תרגול מספר 5 –
מבוא ללמידה מודרכת (Supervised)
ו-KNN

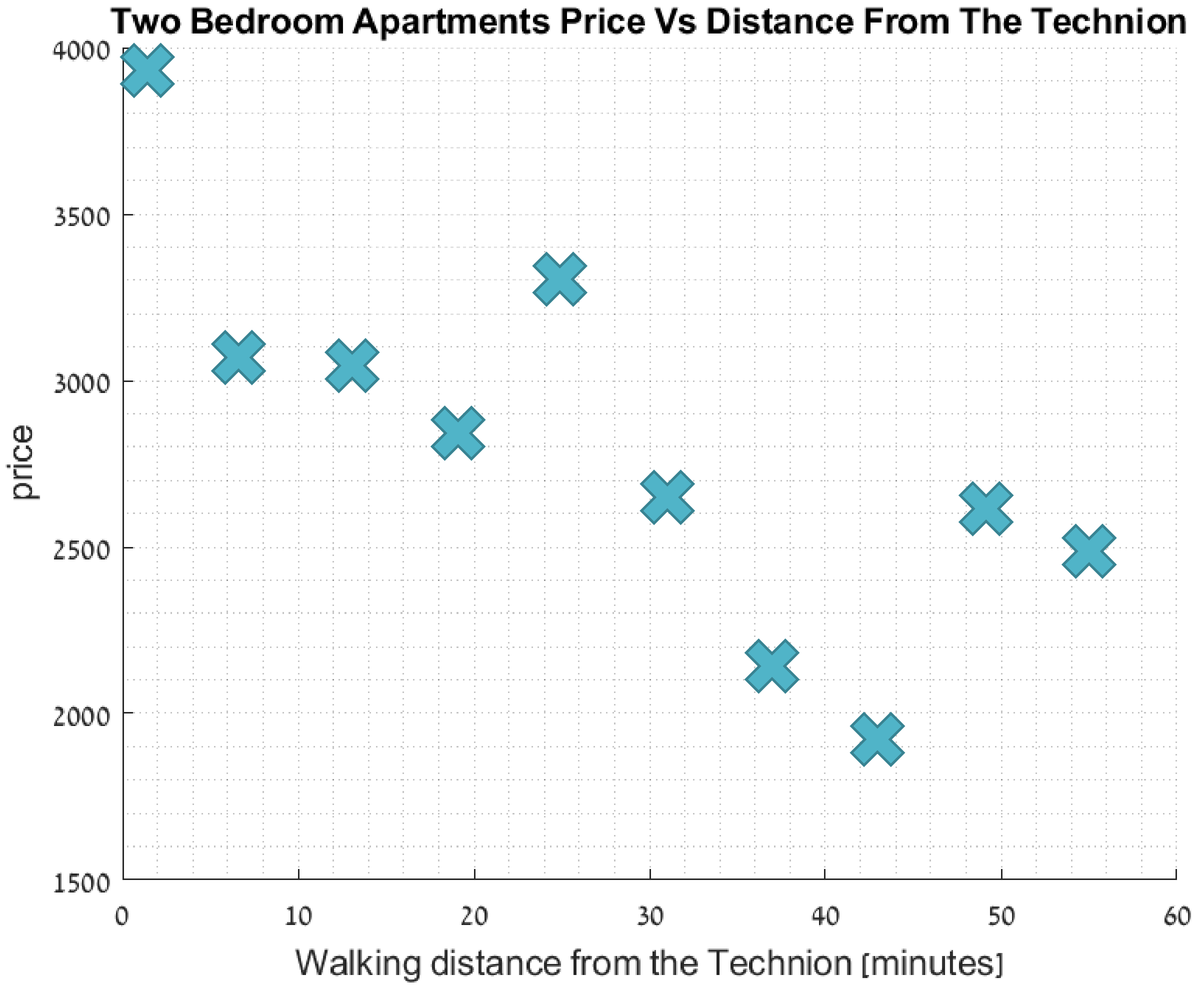
סימונים ותקציר תאוריה

- $D=\{x_i,y_i\}_{i=1}^{n}$ - דוגמאות (דגימות - samples) בלתי תלויות של X עם פרדיקציה ידועה Y.
- X - מרחב הקלט. $x\in X$.
- המטרה היא ללמוד מיפוי (פונקציה, אלגוריתם) f אשר מאפשר חישוב פלט מתאים

  $y=f(x)$ **לכל** קלט אפשרי $x\in X$. למיפוי הנלמד f נקרא "**פונקציית החיזוי**" או "**חזאי**".
- $\mathcal{F}$ - מרחב המיפויים האפשריים. למשל – מיפויים ליניאריים $\mathcal{F}=\{\mathrm{f}(\mathbf{x})=\mathbf{w}^{\mathrm{T}}\mathbf{x}+\mathrm{b}|\forall\mathbf{w}\in\mathbb{R}^{\mathrm{d}},\mathrm{b}\in\mathbb{R}\}$
- מרחב הפלט תלוי במשימה, נבחין בין שתי משימות שונות:

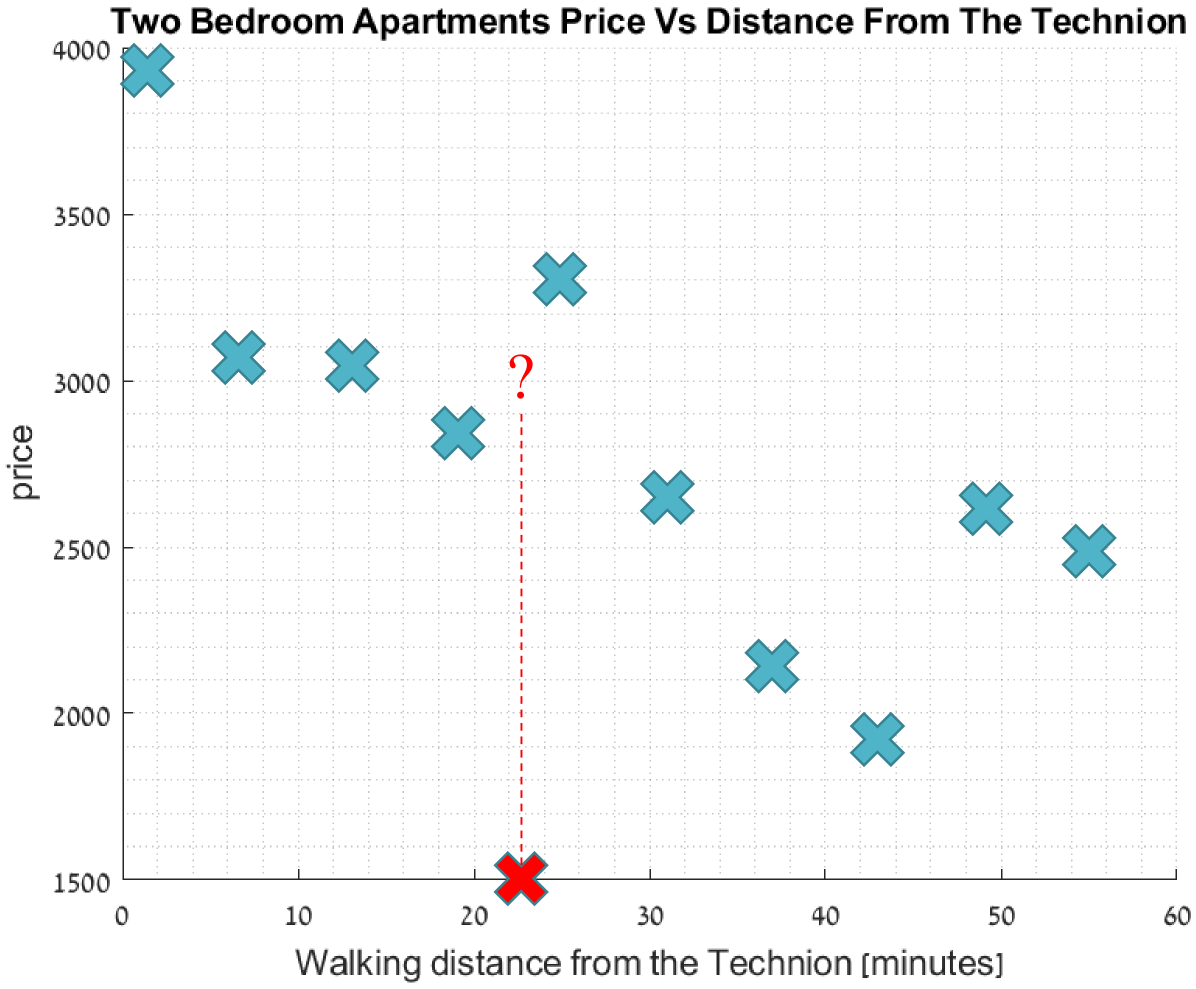
1. רגרסיה – Regression : הפלט y הינו מספר ממשי.
2. סיווג – Classification : הפלט y הינו משתנה קטגורי בעל מספר סופי של ערכים $y=\{1,\dots,K\}$.
   ערכי y נקראים גם מחלקות.

Regression



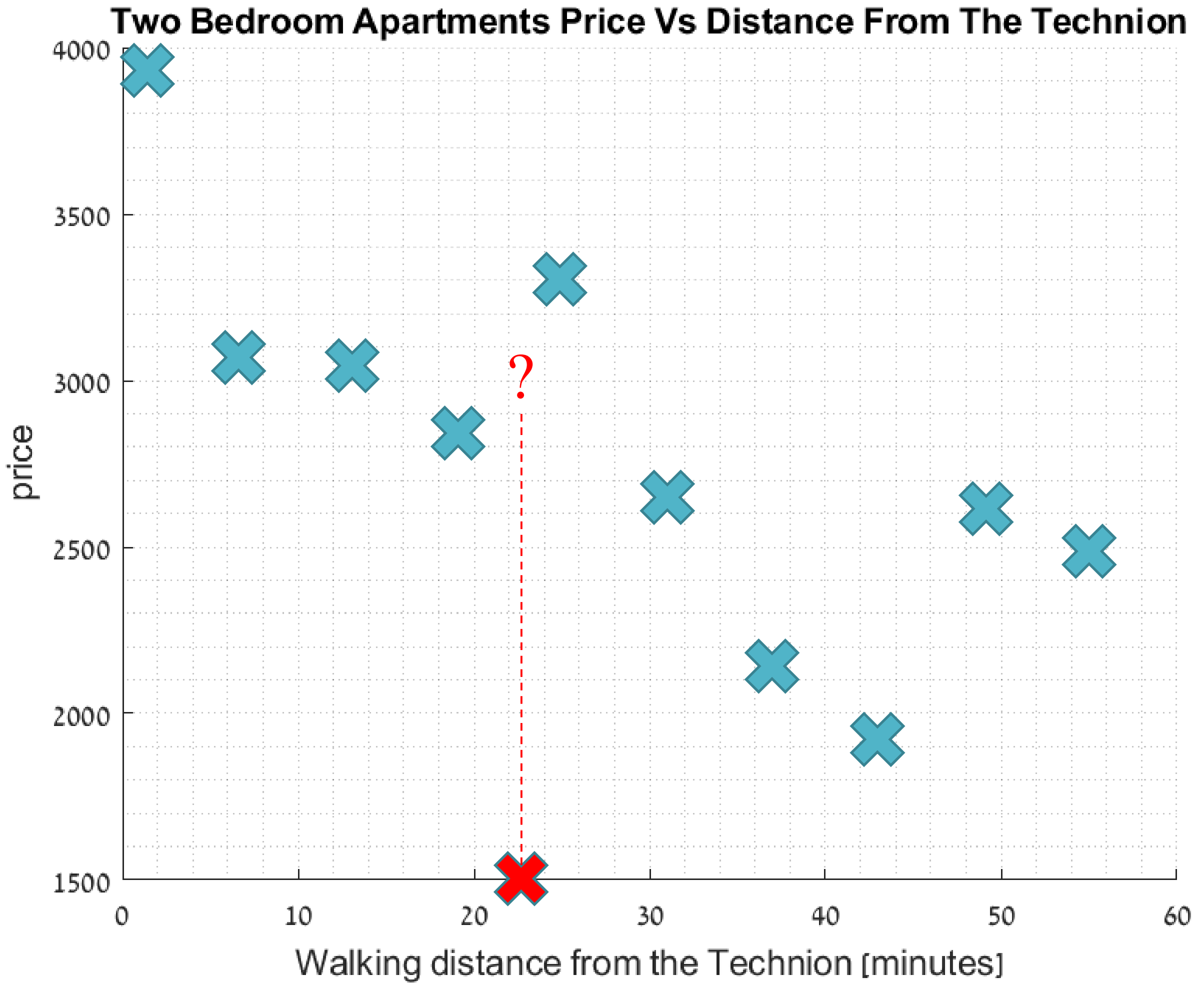
Two Bedroom Apartments Price Vs Distance From The Technion
4000
3500
3000
2500
2000
1500
price
0
10
20
30
40
50
60
Walking distance from the Technion [minutes]

Regression



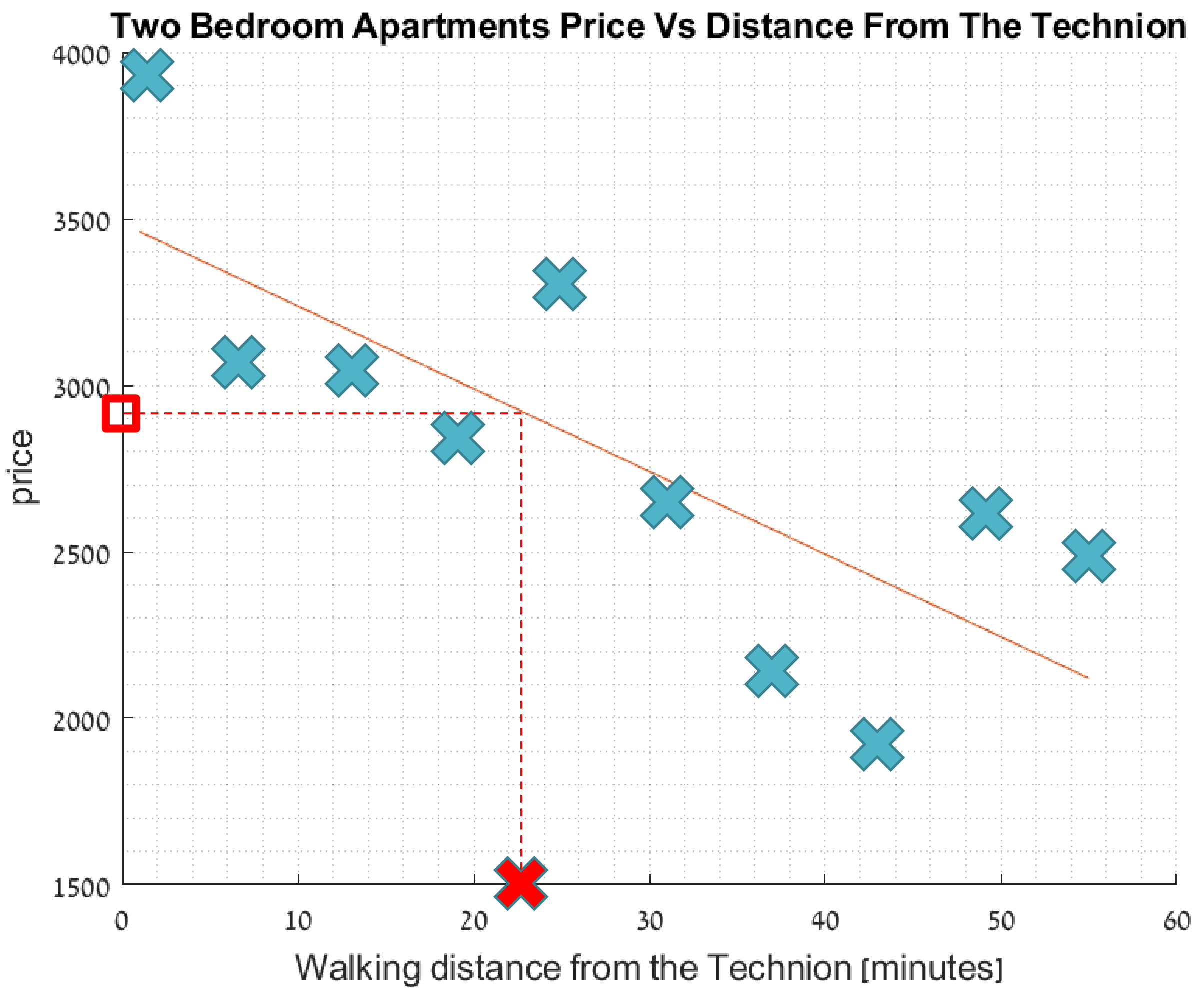
Two Bedroom Apartments Price Vs Distance From The Technion
4000
3500
3000
2500
2000
1500
price
?
0
10
20
30
40
50
60
Walking distance from the Technion [minutes]

Regression



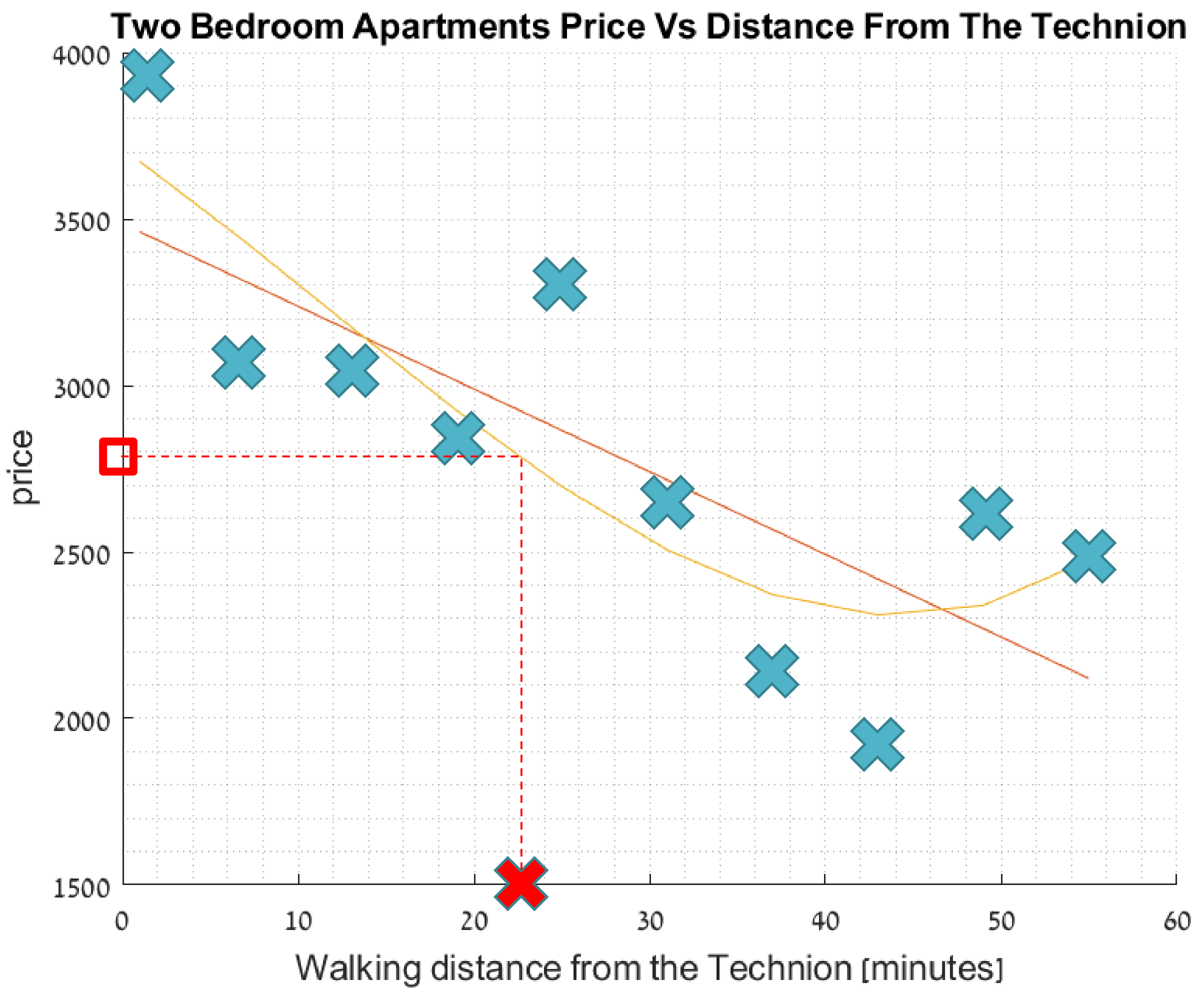
- ההבדל בין קירוב לחיזוי

Regression



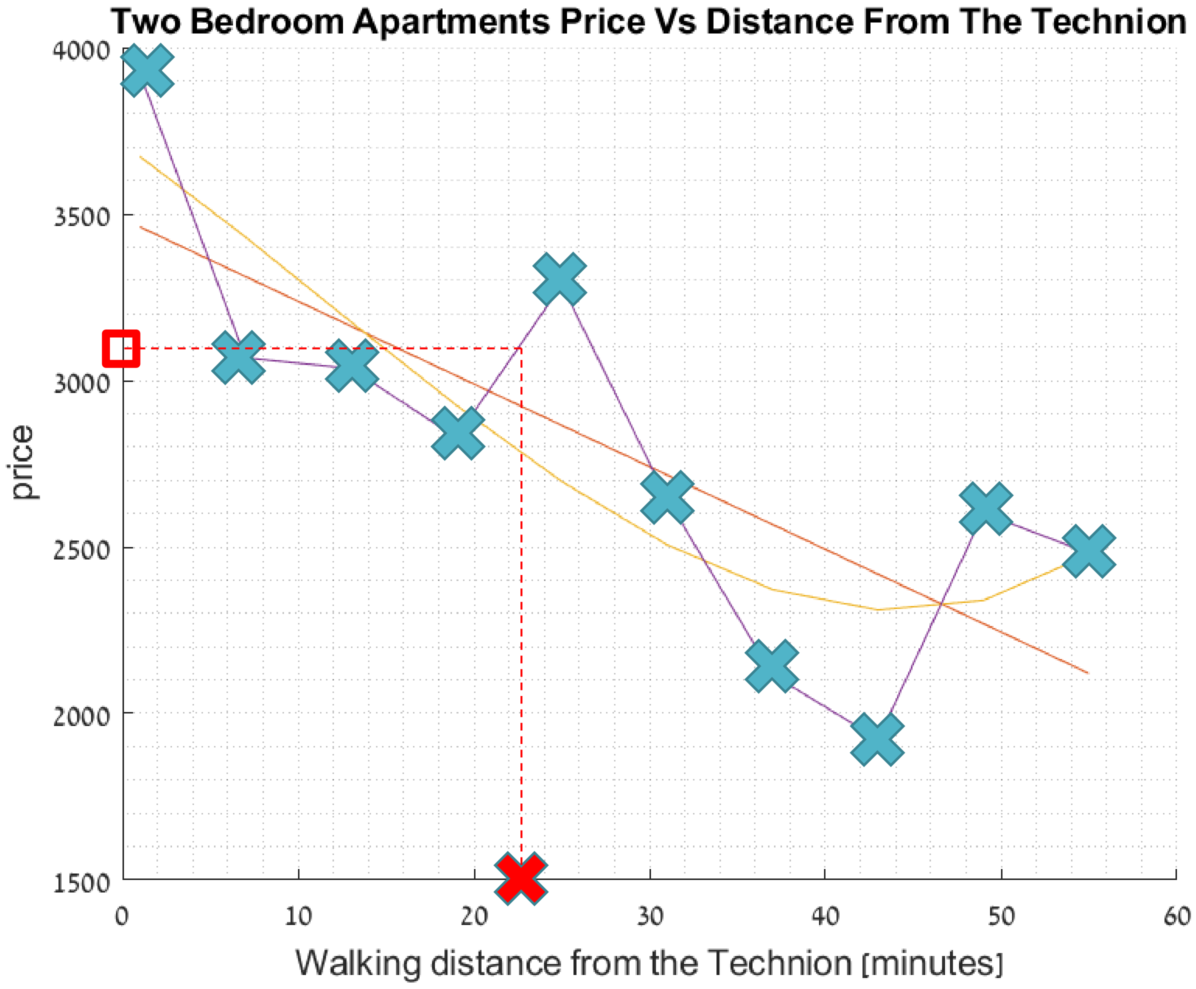
Two Bedroom Apartments Price Vs Distance From The Technion
4000
3500
3000
2500
2000
1500
price
0
10
20
30
40
50
60
Walking distance from the Technion [minutes]

