תרגול מספר 5 – מבוא ללמידה מודרכת ו-KNN

1 תקציר התאוריה

1.1 סימונים

$D=\{x_i,y_i\}_{i=1}^n$ - דוגמאות (דגימות - $samples$) בלתי תלויות של X עם פרדיקציה ידועה Y.

X – מרחב הקלט. $x\in X$.

מרחב הפלט תלוי במשימה. נבחין בין שתי משימות שונות:

Classification –

המטרה היא סיווג דוגמאות לאחת מבין מספר סופי של מחלקות אשר הגדרנו מראש.

לדוגמא : זיהוי חתול וכלב מתמונה.

Ω – מרחב סופי של קטגוריות (מחלקות) $\omega_i\in\Omega,\ i=1,...,N$.

$f:X\to\mathbb{R}$ - חזאי אשר מסווג כל $x\in X$ ל-$y\in\Omega$.

Regression –

המטרה היא מציאת חזאי $f:X\to\mathbb{R}$ אשר מקיימת את הקשר הבא $f(x_i)=y_i$ לכל צמד $x\in X, y\in\mathbb{R}$. הפלט y הינו ערך מסוים שאנו רוצים לחזות באמצעות החזאי f ובהינתן קלט x.

לדוגמא: מתן תחזית לערך מנייה מסוימת על סמך נתוני הבורסה.

$f:X\to\mathbb{R}$ - מסווג כל $x\in X$ ל-$y\in\mathbb{R}$.

1.2 ולידציה

Training Set (סט אימון) – סט דוגמאות מתויג $D_{train} = \{x_i, y_i\}_{i=1}^{n}$ שבאמצעותו האלגוריתם לומד.

Validation Set (סט אימות) – סט דוגמאות מתויג $D_{validation} = \{x_i, y_i\}_{i=1}^{n}$ שבאמצעותו נעריך את טיבם של המודלים, על מנת לבחור ביניהם.

Test Set (סט בחן) – סט דוגמאות מתויג $D_{validation} = \{x_i, y_i\}_{i=1}^{n}$ שבאמצעותו נעריך את ביצועי המודל הסופי שבחרנו. נשים לב שהשימוש בסט זה הינו השלב האחרון בתהליך הלמידה, **ואין להשתמש בו כדי להעריך את ביצועי המודל במהלך הלימוד.**

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: $D_1, \ldots, D_k$,

   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 $\hat{L}_n^{(j)}(M)$

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

סיווג בעזרת אלגוריתם K-NN ($\boldsymbol{k}$ Nearest Neighbours)

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

מערכות לומדות - 046195 | הפקולטה להנדסת חשמל

חורף תש״פ (2020) | הטכניון - מכון טכנולוגי לישראל

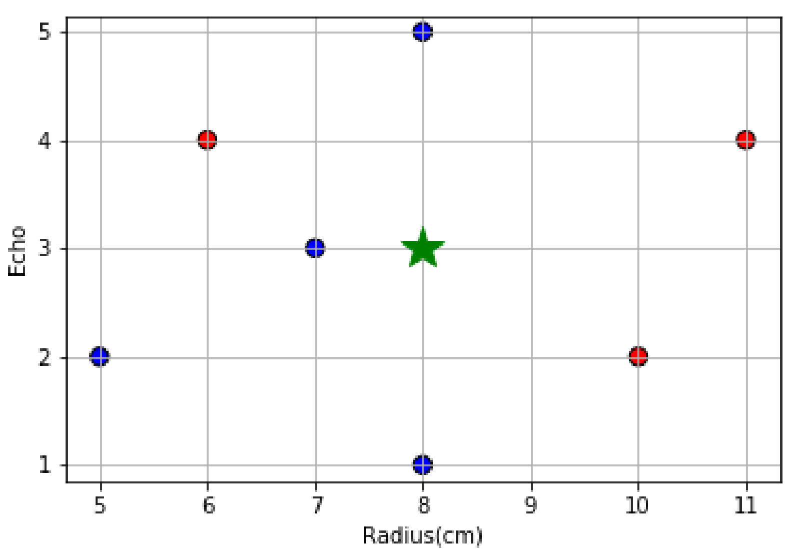
תרגיל 1

סטודנט נבון ניגש לבחור אבטיחים בסופרמרקט. ידוע כי זוהי רק תחילתה של עונת האבטיחים וקיים מספר לא מבוטל של אבטיחי בוסר הסטודנט שם לב כי ניתן לאפיין את האבטיחים ע"פ ההד בהקשה וע"פ קוטר האבטיח. הסטודנט החליט למפות את ניסיון העבר שלו:

