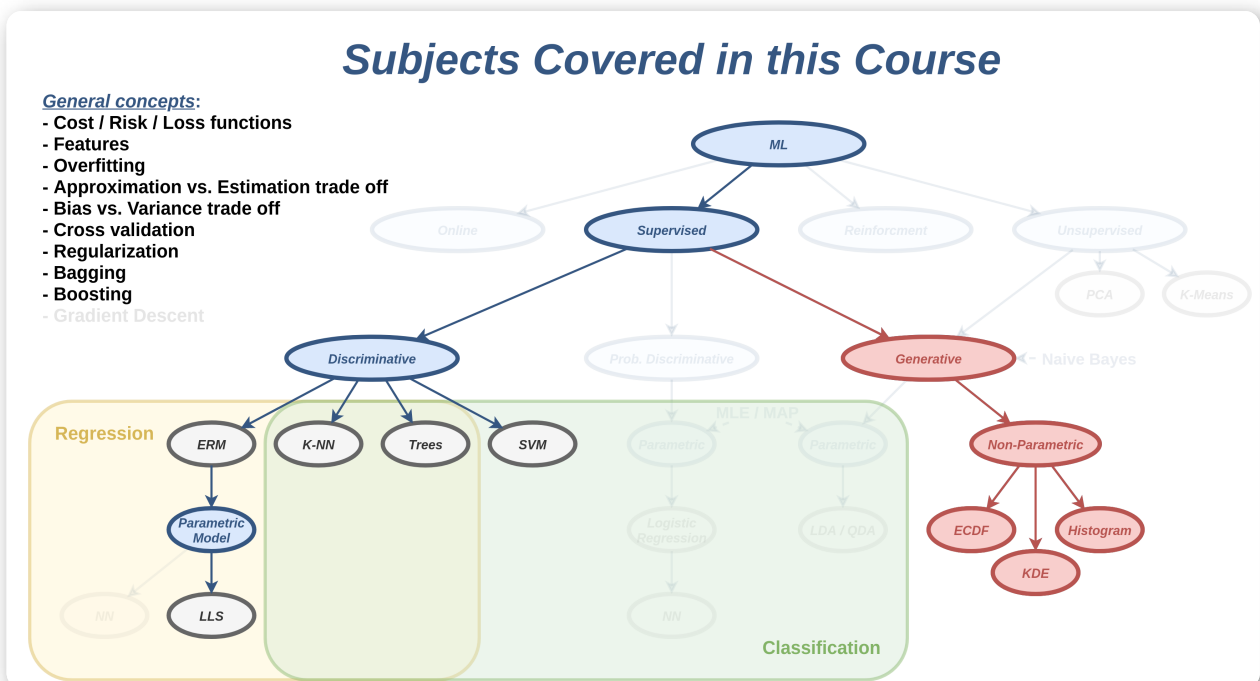


הרצאה 7 - שיערוך פילוג

בשיטות לא פרמטריות

Slides PDF Code

מה נלמד היום



הגישה הגנרטיבית

דיסקרימינטיבי vs. גנרטיבי

עד כה, עסקנו בשיטות לפתרון בעיות supervised learning אשר פעלו תחת הגישה הדיסקרימינטיבית שבה ניסינו באופן ישיר למצוא חזאי אשר יתאים למדגם. בשלושת השבועות הקרובים אנו נכיר גישה אחרת לפתרון בעיות supervised learning אשר נקראת הגישה הגנרטיבית.

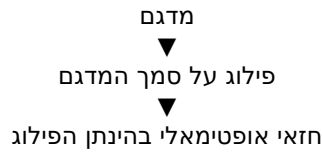
כפי שציינו בעבר, ההבדל העיקרי בין בעיות חיזוי קלאסיות לבעיות supervised learning היא העובדה שאין בידינו את הפילוג של המשתנים האקראיים ובמקום זה יש בידינו מדגם מייצג שלהם. בגישה הגנרטיבית ננסה לגשר על פער זה על ידי שימוש במדגם לצורך שיערוך הפילוג של המשתנים האקראיים. בהינתן הפילוג המשוערך אנו נקבל בעיית חיזוי קלאסית אשר לרוב ניתן לפתרון בצורה פשוטה.

ננסה לתאר את ההבדל בין הגישה הדיסקרימינטיבית לגנרטיבית בעזרת השרטוט הבא:

הגישה הדיסקרימינטיבית



הגישה הגנרטיבית



הגישה הגנרטיבית מקבלת את שמה מהעובדה שהיא מנסה ללמוד את החוקיות אשר יצרה (generate) את הדגימות, בעוד שהשיטה הדיסקרימינטיבית רק מנסה להתאים לכל מדידה תווית מתאימה (discriminate).

הקשר לבעיות unsupervised learning

בקורס זה לא נעסוק כמעט בבעיות unsupervised learning אך כן ננצל ההזדמנות זו בכדי לתאר בקצרה את הקשר של שיטות גנרטיביות לבעיות מסוג זה. בבעיות unsupervised learning המדגם לא מכיל שני סוגי משתנים x ו y , אלא רק סוג בודד. לדוגמא אוסף של תמונות פנים, או אוסף של הקלטות דיבור של אדם מסויים. בבעיות מסוג זה, ננסה לרוב ללמוד מהם התכונות שמאפיינות את הדגימות במדגם. אחת הדרכים הטובות ביותר לתאר את המאפיינים של הדגימות היא על ידי שיערוך של הפילוג שלהם ואכן שיטות גנרטיביות דומות שאלו שנלמד בפרק זה משמשות גם בבעיות unsupervised learning.

שיערוך הפילוג

הבעיה של בניית מודל הסתברותי של משתנים אקראיים מתוך מדגם מכונה **בעיית שיערוך (estimation)**. את המודל ההסתברותי אנו נבטא בעזרת אחת מהפונקציות הבאות:

- פונקציית ההסתברות (probability mass function - PMF)
- פונקציית צפיפות ההסתברות (probability density function - PDF)
- פונקציית הפילוג המצרפית (cumulative distribution function CDF).

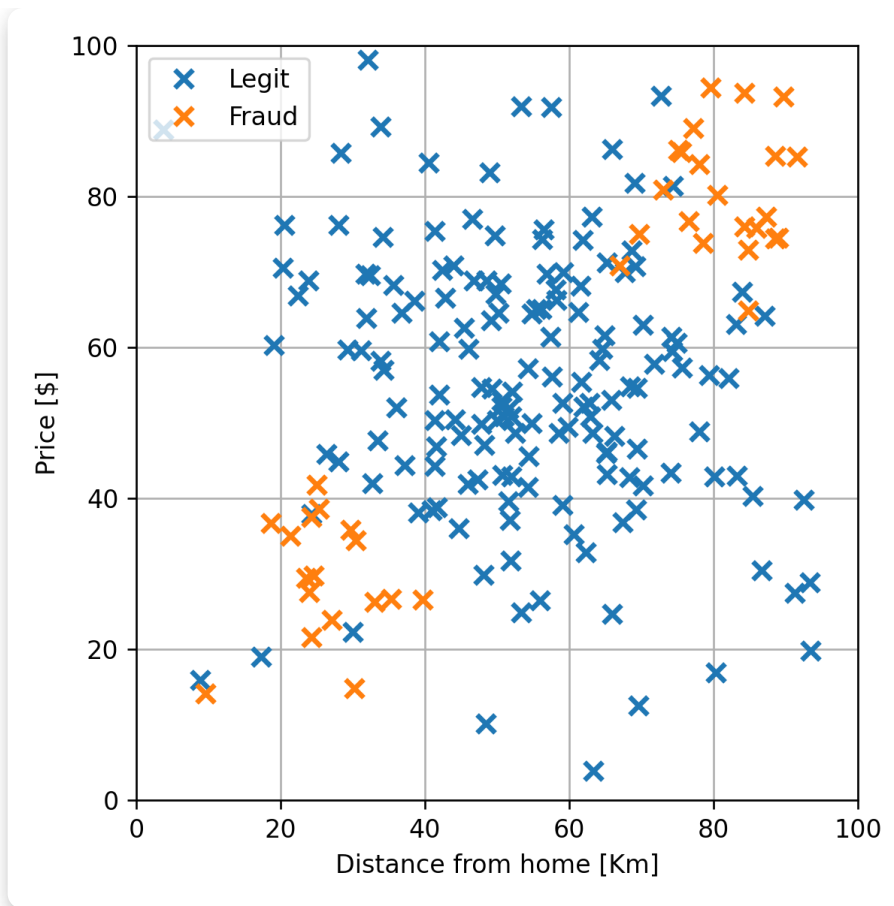
חיזוי ושיערוך

בעיות חיזוי (prediction) ובעיות שיערוך (estimation) קרובים מאד באופי, ובמקרים רבים מבלבלים בין השתיים. ננסה לחדד את ההבדלים בניהם:

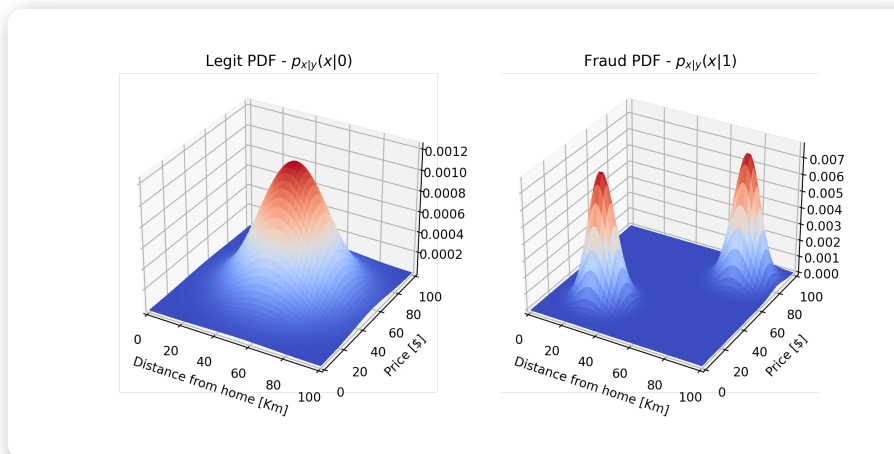
- **בבעיות חיזוי** אנו מועניינים לחזות את ערכו של **משתנה אקראי**, לרוב על סמך משתנה / וקטור אקראי בודד (**דגימה יחידה**).
- **בבעיות שיערוך** אנו מעוניינים לבנות **מודל הסתברותי** של משתנה / משתנים אקראיים לרוב על סמך **הרבה דגימות**.