Regression



Two Bedroom Apartments Price Vs Distance From The Technion
4000
3500
3000
2500
2000
1500
price
0
10
20
30
40
50
60
Walking distance from the Technion [minutes]

Regression


Two Bedroom Apartments Price Vs Distance From The Technion
4000
3500
3000
2500
2000
1500
price
0
10
20
30
40
50
60
Walking distance from the Technion [minutes]

בחירת סדר המודל

- למשל עבור מיפויים ליניאריים $\mathcal{F} = \{\mathrm{f}(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + \mathrm{b} | \forall \mathbf{w} \in \mathbb{R}^{\mathrm{d}}, \mathrm{b} \in \mathbb{R}\}$ בחירת הגודל $d + 1$.
- סדר המודל (למשל, מספר הפרמטרים במודל פרמטרי) קובע, כעיקרון, את גודל קבוצת הפונקציות הכלולות במודל. ככל שסדר המודל גדול יותר, כך קבוצה זו עשירה יותר, ומכילה פונקציות מסובכות יותר.
- סוגיה יסודית בבחירת סדר המודל הינה הניגוד הבא:
  - מודל פשוט מדי (בעל סדר נמוך) לא יאפשר תיאור מדויק של הקשר ה"אמיתי" בין הקלט לפלט.
  - מודל מסובך מדי (בעל סדר גבוה) הוא בעל מספר גדול של דרגות חופש, ולפיכך עלול לדרוש מספר רב של דוגמאות על מנת לבצע הכללה סבירה.

בחירת סדר המודל

פונקציית סיכון (risk):

$$R\left(\hat{f}\right) = E_{(\boldsymbol{x},y)\sim p_{X,Y}}\left(\ell(\hat{f}(\boldsymbol{x}), y)\right)$$

פונקציית סיכון אמפירית:

$$\hat{R}\left(\hat{f}\right) = \frac{1}{N}\sum_{ii} \ell(\hat{f}(\boldsymbol{x_i}), y_i)$$

כאשר, ℓ - פונקציית מחיר כלשהי.

למשל:

$$\ell\left(\hat{f}(\boldsymbol{x}), y\right) = \begin{cases} 0 & \hat{f}(\boldsymbol{x}) = y \\ 1 & otherwise \end{cases}$$

$$\ell\left(\hat{f}(\boldsymbol{x}), y\right) = \left(\hat{f}(\boldsymbol{x}) - y\ \right)^2$$

בחירת סדר המודל - Bias-Variance Tradeoff

- <u>פרוק השגיאה</u>:

$$R\left(\hat{f}\right) = E_{app}(F) + E_{est}\left(\hat{f}, F\right)$$

- כאשר:

$$E_{app}(F) = \min_{f \in F} R(f) \triangleq R(f^*)\ , with\ f^* \triangleq \underset{f \in F}{argmin}\ R(f)$$

$$E_{est}\left(\hat{f}, F\right) = R\left(\hat{f}\right) - R(f^*)$$

F – משפחת היפותזות , מודל

- $E_{app}(F)$ היא שגיאת הקירוב (approximation error), הנקראת גם ההטיה (bias). היא מציינת את הסיכון המינימלי שחזאי <u>כלשהו</u> מתוך המודל F יכול להשיג. גודל זה הינו דטרמיניסטי ואינו תלוי בסדרת הלימוד.
- $E_{est}\left(\hat{f},F\right)$ היא שגיאת השערוך (estimation error). גודל זה הינו הפרש הסיכון עבור החזאי האופטימאלי מתוך המודל – דהיינו f^* לבין החזאי שנבחר על ידי אלגוריתם הלימוד, דהיינו $\hat{f} = \hat{f}(D)$.

בחירת סדר המודל - Bias-Variance Tradeoff

- פרוק השגיאה:

$$R(\hat{f}) = E_{app}(F) + E_{est}(\hat{f}, F)$$

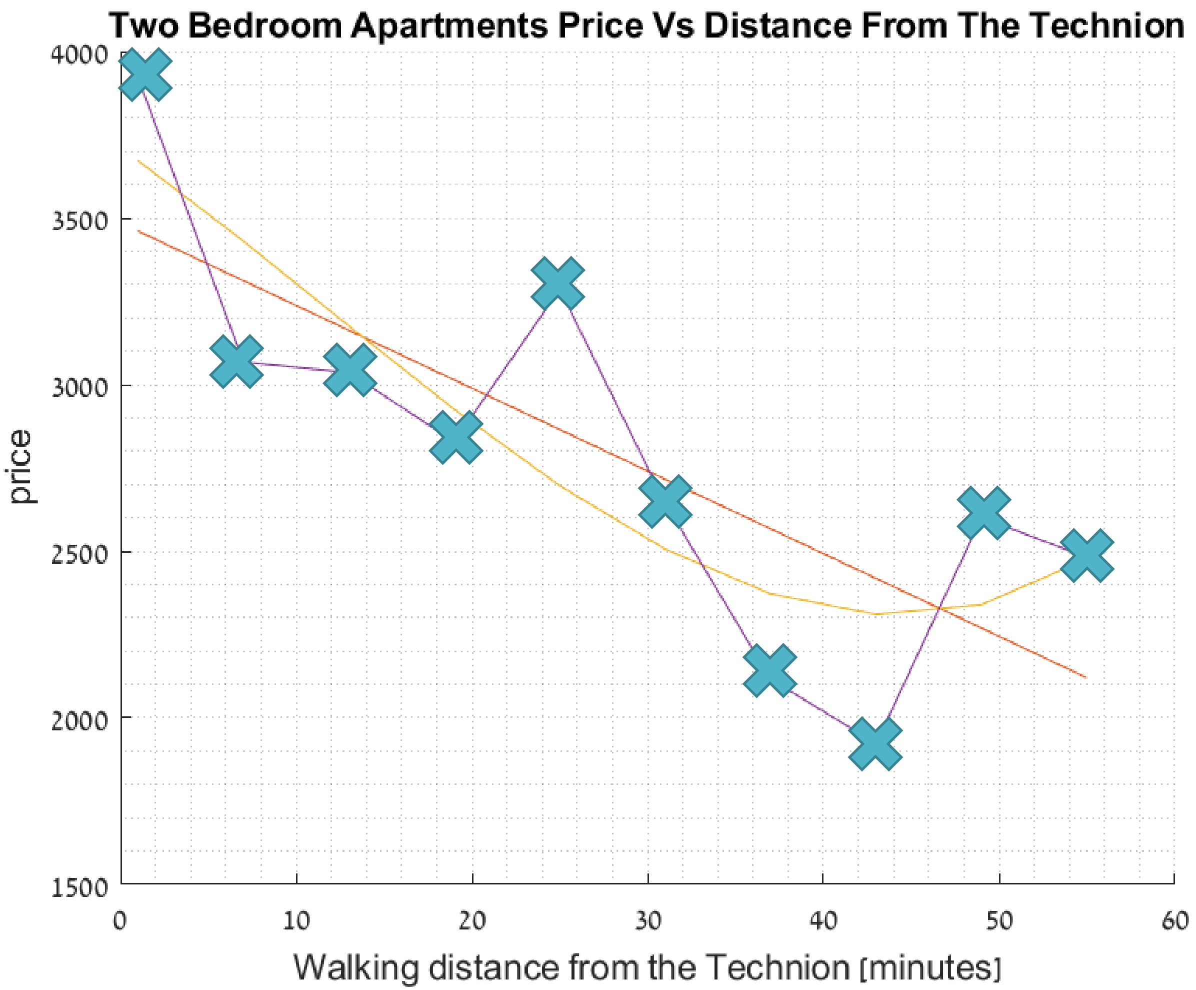
- כאשר:

$$E_{app}(F) = \min_{f \in F} R(f) \triangleq L(f^*) \text{ , with } f^* \triangleq \underset{f \in F}{\operatorname{argmin}} R(f)$$

$$E_{est}(\hat{f}, F) = R(\hat{f}) - R(f^*)$$

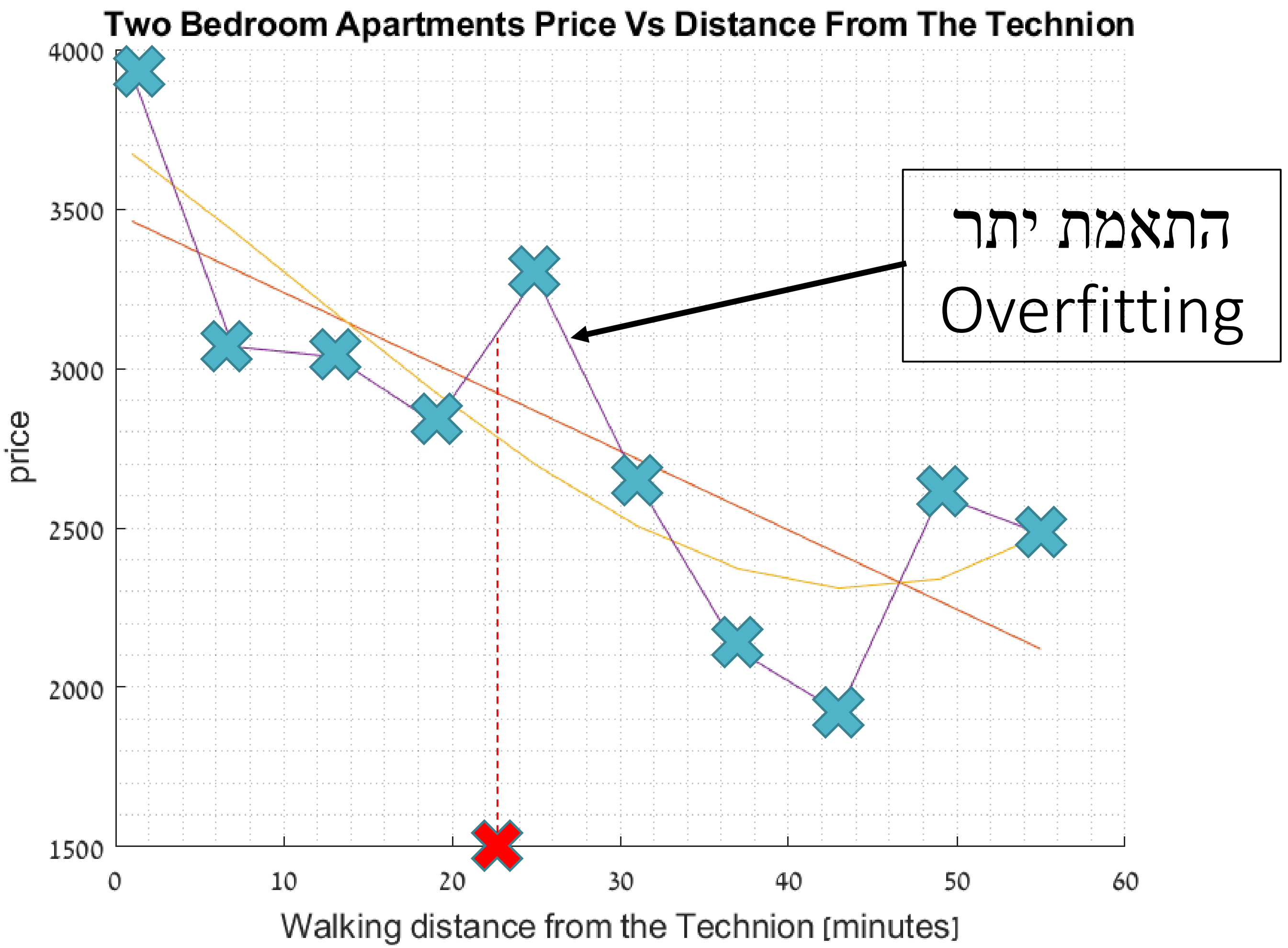
Bias גבוה	מודל פשוט מדי (בעל סדר נמוך) לא יאפשר תיאור מדויק של הקשר ה"אמיתי" בין הקלט לפלט.	$E_{app}(F)$ היא שגיאת הקירוב (approximation error), הנקראת גם ההטיה (bias). היא מציינת את הסיכון המינימלי שחזאי כלשהו מתוך המודל F יכול להשיג.
Variance גבוה	מודל מסובך מדי (בעל סדר גבוה) הוא בעל מספר גדול של דרגות חופש, ולפיכך עלול לדרוש מספר רב של דוגמאות על מנת לבצע הכללה סבירה.	$E_{est}(\hat{f},F)$ היא שגיאת השערוך (estimation error). גודל זה הינו הפרש הסיכון עבור החזאי האופטימאלי מתוך המודל – דהיינו f^* לבין החזאי שנבחר על ידי אלגוריתם הלימוד, דהיינו $\hat{f} = \hat{f}(D)$.

Regression



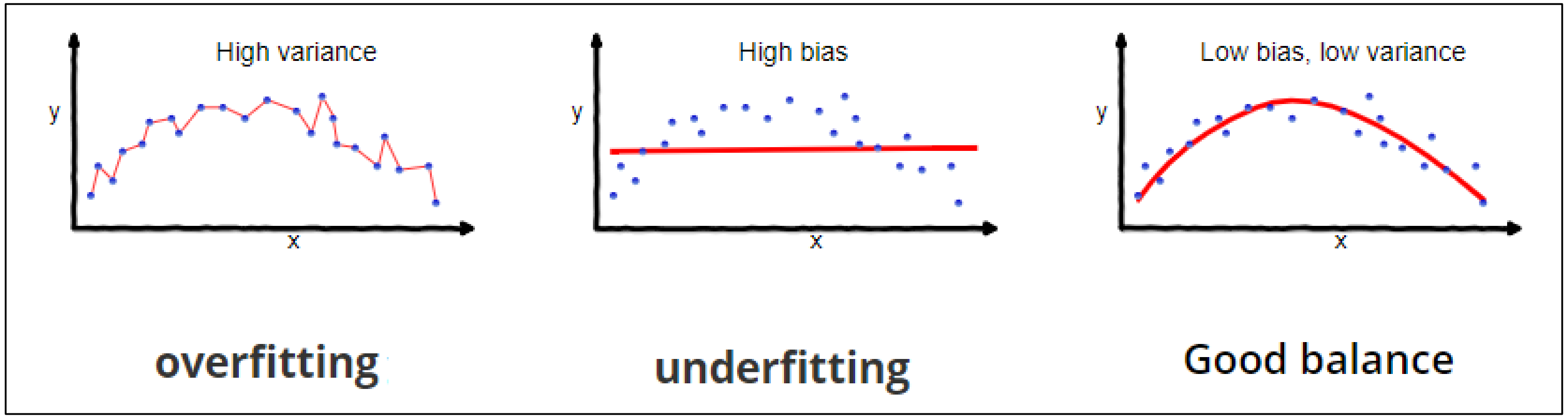
Bias גבוה	מודל פשוט מדי (בעל סדר נמוך) לא יאפשר תיאור מדויק של הקשר ה"אמיתי" בין הקלט לפלט.	$E_{app}(F)$ היא שגיאת הקירוב (approximation error), מציינת את הסיכון המינימלי שחזאי <u>כלשהו</u> מתוך המודל F יכול להשיג.
Variance גבוה	מודל מסובך מדי (בעל סדר גבוה) הוא בעל מספר גדול של דרגות חופש, ולפיכך עלול לדרוש מספר רב של דוגמאות על מנת לבצע הכללה סבירה.	$E_{est}(\hat{f},F)$ היא שגיאת השערוך (estimation error). גודל זה הינו הפרש הסיכון עבור החזאי האופטימאלי מתוך המודל לבין החזאי שנבחר על ידי אלגוריתם הלימוד, דהיינו $\hat{f}=\hat{f}(D)$.

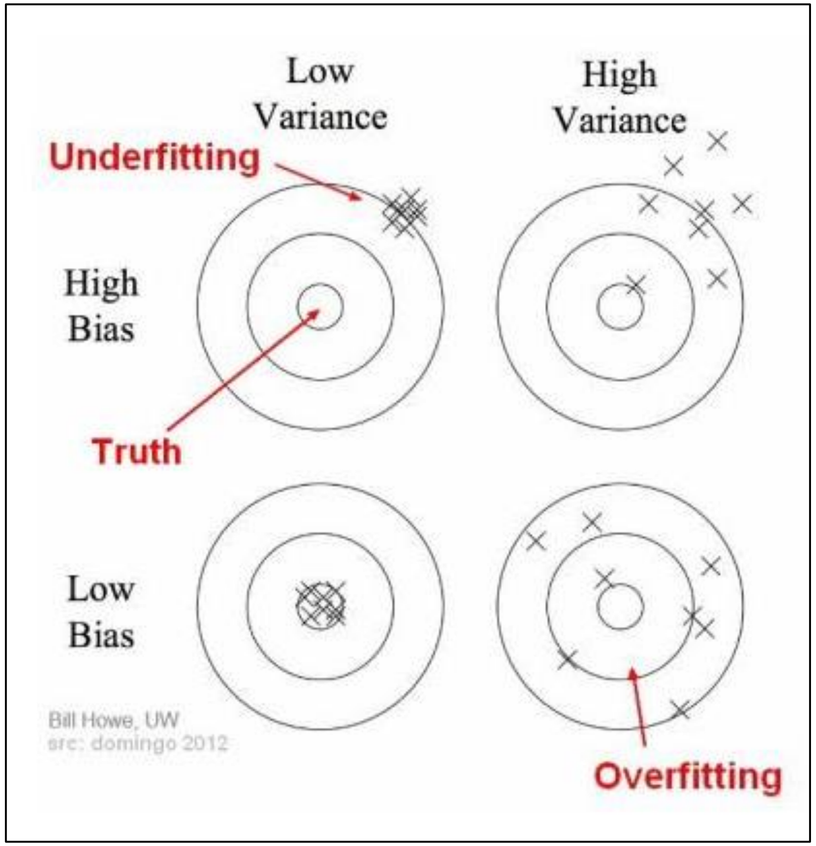
Regression



Bias גבוה	מודל פשוט מדי (בעל סדר נמוך) לא יאפשר תיאור מדויק של הקשר ה"אמיתי" בין הקלט לפלט.	$E_{app}(F)$ היא שגיאת הקירוב (approximation error), מציינת את הסיכון המינימלי שחזאי <u>כלשהו</u> מתוך המודל F יכול להשיג.
Variance גבוה	מודל מסובך מדי (בעל סדר גבוה) הוא בעל מספר גדול של דרגות חופש, ולפיכך עלול לדרוש מספר רב של דוגמאות על מנת לבצע הכללה סבירה.	$E_{est}(\hat{f},F)$ היא שגיאת השערוך (estimation error). גודל זה הינו הפרש הסיכון עבור החזאי האופטימאלי מתוך המודל לבין החזאי שנבחר על ידי אלגוריתם הלימוד, דהיינו $\hat{f} = \hat{f}(D)$.