1. הד חזק (עוצמה 1), רדיוס 8 ס"מ – מתוק
2. הד בינוני (עוצמה 2), רדיוס 10 ס"מ – חמוץ
3. הד בינוני (עוצמה 2), רדיוס 5 ס"מ – מתוק
4. הד חלש (עוצמה 3), רדיוס 7 ס"מ – מתוק
5. הד רפה (עוצמה 4), רדיוס 6 ס"מ – חמוץ
6. הד רפה (עוצמה 4), רדיוס 11 ס"מ – חמוץ
7. הד עמום (עוצמה 5), רדיוס 8 ס"מ – מתוק

הסטודנט מחזיק בידו האבטיח בעל הד חלש רדיוס 8 ס"מ. האם סביר שהאבטיח מתוק או חמוץ?

א. בדקו את תוצאות ה-classification עבור k-nearest neighbors, כאשר K=1,3.
ב. בצע Cross Validation להערכת טיב המודל, עבור K=1,3 ו-7 קבוצות (Leave-one-out Cross Validation). באיזה מסווג נבחר?
ג. מה יקרה אם נבחר את k להיות בגודל ה- dataset.
ד. סטודנטית נבונה (אף יותר!) העירה לסטודנט כי קוטר האבטיח אינו משנה, וכי עליו להתייחס אך ורק להד. חזרו על התהליך במקרה זה.



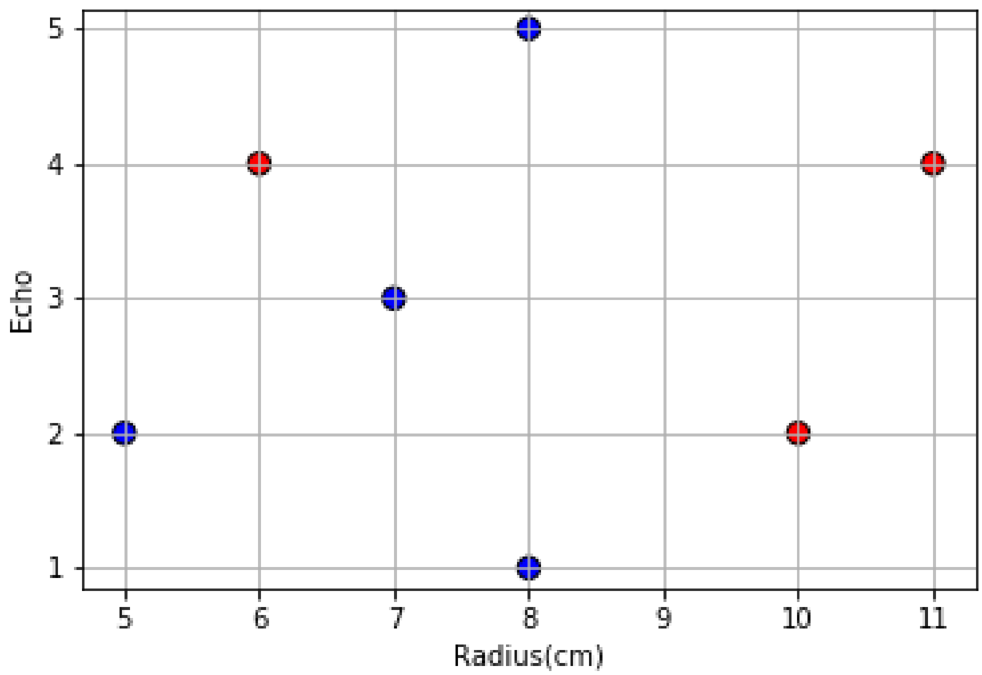
<u>פתרון</u>

א. נמפה את הנתונים על גרף, כאשר ציר x הינו הד האבטיח מ-1 (חזק) ל-5 (עמום).
כל נקודה ניתן לרשום כ $(x_i;w_i)$ כאשר $i=1,...,7$, הנקודה הנבדקת היא $(\bar{x},\bar{w})$.

בבדיקה ישירה במרחק אוקלידי: השכן הקרוב ביותר הוא x_4 ולכן נעניק לנקודה הנבדקת $\bar{w}=w_4=sweet$.