דוגמא לבעיית שיערוך

נסתכל לדוגמא על המדגם של הונאות אשראי מהרצאה הקודמת:



היינו מעוניינים לבנות על סמך מדגם זה את הפילוג של המשתנים האקראיים. לדוגמא היינו רוצים למצוא פונקציות כדוגמאת אלה אשר יתארו את הפילוג של הדגימות החוקיות ושל הונואות:



בשלושת ההרצאות הקרובות אנו נעסוק בשאלה של כיצד לשערך פילוגים מסוגים אלו מתוך המדגם, וכיצד ניתן לבנות על סמך שיערוכים אלו את פונקציית החיזוי.

שיערוך של פונקציות פילוג בשיטות א-פרמטריות

בהרצאה הקרובה נעסוק בשיטות שיערוך אשר מכונות שיטות לא פרמטריות או א-פרמטריות, מהות השם תהיה ברורה יותר אחרי שנציג בהרצאה הבאה את הנושא של שיטות פרמטריות.

שיערוך ההסתברות של מאורע

נתחיל בבעיה פשוטה. ננסה לשערך את ההסתברות להתרחשות של מאורע מסוים על סמך מדגם.

דוגמא

נניח שיש בידינו את המדגם הבא של מדידות של זמני נסיעה (בדקות) מחיפה לתל אביב על כביש החוף:

$$\mathcal{D} = \{x^{(i)}\} = \{55, 68, 75, 50, 72, 84, 65, 58, 74, 66\}$$

ברצונינו לשערך את ההסתברות של המאורע שנסיעה מסוימת תיקח פחות משעה, $A = \{x < 60\}$. המשערך הטבעי ביותר לבעיה זו הינו משערך אשר שווה למספר הפעמים היחסי שמאורע זה קרה במדגם הנתון. בדוגמא זו יש 3 מתוך 10 נסיעות שבהן זמן הנסיעה היה קצר משעה, לכן נשערך שההסתברות של מאורע זה הינה:

$$\Pr(A) \approx \hat{p}_{A,\mathcal{D}} = 0.3$$

בדומה לסימון בבעיות חיזוי, נשתמש בסימון "כובע" לציון גודל שאותו אנו חוזים / משערים באופן אמפירי (על סמך מדגם). בנוסף אנו נקפיד לצייין את העובדה שמשערך תלוי במדגם שבו השתמשנו על ידי הוספת \mathcal{D} מתחת למשערך.

שיטת שיערוך זו מכונה **מדידה אמפירית (empirical measure)** או משערך הצבה. נרשום את המשערך בצורה פורמלית.

מדידה אמפירית / משערך הצבה (empirical measure)

בהינתן מדגם מסוים $\mathcal{D} = \{x^{(i)}\}_{i=0}^N$, המדידה האמפירית, $\hat{p}_{A,\mathcal{D}}$, הינה שיערוך של ההסתברות, $\Pr(A)$, והיא מחושבת באופן הבא:

$$\hat{p}_{A,\mathcal{D}} = \frac{1}{N} \sum_{i=1}^N I\{x^{(i)} \in A\}$$

נוכל כעת להשתמש בשיטה זו על מנת לנסות ולשערך את הפילוג של משתנים אקראיים.

שיערוך פונקציית ההסתברות של משתנה אקראי דיסקרטי

שיערוך פונקציית ההסתברות (ה PMF) של משתנים אקראיים דיסקרטיים הוא לרוב משימה פשוטה. נסתכל על שתי דוגמאות:

דוגמא 1 - משתנה בינארי

יש בידינו מטבע לא הוגן (כזה שההסתברות שיפול על עץ או על פלי היא לא חצי-חצי). נסמן את תוצאת ההטלה של המטבע ב x כך ש 1 מצייין עץ ו 0 מצייין פלי. בכדי לקבוע את ה PMF של x הטלנו את המטבע 10 פעמים וקיבלנו:

$$\mathcal{D} = \{x^{(i)}\} = \{0, 0, 0, 0, 1, 0, 0, 1, 0, 0\}$$

גם פה הפתרון הטבעי הוא לשערך את הסתברות לקבל כל ערך של x על פי השכיחות של אותו ערך במדגם. זאת אומרת כיוון שמתוך ה 10 דגימות יש 2 פעמים את הערך 1 ו 8 פעמים את הערך 0 נשערך את ה PMF להיות:

$$p_x(x) \approx \hat{p}_{x,\mathcal{D}}(x) = \begin{cases} 0.8 & 0 \\ 0.2 & 1 \end{cases}$$

למעשה אנו משתמשים כאן במדידה אמפירית של המאורע ש $\{x = x\}$ לשיערוך של כל אחד מהערכים.

דוגמא 2 - משתנה לא בינארי

את אותו השיערוך נוכל כמובן לבצע גם על משתנים דיסקרטיים אשר יכולים לקבל מספר כל שהוא של ערכים. דוגמא נסתכל על בעיה דומה עם קוביה לא הוגנת שב 10 הטלות שלה התקבלו הדגימות הבאות:

$$\mathcal{D} = \{x^{(i)}\} = \{3, 2, 5, 1, 2, 6, 2, 5, 5, 3\}$$

גם כאן נשערך את ההסתברות לקבל כל ערך לפי השכיחות שלו במדגם:

$$p_x(x) \approx \hat{p}_{x,D}(x) = \begin{cases} 0.1 & 1 \\ 0.3 & 2 \\ 0.2 & 3 \\ 0 & 4 \\ 0.3 & 5 \\ 0.1 & 6 \end{cases}$$

ניסוח פורמאלי

בהינתן מדגם מסוים $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=0}^N$, נוכל לשערך את ה PMF של משתנה / וקטור אקראי דיסקרטי באופן הבא:

$$\hat{p}_{x,D}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N I\{\mathbf{x}^{(i)} = \mathbf{x}\}$$

שימו לב שמובטח לנו שנקבל פונקציית הסתברות חוקית (חיובית שהסכום עליה שווה ל 1).

שיערוך פונקציית הפילוג המצרפי של משתנה אקראי

(ECDF (Empirical Cumulative Distribution Function

נזכור כי פונקציית הפילוג המצרפי (ה CDF) מוגדרת באופן הבא:

$$F_x(\mathbf{x}) = \Pr(\{x_j \leq x_j \forall j\})$$

נוכל אם כן לשערך גודל זה על ידי שימוש במדידה האמפירית בעבור המאורע של $A = \{x_j \leq x_j \forall j\}$ באופן הבא:

$$\hat{F}_{x,D}(\mathbf{x}) = \hat{p}_{A,D} = \frac{1}{N} \sum_{i=1}^N I\{x_j \leq x_j \forall j\}$$

משערך זה נקרא (empirical cumulative distribution function) (ECDF).

דוגמא

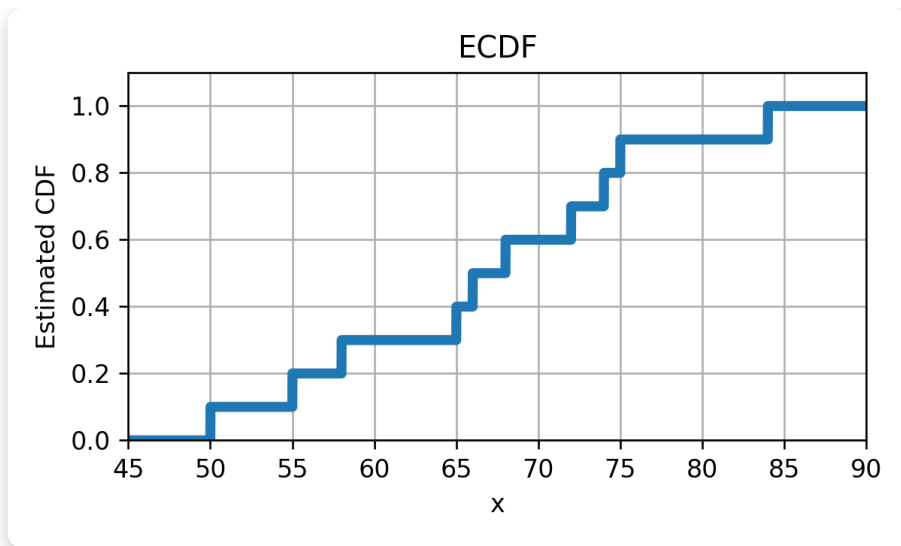
נשערך את פונקציית הפילוג המצרפי של המדגם של 10 זמני הנסיעה בכביש החוף

$$\mathcal{D} = \{\mathbf{x}^{(i)}\} = \{55, 68, 75, 50, 72, 84, 65, 58, 74, 66\}$$