Bias-Variance Tradeoff





Pictures taken from: https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

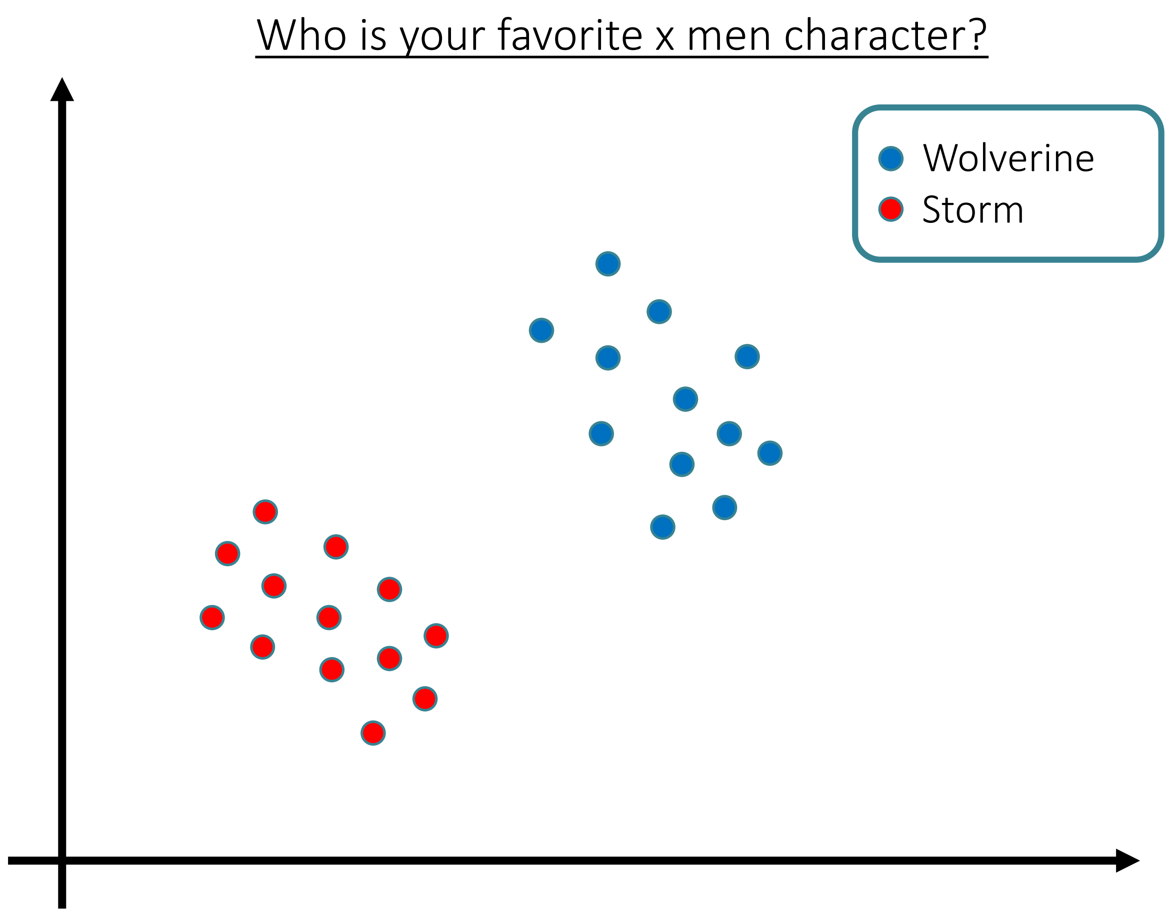
סימונים ותקציר תאוריה

- $D=\{x_i,y_i\}_{i=1}^{n}$ - דוגמאות (דגימות - samples) בלתי תלויות של X עם פרדיקציה ידועה Y.
- X - מרחב הקלט. $x\in X$.
- המטרה היא ללמוד מיפוי (פונקציה, אלגוריתם) f אשר מאפשר חישוב פלט מתאים

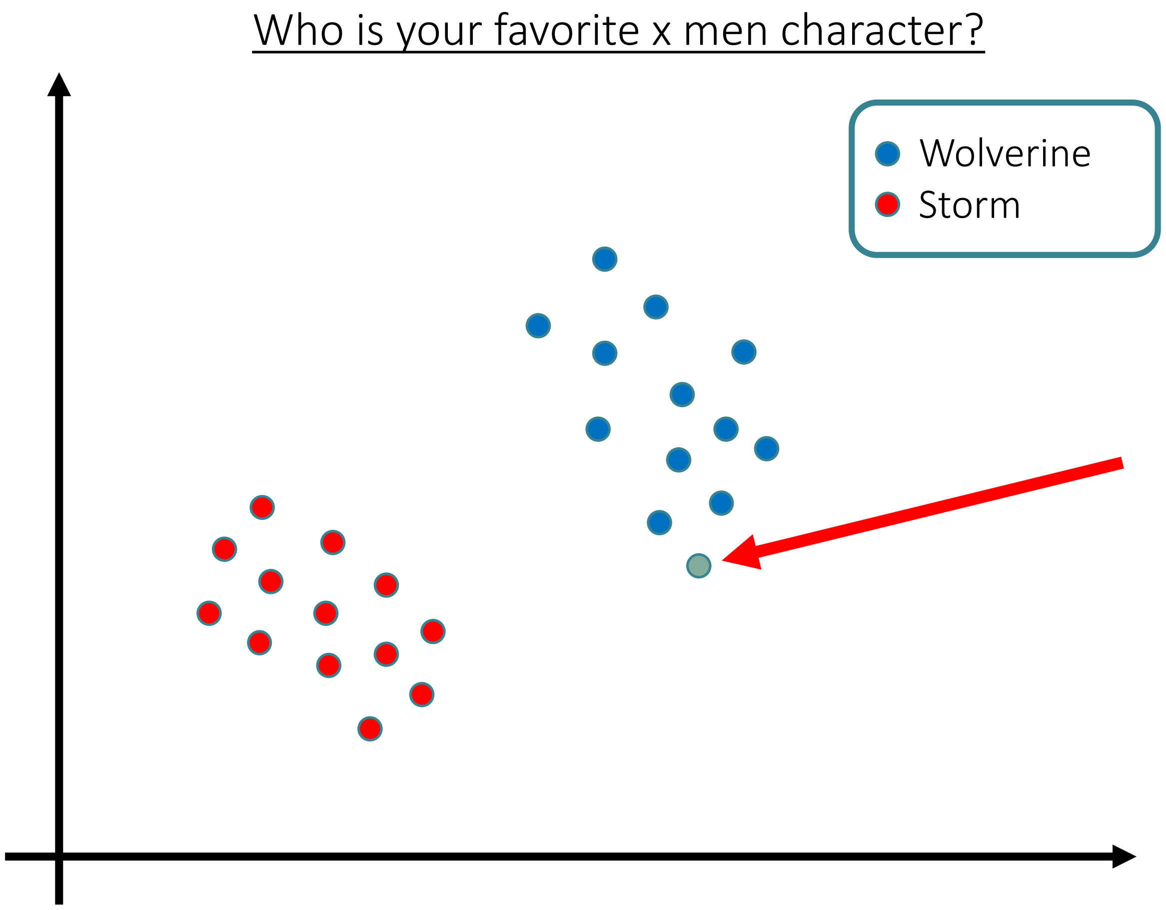
  $y=f(x)$ **לכל** קלט אפשרי $x\in X$. למיפוי הנלמד f נקרא "**פונקציית החיזוי**" או "**חזאי**".
- $\mathcal{F}$ - מרחב המיפויים האפשריים. למשל – מיפויים ליניאריים $\mathcal{F}=\{\mathrm{f}(\mathbf{x})=\mathbf{w}^{\mathrm{T}}\mathbf{x}+\mathrm{b}|\forall\mathbf{w}\in\mathbb{R}^{\mathrm{d}},\mathrm{b}\in\mathbb{R}\}$
- מרחב הפלט תלוי במשימה. נבחין בין שתי משימות שונות:

1. רגרסיה – Regression : הפלט y הינו מספר ממשי.
2. סיווג – Classification : הפלט y הינו משתנה קטגורי בעל מספר סופי של ערכים $y=\{1,\ldots,K\}$. ערכי y נקראים גם מחלקות.

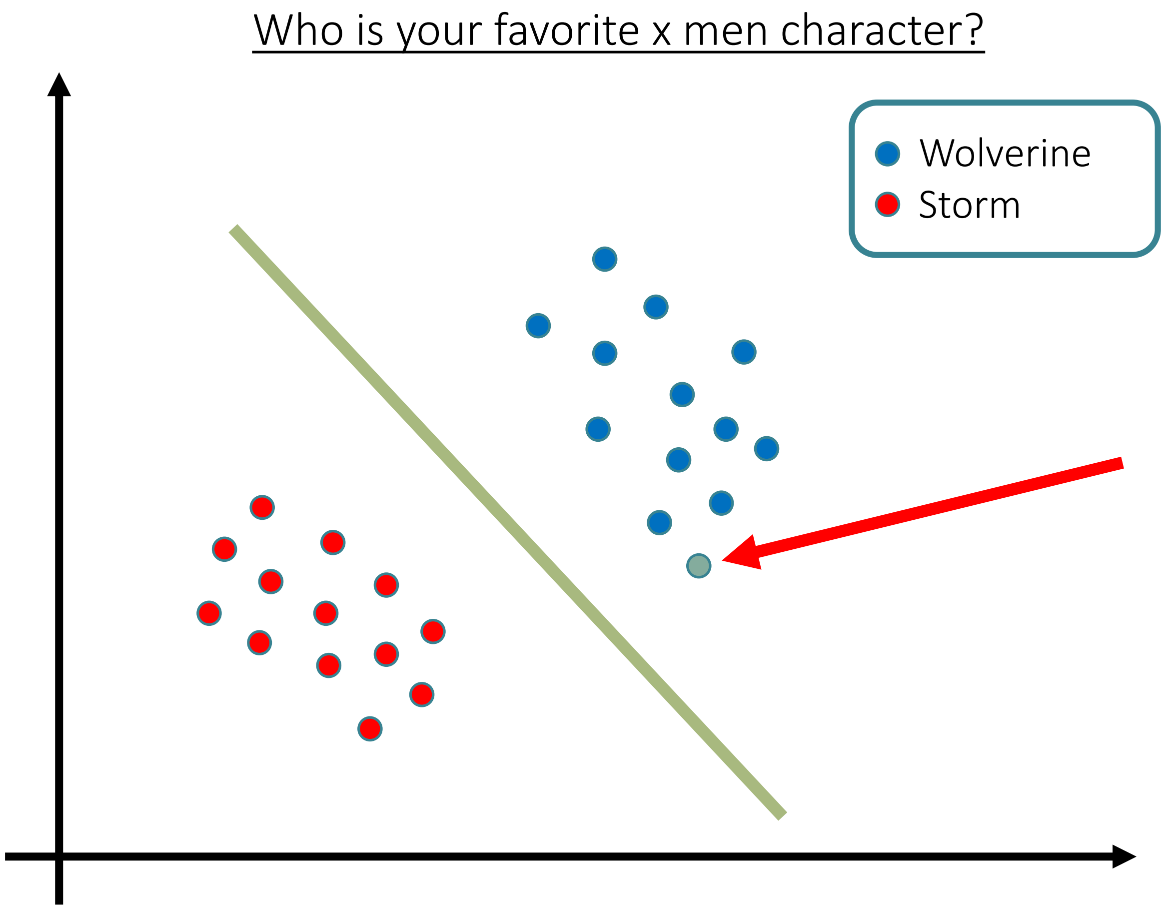
Classification



Classification

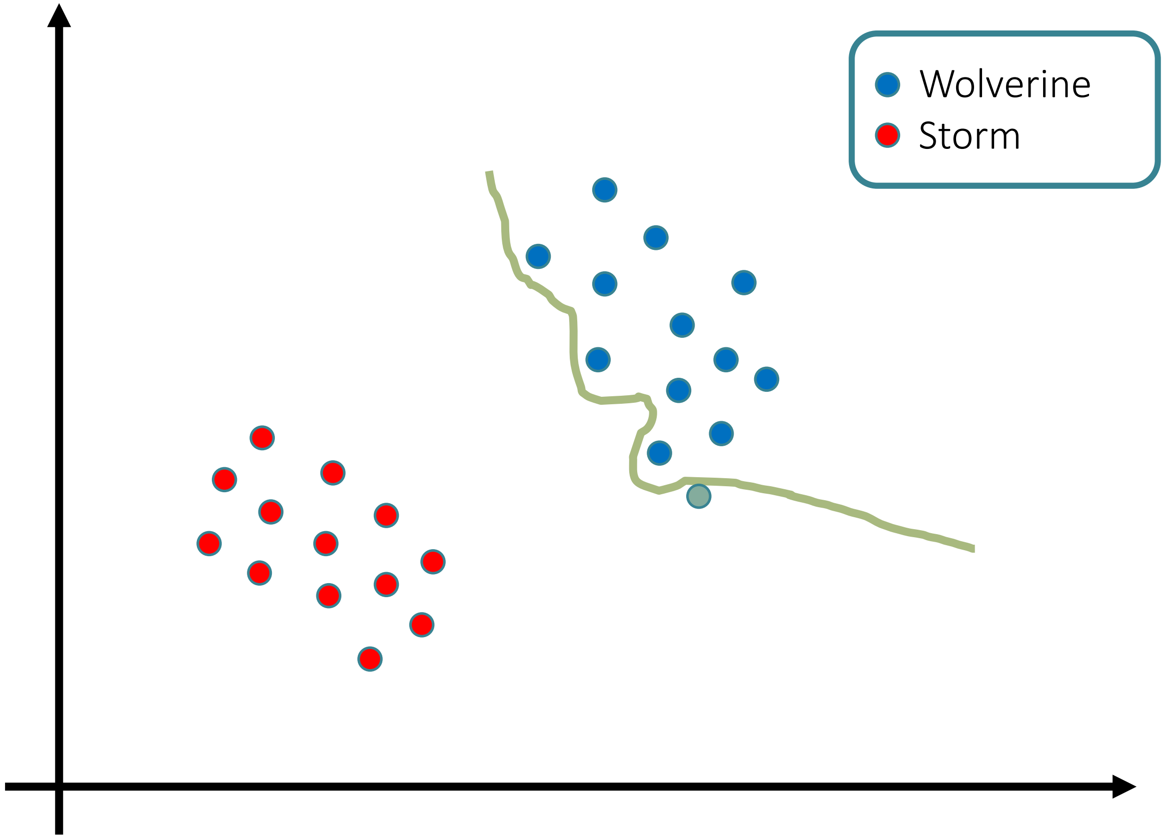


Classification



Classification

Who is your favorite x men character?

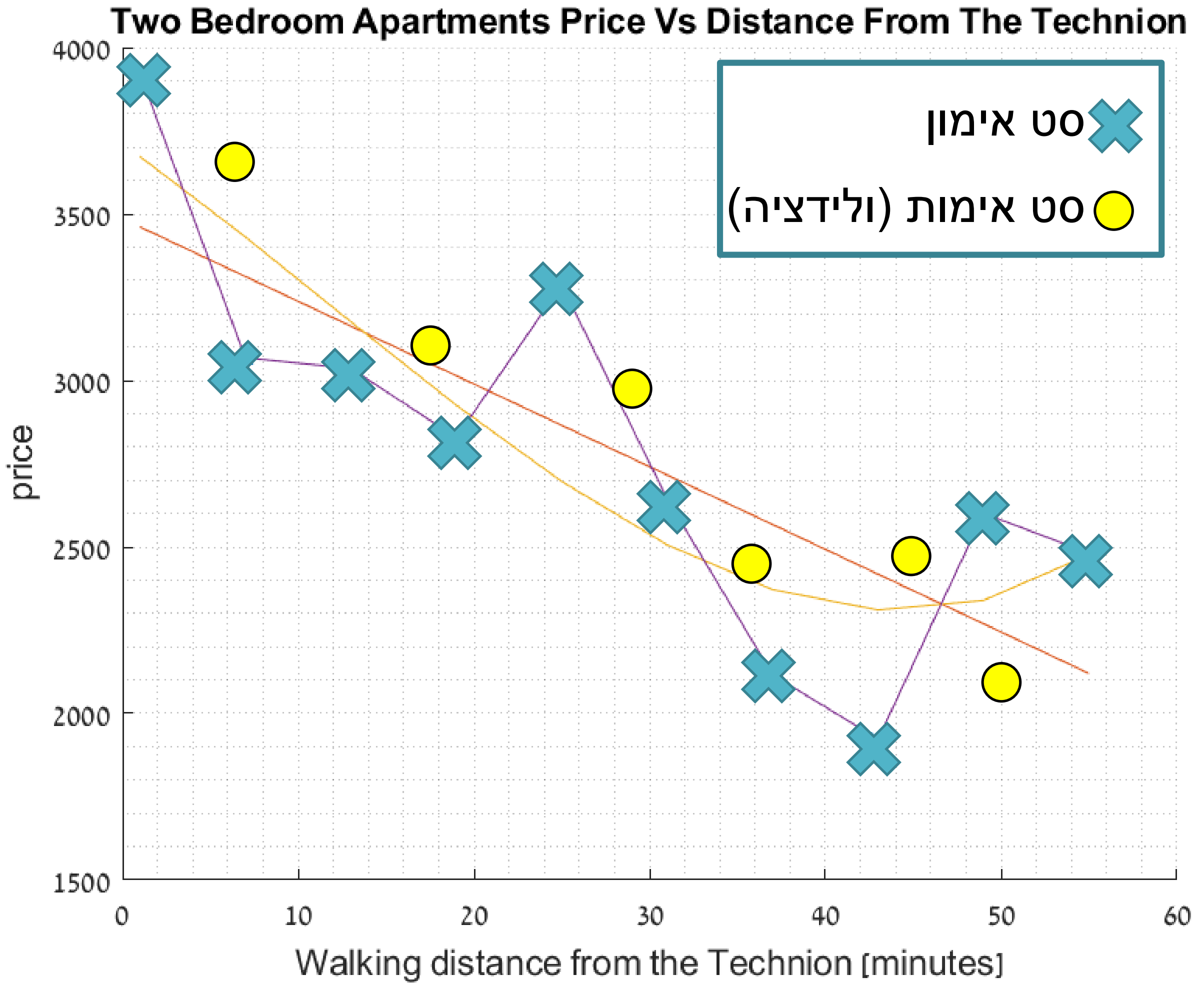


תהליך הלימוד (ולידציה)

- נחלק את ה-Data שיש לנו לשלוש קבוצות:

1. Training set (סט אימון) – סט דוגמאות מתויג $D = \{x_i, y_i\}_{i=1}^{n}$ שבאמצעותו האלגוריתם לומד.
2. Validation set (סט אימות) - סט דוגמאות מתויג $D = \{x_i, y_i\}_{i=1}^{n}$ שבאמצעותו נעריך את טיבם של המודלים על מנת לבחור ביניהם.
3. Test set (סט בוחן) - סט דוגמאות מתויג $D = \{x_i, y_i\}_{i=1}^{n}$ שבאמצעותו נעריך את ביצועי המודל הסופי שבחרנו.
   - הערה: השימוש בסט זה הינו השלב האחרון בתהליך הלמידה, ואין להשתמש בו כדי להעריך את ביצועי המודל במהלך הלימוד.

Regression



Two Bedroom Apartments Price Vs Distance From The Technion
סט אימון
סט אימות (ולידציה)
price
4000
3500
3000
2500
2000
1500
0
10
20
30
40
50
60
Walking distance from the Technion [minutes]

K-fold Cross-Validation

- במקרים בהם ה- Data הניתן לנו הוא מוגבל, לא נרצה לבזבז Data על ידי הקצאתו ל- Validation Set. שיטה זו מאפשרת לקבל הערכה לשגיאת שערוך.

input : $D=\{x_i, y_i\}_{i=1}^{n}$, integer k, learning algorithm A, model M

1. Create k data partitions: $D_1,...,D_k$,

$$\text{s.t.} \quad \bigcap_{j=1...k} D_j = \phi, \quad \bigcup_{j=1...k} D_j = D, \quad \forall j,l \ \left|D_j\right| \simeq \left|D_l\right|$$