שלושת השכנים הקרובים ביותר הם $\bar{w}=w_4=sweet$ וע"פ הצבעת רוב $\bar{w}=sour$.

ב. כאמור ה- Data שלנו הינו:



נזכור שבמצב שהמרחקים שווים, נבחר צבע באקראי. במקרה זה נניח שתמיד שובר השוויון הינו כחול, שכן יש לנו יותר נקודות כחולות.

נבדוק כל אחת מהנקודות:

$\underbrace{(5,2),blue}_{X,Y}: K=1 \to \underbrace{blue(random)}_{prediction},\ K=3 \to \underset{prediction}{blue}$

$(6,4),red: K=1 \to blue,\ K=3 \to blue$

$(7,3),blue: K=1 \to red\ ,\ K=3 \to blue$

$(8,1),blue: K=1 \to blue(random)\ ,\ K=3 \to blue$

$(8,5),blue: K=1 \to blue(random),\ K=3 \to red$

$(10,2),red: K=1 \to blue\ ,\ K=3 \to blue$

$(11,4),red: K=1 \to red\ ,\ K=3 \to blue$

נחשב את שגיאת ה- CV:

עבור K=1, מספר השגיאות הוא 3, לכן שגיאת ה- CV הינה: $\hat{L}_n^1 = \frac{3}{7}$

עבור K=3, מספר השגיאות הינו 4, לכן שגיאת ה- CV הינה $\hat{L}_n^1 = \frac{4}{7}$.

קיבלנו שעבור k=1 שגיאת המסווג קטנה יותר מאשר k=3, לכן – במקרה זה נבחר במסווג בעל k=1.

נשים לב שבמקרה זה תוצאות שני המסווגים גרועות ביותר.

ג. במצב כזה ההחלטה שלנו תקבע ישירות לפי באיזה מהמחלקות יש יותר דוגמאות (במקרה זה, לכל דוגמא חדשה יתקבל הצבע הכחול – כל האבטיחים מתוקים!)

Bias-Variance Tradeoff:

כפי שנלמד בהרצאה, ניתן לרשום את שגיאת הלימוד באופן הבא:

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

F - משפחת המודלים ממנה נבחר את המודל הטוב ביותר.

$\hat{f}$ - הינו המודל הנלמד מתוך D.

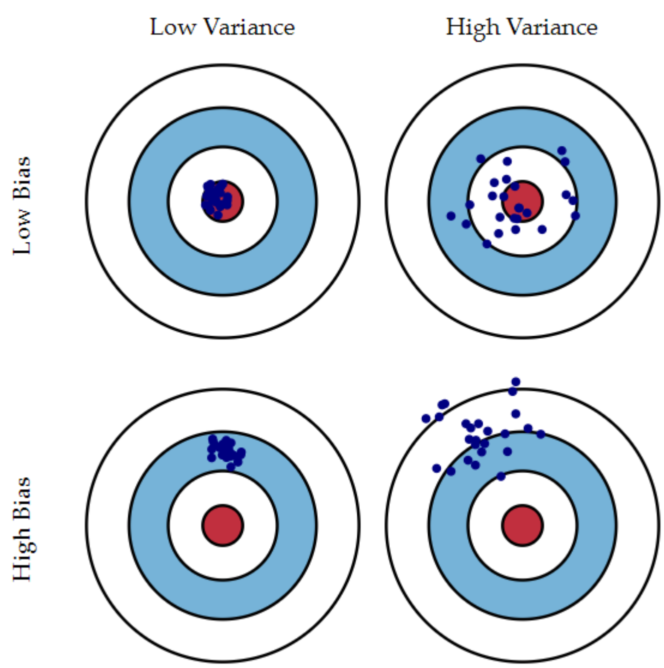
המונח Variance מתייחס לעושר של משפחת המודלים F. בהינתן D סופית, ככל שמשפחת המודלים גדולה יותר, ניתן לבחור מודל אשר מתאר את ה Data טוב יותר. לכן, שגיאת הקירוב $E_{app}\left(F\right)$ תקטן. מצב זה רצוי, שכן אנחנו מקטינים את אחד משני חלקי השגיאה.