משערך ה ECDF של x יהיה במקרה זה:

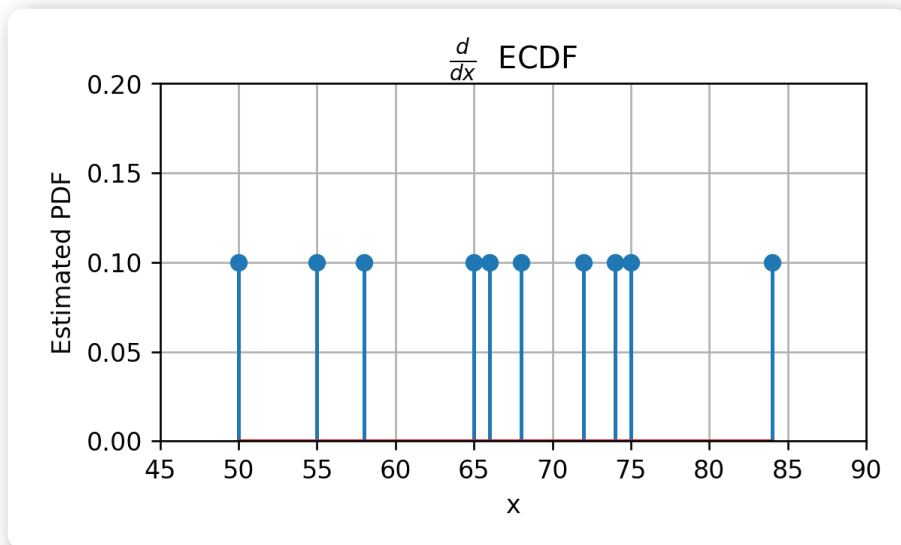
$$\hat{F}_{x,D}(\mathbf{x}) = \begin{cases} 0 & x < 50 \\ 0.1 & 50 \leq x < 55 \\ 0.2 & 55 \leq x < 58 \\ 0.3 & 58 \leq x < 65 \\ 0.4 & 65 \leq x < 66 \\ 0.5 & 66 \leq x < 68 \\ 0.6 & 68 \leq x < 72 \\ 0.7 & 72 \leq x < 74 \\ 0.8 & 74 \leq x < 75 \\ 0.9 & 75 \leq x < 84 \\ 1 & 84 \leq x \end{cases}$$

זוהי למעשה פונקציה קבועה למקוטעין אשר נראית כך:



הבעיה עם ECDF

הבעיה העיקרית עם משעריך ה-ECDF הינה שהוא מייצר פונקציה שהיא קבועה למקוטעין, כאשר בעבור משתנים רציפים היינו מצפים לפונקציה רציפה אשר עולה בהדרגה מ 0 ל 1. אחד הבעיות העיקריות עם העובדה שהפונקציה אינה רציפה הינה פונקציית ה-PDF המתקבלת מתוך נסיון לגזור את ה-ECDF. פונקציית ה-PDF שנקבל תהיה מורכבת מאוסף של פונקציות דלתא:



ופונקציה כזו היא לא מאד שימושית.

היסטוגרמה

היסטוגרמה היא מעין נסיון לשערך פילוג של משתנה רציף על ידי כך שנעשה לו קוונטיזציה. בשיטה זו נחלק את טווח הערכים שמשתנה אקראי יכול לקבל למספר סופי של חלקים המכונים bins (תאים). אחרי חלוקה זו נשתמש במדידה אמפירית על מנת לשערך את ההסתברות להימצא בכל תא.

דוגמא

לדוגמא בעבור המקרה של של זמני הנסיעה, נוכל לחלק את התחום ל 5 קטעים:

[45, 54), [54, 63), [63, 72), [72, 81), [81, 90]

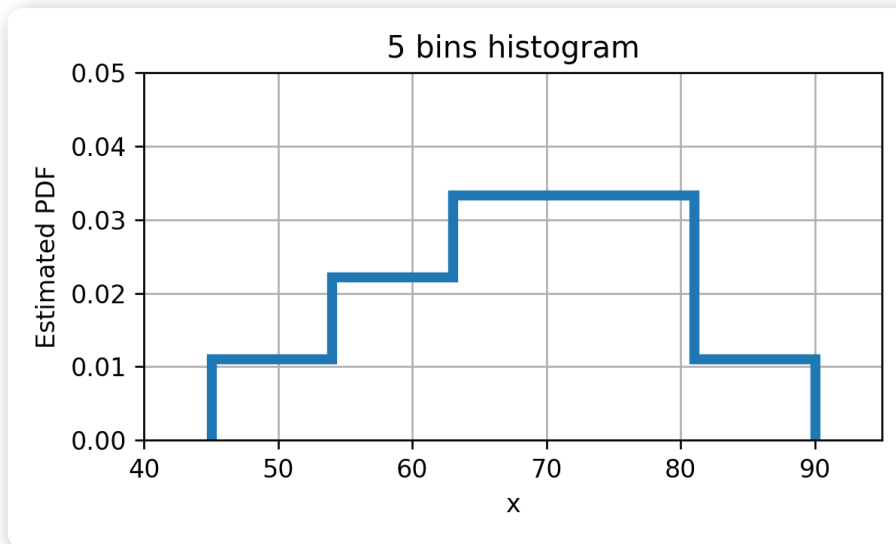
הבחירה של ה bins נעשתה כך שהם יכסו את כל התחום ולא תהיה בניהם חפיפה (כולל בקצוות ה bin). באופן כללי ניתן לבחור את ה bins בכל צורה שהיא כל עוד הם מקיימים את אותם שני תנאים של כיסוי מלא וחוסר חפיפה.

נחשב את ההסתברות להיות בכל bin בעזרת המדידה האמפירית:

$$\begin{aligned}\hat{p}_{\{45 \leq x < 54\}, \mathcal{D}} &= 0.1 \\ \hat{p}_{\{54 \leq x < 63\}, \mathcal{D}} &= 0.2 \\ \hat{p}_{\{63 \leq x < 72\}, \mathcal{D}} &= 0.3 \\ \hat{p}_{\{72 \leq x < 81\}, \mathcal{D}} &= 0.3 \\ \hat{p}_{\{81 \leq x < 90\}, \mathcal{D}} &= 0.1\end{aligned}$$

בכדי להפוך את ההסתברויות של המאורעות האלה לצפיפות הסתברות נרצה "למרוח" את ההסתברות שקיבלנו להיות ב bin מסויים באופן אחיד על פני ה bin. זאת אומרת שצפיפות ההסתברות בכל נקודה ב bin תהיה ההסתברות להימצא ב bin חלקי גודל ה bin. נקבל אם כן את פונקציית הצפיפות הפילוג הבאה:

$$\hat{p}_{x, \mathcal{D}}(x) = \begin{cases} \frac{1}{\text{size of bin } 1} \hat{p}_{\{x \text{ in bin } 1\}, \mathcal{D}} & x \text{ in bin } 1 \\ \vdots \\ \frac{1}{\text{size of bin } B} \hat{p}_{\{x \text{ in bin } B\}, \mathcal{D}} & x \text{ in bin } B \end{cases}$$



ניסוח פורמאלי

בהינתן מדגם מסויים $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=0}^N$, ההיסטוגרמה הינה שיערוך של ה PDF של משתנה / וקטור אקראי והיא מחושבת באופן הבא:

1. מחלקים את תחום הערכים של x יכול לקבל ל bins (תאים) לא חופפים אשר מכסים את כל התחום.
2. לכל bin משערכים את ההסתברות של המאורע שבו x יהיה בתוך התא.
3. הערך של פונקציית הצפיפות בכל תא תהיה ההסתברות המשוערכת להיות בתא חלקי גודל התא.

נרשום זאת בעבור המקרה של משתנה אקראי סקלרי. נסמן ב B את מספר התאים וב l_b ו r_b את הגבול השמאלי והימני בהתאמה של התא ה b . ההסטוגרמה תהיה נתונה על ידי:

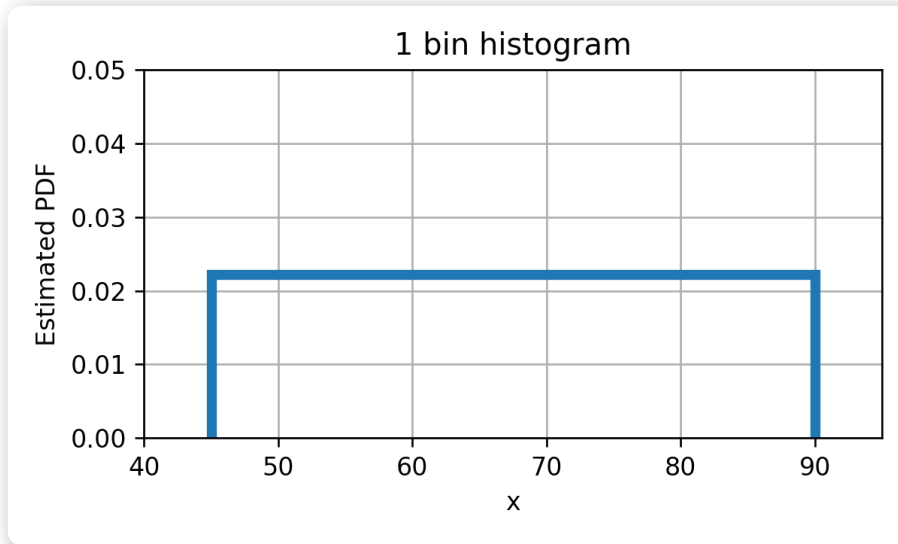
$$\hat{p}_{x, \mathcal{D}}(x) = \begin{cases} \frac{1}{N(r_1 - l_1)} \sum_{i=1}^N I\{l_1 \leq x^{(i)} < r_1\} & l_1 \leq x < r_1 \\ \vdots \\ \frac{1}{N(r_B - l_B)} \sum_{i=1}^N I\{l_B \leq x^{(i)} < r_B\} & l_B \leq x < r_B \end{cases}$$

לבחירת ה bins יש השפעה גדולה על איכות השיערוך שנקבל. ננסה להבין את השיקולים בבחירת ה bins.