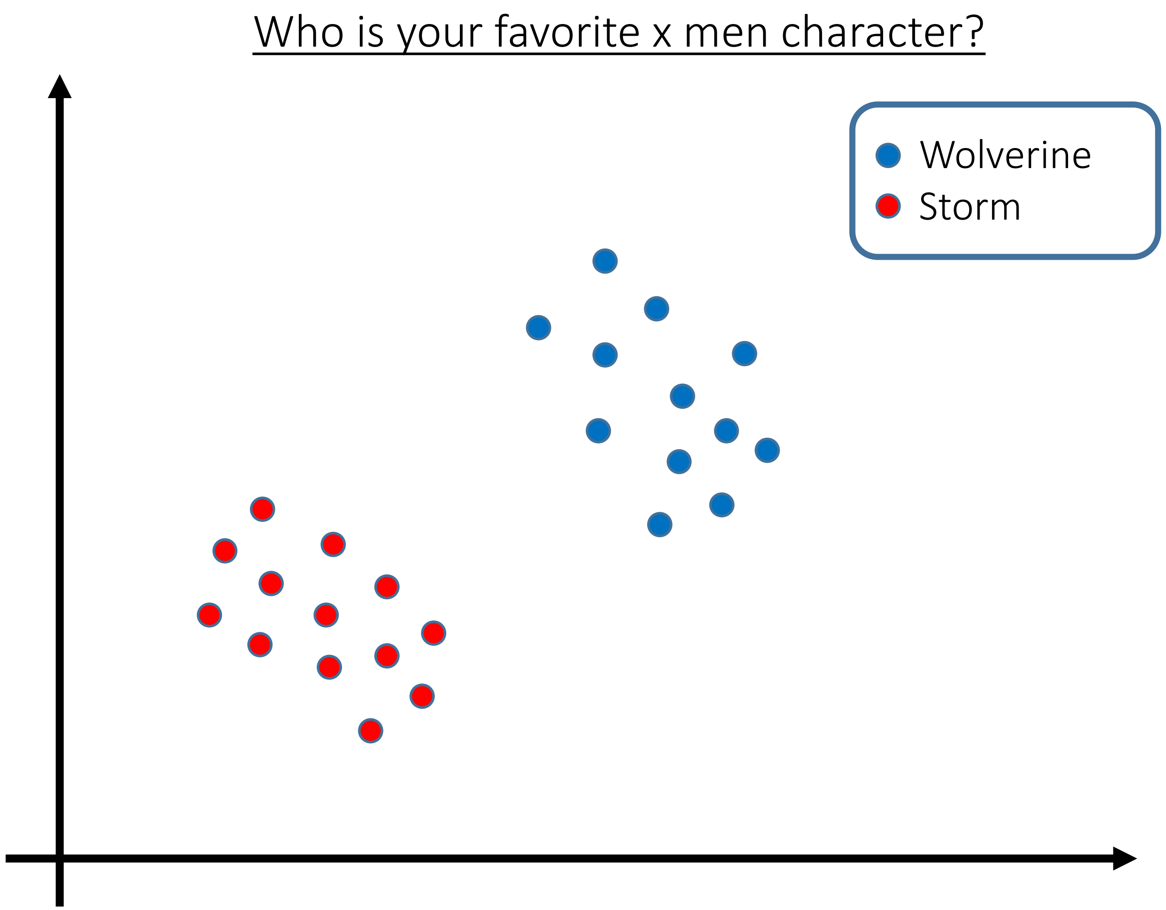
2. For $j=1,...,k$

   2.1 Fit model M by algorithm A with data $\{\mathrm{D} \setminus \mathrm{D}_j\}$

   2.2 Calculate $\hat{R}_n^{(j)}(M)$

3. Return $\hat{R}_n = \frac{1}{k}\sum_{j=1...k} \hat{R}_n^{(j)}$

K-fold Cross-Validation



K-fold Cross-Validation

input : $D = \{x_i, y_i\}_{i=1}^{n}$, integer k, learning algorithm A, model M

1. Create k data partitions: $D_1, \ldots, D_k$,

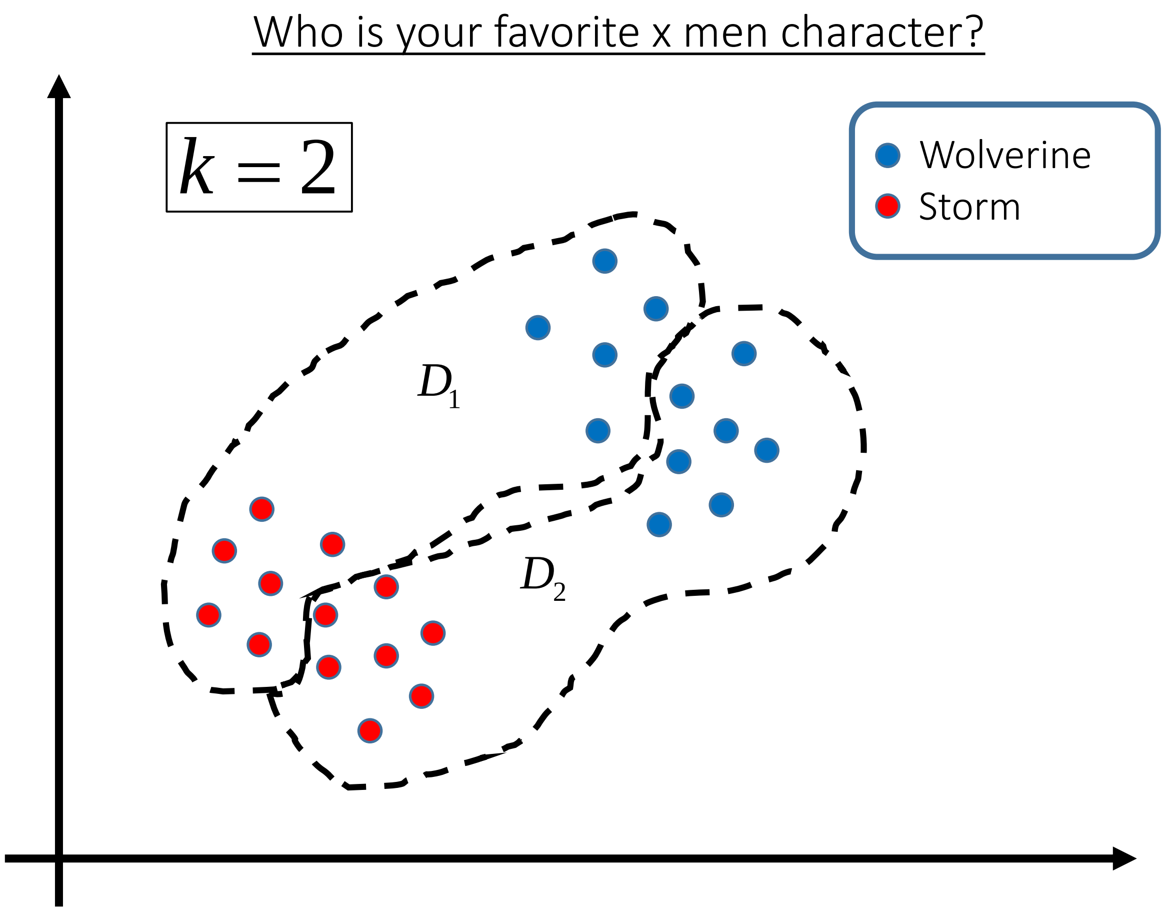
   s.t. $\bigcap_{j=1\ldots k} D_j = \phi, \quad \bigcup_{j=1\ldots k} D_j = D, \quad \forall j,l \ |D_j| \simeq |D_l|$

2. For $j = 1, \ldots, k$

   2.1 Fit model M by algorithm A with data $\{D \setminus D_j\}$

   2.2 Calculate $\hat{R}_n^{(j)}(M)$

3. Return $\hat{R}_n = \frac{1}{k} \sum_{j=1\ldots k} \hat{R}_n^{(j)}$



K-fold Cross-Validation

input : $D = \{x_i, y_i\}_{i=1}^{n}$, integer k, learning algorithm A, model M

1. Create k data partitions: $D_1, \ldots, D_k$,

   s.t. $\bigcap_{j=1\ldots k} D_j = \phi, \quad \bigcup_{j=1\ldots k} D_j = D, \quad \forall j,l \; |D_j| \simeq |D_l|$

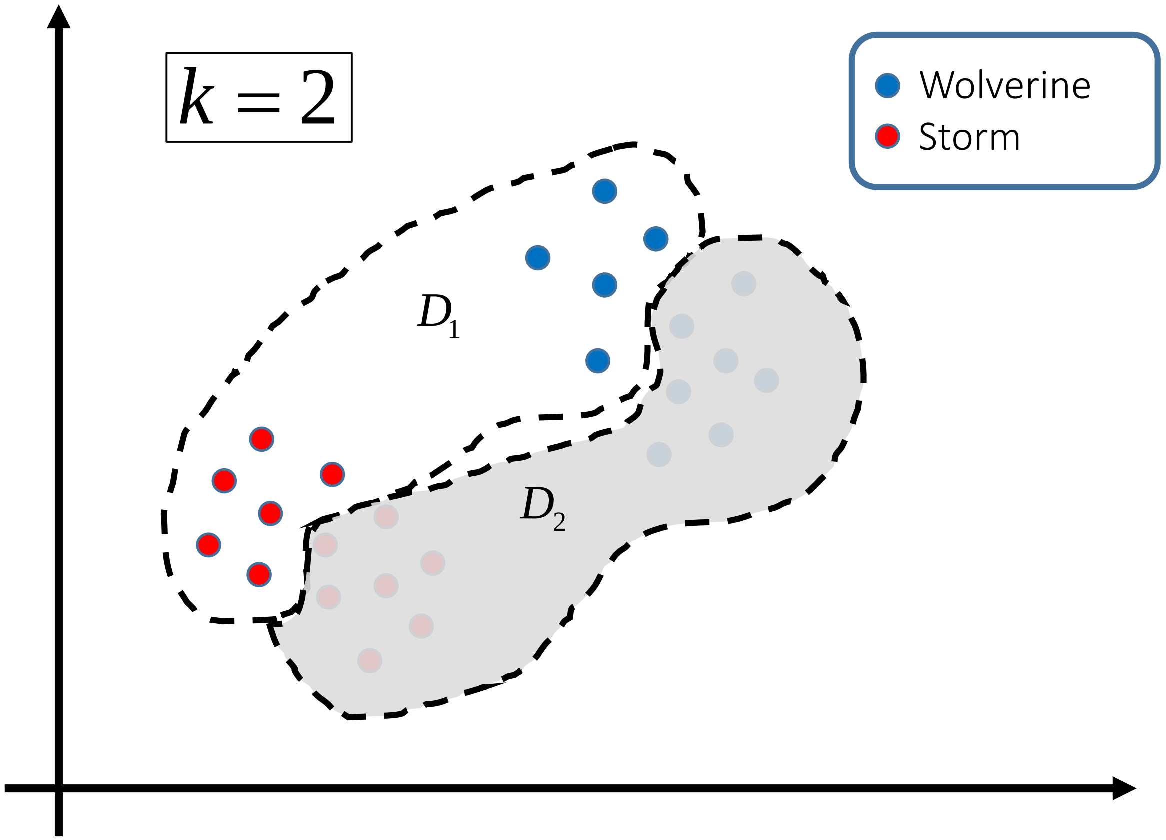
2. For $j = 1, \ldots, k$

   2.1 Fit model M by algorithm A with data $\{D \setminus D_j\}$

   2.2 Calculate $\hat{R}_n^{(j)}(M)$

3. Return $\hat{R}_n = \frac{1}{k} \sum_{j=1\ldots k} \hat{R}_n^{(j)}$

Who is your favorite x men character?



K-fold Cross-Validation

input : $D = \{x_i, y_i\}_{i=1}^{n}$, integer k, learning algorithm A, model M

1. Create k data partitions: $D_1, \ldots, D_k$,

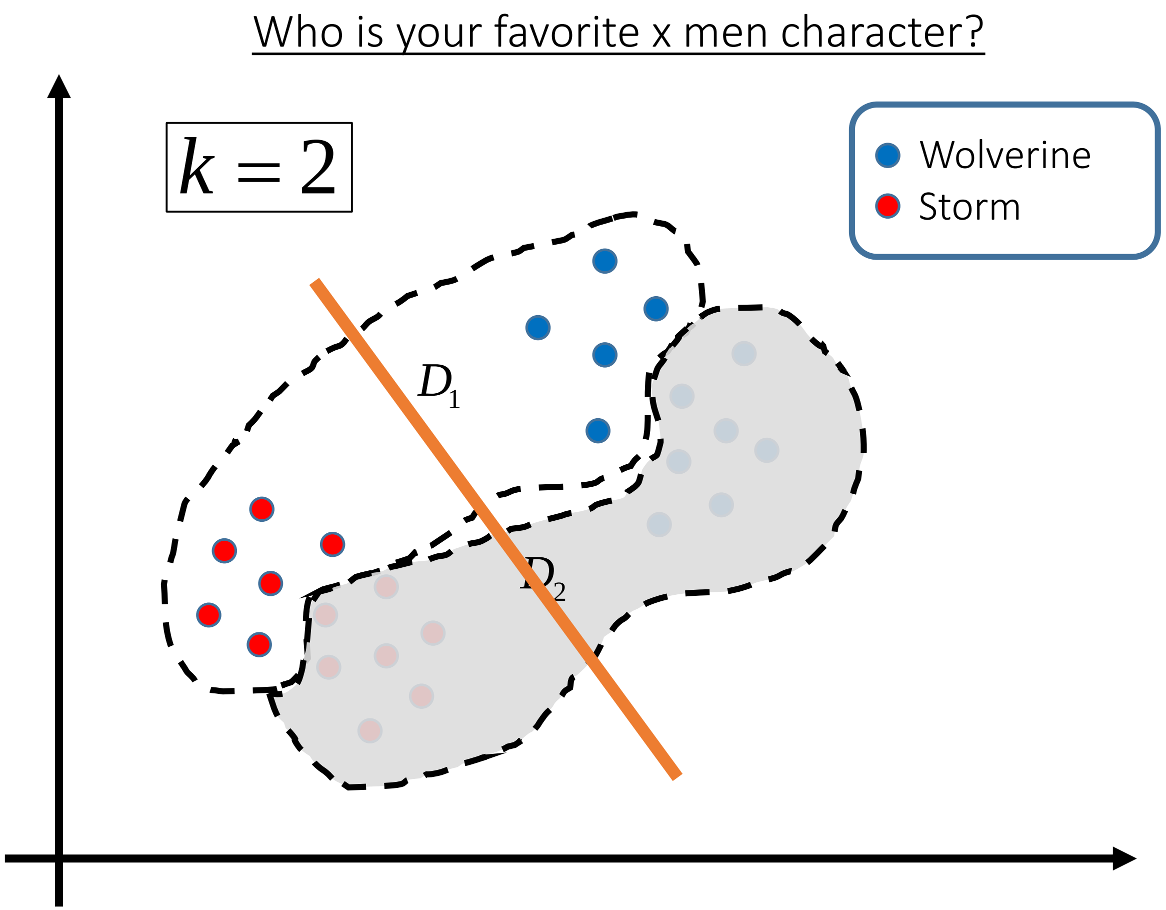
   s.t. $\bigcap_{j=1\ldots k} D_j = \phi, \quad \bigcup_{j=1\ldots k} D_j = D, \quad \forall j,l \; |D_j| \simeq |D_l|$

2. For $j = 1, \ldots, k$

   2.1 Fit model M by algorithm A with data $\{\mathrm{D} \setminus \mathrm{D}_j\}$

   2.2 Calculate $\hat{R}_n^{(j)}(M)$

3. Return $\hat{R}_n = \frac{1}{k} \sum_{j=1\ldots k} \hat{R}_n^{(j)}$



K-fold Cross-Validation

input : $D = \{x_i, y_i\}_{i=1}^{n}$, integer k, learning algorithm A, model M

1. Create k data partitions: $D_1, \ldots, D_k$,

   s.t. $\bigcap_{j=1 \ldots k} D_j = \phi, \quad \bigcup_{j=1 \ldots k} D_j = D, \quad \forall j,l \; |D_j| \simeq |D_l|$

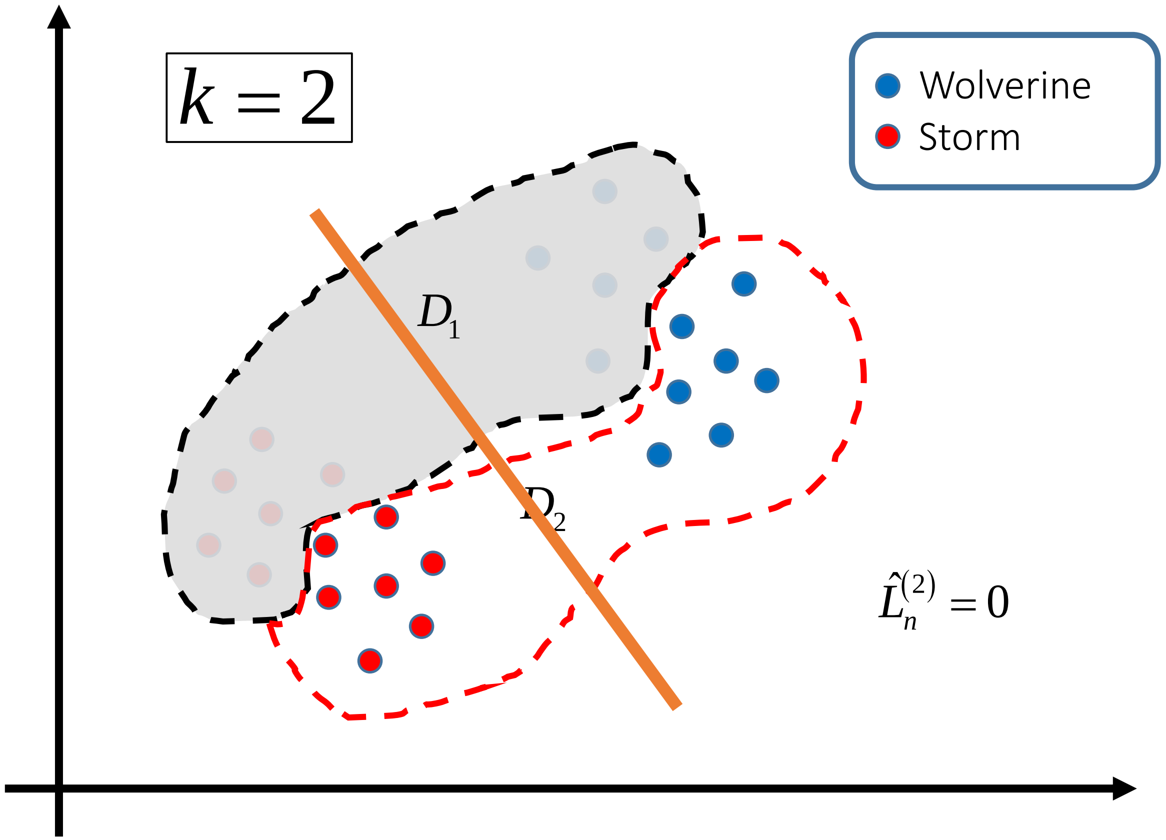
2. For $j = 1, \ldots, k$

   2.1 Fit model M by algorithm A with data $\{D \setminus D_j\}$

   2.2 Calculate $\hat{R}_n^{(j)}(M)$

3. Return $\hat{R}_n = \frac{1}{k} \sum_{j=1 \ldots k} \hat{R}_n^{(j)}$

Who is your favorite x men character?



K-fold Cross-Validation

input : $D = \{x_i, y_i\}_{i=1}^{n}$, integer k, learning algorithm A, model M

1. Create k data partitions: $D_1, \dots, D_k$,

   s.t. $\bigcap_{j=1\dots k} D_j = \phi, \quad \bigcup_{j=1\dots k} D_j = D, \quad \forall j,l \ |D_j| \simeq |D_l|$

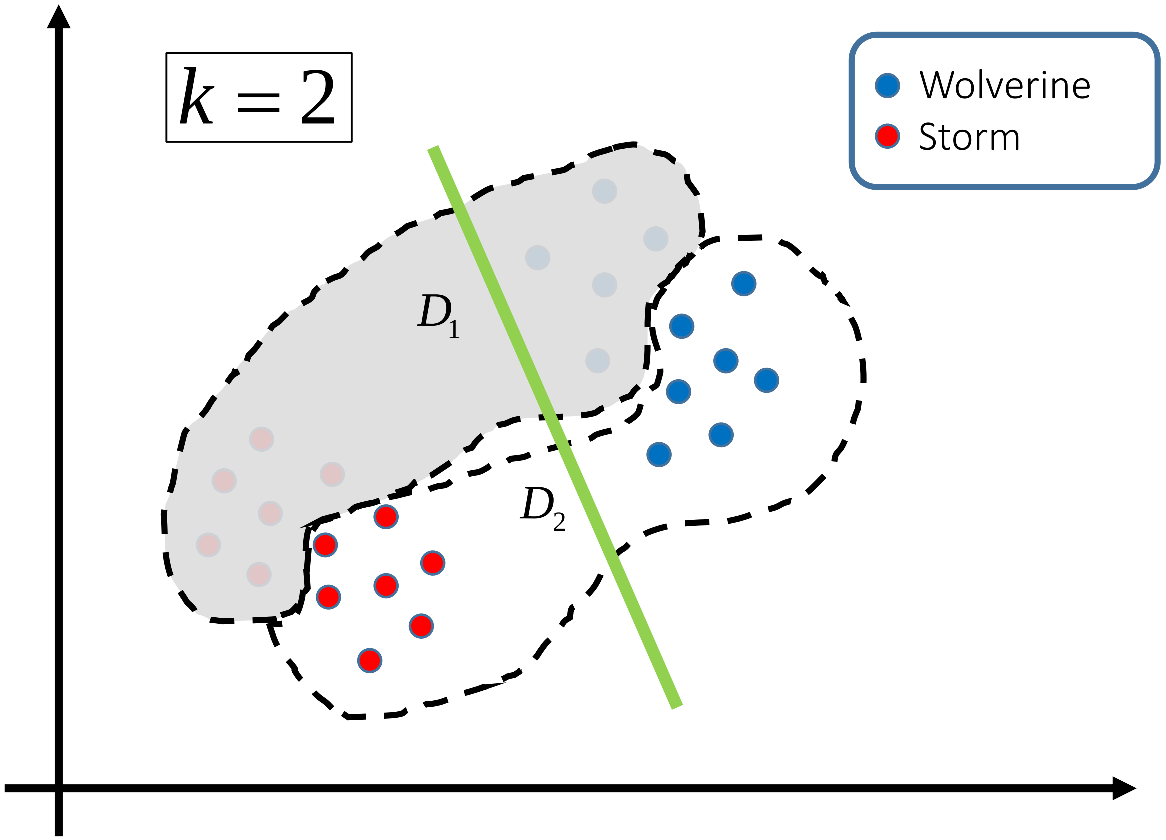
2. For $j = 1, \dots, k$

   2.1 Fit model M by algorithm A with data $\{D \setminus D_j\}$

   2.2 Calculate $\hat{R}_n^{(j)}(M)$

3. Return $\hat{R}_n = \frac{1}{k} \sum_{j=1\dots k} \hat{R}_n^{(j)}$

Who is your favorite x men character?



K-fold Cross-Validation

input : $D = \{x_i, y_i\}_{i=1}^{n}$, integer k, learning algorithm A, model M

1. Create k data partitions: $D_1, \ldots, D_k$,

   s.t. $\bigcap_{j=1\ldots k} D_j = \phi, \quad \bigcup_{j=1\ldots k} D_j = D, \quad \forall j,l \; |D_j| \simeq |D_l|$

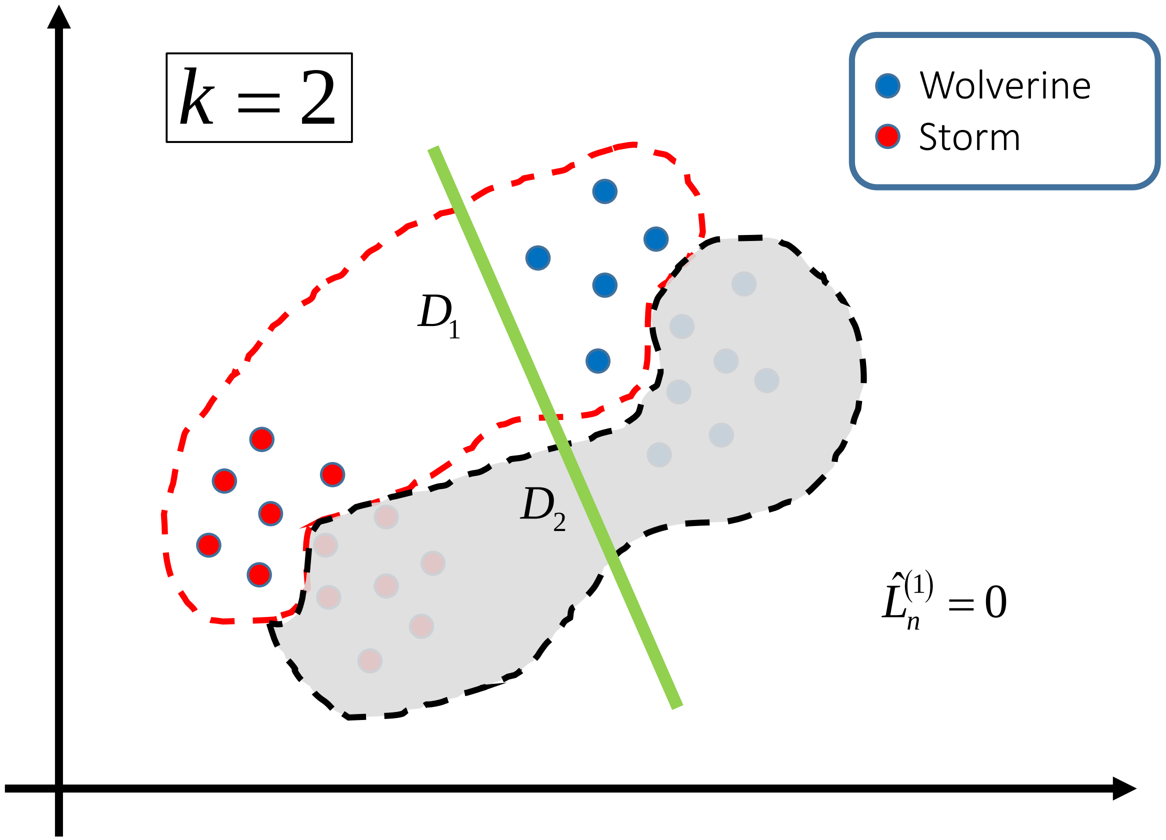
2. For $j = 1, \ldots, k$

   2.1 Fit model M by algorithm A with data $\{\mathrm{D} \setminus \mathrm{D}_j\}$

   2.2 Calculate $\hat{R}_n^{(j)}(M)$

3. Return $\hat{R}_n = \frac{1}{k} \sum_{j=1\ldots k} \hat{R}_n^{(j)}$

Who is your favorite x men character?



