תרגיל בית 1:

מגישים:

איתי גיא – 305104184

אורי בן יצחק – 066374737

שאלה 1:

1. בשאלה זו μ איננו ידוע ו-σ^2 כן, לכן:

$$\boldsymbol{Bias} = (E(\hat{\mu}) - \mu) = (\mathbf{1} - \boldsymbol{\mu})$$

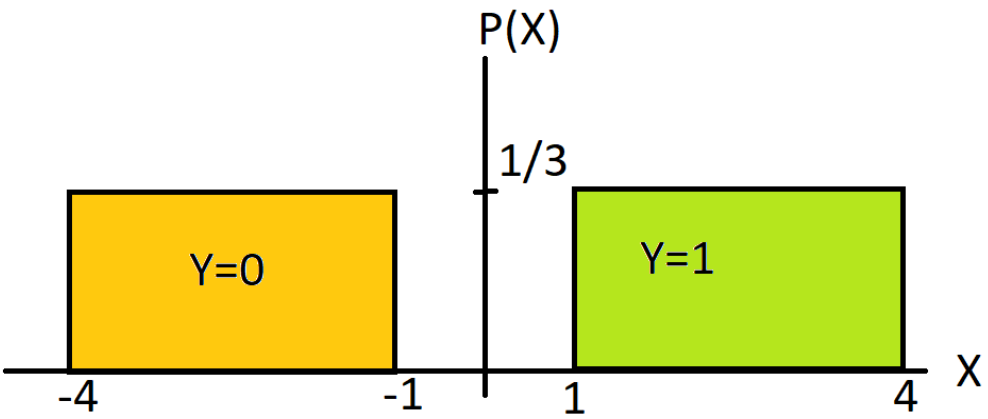
$$\begin{aligned}\boldsymbol{Var}(\hat{\boldsymbol{\mu}}) &= E((\hat{\mu} - \mu)^2) - (E(\hat{\mu}) - \mu)^2 = E((1-\mu)^2) - (1-\mu)^2 \\ &= E(1 + \mu^2 - 2\mu) - (1 + \mu^2 - 2\mu) \\ &= 1 + E(\mu^2) - 2E(\mu) - 1 - \mu^2 + 2\mu = E(\mu^2) - \mu^2 = \mathbf{0}\end{aligned}$$

2. א.

$$\boldsymbol{P(X|Y=0)} = \frac{1}{-1-(-4)} = \frac{\mathbf{1}}{\mathbf{3}}$$

$$\boldsymbol{P(X|Y=1)} = \frac{1}{4-1} = \frac{\mathbf{1}}{\mathbf{3}}$$

ציור של מרחב ההתפלגות:



ב.

אין חיתוך כלל בין המחלקות ולכן בהינתן דוגמא חדשה לסווג אותה באופן מדויק לפי ההשתייכות שלה לאינטרוול מסויים על ציר ה-X.

ג.

$$\begin{aligned}P(X) &= P(X|Y=0)P(Y=0) + P(X|Y=1)P(Y=1) \\ &= \frac{1}{4} * \frac{1}{2} + \frac{1}{4} * \frac{1}{2} = \frac{2}{8} = \frac{1}{4}\end{aligned}$$

$$P(error) = P(X) * P(error|X) = \frac{1}{4} * \frac{1}{2} = \mathbf{\frac{1}{8}}$$

ד.

יש אינסוף נקודות ולכן ניתן לחשב תוחלת וסטיית תקן לכל מחלקה.