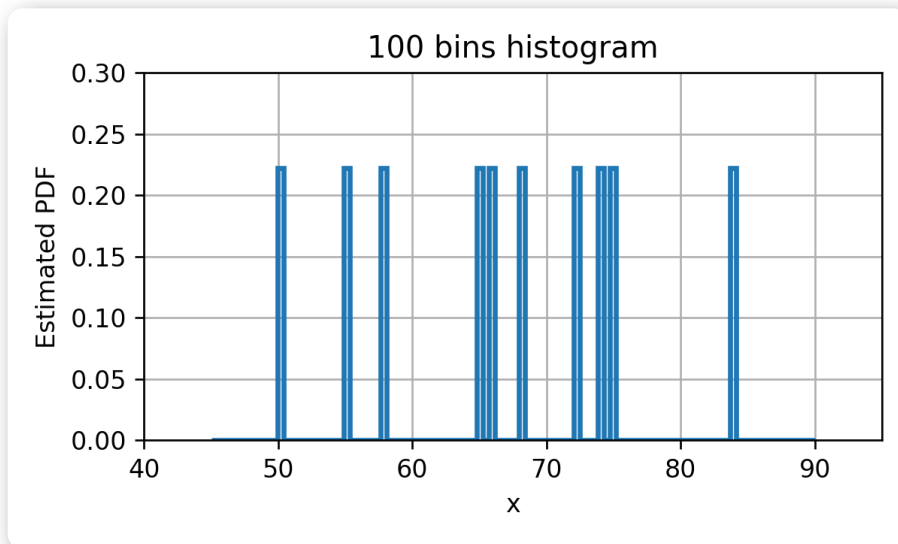
underfitting ו Overfitting של היסטוגרמה

דוגמא

נסתכל על שני מקרים קיצוניים. בעבור בחירה של bin יחיד אשר מכסה את כל התחום, נקבל את ההיסטוגרמה הבאה:



בעבור חלוקה של התחום ל 100 bins בעלי גודל אחיד, נקבל את ההיסטוגרמה הבאה:



כאשר מספר התאים מאד קטן היכולת שלנו לקרב את ה PDF האמיתי תהיה מאד מוגבלת ולכן נקבל PDF משוערך שמאד שונה מה PDF האמיתי. זהו למעשה מקרה קלאסי של underfitting שבו אנו משתמשים במודל מוגבל אשר יכול ללמוד רק מאפיינים מאד גסים של המדגם וניתן לשפר את התוצאה על ידי שימוש במודל בעל ביטוי גדולה יותר.

מצד שני כאשר מספר ה bin מאד גדול ההיסטוגרמה תתאר בצורה טובה את הפילוג של הדגימות **הספציפיות שבמדגם** אך כנראה שפילוג זה לא יתאר בצורה טובה את הפילוג של מדגם אקראי אחר, או לחילופין את הפילוג האמיתי של המשתנה האקראי. זהו מקרה קלאסי של overfitting.

גם כאן החלוקה האופטימלית ל bins, שתייצר את פונקציית ההיסטוגרמה הקרובה ביותר ל PDF האמיתי, תהיה לרוב איזו שהיא נקודת ביניים בין חלוקה למספר גדול של bins אשר תיצור overfitting לבין חלוקה למספר קטן של bins אשר תייצר underfitting.

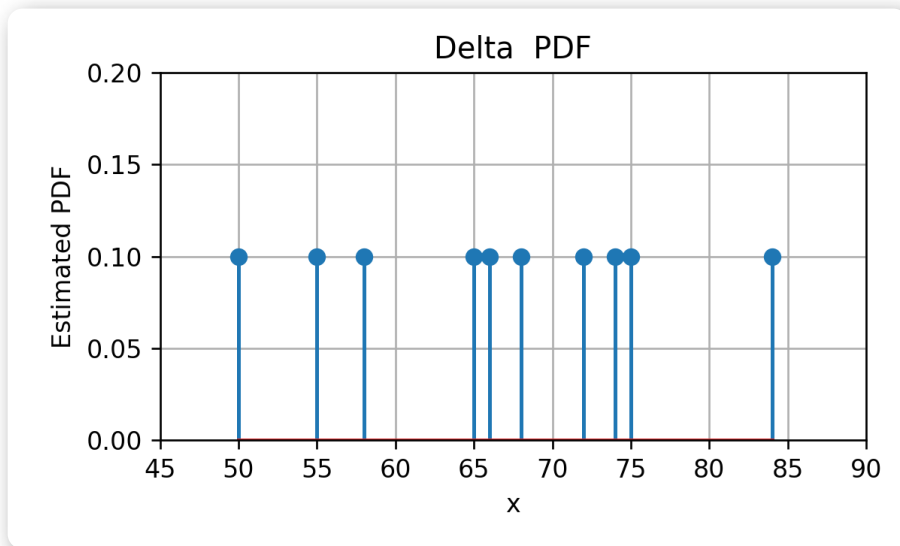
הערה: בפסקה האחרונה ציינו את פונקציית ההיסטוגרמה הקרובה ביותר ל PDF האמיתית, כאשר למעשה לא הגדרנו מדד למרחק בין פונקציות צפיפות. מסתבר שזהו נושא חשוב ויש הרבה דרכים לעשות זאת, אך בקורס זה לא נספיק לעסוק בו ונסתפק בהערכה איכותית של השיערוך ולא בהערכה כמותית.

בחירת התאים

בחירה מקובלת של החלוקה לתאים הינה החלוקה של התחום ל k תאים אחידים בגודלם. נשאר אם כן לבחור את k . מכיוון שה k האופטימאלי ישתנה מבעיה לבעיה, נאלץ לרוב לבחור אותו בעזרת ניסוי וטעיה. אך אם זאת, ישנם מספר כללי אצבע אשר במרבית המקרים יתנו תוצאה לא רעה. הכלל הנפוץ ביותר הינו לבחירה של k הינה שורש מספר הדגימות במדגם (מעוגל כלפי מעלה): $\lceil \sqrt{N} \rceil$.

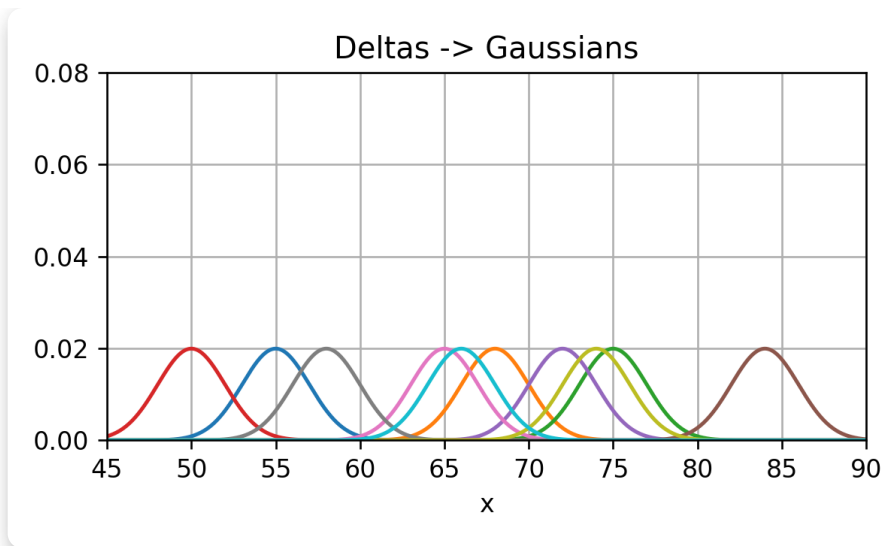
(Kernel Density Estimation (KDE

נציג כעת שיטה נוספת מאד פופולרית לשיערוך פונקציית pdf מתוך מדגם המכונה (kernel density estimation (KDE. בכדי להבין איך השיטה פועלת נתחיל מ PDF שבו אנו ממקמים פונקציית דלתא בגובה $\frac{1}{N}$ בכל נקודה אשר מופיעה במדגם. לדוגמא, בעבור 10 הדגימות של זמני הנסיעה בכביש החוף נקבל:

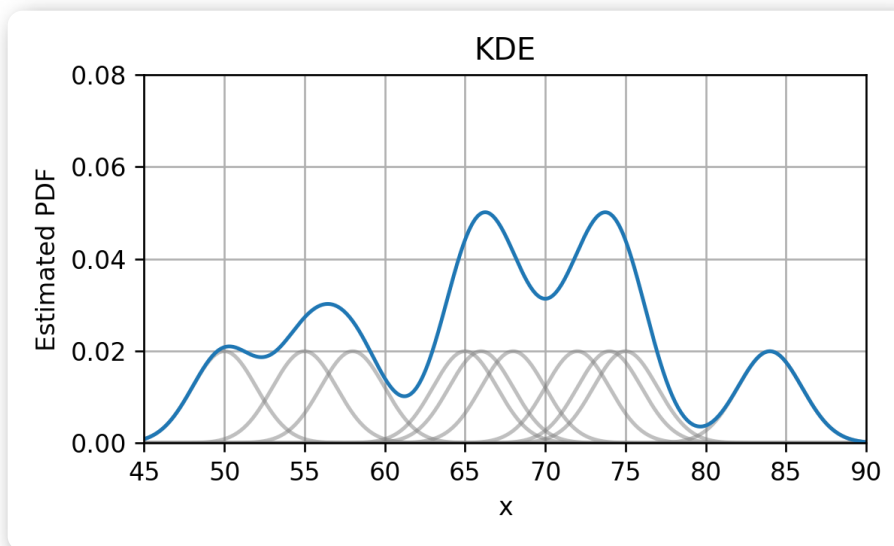


ראינו קודם כי שיטה אחת שבה PDF שכזה מתקבל הינה מתוך נסיון לגזור את פונקציית ה ECDF.