K-fold Cross-Validation

input : $D = \{x_i, y_i\}_{i=1}^{n}$, integer k, learning algorithm A, model M

1. Create k data partitions: $D_1, \ldots, D_k$,

   s.t. $\bigcap_{j=1\ldots k} D_j = \phi, \quad \bigcup_{j=1\ldots k} D_j = D, \quad \forall j,l \; |D_j| \simeq |D_l|$

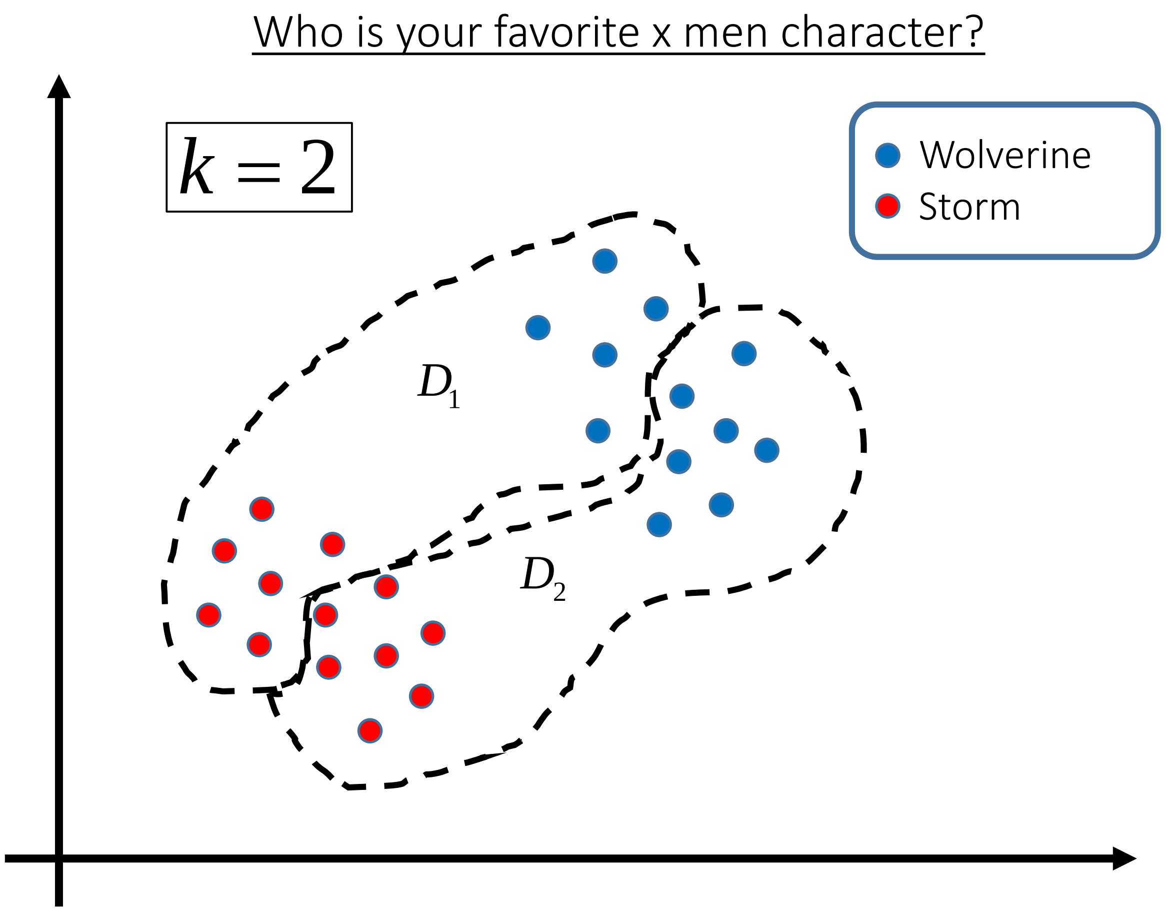
2. For $j = 1, \ldots, k$

   2.1 Fit model M by algorithm A with data $\{D \setminus D_j\}$

   2.2 Calculate $\hat{R}_n^{(j)}(M)$

3. Return $\hat{R}_n = \frac{1}{k} \sum_{j=1\ldots k} \hat{R}_n^{(j)}$

$$\hat{R}_n = 0$$



K-fold Cross-Validation

input : $D = \{x_i, y_i\}_{i=1}^{n}$, integer k, learning algorithm A, model M

1. Create k data partitions: $D_1, \ldots, D_k$,

   s.t. $\bigcap_{j=1\ldots k} D_j = \phi, \quad \bigcup_{j=1\ldots k} D_j = D, \quad \forall j,l \ |D_j| \simeq |D_l|$

2. For $j = 1, \ldots, k$

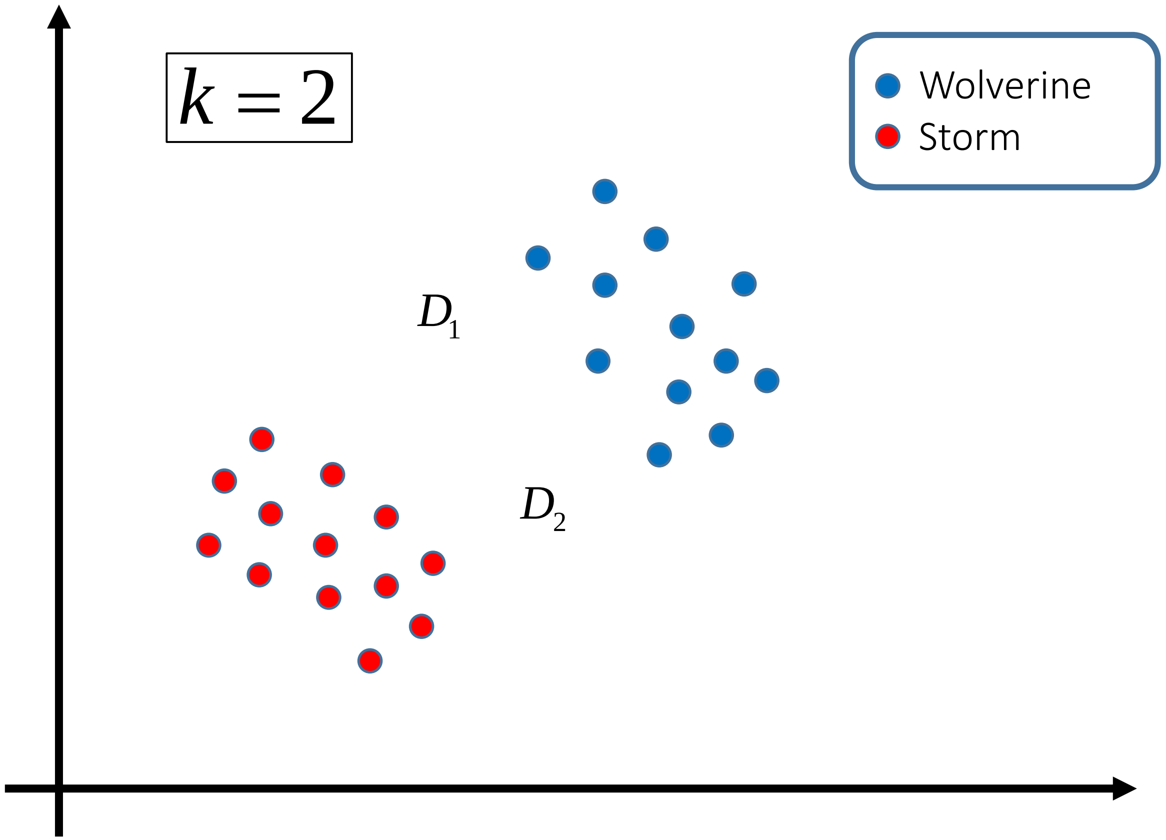
   2.1 Fit model M by algorithm A with data $\{D \setminus D_j\}$

   2.2 Calculate $\hat{R}_n^{(j)}(M)$

3. Return $\hat{R}_n = \frac{1}{k} \sum_{j=1\ldots k} \hat{R}_n^{(j)}$

$$\hat{R}_n = 0$$

Who is your favorite x men character?



K-fold Cross-Validation

input : $D = \{x_i, y_i\}_{i=1}^{n}$, integer k, learning algorithm A, model M

1. Create k data partitions: $D_1, \ldots, D_k$,

   s.t. $\bigcap_{j=1\ldots k} D_j = \phi, \quad \bigcup_{j=1\ldots k} D_j = D, \quad \forall j,l \; |D_j| \simeq |D_l|$

2. For $j = 1, \ldots, k$

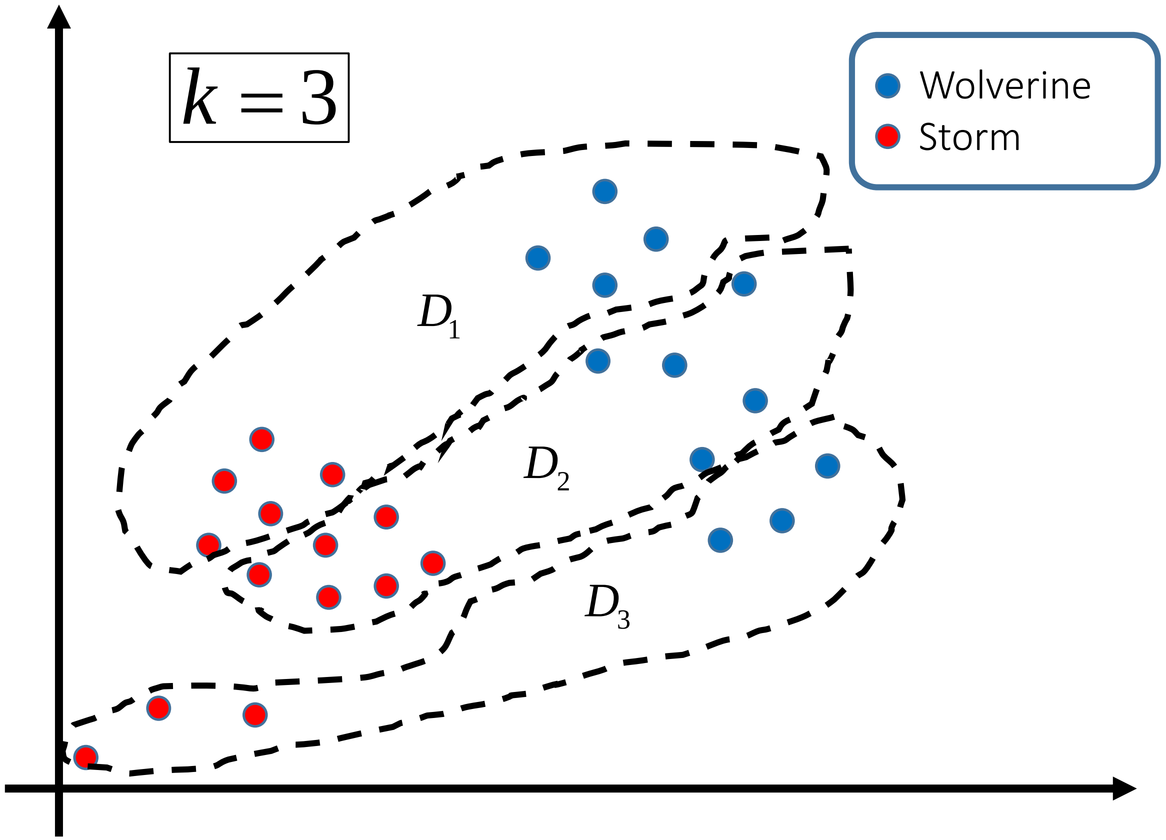
   2.1 Fit model M by algorithm A with data $\{D \setminus D_j\}$

   2.2 Calculate $\hat{R}_n^{(j)}(M)$

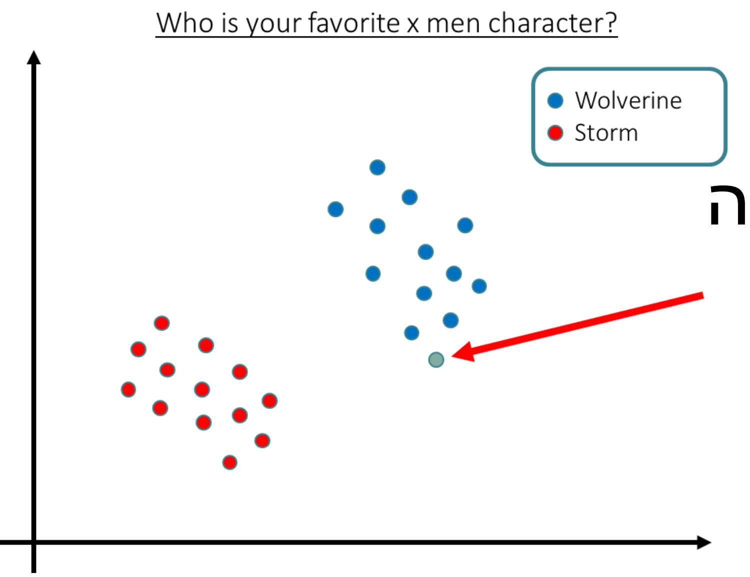
3. Return $\hat{R}_n = \frac{1}{k} \sum_{j=1\ldots k} \hat{R}_n^{(j)}$

* Repeated cross-validation

Who is your favorite x men character?



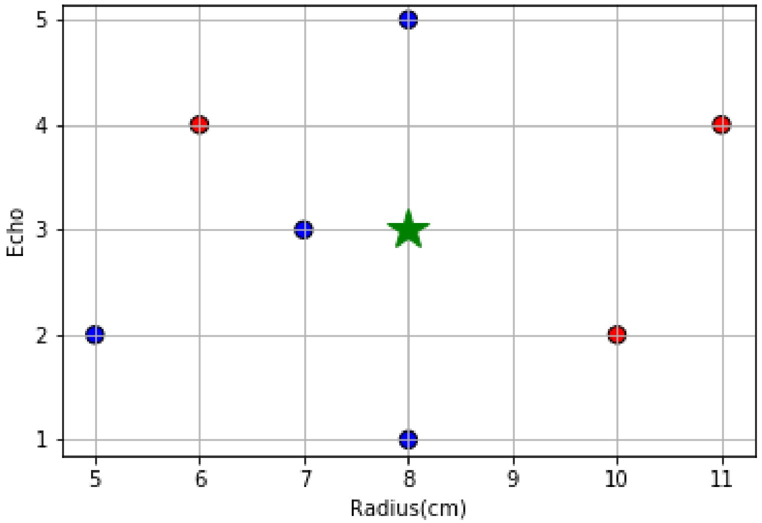
סיווג בעזרת אלגוריתם K-NN (k Nearest Neighbours)



1. מצא את K השכנים הקרובים ביותר לנקודה החדשה.
2. מצא לאיזו קבוצה שייכים רוב השכנים. הנקודה החדשה שייכת לקבוצה זו.
   1. במקרה של שוויון בשלב 2, השווה סכום מרחקים. הנקודה החדשה שייכת לקבוצה בעלת הסכום המינימלי.
      1. במקרה של שוויון בשלב 2.1, בחר אקראית.

תרגיל 1

נסה לאפיין האם אבטיח מתוק או חמוץ על פי ההד בהקשה עליהם והקוטר שלהם.



נסמן: R – רדיוס, E – הד

יש לנו את המדידות האלו מהעבר-

$Sweet: (R, E) = (8,1), (5,2), (7,3), (8,5)$

$Sour: (R, E) = (10,2), (6,4), (11,4)$

הסטודנט מחזיק בידו האבטיח בעל המאפיינים $(R,E) = (8,3)$.

האם סביר שהאבטיח מתוק או חמוץ?

1. בדקו את תוצאות ה-classification עבור k-nearest neighbors, כאשר K=1,3.
2. בצע Cross Validation להערכת טיב המודל, עבור K=1,3 ו-7 קבוצות (Leave-one-out Cross Validation). באיזה מסווג נבחר?
3. מה יקרה אם נבחר את k להיות בגודל ה- dataset.

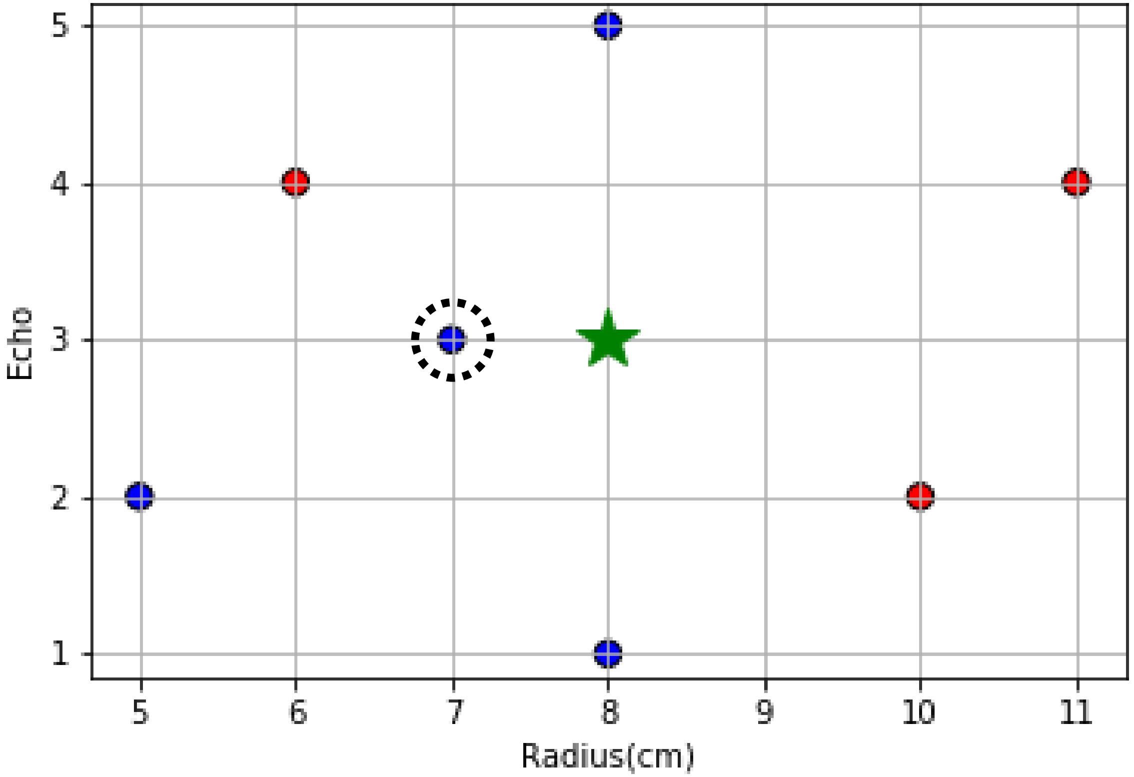
תרגיל 1 - פתרון

1. בדקו את תוצאות ה-classification עבור k-nearest neighbors, כאשר K=1,3.

<u>פתרון:</u> עבור $K = 1$:

- Sweet
- Sour

$\hat{y} = sweet$



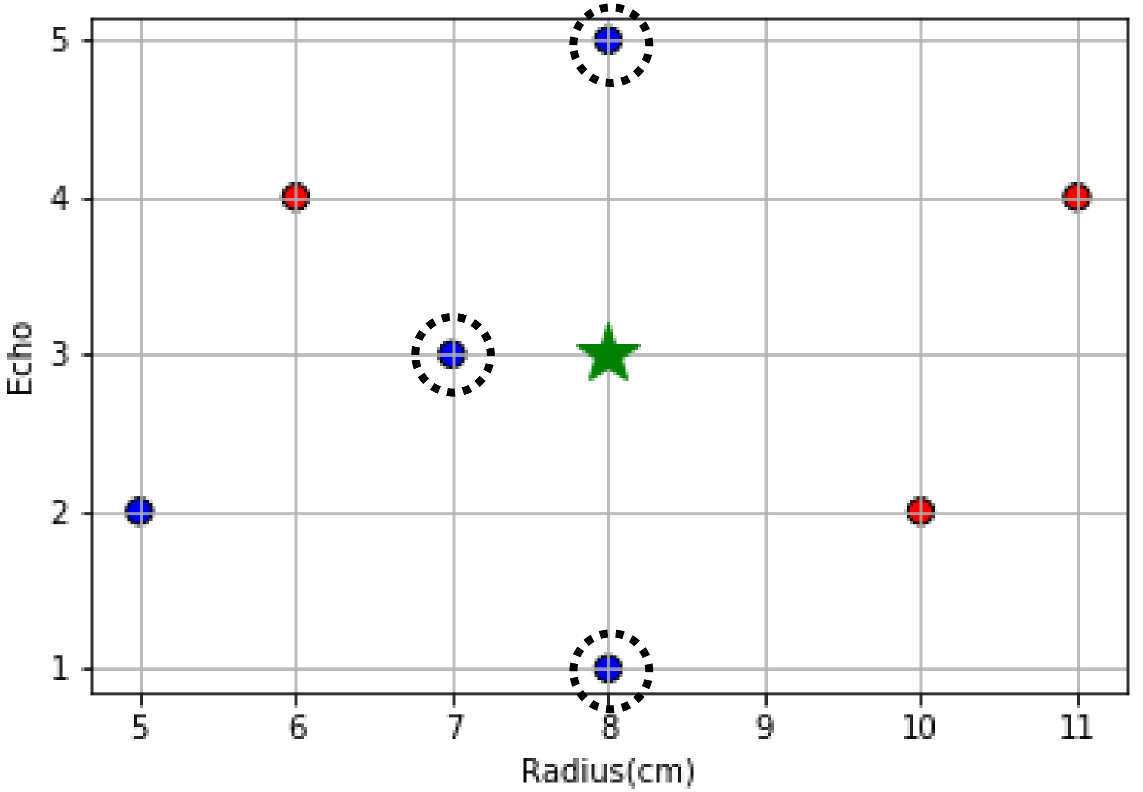
תרגיל 1 - פתרון

1. בדקו את תוצאות ה-classification עבור k-nearest neighbors, כאשר K=1,3.