עם זאת, ככל שבחירת המודל $\hat{f}$ תהיה קרובה יותר ב- Data, שגיאת השערוך $E_{est}\left(\hat{f},F\right)$ גדלה, שכן בחירת מודל אשר מתאים בייתר פירוט ל- DATA תסביר גם את הרעש שבו. כאן, בא לידי הביטוי ה- Tradeoff:

מודל פשוט מקשה עלינו למצוא התאמה ל- DATA, אך מאפשר הכללה לדוגמאות חדשות.

לעומתו, מודל מסובך מאפשר התאמה טובה ל- DATA, אך מביא ל- Overfitting, כלומר לפגיעה בשגיאת השערוך.

נציג דוגמא ויזואלית אשר מסבירה בצורה קונספטואלית את ה- Bias-Variance Tradeoff:



הפגיעה במרכז המטרה מציג ה- Bias (שגיאת הקירוב), הפיזור מציג את ה- Variance (שגיאת השערוך).

דוגמא מסכמת:

בשאלה זו ננסה לחזות את בחירתו של אזרח אמריקאי באמצעות אלגוריתם K-NN.

לשם הפשטות, נניח כי כל אזרח מיוצג על ידי שני מאפיינים:

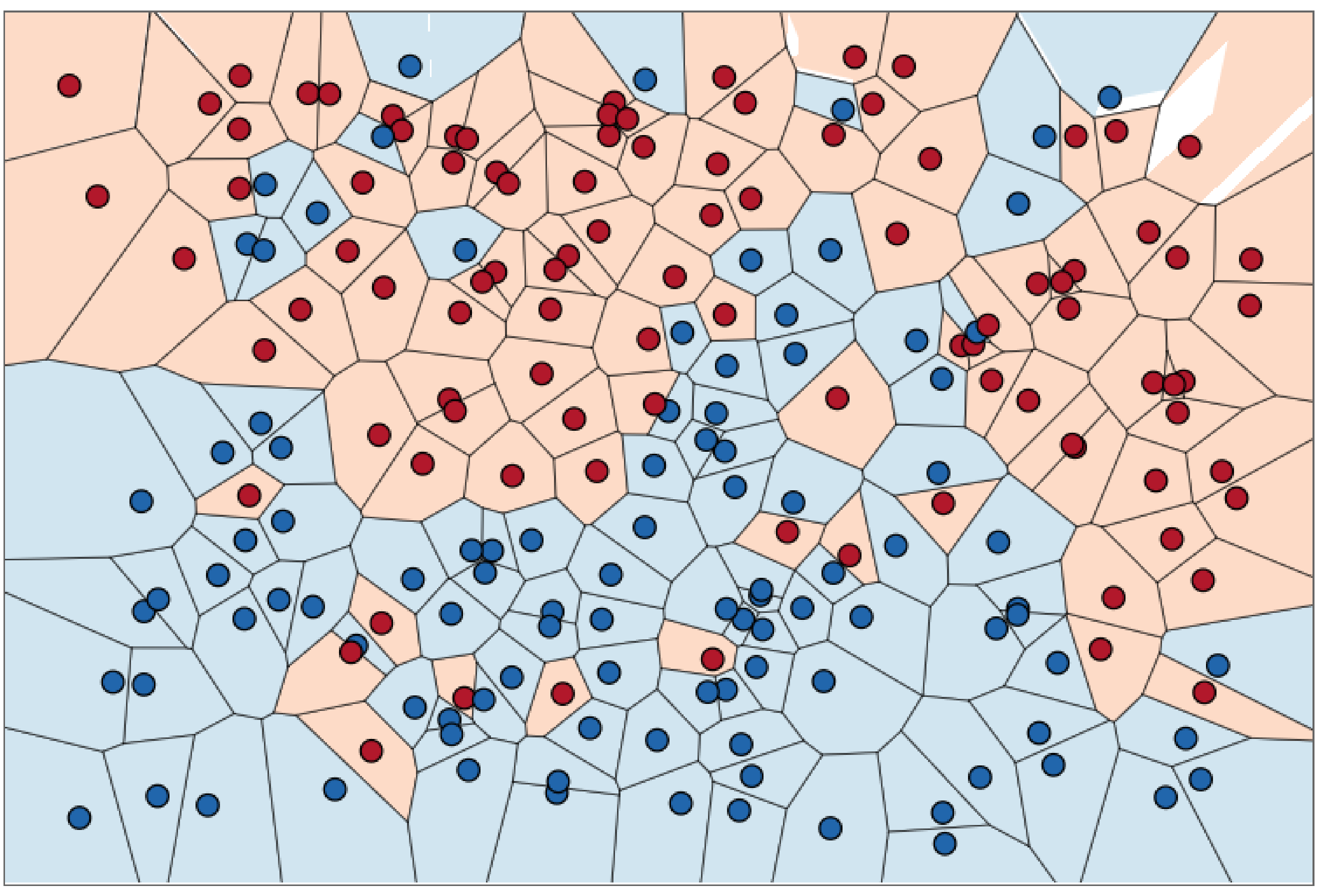
מצבו הכלכלי (הציר האופקי - x) וקרבתו לדת (הציר האנכי – y).