ככדי להפוך את הפילוג הזה ליותר סימפטי ננסה "למרוח" את פונקציות הדלתא על ידי החלפתם בפונקציות בעלות רוחב גדול מ-0 (בניגוד לרוחב 0 של פונקציות הדלתא). לדוגמא, בחירה נפוצה להחלפה שכזו היא החלפה של כל פונקציית דלתא בגאוסיאן:



הפונקציות שבהם אנו מחליפים את פונקציות הדלתא מכונות **פונקציות גרעין (kernel)** או **Parzen window** ומקובל לסמנם ב $\phi(\mathbf{x})$. לאחר ההחלפה של הדלתאות בפונקציות הגרעין, נסכום את כל פונקציות הגרעין שקיבלנו לקבלת ה PDF המשוער:



אם כן, משערך ה KDE נתון על ידי:

$$\hat{p}_{\mathbf{x},\phi,\mathcal{D}}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \phi(\mathbf{x} - \mathbf{x}^{(i)})$$

הערה למי שלקח קורסים בעיבוד אותות: למעשה אנו מבצעים קונבולוציה בין פונקציית הדלתאות לבין פונקציית גרעין כל שהיא. לרוב אנו נרצה שהגרעין ישמש כמעין low pass filter שמטרתו להחליק את פונקציית הדלתאות.

הערה: תנאי מספיק והכרחי בכדי שנקבל PDF חוקי, הינו שפונקציית הגרעין תהיה בעצמה PDF חוקי. זאת אומרת שהיא חייבת להיות חיוביות ושהאינטגרל עליה יהיה שווה ל 1.

הוספת פרמטר רוחב

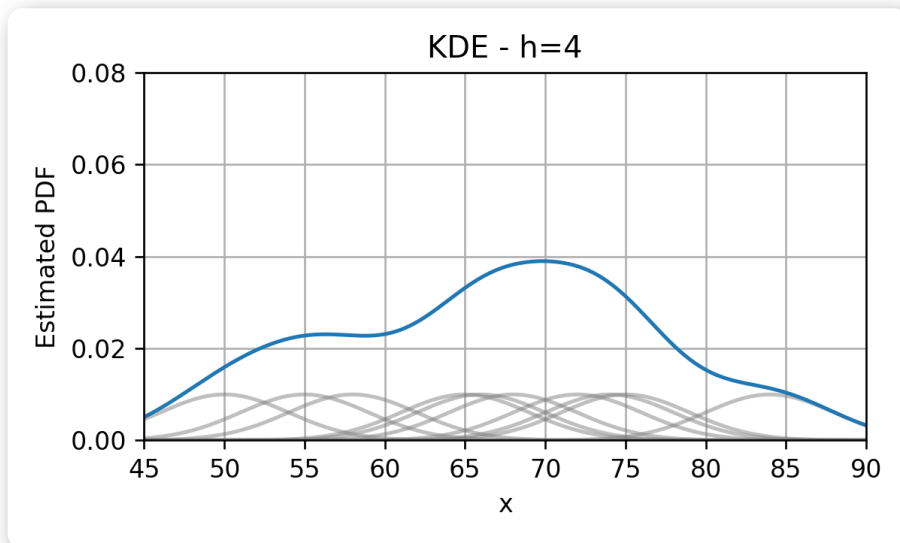
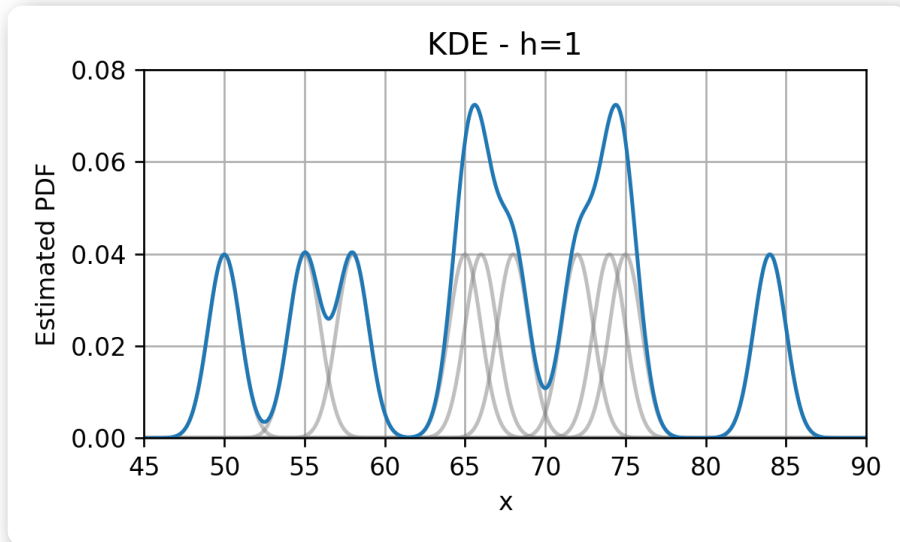
מקובל להוסיף לפונקציות הגרעין פרמטר h אשר שולט ברוחב שלה באופן הבא:

$$\phi_h(\mathbf{x}) = \frac{1}{h^D} \phi\left(\frac{\mathbf{x}}{h}\right)$$

החלוקה ב h^D היא על מנת לשמור על הנרמול של הפונקציה. כאשר D הוא המימד של \mathbf{x} .

בתוספת פרמטר זה המשערך יהיה:

$$\hat{p}_{\mathbf{x},\phi,h,D}(\mathbf{x}) = \frac{1}{Nh^D} \sum_{i=1}^N \phi\left(\frac{\mathbf{x} - \mathbf{x}^{(i)}}{h}\right)$$



בדומה לבחירה של מספר התאים בהיסטוגרמה גם כאן רוחב הגרעין ישלוט במידת ה overfitting. בעבור h גדול נקבל underfitting ובעבור h קטן נקבל overfitting.

פונקציות גרעין נפוצות

שתי הבחירות הנפוצות ביותר לפונקציית הגרעין הינן:

1. חלון מרובע:

$$\phi_h(\mathbf{x}) = \frac{1}{h^D} I\{|x_j| \leq \frac{h}{2} \quad \forall j\}$$

כלל אצבע עבור חלון ריבועי הוא לבחור בצורה אדפטיבית את גודל החלון כך שיקלול מספר נתון (k) של דגימות מסביב לנקודה הנחקרת. בחירה סבירה הינה $k \propto \sqrt{N}$, בדומה למה שעשינו בהיסטוגרמות.

$$\phi_\sigma(\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^D}} \exp\left(-\frac{\|\mathbf{x}\|_2^2}{2\sigma^2}\right)$$

כלל אצבע לבחירת רוחב הגרעין במקרה הגאוסית הסקלרית הינו $\sigma = \left(\frac{4 \cdot \text{std}(\mathbf{x})^5}{3N}\right)^{\frac{1}{5}} \approx 1.06 \text{std}(\mathbf{x})N^{-\frac{1}{5}}$, כאשר $\text{std}(\mathbf{x})$ הינה הסטיית תקן של \mathbf{x} (אשר לרוב תהיה משוערכת גם היא מתוך המדגם)