<u>פתרון:</u> עבור $K = 3$:

- Sweet
- Sour

$\hat{y} = sweet$

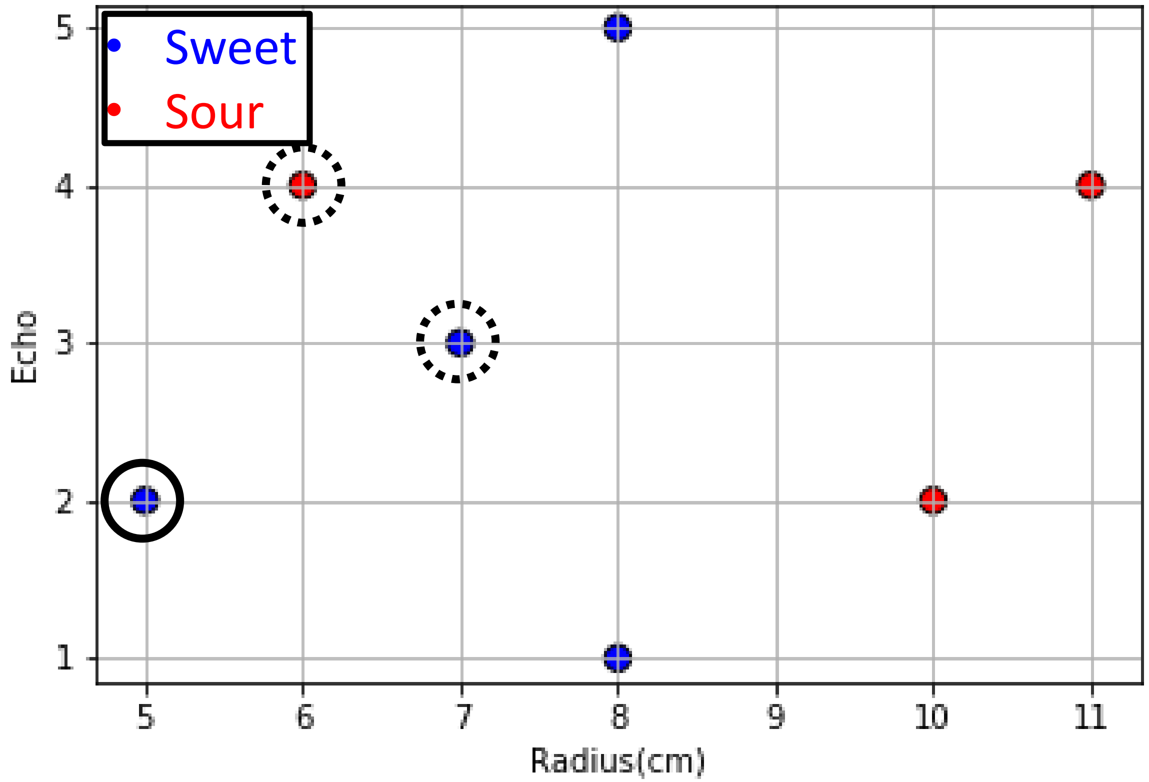


תרגיל 1 - פתרון

2. בצע Cross Validation להערכת טיב המודל, עבור K=1,3 ו-7 קבוצות (Leave-one-out Cross Validation). באיזה מסווג נבחר?

<u>פתרון:</u>

$$\underbrace{(5,2),blue}_{X,Y}: K=1 \rightarrow \underbrace{blue(random)}_{prediction}$$



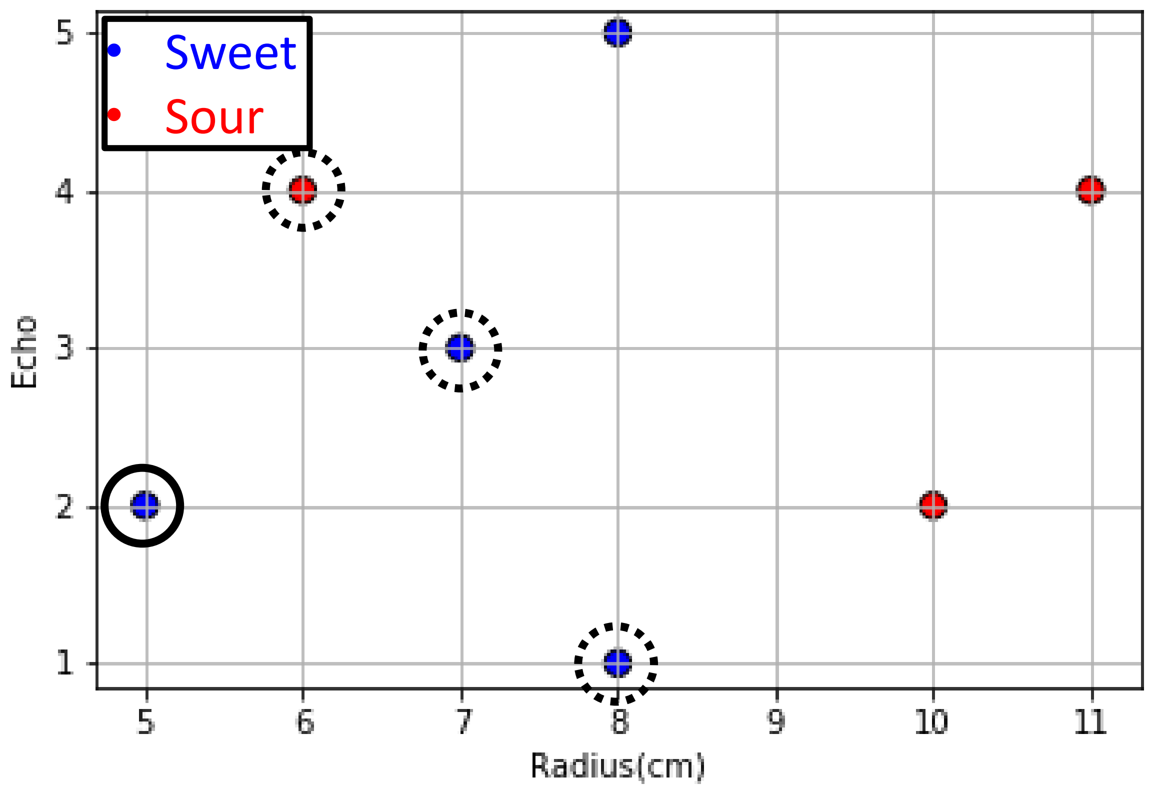
תרגיל 1 - פתרון

2. בצע Cross Validation להערכת טיב המודל, עבור K=1,3 ו-7 קבוצות (Leave-one-out Cross Validation). באיזה מסווג נבחר?

<u>פתרון:</u>

$$\underbrace{(5,2)}_{X,Y}, blue : K = 1 \rightarrow \underbrace{blue(random)}_{prediction},$$

$$K = 3 \rightarrow \underbrace{blue}_{prediction}$$



תרגיל 1 - פתרון

2. בצע Cross Validation להערכת טיב המודל, עבור K=1,3 ו-7 קבוצות (Leave-one-out Cross Validation). באיזה מסווג נבחר?

<u>פתרון:</u>

	$\boldsymbol{K=1}$	$\boldsymbol{K=3}$
$(5,2), blue$	$blue$ ✓	$blue$ ✓
$(6,4), red$	$blue$ ✗	$blue$ ✗
$(7,3), blue$	red ✗	$blue$ ✓
$(8,1), blue$	$blue$ ✓	$blue$ ✓
$(8,5), blue$	$blue$ ✓	red ✗
$(10,2), red$	$blue$ ✗	$blue$ ✗
$(11,4), red$	red ✓	$blue$ ✗

תרגיל 1 - פתרון

2. בצע Cross Validation להערכת טיב המודל, עבור K=1,3 ו-7 קבוצות (Leave-one-out Cross Validation). באיזה מסווג נבחר?

<u>פתרון:</u>

נחשב את שגיאת ה- CV:

עבור $K=1$, מספר השגיאות הוא 3, לכן שגיאת ה- CV הינה:

$$K=1:\ \hat{L}_n=\frac{3}{7}$$

עבור $K=3$, מספר השגיאות הוא 4, לכן שגיאת ה- CV הינה:

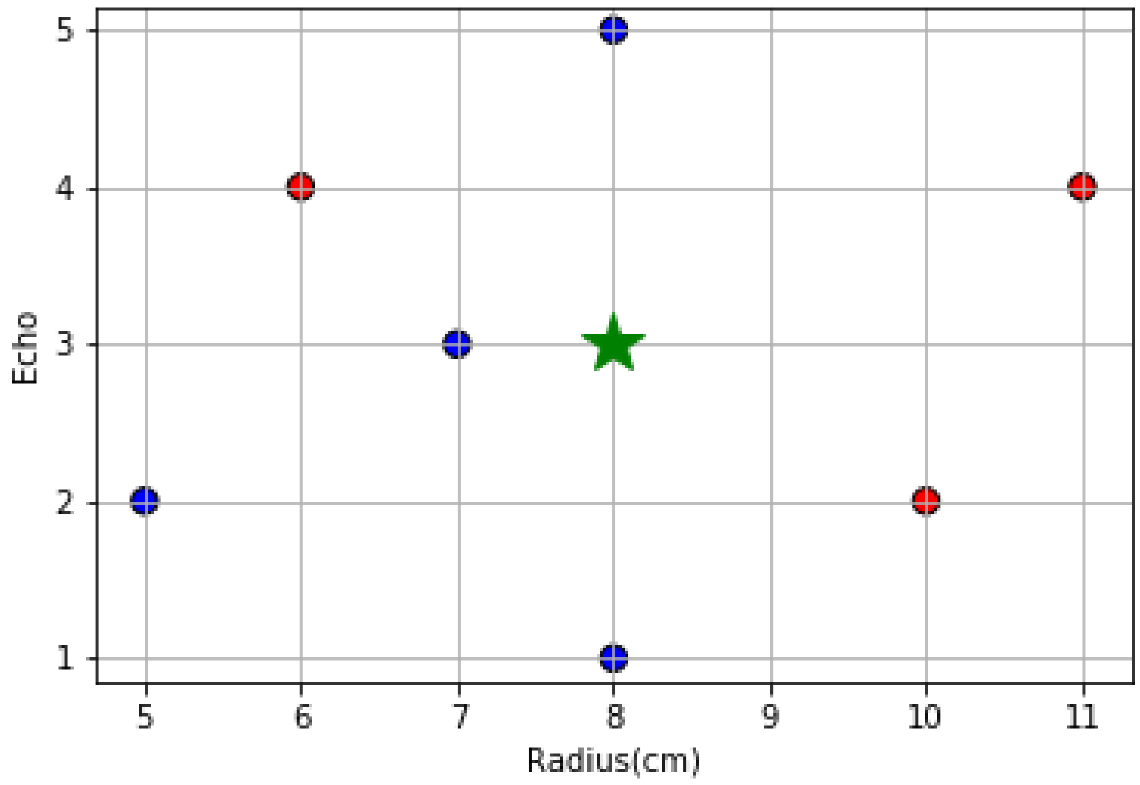
$$K=3:\ \hat{L}_n=\frac{4}{7}$$

$$\hat{L}_n=\frac{1}{k}\sum_{j=1\ldots k}\hat{L}_n^{(j)}$$

תרגיל 1 - פתרון

3. מה יקרה אם נבחר את k להיות בגודל ה- dataset.

- Sweet
- Sour



חזרה ל-KNN+Bias-Variance Tradeoff

- בשאלה זו ננסה לחזות את בחירתו של אזרח אמריקאי באמצעות אלגוריתם K-NN.
- לשם הפשטות, נניח כי כל אזרח מיוצג על ידי שני מאפיינים:
- מצבו הכלכלי (הציר האופקי - x) וקרבתו לדת (הציר האנכי – y).
- בסימונים שלמדנו:

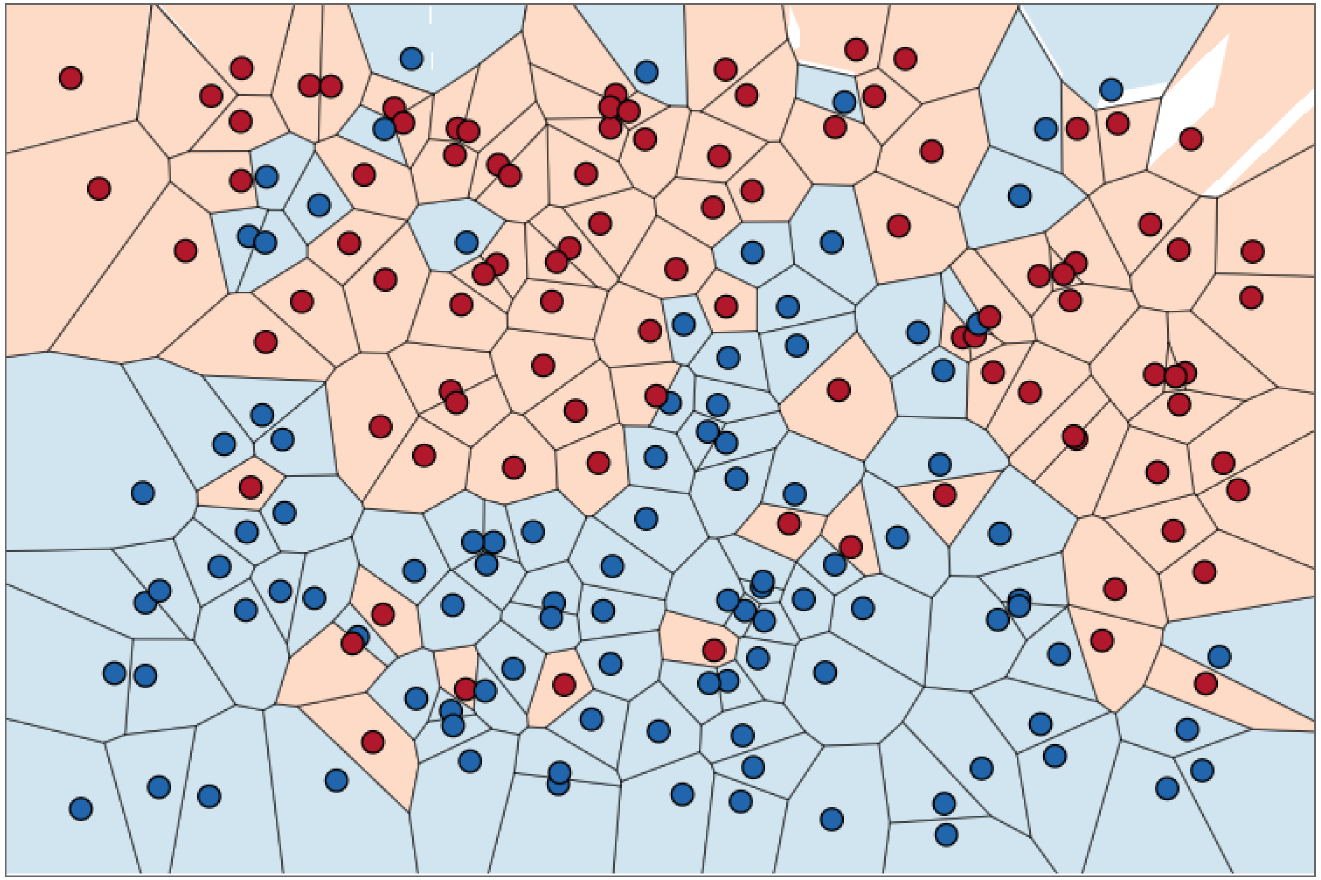
$$x = (wealth, religiousness) \in X = \mathbb{R}^2_+$$

$$y \in \{0,1\} = \{\text{Republican}, \text{Democrat}\}$$

Pictures taken from: http://scott.fortmann-roe.com/docs/BiasVariance.html

חזרה ל-KNN+Bias-Variance Tradeoff

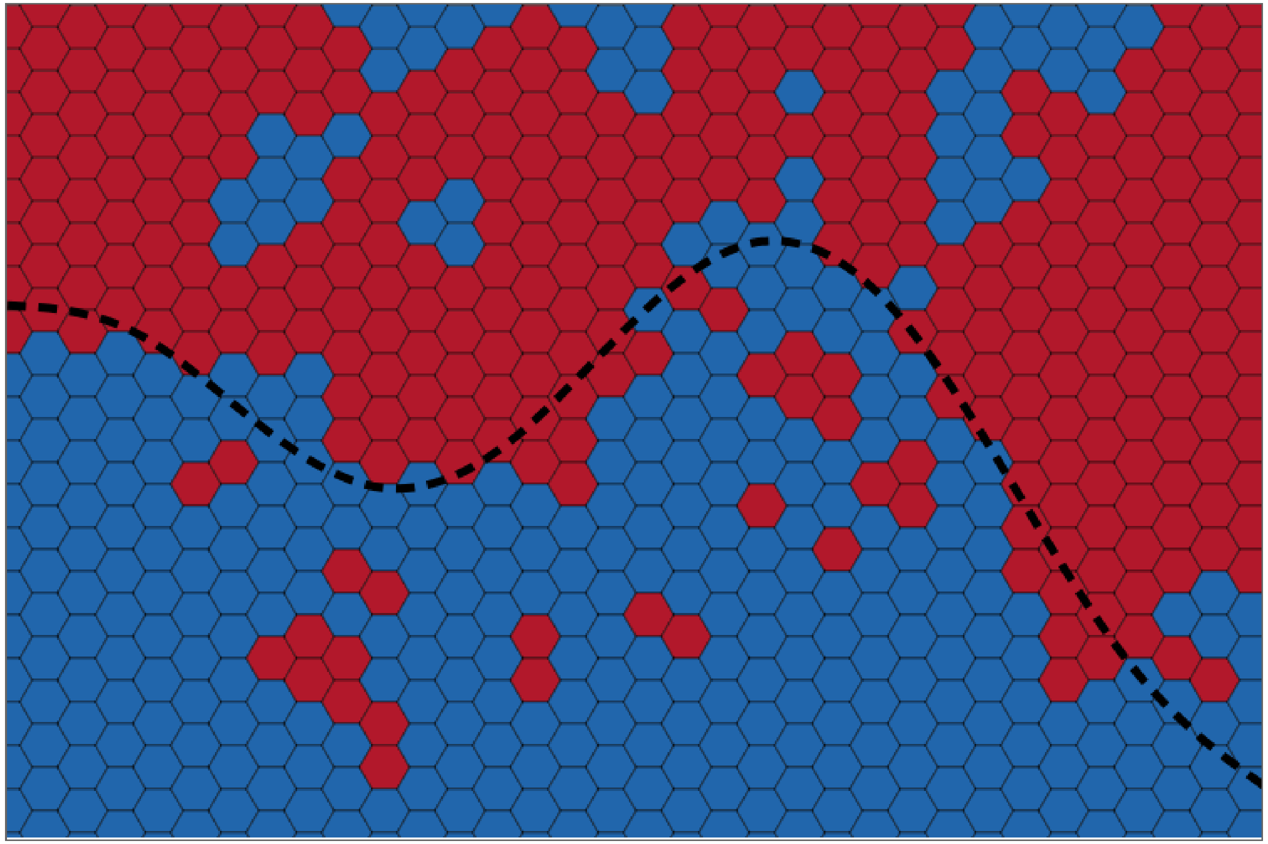
ה- Dataset (מיוצג ע"י נקודות) ומשטחי ההחלטה שנובעים מאלגוריתם 1-nn:



Pictures taken from: http://scott.fortmann-roe.com/docs/BiasVariance.html

חזרה ל-KNN+Bias-Variance Tradeoff

משטחי ההחלטה שנובעים מאלגוריתם 1-nn:

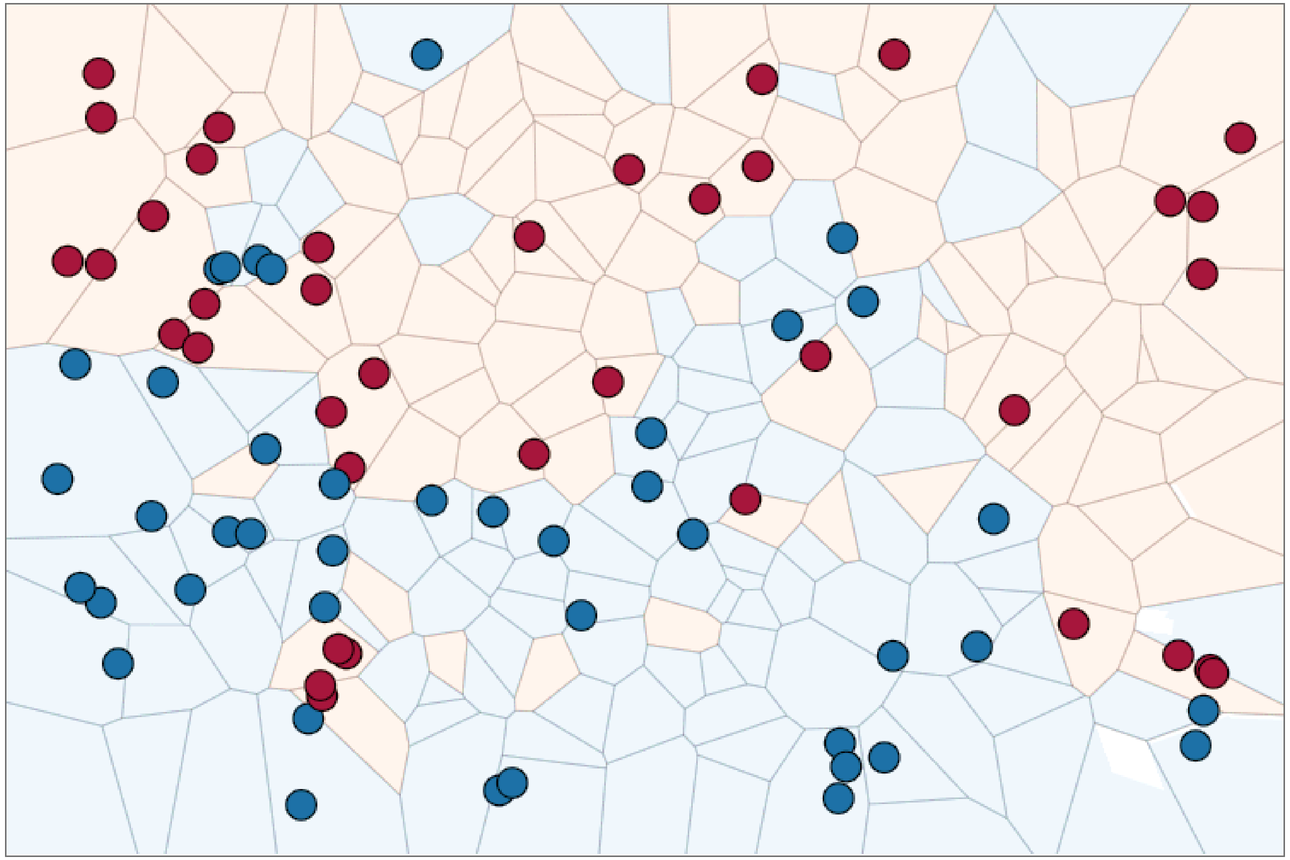


Pictures taken from:
http://scott.fortmann-roe.com/docs/BiasVariance.html

k-Nearest Neighbors: 1

חזרה ל-KNN+Bias-Variance Tradeoff

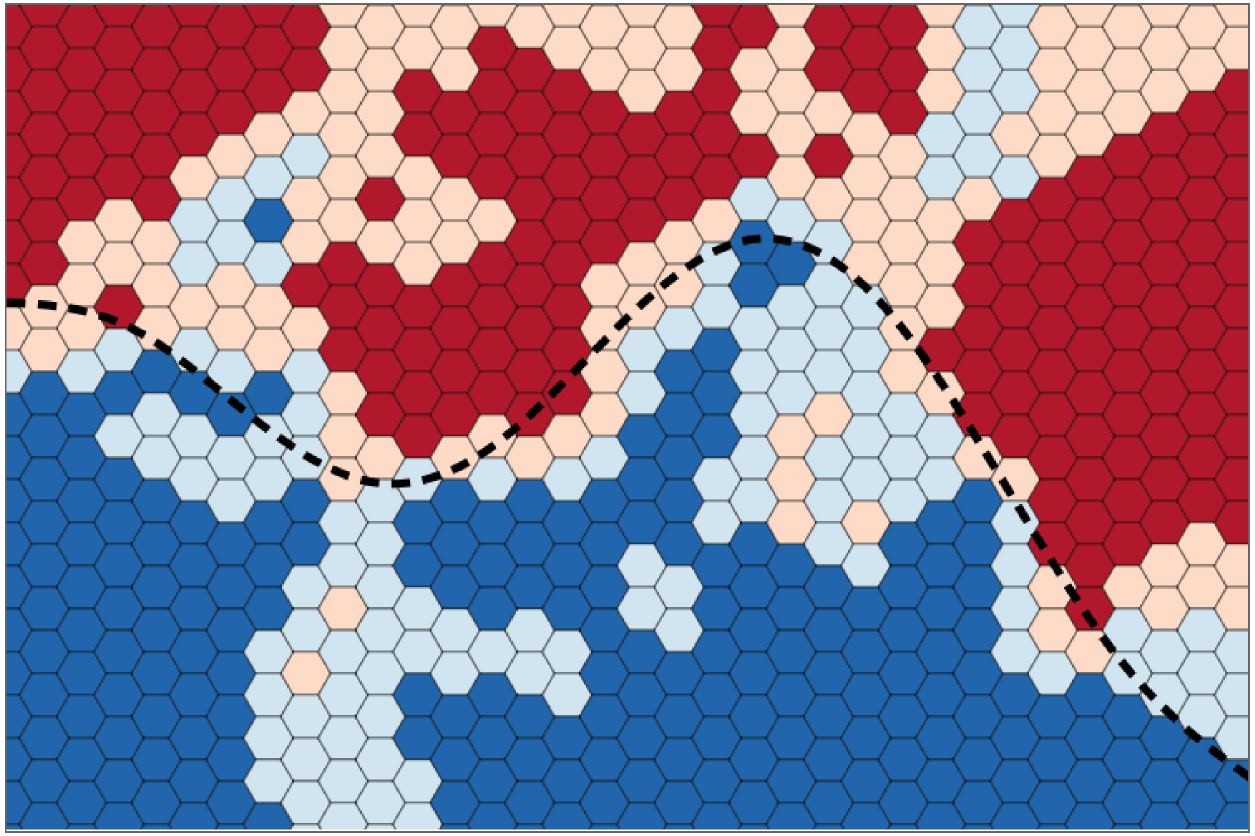
משטחי ההחלטה שנובעים מאלגוריתם 1-nn על סט האימון שלנו + סט בוחן כלשהו:



Pictures taken from:
http://scott.fortmann-roe.com/docs/BiasVariance.html

חזרה ל-KNN+Bias-Variance Tradeoff

משטחי ההחלטה שנובעים מאלגוריתם 3-nn:

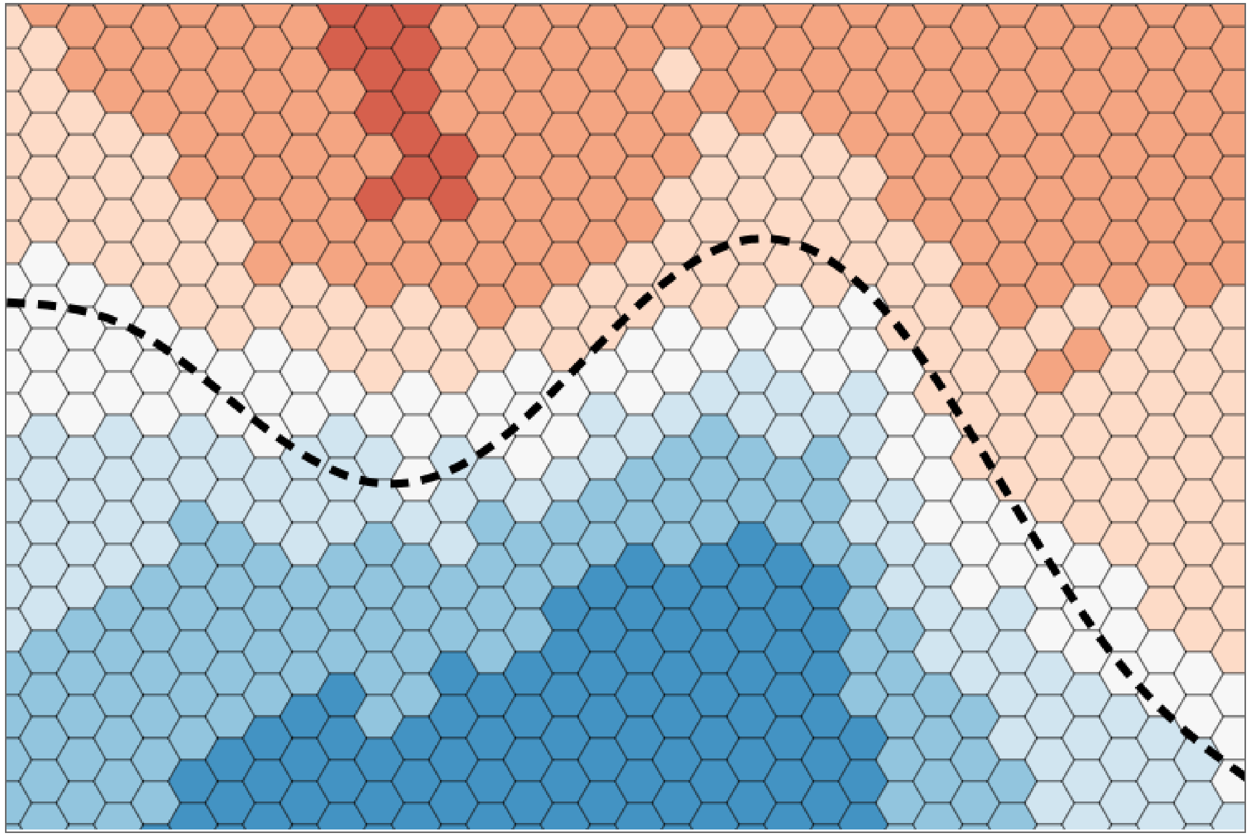


k-Nearest Neighbors: 3

Pictures taken from:
http://scott.fortmann-roe.com/docs/BiasVariance.html

חזרה ל-KNN+Bias-Variance Tradeoff

משטחי ההחלטה שנובעים מאלגוריתם 50-nn:

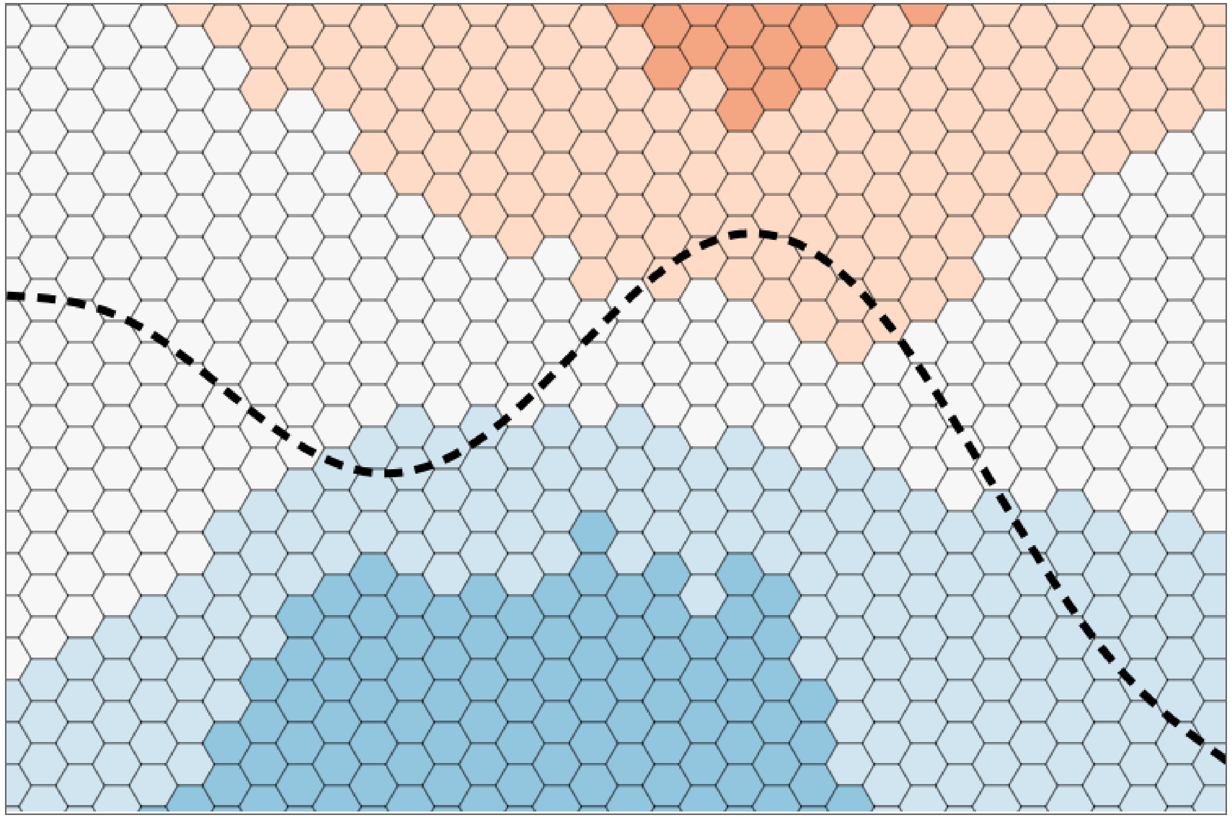


k-Nearest Neighbors: 50

Pictures taken from:
http://scott.fortmann-roe.com/docs/BiasVariance.html

חזרה ל-KNN+Bias-Variance Tradeoff

משטחי ההחלטה שנובעים מאלגוריתם 100-nn:



k-Nearest Neighbors: 100

Pictures taken from:
http://scott.fortmann-roe.com/docs/BiasVariance.html

Breast Cancer Wisconsin - Dataset

שיטה נפוצה היום להבחנה של סרטן היינה בשיטת **Fine-needle aspiration.**

בשיטה זאת נלקחת דגימה של הרקמה ובעזרת מחט ומבוצעת אנליזה בעזרת מיקרוסקופ על מנת לאבחן שני מקרים:

- רקמה סרטנית – Maligant
- רקמה לא סרטנית - Benign

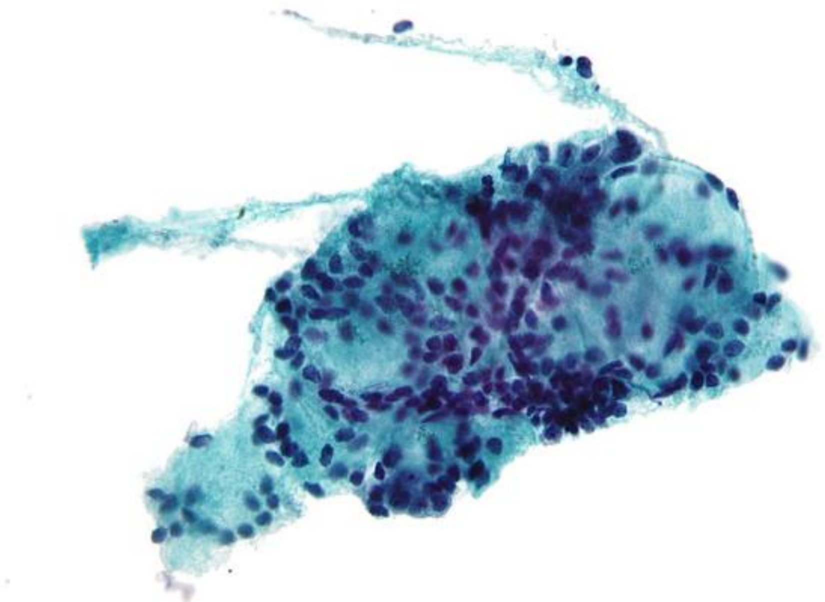


Image taken from Wikipedia

מבנה ה - DataSet

- המידע מאופיין על ידי 30 ערכים נומריים כגון שטח רקמה ממוצעת.
- המידע חולץ מתוך 569 דגימות שונות.
- כל דגימה מתויגת אם הרקמה סרטנית או לא.

Dataset זה בנוי למשימת קלסיפיקציה בינארית – Binary Classification

מבנה ה - DataSet

- 10 הרשומות הראשונות (לא באופן מלא) של ה- Dataset נראות כך:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	...
5	843786	M	12.45	15.70	82.57	477.1	0.12780	0.17000	0.15780	0.08089	...
6	844359	M	18.25	19.98	119.60	1040.0	0.09463	0.10900	0.11270	0.07400	...
7	84458202	M	13.71	20.83	90.20	577.9	0.11890	0.16450	0.09366	0.05985	...
8	844981	M	13.00	21.82	87.50	519.8	0.12730	0.19320	0.18590	0.09353	...
9	84501001	M	12.46	24.04	83.97	475.9	0.11860	0.23960	0.22730	0.08543	...

10 rows × 32 columns

אנליזה ראשונית

לשם פשטות, רק לצורך הצגת התוצאות נבחן את שני המאפיינים הבאים (features) :

- **Radius_mean** – הרדיוס הממוצע של התאים בדגימה.
- **Texture_mean** – סטיית תקן הממוצעת של רמות האפור של התאים בדגימה.

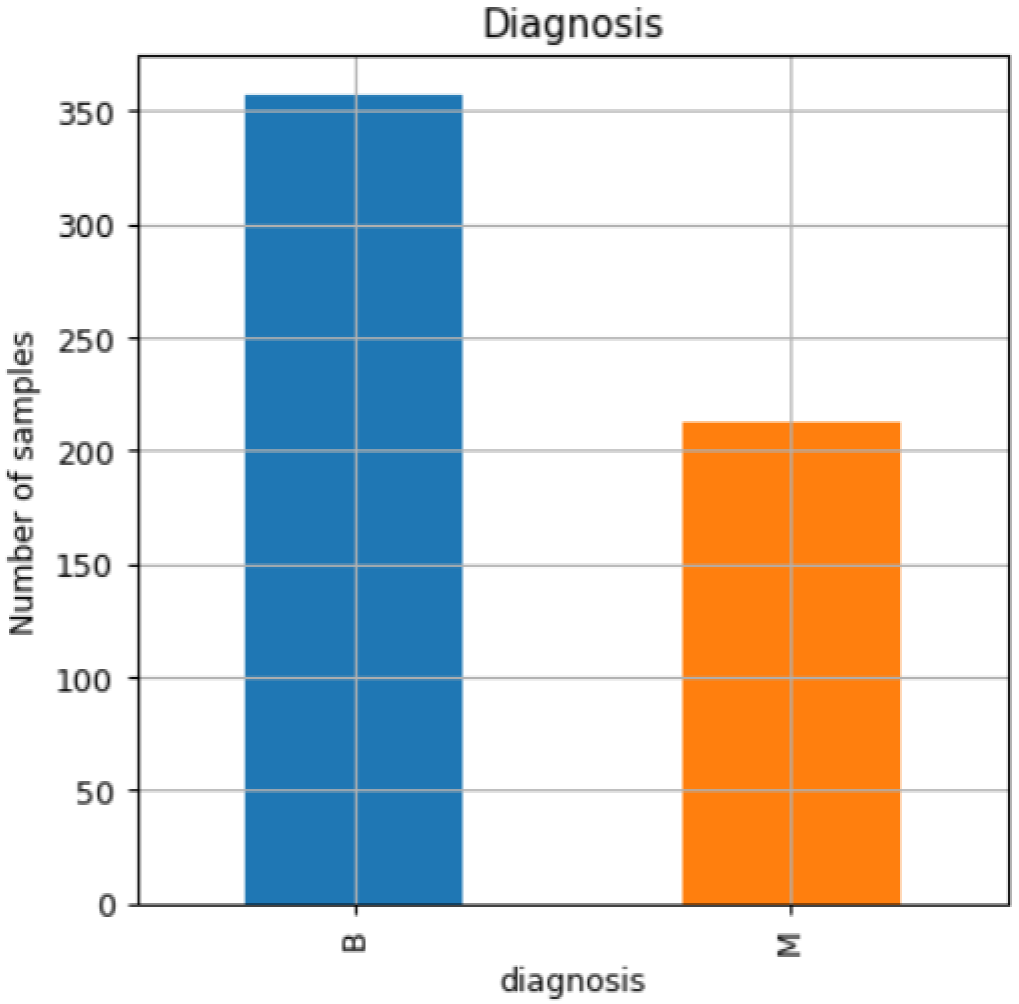
$$x_i = \{\textbf{Radius_mean, Texture_mean}\}$$

והתיוג - y_i:

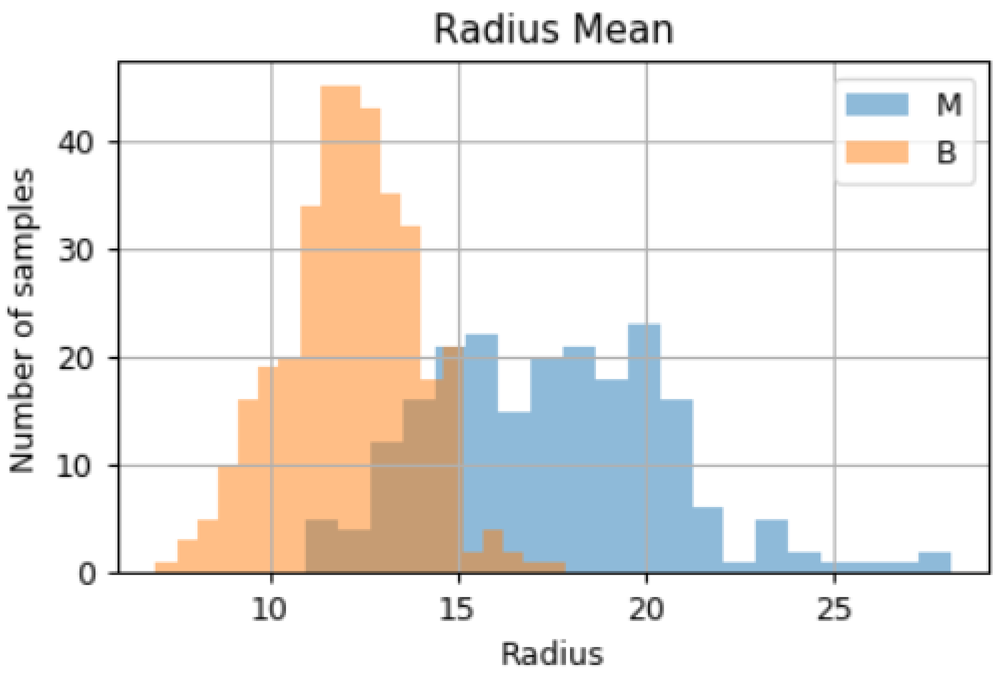
- **Diagnosis** – מציין את התיוג, האם הרקמה סרטנית.

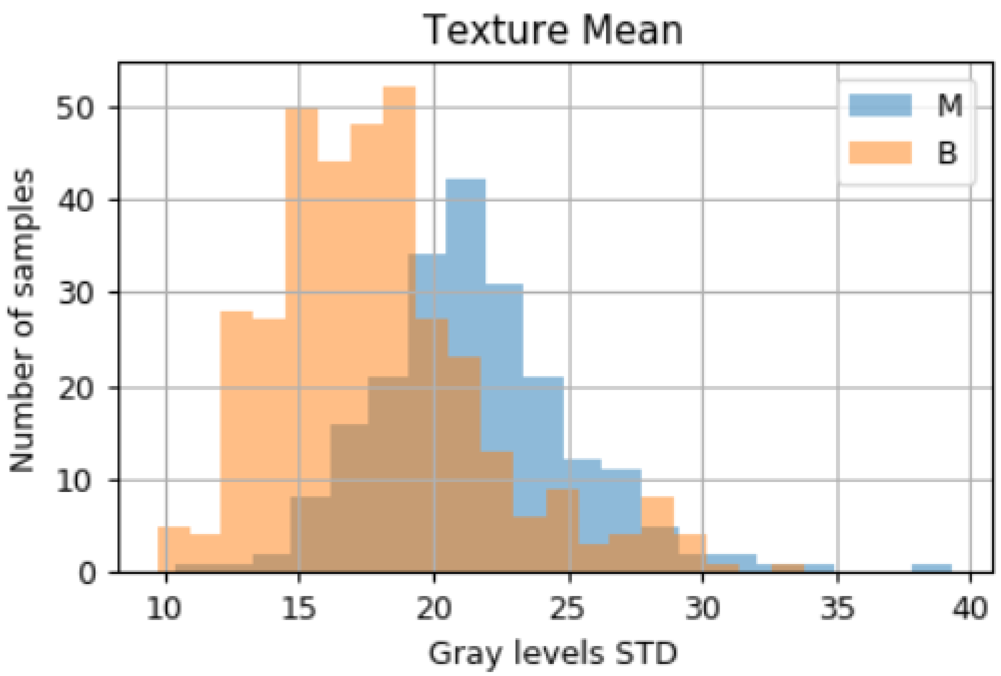
<u>הערה:</u> בפועל מבצעים לימוד על כל המאפיינים.