בסימונים שלמדנו:

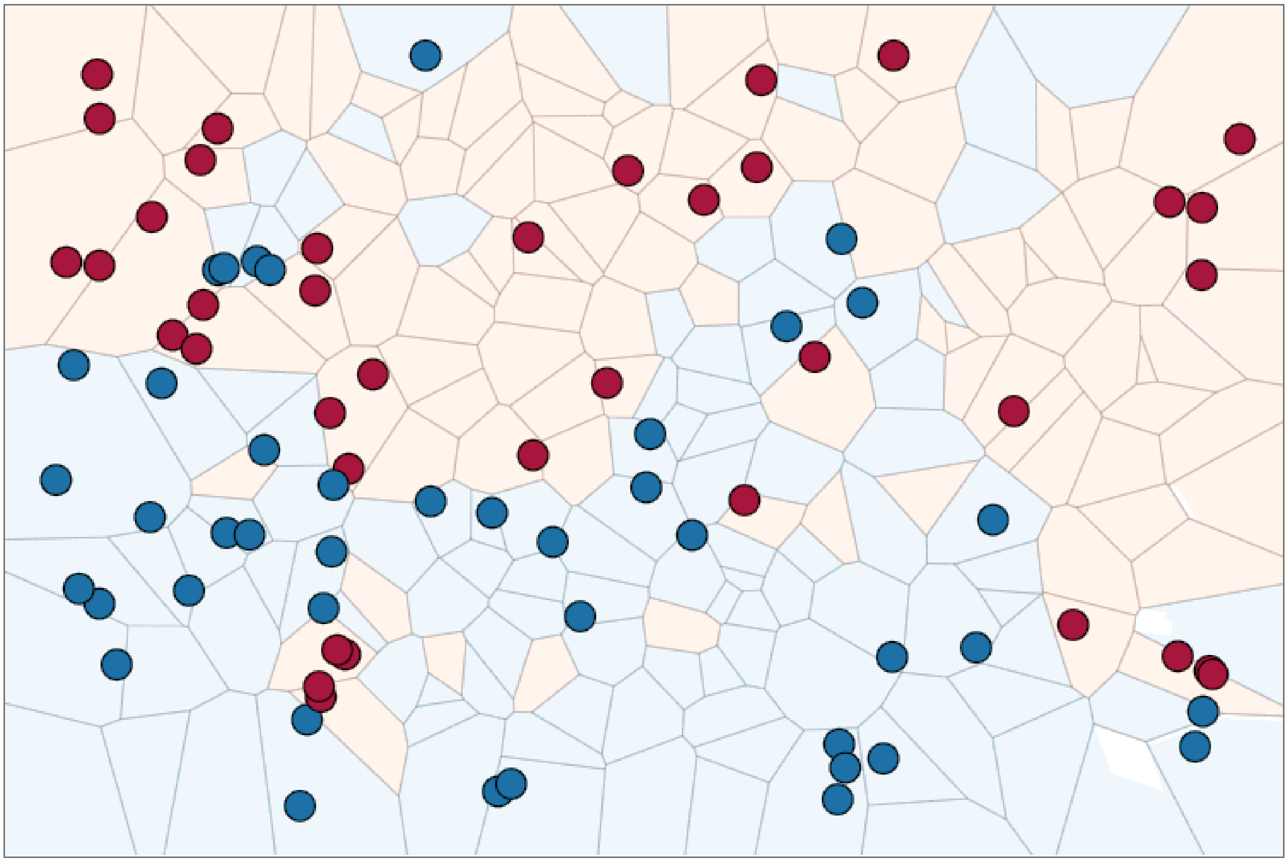
$$x=\left(wealth,religiousness\right)\in X=\mathbb{R}_{+}^{2}$$