שיערוך של פילוגים מעורבים

במקרים רבים אנו נרצה לשיערוך פילוגים אשר מערבים משתנים רציפים ומשתנים בדידים. נניח לדוגמה שאנו רוצים לשיערוך את הפילוג המשותף של \mathbf{x} ו y כאשר \mathbf{x} הוא משתנה רציף ו y הוא משתנה בדיד. במקרים כאלה נוח לפרק את פונקציית הפילוג המשותף באופן הבא:

$$p_{\mathbf{x},y}(\mathbf{x}, y) = p_{\mathbf{x}|y}(\mathbf{x}|y)p_y(y)$$

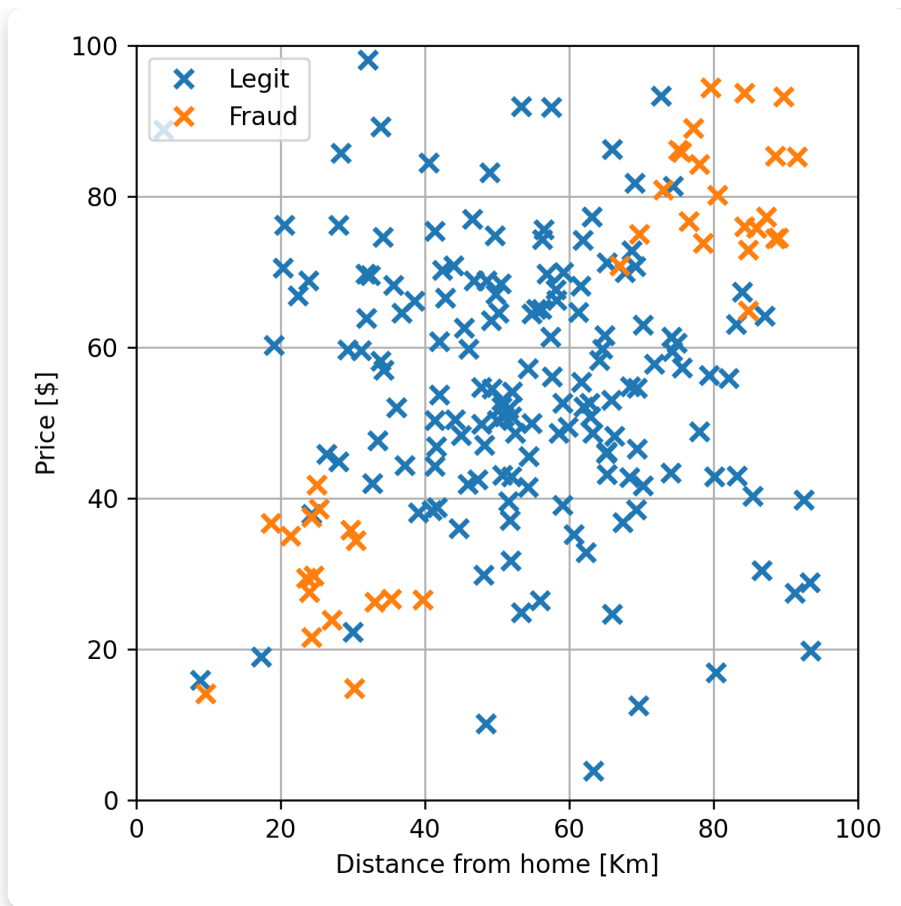
ואז להפריד את בעיית השיערוך לשני חלקים:

1. השיערוך של $p_y(y)$ - שיערוך זה יהיה לרוב פשוט שכן שיערוך זה לא תלוי כלל בערכו של \mathbf{x} , ו y הוא משתנה אקראי דיסקרטי שאותו קל יותר לשיערוך.
2. השיערוך של $p_{\mathbf{x}|y}(\mathbf{x}|y)$ - כאן יהיה לרוב נוח לפצל את השיערוך למספר שיערוכים שונים בעבור כל ערך אפשרי של y . זאת אומרת $p_{\mathbf{x}|y}(\mathbf{x}|1)$, $p_{\mathbf{x}|y}(\mathbf{x}|2)$, וכו'. הדרך לעשות זאת היא על ידי פיצול של המדגם על פי הערכים של y ושיערוך הפילוג של $\mathbf{x}|y$ בנפרד על כל חלק של המדגם.

הצורך לשיערוך פילוגים משותפים מופיע לדוגמה בבעיות סיווג שבהם התוויות y הם דיסקרטיות והמידות \mathbf{x} הם רציפות.

דוגמא

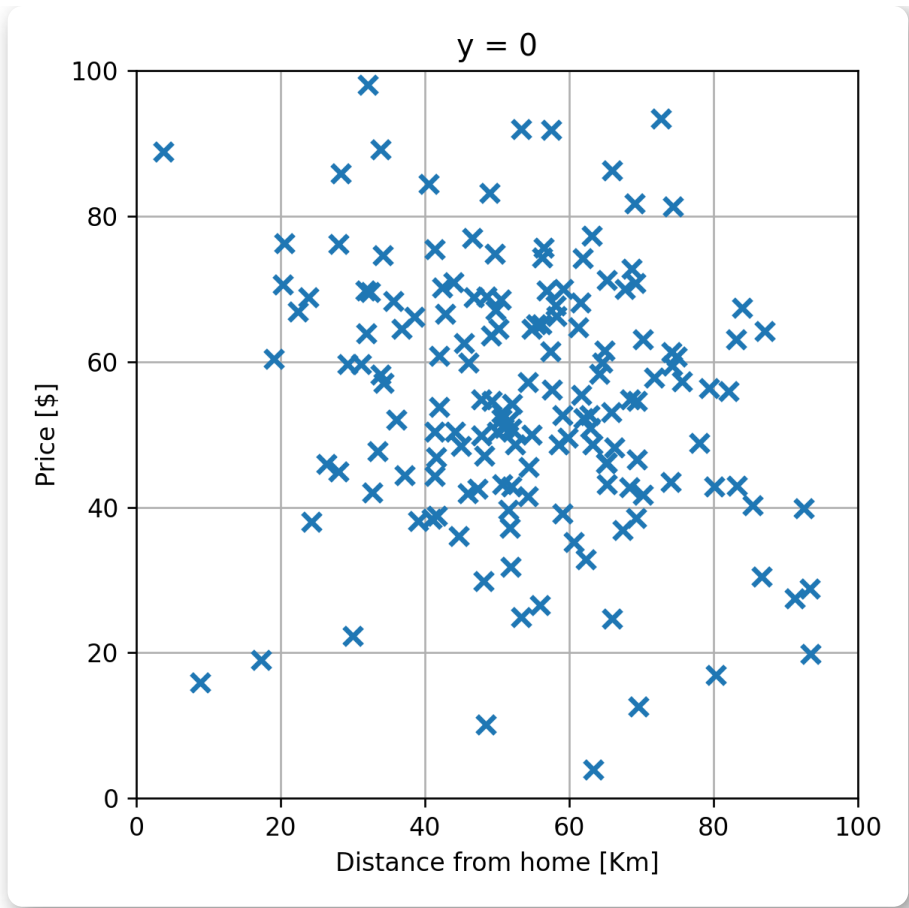
נחזור לדוגמא של הונאות האשראי:



נתחיל אם כן בשיערוך של הפילוג של התוויות. כפי שציינו, מכיוון ש y בדיד נוכל לשערך בפשטות את ה PMF שלו על פי השכיחות של כל אחד מהערכים 0 ו 1 במדגם. מכיוון שמתוך ה 200 עסקאות שיש במדגם (ב train set) ישנם 160 עסקאות חוקיות ($y = 0$) ו 40 עסקאות שחשודות כהונאה ($y = 1$) השיערוך של ה PMF של y יהיה:

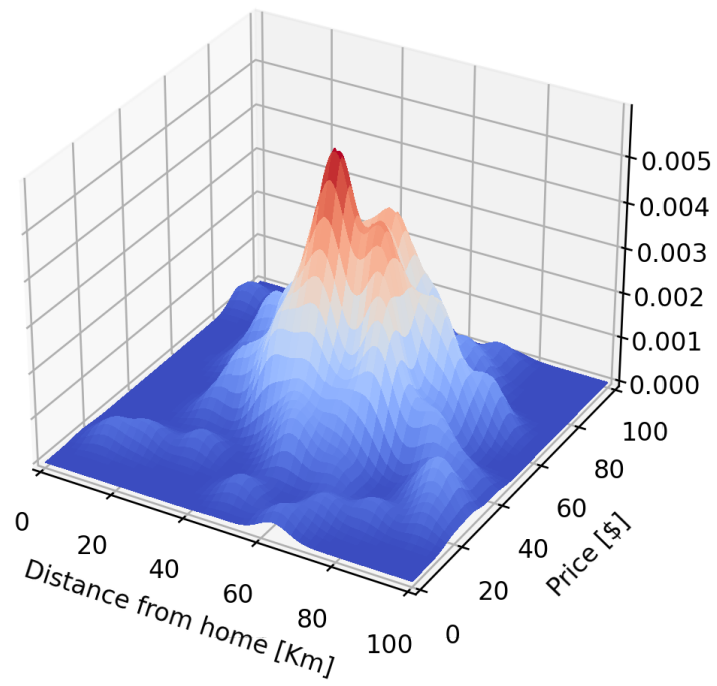
$$\hat{p}_{y,D}(y) = \begin{cases} \frac{160}{200} & 0 \\ \frac{40}{200} & 1 \end{cases} = \begin{cases} 0.8 & 0 \\ 0.2 & 1 \end{cases}$$

נמשיך לשיערוך של $p_{x|y}(\mathbf{x}|y)$. נשערך בנפרד את $p_{x|y}(\mathbf{x}|0)$ ואת $p_{x|y}(\mathbf{x}|1)$. נתחיל מ $p_{x|y}(\mathbf{x}|0)$. בשביל לשערך פילוג זה נסתכל רק על הדגימות השייכות של $y = 0$:



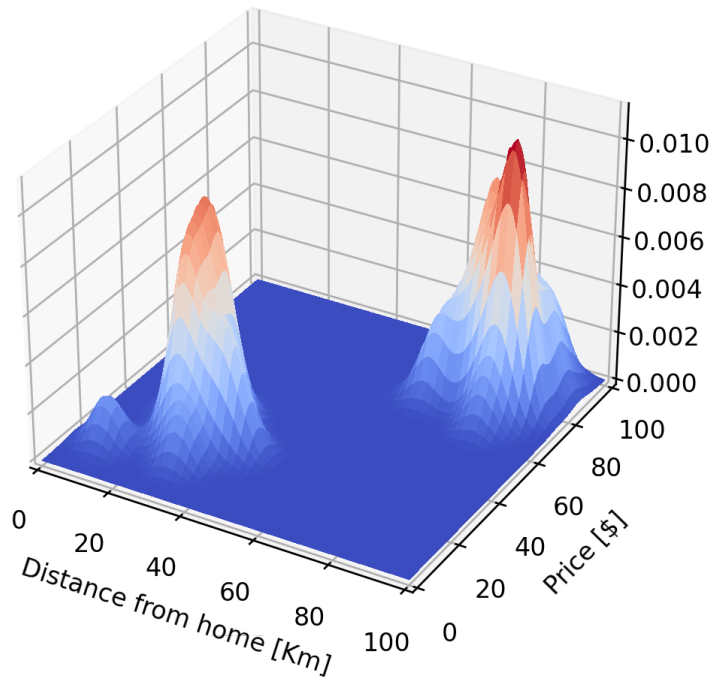
נשתמש ב KDE על מנת לשערך את פונקציית הפילוג של מדגם זה:

Legit KDE - $\hat{p}_{x|y, D}(x|0)$



באופן דומה נשערך גם את $p_{x|y}(x|1)$:

Fraud KDE - $\hat{p}_{x|y, D}(x|1)$



שלושת הפילוגים ששיערכנו, $p_{x|y}(\mathbf{x}|1)$ ו $p_y(y)$, $p_{x|y}(\mathbf{x}|0)$ מרכיבים למעשה את הפילוג המשותף המלא של \mathbf{x} ו y . זאת מכיוון שבעבור כל צמד ערכים של \mathbf{x} ו y נוכל לחשב את הפילוג המשותף שלהם על פי:

$$p_{\mathbf{x},y}(\mathbf{x}, y) = p_{\mathbf{x}|y}(\mathbf{x}|y)p_y(y)$$

שימוש בפילוג המשוערך לפתרון בעיות supervised learning

נחזור כעת לסיבה שבגללה אנו רוצים לנסות לשערך את פונקציית הפילוג של משתנים אקראיים. כפי שצינו קודם, בכדי לפתור בעיות supervised learning בגישה הגנרטיבית נרצה לשערך את פונקציית הפילוג על מנת שנוכל לבנות על פיה את פונקציית החיזוי. נזכיר כי בעבור פונקציות המחיר הנפוצות אנו כבר יודעים מהו החזאי האופטימאלי בהינתן הפילוג:

• **MSE**: התוחלת המותנית:

$$h^*(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$$

• **MAE**: החציון של הפילוג המותנה:

$$h^*(\mathbf{x}) = y_{\text{median}} \quad \text{s.t.} \quad F_{y|\mathbf{x}}(y_{\text{median}}|\mathbf{x}) = 0.5$$

(כאשר $F_{y|\mathbf{x}}$ היא פונקציית הפילוג המצרפי של y בהינתן \mathbf{x} .)

• **Misclassification rate**: הערך הכי סביר (ה mode):

$$h^*(\mathbf{x}) = \arg \max_y p_{y|\mathbf{x}}(y|\mathbf{x})$$

לכן במקרים אלו כל שעלינו לעשות זה להציב את הפילוג שמצאנו לביטוי לחזאי האופטימאלי.

בעבור הפילוג שמצאנו על פי המדגם של הונאות האשראי נחפש את החזאי אשר ממזער את ה misclassification rate. אנו יודעים כי חזאי זה נתון על ידי:

$$h(\mathbf{x}) = \arg \max_y p_{y|\mathbf{x}}(y|\mathbf{x})$$

במקרה הבינארי חזאי זה שווה ל:

$$h(\mathbf{x}) = \begin{cases} 1 & p_{y|\mathbf{x}}(1|\mathbf{x}) > p_{y|\mathbf{x}}(0|\mathbf{x}) \\ 0 & \text{else} \end{cases}$$

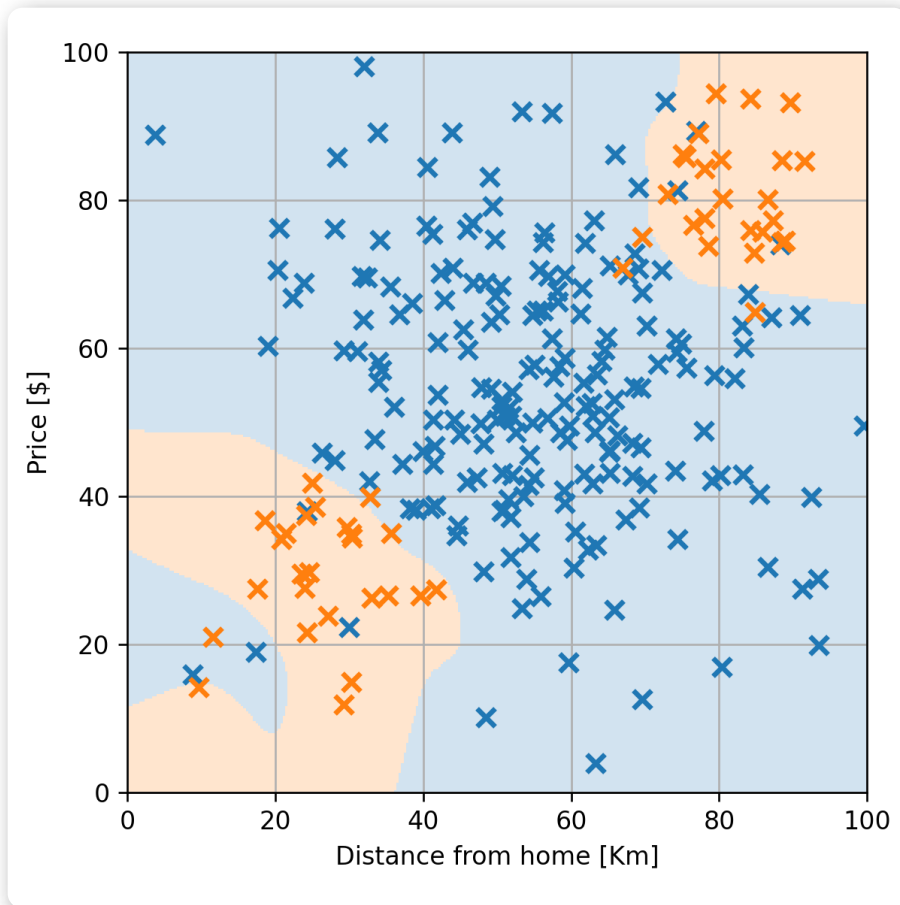
את $p_{y|\mathbf{x}}(y|\mathbf{x})$ נוכל לחשב מתוך הפילוג המשותף באופן הבא:

$$p_{y|\mathbf{x}}(y|\mathbf{x}) = \frac{p_{\mathbf{x},y}(\mathbf{x}, y)}{p_{\mathbf{x}}(\mathbf{x})} = \frac{p_{\mathbf{x}|y}(\mathbf{x}|y)p_y(y)}{p_{\mathbf{x}}(\mathbf{x})}$$

(זהו למעשה חוק בייס). אם כן, בכדי לבדוק האם עסקה מסוימת הינה הונאה או לא, עלינו לבדוק האם:

$$\begin{aligned} & p_{y|\mathbf{x}}(1|\mathbf{x}) > p_{y|\mathbf{x}}(0|\mathbf{x}) \\ \Leftrightarrow & \frac{p_{\mathbf{x}|y}(\mathbf{x}|1)p_y(1)}{p_{\mathbf{x}}(\mathbf{x})} > \frac{p_{\mathbf{x}|y}(\mathbf{x}|0)p_y(0)}{p_{\mathbf{x}}(\mathbf{x})} \\ \Leftrightarrow & p_{\mathbf{x}|y}(\mathbf{x}|1)p_y(1) > p_{\mathbf{x}|y}(\mathbf{x}|0)p_y(0) \end{aligned}$$

אם נציב את פונקציות הפילוג ששיערכנו קודם לכן ונקבל את החזאי הבא:



ה misclassification rate של חזאי זה על ה test set הינו 0.12.

ה bias וה variance של משערך

בדומה לחזאים שבינו בגישה הגנרטיבית, גם המשערכים שתיארנו כאן תלויים בצורה חזקה במדגם שאיתו אנו עובדים. לכן, בדומה לאנליזה שעשינו כאשר דיברנו על ה bias-variance tradeoff, גם כאן נוכל להסתכל על האקראיות של השיערוך הנובעת מהאקראיות של המדגם.

נשתמש שוב בסימון $\mathbb{E}_{\mathcal{D}}$ בכדי לסמן תוחלת על פני הפילוג של המדגם. בעזרת תוחלת זו נגדיר את המושגים של ה bias וה variance של משערוך מסויים:

Bias

בעבור שיערוך של גודל כל שהוא z בעזרת משערוך $\hat{z}_{\mathcal{D}}$, ה bias (היסט) של השיערוך מוגדר כ:

$$\text{Bias}(\hat{z}) = \mathbb{E}_{\mathcal{D}}[\hat{z}_{\mathcal{D}}] - z$$

כאשר ההטיה שווה ל-0, אנו אומרים שהמשערוך אינו מוטא (Unbiased).