מספר התאים הסרטניים והלא סרטניים

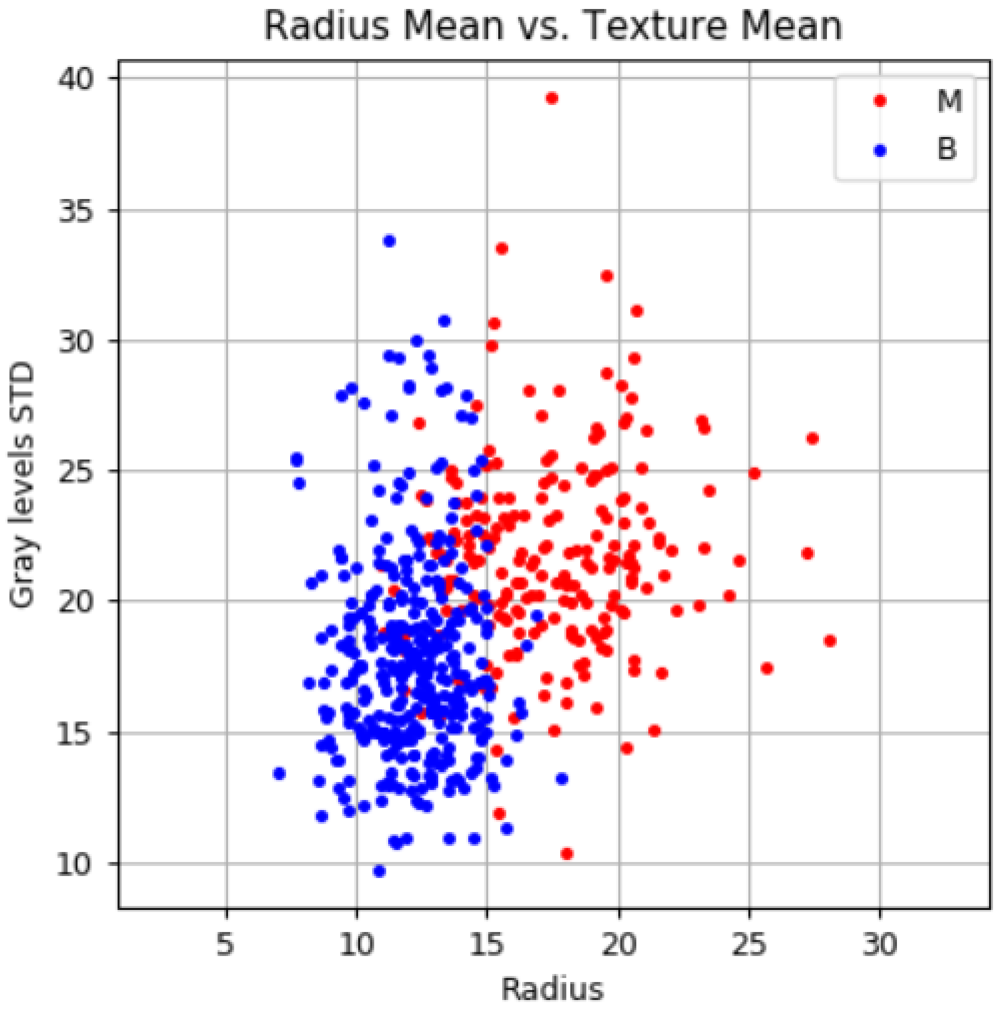


ההתפלגות של המאפיינים





ההתפלגות של המאפיינים



הערכת ביצועים וחלוקת הdata

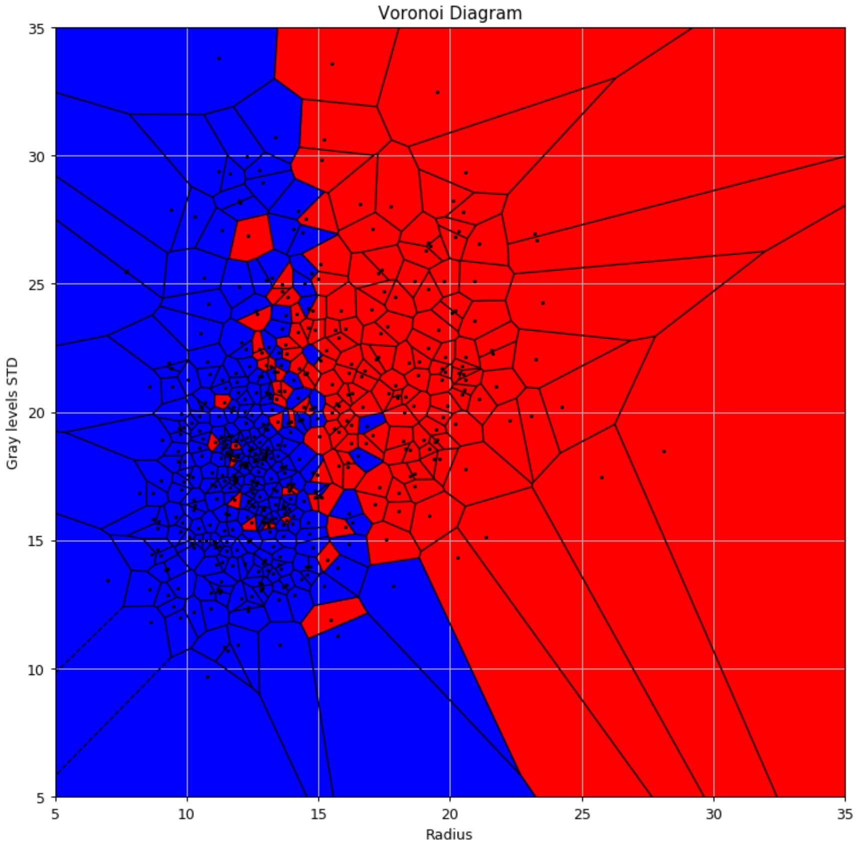
- Misclassification Rate:

$$R\{h,\{\boldsymbol{x},y\}\} = \frac{1}{N}\sum_i I\{h(\boldsymbol{x}_i) \neq y_i\}$$

- נחלק את ה Dataset לשניים
  - Train set שהיינו 80% מכלל ה Dataset
  - Test set שהיינו 20% מכלל ה Dataset

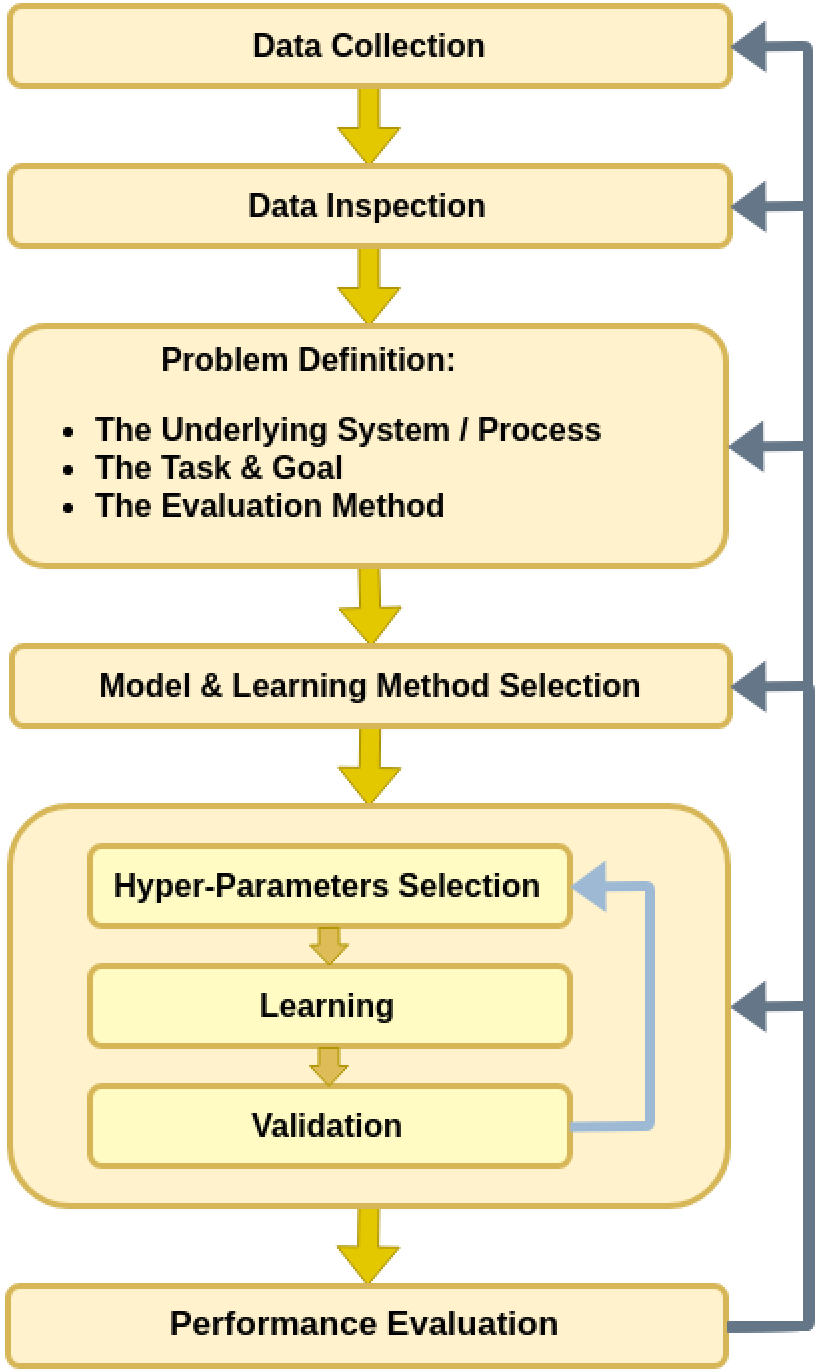
Model: 1-NN

נאמן את המודל השכן הקרוב ונציג את אזורי ההחלטה באמצעות פונקציית Voroni מחבילת SciPy.



הערכת ביצועים:

Test set risk: 0.14



שיפור ביצועים

מה ניתן לעשות על מנת לשפר ביצועים?

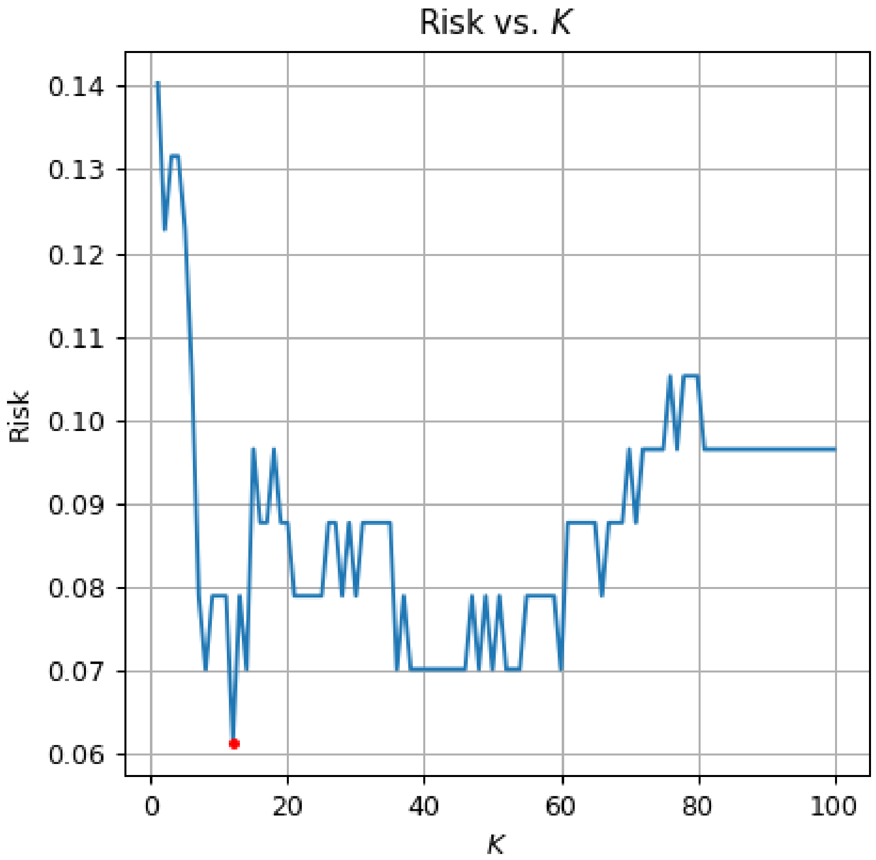
- ניתן לשנות את ההיפר-פרמטרים, כלומר את משפר בשכנים - K.

שיטות לכיוון היפר-פרמטרים:

- Brute Force / Grid Search
- ניסוי וטעיה
- Cross-validation

Grid Search

הערכת ביצועים:
Test set risk: 0.061
$K = 12$



Train-Validation-Test split / 3-fold split

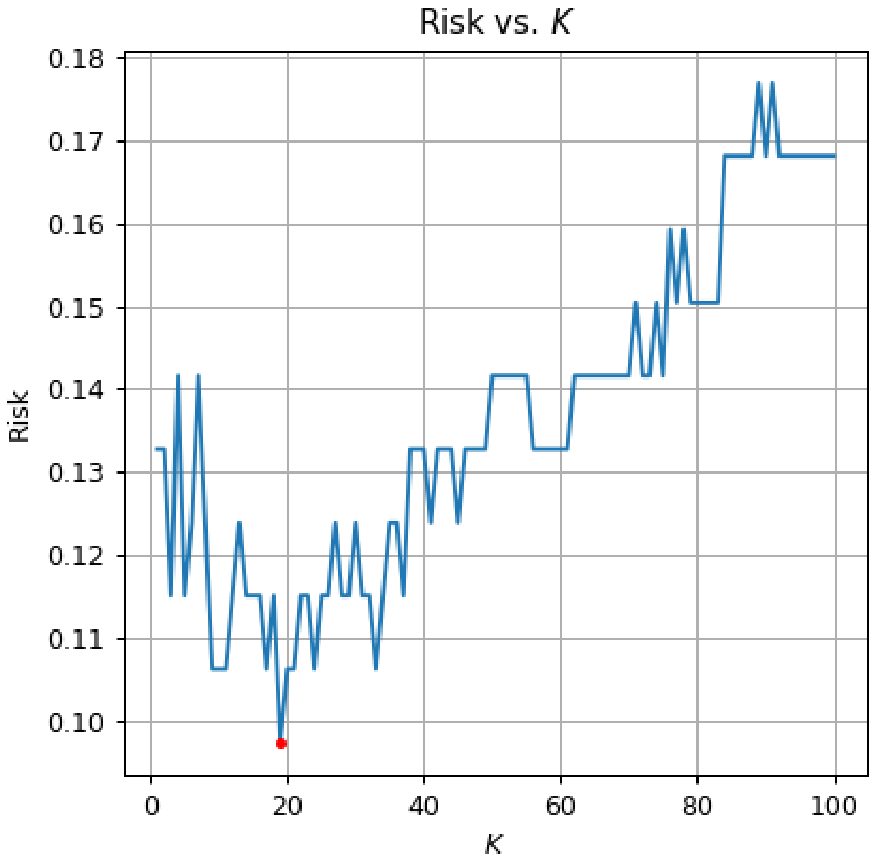
על מנת להימנע מהתאמת ייתר (over fitting) נחלץ מתוך הdataset מקטע לוולידציה.

נהוג לחלק ל:

- 60% train set
- 20% validation set
- 20% test set

הערה: ביצוע כיוון פרמטרים על הtest set מוביל לover fitting, והערכת ביצועים מוטעית

3-fold split



הערכת ביצועים:
Test set risk: 0.097
$\boldsymbol{K = 19}$

Cross Validation

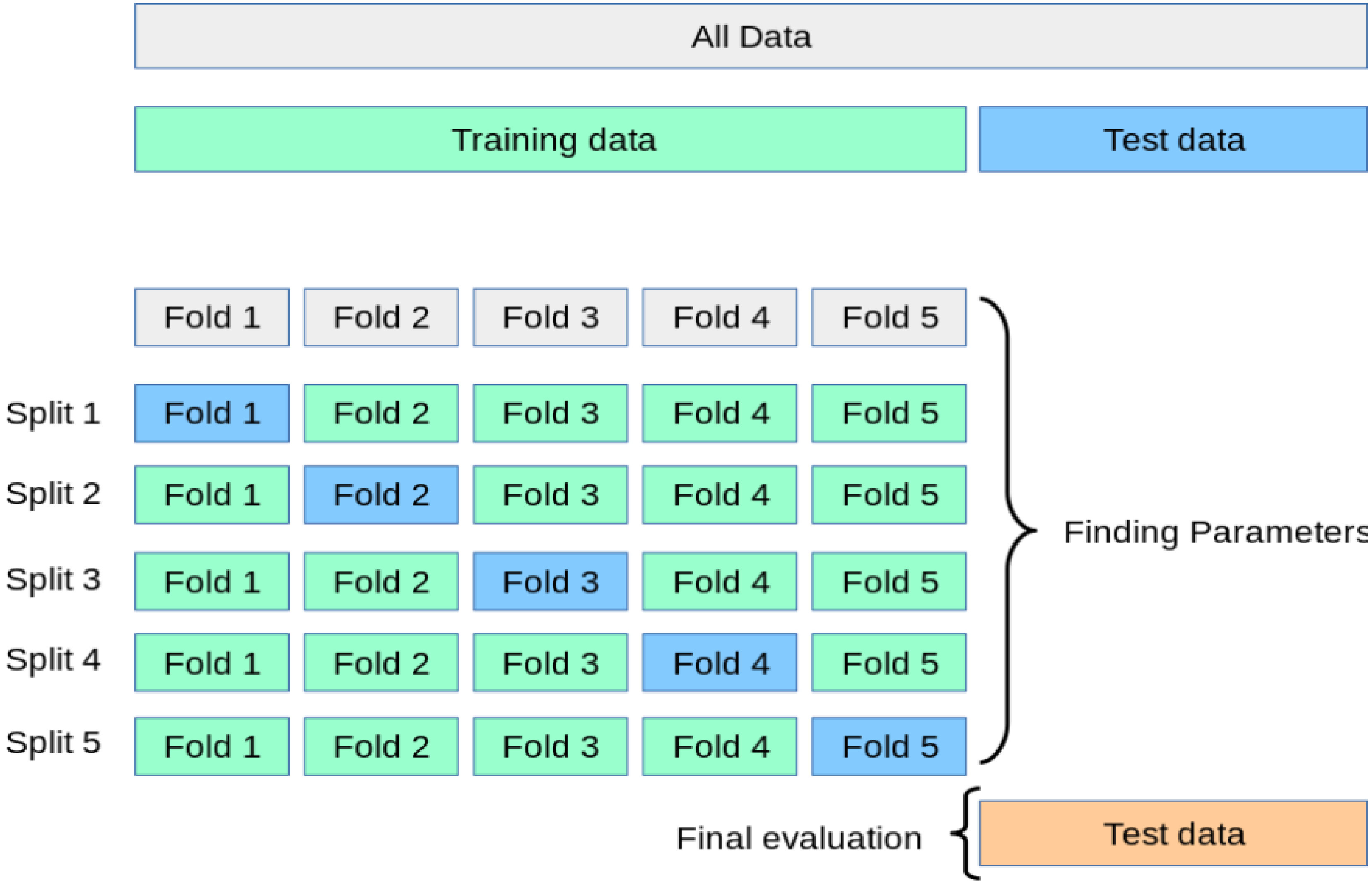
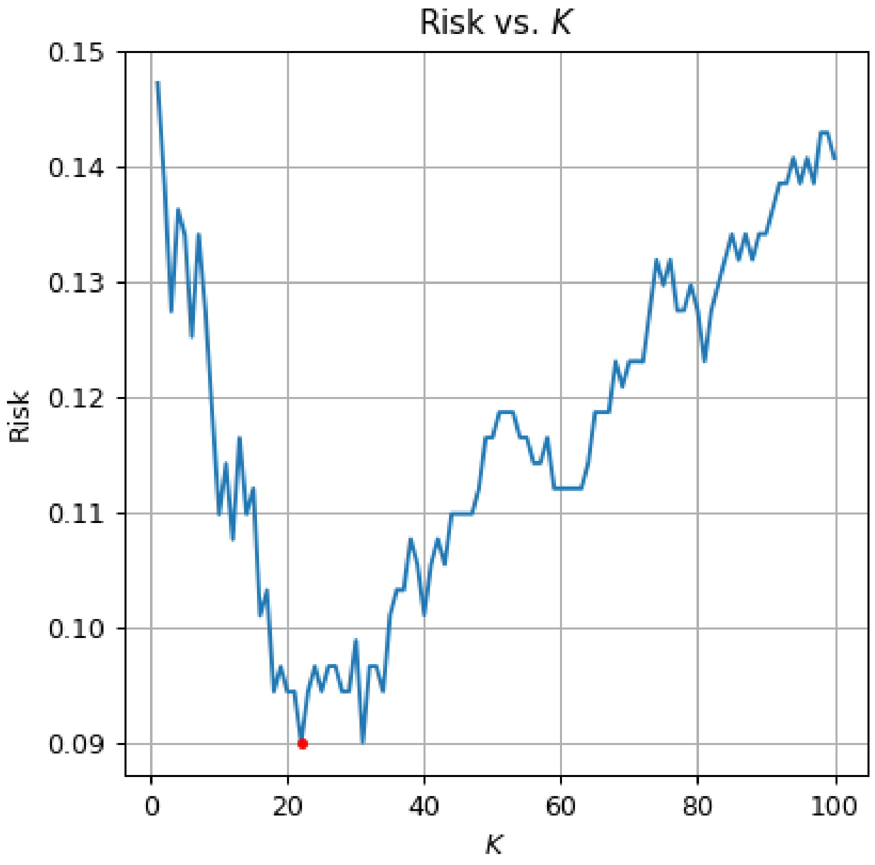


Image taken from SciKit Learn

Cross Validation

הערכת ביצועים:
Test set risk: 0.079
$\boldsymbol{K = 22}$