<u>תוחלות:</u>

$$E(X|Y=1) = \int_1^4 x * P(X=x|Y=1)$$
$$= \frac{1}{3}\int_1^4 x = \frac{1}{3}\frac{x^2}{2}|_1^4 = \frac{16}{6} = \sim 2.6$$

באופן סימטרי:

$$E(X|Y=0) = \sim -2.6$$

<u>שונויות:</u>

$$Var(X|Y=1) = E(X^2|Y=1) - E(X|Y=1)^2$$
$$= E(X^2|Y=1) - 6.76$$
$$= \int_1^4 x^2 * P(X=x|Y=1) - 6.76 = \frac{1}{3}\frac{x^3}{3}|_1^4 - 6.76$$
$$= \sim 0.24$$

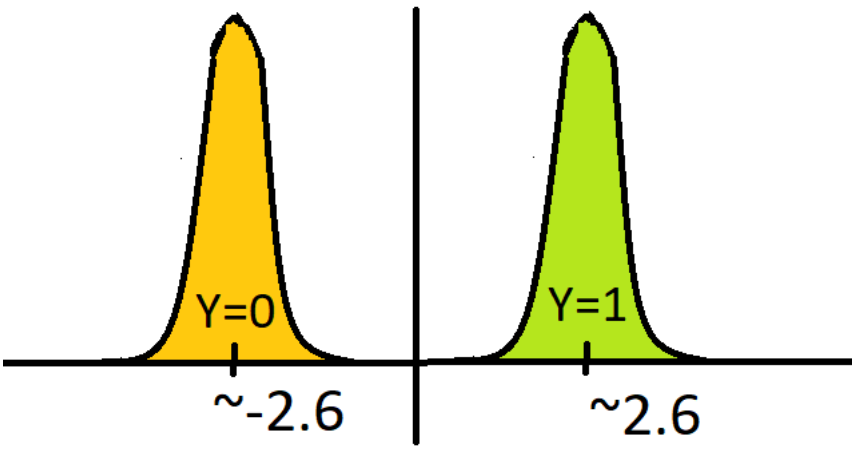
באופן סימטרי:

$$Var(X|Y=0) = \sim 0.24$$

לפיכך, קיבלנו ש:

$$Y = 1 \sim N(\sim 2.6, 0.24) \; and \; Y = 0 \sim N(\sim -2.6, 0.24)$$

ולכן נקבל את הגרף הבא:



מהגרף ניתן להסיק שלמרות שיש לנו מודל לא מדויק עדיין ניתן לאמר שכאשר דוגמת הקלט קטנה מ-0 היא תשתייך באופן מדויק ל-$Y = 0$

ואילו כאשר דוגמת הקלט גדולה מ-0 היא תשתייך באופן מדויק ל-$Y = 1$ ולכן $P(error) = 0$ ולא יהיו לנו טעויות סיווג.

<u>שאלה 2:</u>

נתונים:

$$R(Overdraw|Good\ Credit) = 0.01$$
$$R(Overdraw|Bad\ Credit) = 0.1$$

ידוע כי: $P(New\ customer\ belong\ to\ Good\ Credit) = 0.7$

1. ידוע שללקוח היתה משיכת יתר בחודש הראשון.

$$P(Good\ Credit|Overdraw) = \frac{P(Overdraw|Good\ Credit)P(Good\ Credit)}{P(Overdraw)} = \frac{0.01 * 0.7}{0.037} = \mathbf{0.18}$$

אם הלקוח משתייך לקבוצת הטובים אז בסיכוי יחסית גבוה לזה שהגדיר הבנק הוא יהיה במינוס

2. ידוע שללקוח לא היתה משיכת יתר בחודש השני.

$$P(Good\ Credit|Overdraw\ I, not\ Overdraw\ II) = \frac{P\big((Overdraw\ I \cap not\ Overdraw\ II)|Good\ Credit)P(Good\ Credit)\big)}{P(Overdraw\ I \cap not\ Overdraw\ II)} = \frac{0.01 * 0.99 * 0.7}{0.03393} = \frac{0.00693}{0.03393} = \mathbf{0.204}$$

