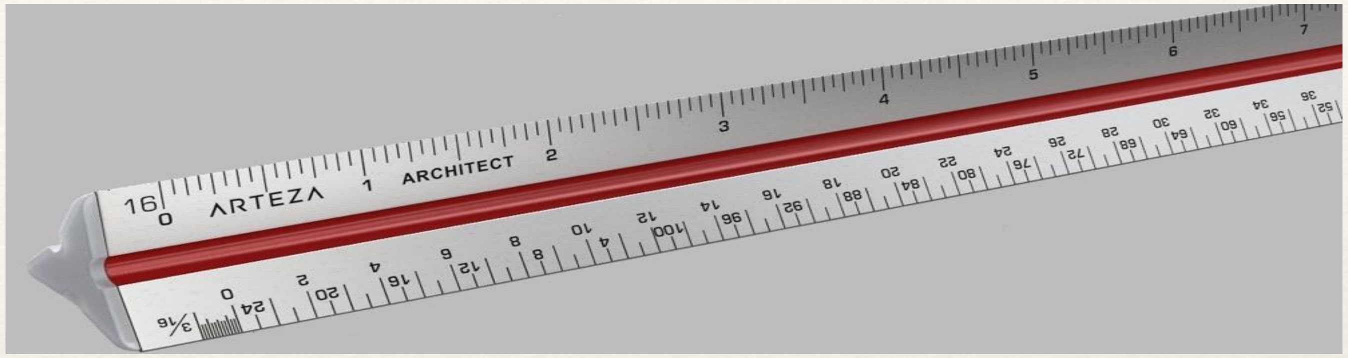
Sepal length	Sepal width	Petal length	Petal width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2

עיבוד מידע (data processing)
- סילום (Scaling)

מושגים:

- סילום (scaling)
- t-distribution standardization
- minmax normalization

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

- ▪ הופכים את ההתפלגות להתפלגות t

<u>minmax normalization</u> – הסילום מתבצע כך שערכי המשתנה יהיו בין מינימום למקסימום חדשים.

- [0,1] - המינימום והמקסימום החדשים, הינם 0 ו-1 בהתאמה
- [-1,1] - המינימום והמקסימום החדשים, הינם -1 ו-1 בהתאמה

מדוע משתמשים בסילום?

בעיקר כדי לא לתת עדיפות למאפיין אחד, על פני האחר, בגלל סולם ערכים שונה (פרטים נוספים בהמשך)

t-distribution standardization - סילום

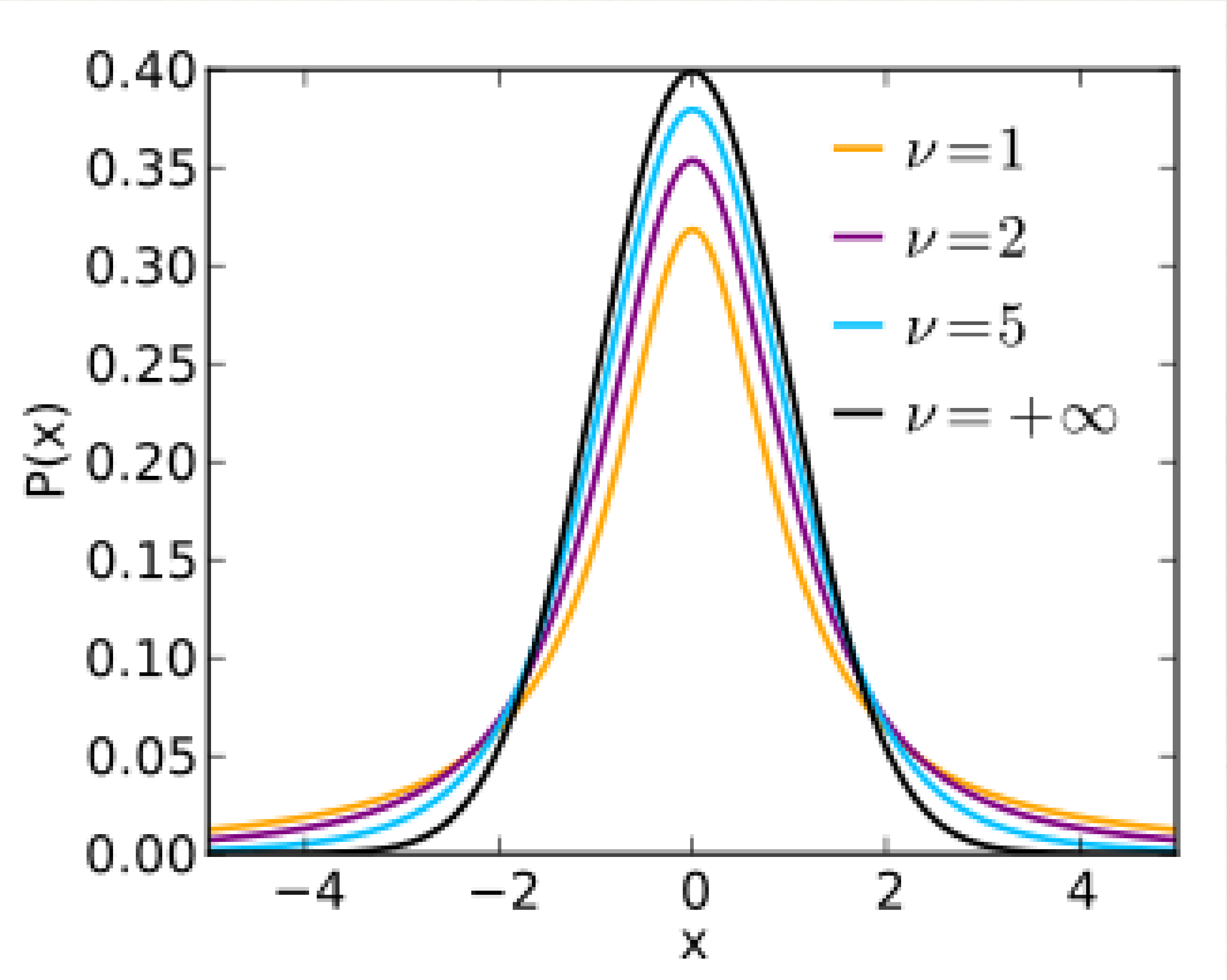
התפלגות t – שואפת להתפלגות z, כאשר גודל המדגם שואף לאינסוף

- 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 ומחלקים בסטיית התקן במדגם (s)

- בדומה לתחומים מדעיים שונים, הקירוב לנורמלי, בעזרת התפלגות t, מאוד מקובל (נרחיב בהמשך)



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]

- בלמידת מכונה בכלל, ולאלגוריתמי למידה מסוימים בפרט, ישנו יתרון, ולעיתים אף צורך במעבר לטווחים אלו.

שאלות

שאלה 1: מהו feature set? ומהו feature vector?

שאלה 2: את מה משנה פעולת הסילום?

שאלה 3: איך מחשבים Standardization t-distribution?

שאלה 4: מה מבצע Minmax normalization?