Variance

ה variance (שונות) של המשערוך יהיה:

$$\text{Var}(\hat{z}) = \mathbb{E}_{\mathcal{D}}[(\hat{z}_{\mathcal{D}} - \mathbb{E}_{\mathcal{D}}[\hat{z}_{\mathcal{D}}])^2] = \mathbb{E}_{\mathcal{D}}[\hat{z}_{\mathcal{D}}^2] - \mathbb{E}_{\mathcal{D}}[\hat{z}_{\mathcal{D}}]^2$$

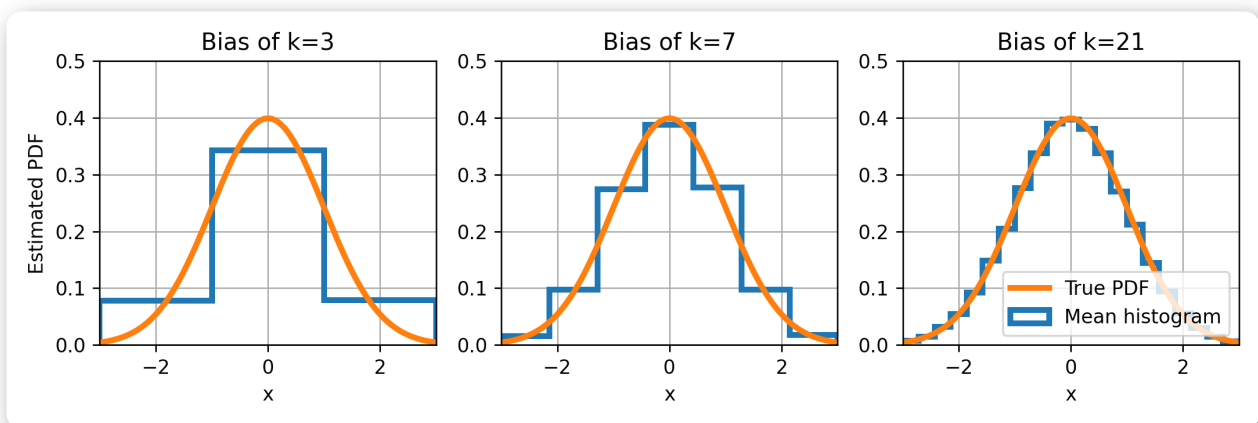
אנו נהיה מעוניינים כמובן במשערוך שגם ה bias וגם ה variance שלו קטנים.

מלבד במקרים מאד מנוונים לרוב לא נוכל באמת לחשב את הגדלים האלה. השימוש העיקרי שלנו בהם יהיה בכדי לנסות ולהבין כיצד שינוי מסויים בשיטת ישפיע על איכות השיערוך מתוך ההבנה של האם הוא מקטין או מגדיל את הגדלים האלה.

דוגמא: אנליזה של מספר ה bins במונחים של bias ו variance

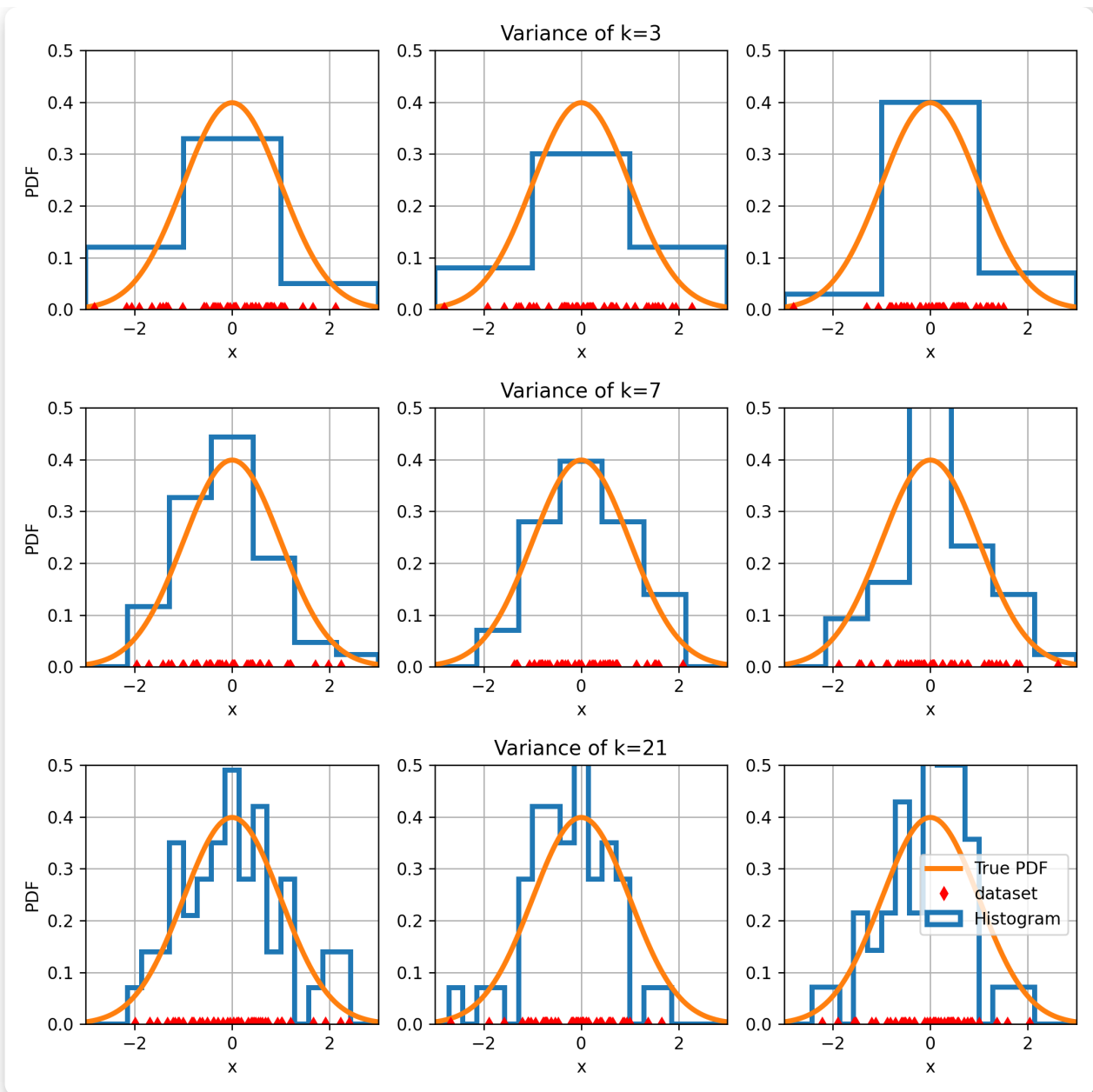
כפי שצינו קודם בעבור היסטוגרמה עם k נמוך אנו נקבל תופעה של underfitting ובעבור k גדול נקבל overfitting. נראה איך זה מתקשר ל bias וה variance של המשערוך.

לצורך הדוגמא ננסה לשערך את ה PDF של משתנה אקראי עם פילוג נורמאלי (גאוס). נעשה זאת בעזרת היסטוגרמות בעלות 3, 7 ו 21 bins. נתחיל בבחינה של ה bias של היסטוגרמות, לשם כך נשרטט את היסטוגרמה הממוצעת לצד ה PDF האמיתי. בדוגמאות מסוג זה, בהם אנו מייצרים את המדגם בצורה מלאכותית, ניתן לחשב בקירוב את היסטוגרמה הממוצעת על ידי מיצוע על מספר גדול של מדגמים או לחילופין (ספציפית במקרה הזה) ניתן לקחת מדגם מאד גדול (לא לכל משערוך זה יהיה נכון).



ה bias בגרפים אלו הוא ההפרש בין היסטוגרמה הממוצעת ל PDF האמיתי (ההפרש בין הקו הכחול לכתום). ניתן לראות שככל שמספר ה bins גדל כך היסטוגרמה הממוצעת מתקרבת ל PDF האמיתי, ניתן אם כך להסיק שבעבור מקרה זה, ה bias של היסטוגרמה קטן ככל שמספר ה bins גדל.

נבחן כעת את ה variance של היסטוגרמה בעבור כל אחת מהבחירות של כמות ה bins. לשם כך נקח כמה מדגמים שונים ונחשב את היסטוגרמה של כל אחד מהם. נסתכל עד כמה משתנה היסטוגרמה בין מדגם למדגם. אנו מצפים שבעבור מקרים שבהם ה variance נמוך השינויים יהיו קטנים ובעבור variance גבוה השינויים יהיו גדולים.



בכל שורה בגרף הזה אנו מגרילים שלושה מדגמים שונים (הנקודות האדומות בתחתית של כל גרף) ומחשבים להם את ההיסטוגרמה. ניתן לראות כי בעבור שלושה bins (השורה הראשונה) אנו מקבלים בערך את אותה התוצאה בעבור כל שלושת המדגמים. מנגד ניתן לראות כי בעבור 21 bins ישנם הבדלים מאד גדולים בין התוצאות המתקבלות בעבור כל אחד מהמדגמים. המשמעות אם כן הינה שבמקרה זה ה variance של ההיסטוגרמה גדל ככל שאנו מגדילים את כמות ה bins.

ראינו עם כן, שבדומה לחזאים שבנינו בגישה הדיסקרימינטיבית, גם בהיסטוגרמה ישנו bias-variance tradeoff וגם כאן אנו נחפש את נקודת האופטימום שמוצאת איזון בין השניים.