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

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

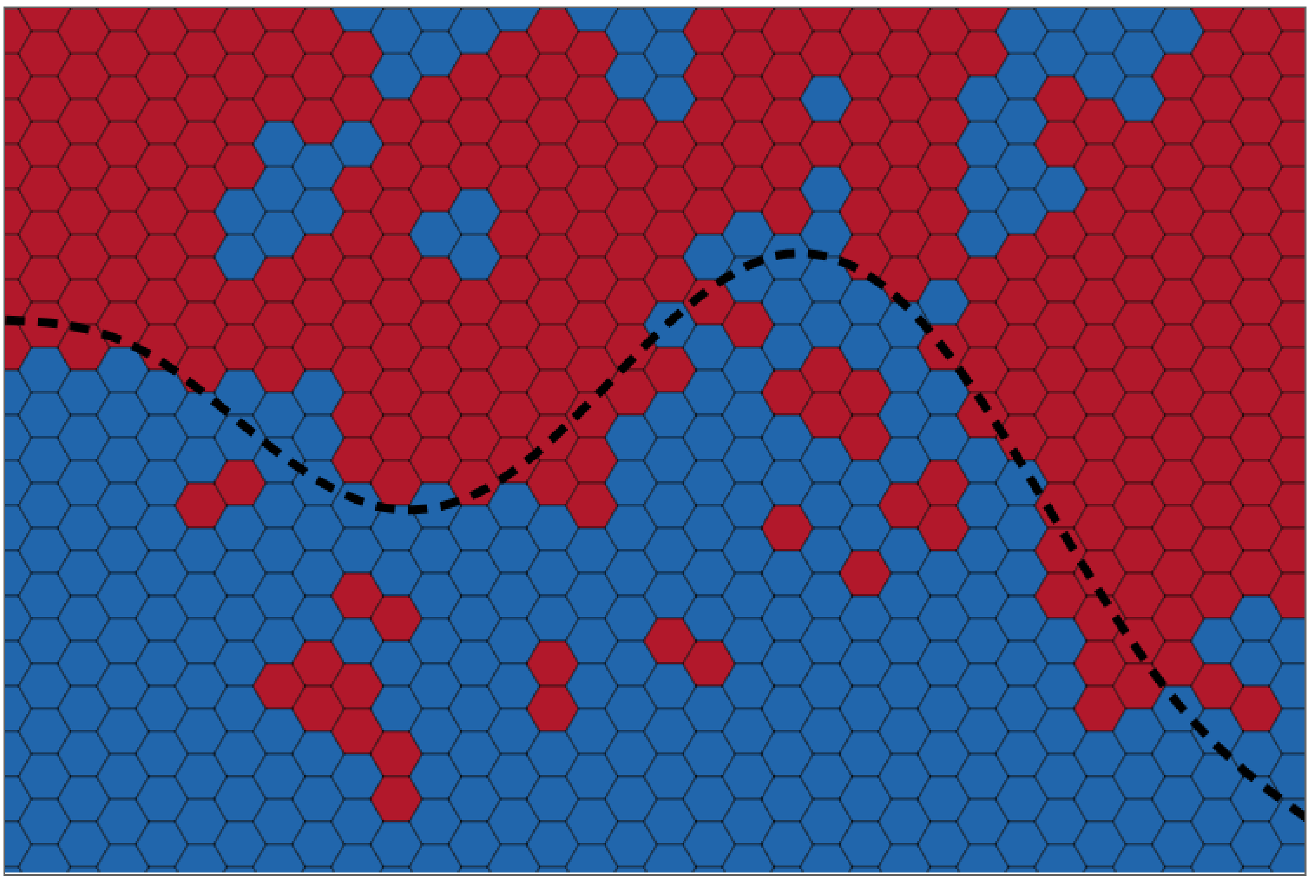


כעת, נבצע פרדיקציה לסט בחן לדוגמא:



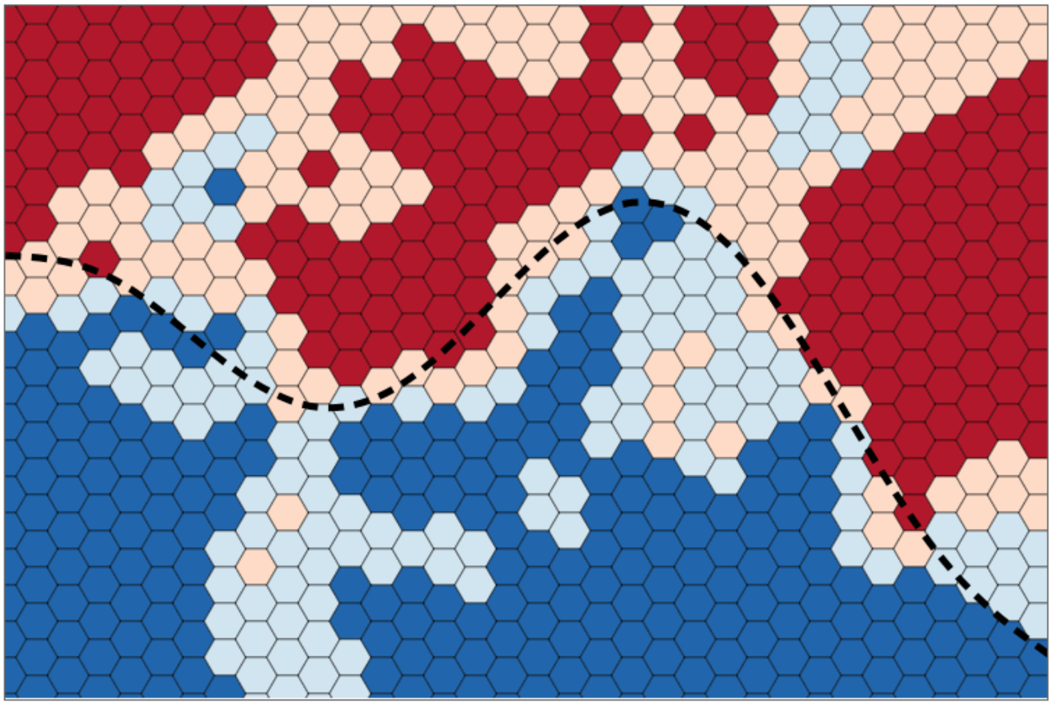
שאלה: איזה משפחה F יותר עשירה? תוצאת סיווג עבור 1-nn או עבור 50-nn?

נראה זאת ויזואלית:
נסתכל על סט אימון גדול מאוד ונניח שהוא מתפזר בצורה אחידה על מרחב המאפיינים.