הלקום גם כאן בסיכוי אך יותר גבוה יהיה במינוסץ

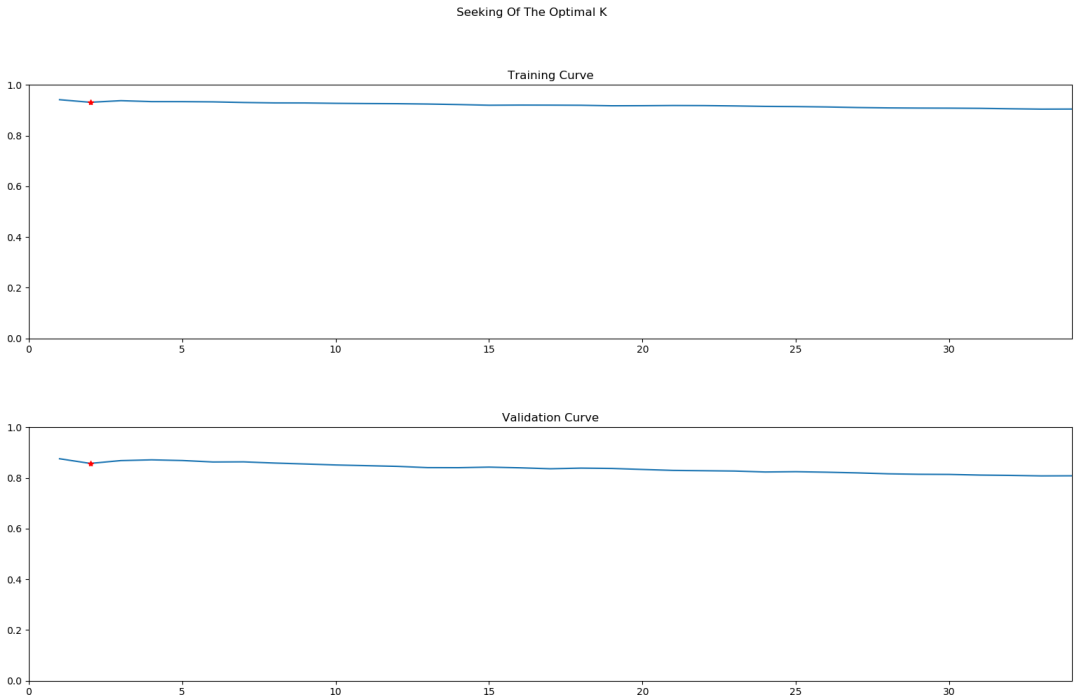
מסקנה – הבנק טעה בהערכתו המקורית כלפי אותו הלקוח.

<u>תרגיל רטוב:</u>

לאחר חיפוש היוריסטי, אחוז הסיווג המקסימאלי הוא $\mathbf{0.901}$ שמתקבל לאחר ערבוב של כל 6000 דוגמאות הקלט.

1. לאחר חיפוש לאחר ה-K האופטימלי תוך התחשבות ב-$Training - Set$ וגם ב-$Validation - Set$ מצאנו ש-$K = 1$:

```
=== Setup Environment ===
=== Setting Train-Set|Validation-Set [80%:20%] ===
=== Found: .\\results ===
=== Found: .\\dataset ===
=== Loading Curves ===
=== Seeking The Optimal K ===
=== Plotting Curves ===
=== Saving Curves As PNG Format ===
=== Validation Accuracy: 0.8783333333333333 ===
=== Done ===
=== Executing Testing ===
=== Loading Data ===
=== Mapping Indices From Dataset ===
=== Found: 1000 Indices ===
=== Executing Test With K=1 ===
=== Test Accuracy: 0.916 ===
=== Done ===
```



2. הדיוק של ה-$ValidationSet$ נמוך במעט מזה של ה-$TestSet$ מכיוון שביצענו ערבוב לכל תמונות הקלט ואז פיצלנו אותם ל-80% אימון ו-20% ולידציה. ככל הנראה לפי הגרף, האימון והולידציה הגיעו ל-k הכי טוב ששניהם יחד יכולים לתת ובכך בוצעה הכללה לקבוצת הבדיקה בצורה טובה יותר.
לאחר בדיקה (ללא שמירה) של הנתונים אם לא מערבבים מגיעים לשגיאה של 0.005 בין הולידציה לקבוצת הבדיקה אך מפאת שכל הרצה היא כ-10 שעות ויתרנו על כך.