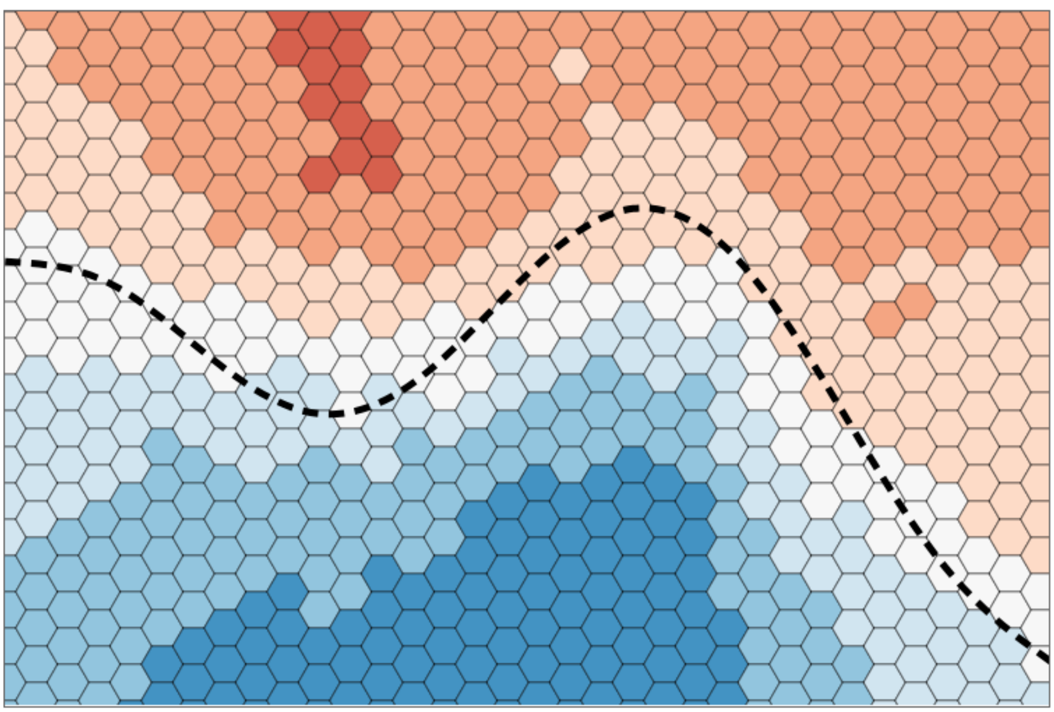


k-Nearest Neighbors: 1

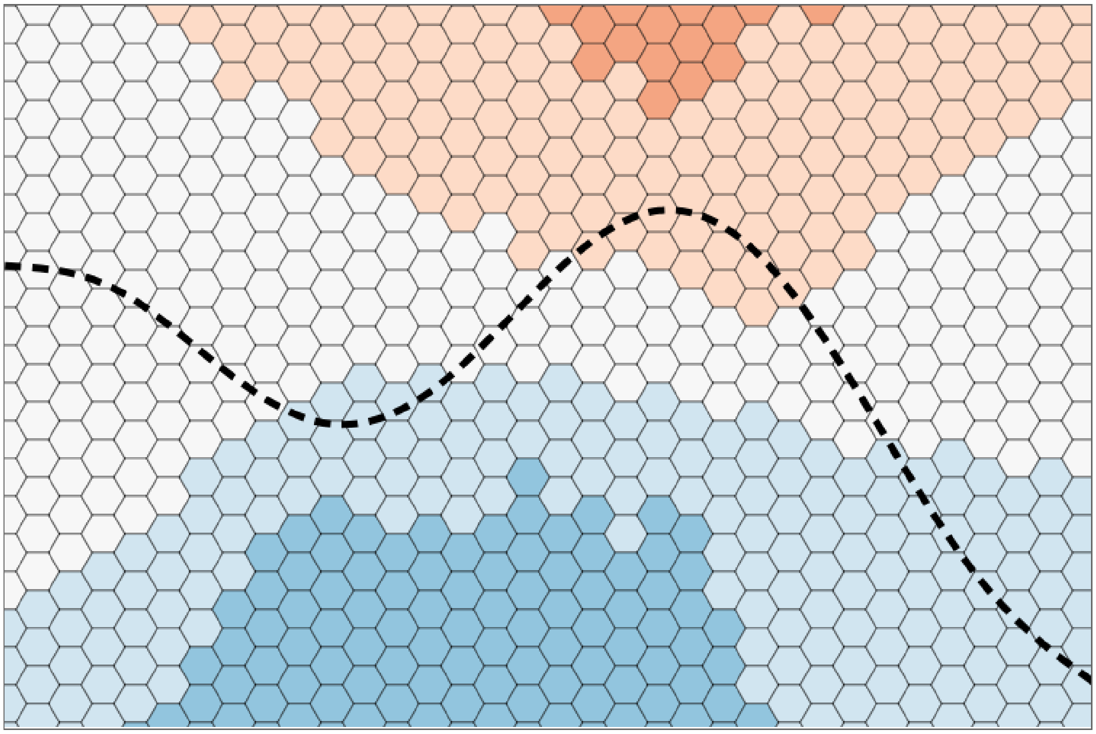
הקו השחור מייצג את קו ההחלטה שלפיו יוצרו הדוגמאות, לפני הוספה של רעש תיוג אקראי.



k-Nearest Neighbors: 3



k-Nearest Neighbors: 50



k-Nearest Neighbors: 100

ניתן לראות שככל שנגדיל את מספר השכנים עליהם נסתכל בעת הסיווג משטחי ההחלטה יהפכו ליותר ויותר חלקים, כלומר המודלים יהיו פשוטים יותר. במצב זה, אנחנו מאבדים עם הגדלת k אזורים מיוחדים וקטנים. יש לכך יתרון במידה ואזורים אלו נובעים מרעש, אך ייתכן Data שאינו פריד לינארית וקיימים בו "מובלעות" החלטה.
עם זאת, מידת הביטחון שלנו בפרדיקציות אשר קרובות לאזור ההחלטה ירדו, מאחר שיהיו שכנים רבים באופן יחסי מהמחלקה השנייה.

מה יקרה אם k=n?
כל נקודה חדשה תסווג למחלקה השכיחה יותר בסט האימון. אם שני הסטים זהים בגודלם, כל נקודה תסווג באופן אקראי לחלוטין.

הערה:
התמונות נלקחו מאתר http://scott.fortmann-roe.com/docs/BiasVariance.html, מומלץ לקרוא את ההסבר המלא.