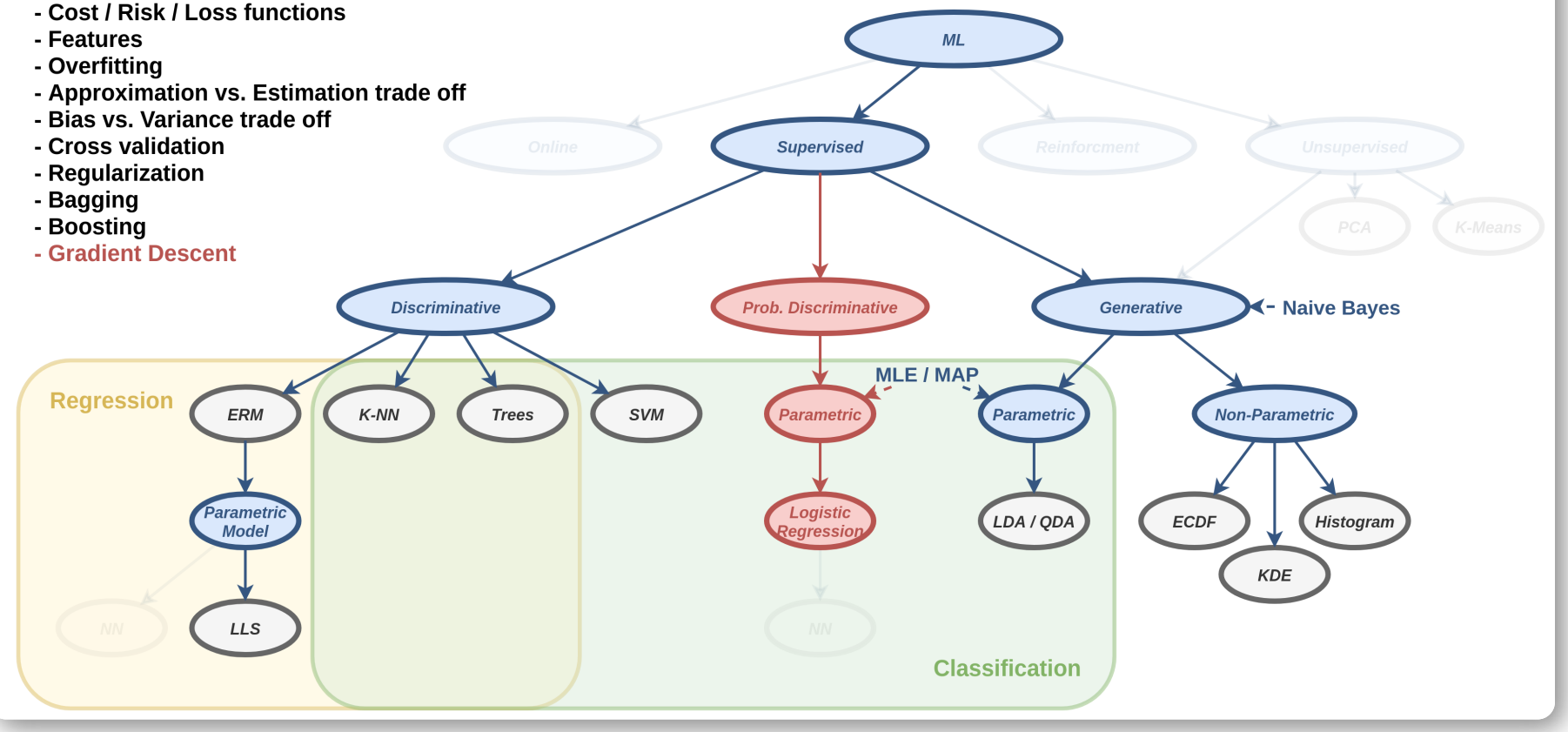
הרצאה 9 - גישה דיסקרימינטיבית הסתברותית

PDF

מה נלמד היום

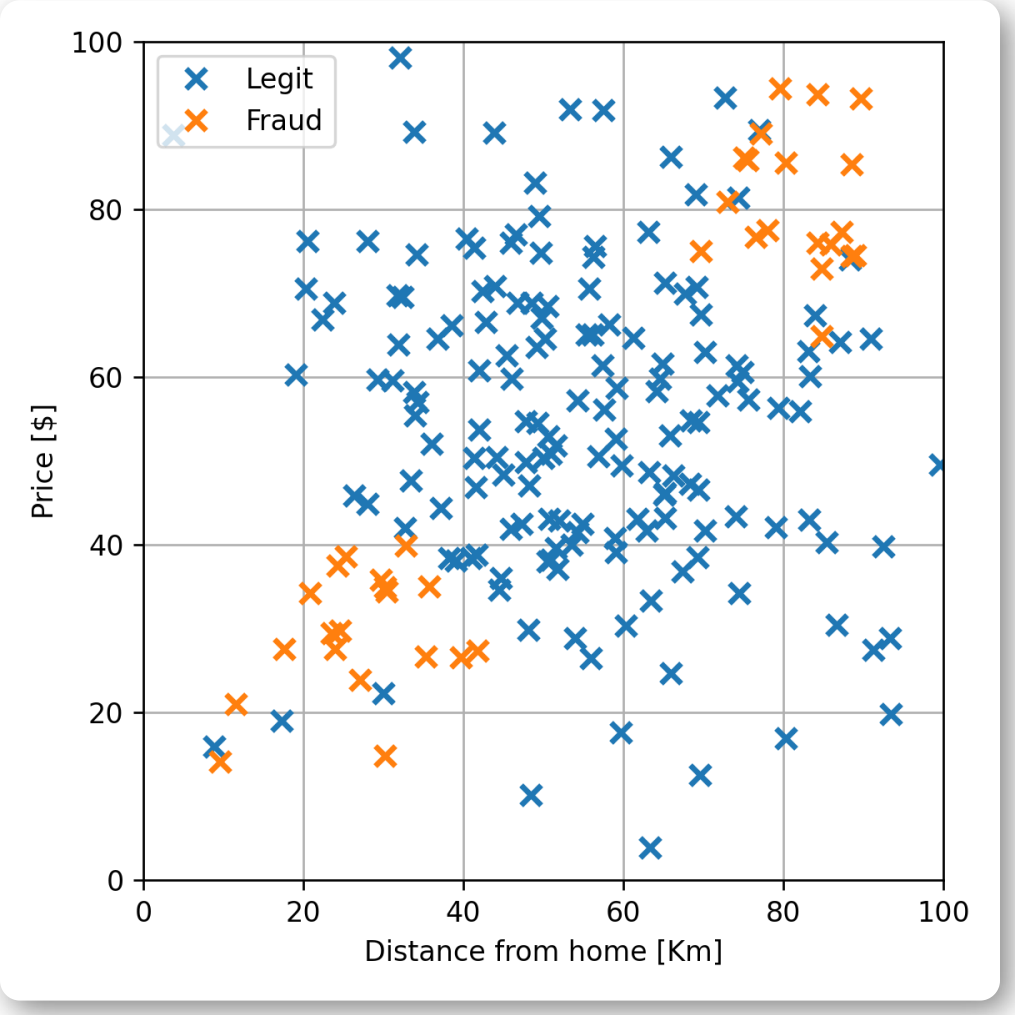
Subjects Covered in this Course

General concepts:
- Cost / Risk / Loss functions
- Features
- Overfitting
- Approximation vs. Estimation trade off
- Bias vs. Variance trade off
- Cross validation
- Regularization
- Bagging
- Boosting
- Gradient Descent



דוגמא לבעיה בגישה הגנרטיבית פרמטרית

נסתכל שוב על הבעיה של חיזוי עסקאות שחשודות כהונאות:



התאמה של מודל QDA

$$p_{\mathrm{y}}(0) = \frac{|\mathcal{I}_0|}{N} = 0.81$$

$$p_{\mathrm{y}}(1) = \frac{|\mathcal{I}_1|}{N} = 0.19$$

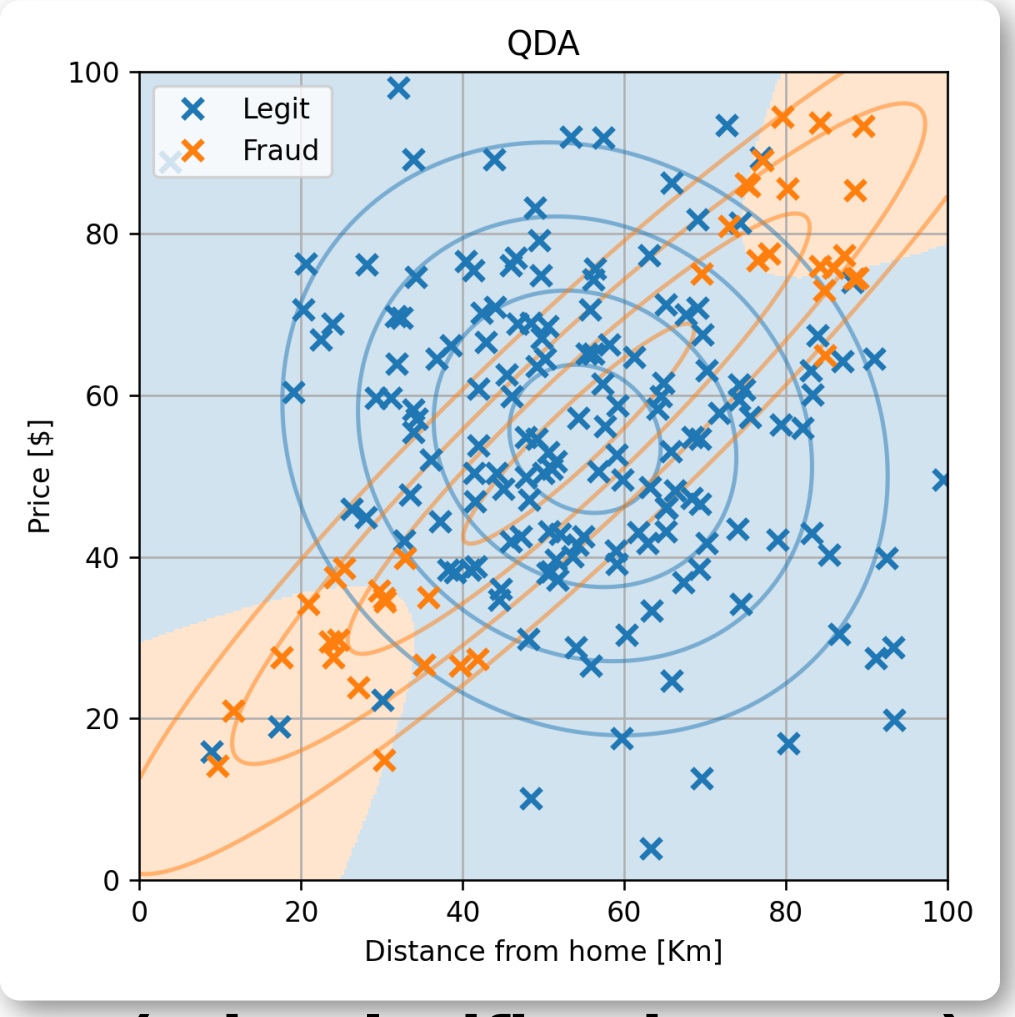
$$\boldsymbol{\mu}_0 = \frac{1}{|\mathcal{I}_0|} \sum_{i \in \mathcal{I}_0} \boldsymbol{x}^{(i)} = [55.1, 54.6]^\top$$

$$\boldsymbol{\mu}_1 = \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \boldsymbol{x}^{(i)} = [54.4, 55.2]^\top$$

$$\Sigma_0 = \frac{1}{|\mathcal{I}_0|} \sum_i \left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}}\right)\left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}}\right)^T = \begin{bmatrix} 350.9 & -42.9 \\ -42.9 & 336 \end{bmatrix}$$

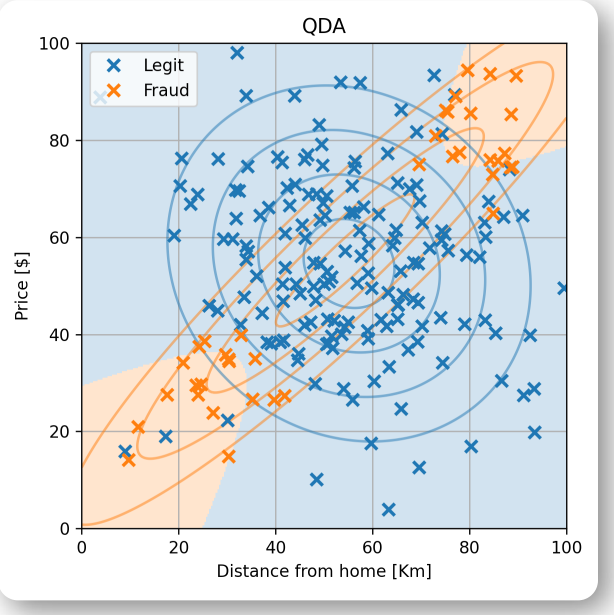
$$\Sigma_1 = \frac{1}{|\mathcal{I}_1|} \sum_i \left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}}\right)\left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}}\right)^T = \begin{bmatrix} 817.9 & 730.5 \\ 730.5 & 741.7 \end{bmatrix}$$

התאמה של מודל QDA



שגיאת החיזוי (miscalssification rate) על ה test set הינה 0.08.

הבעיה של הגישה הגנרטיבית פרמטרית

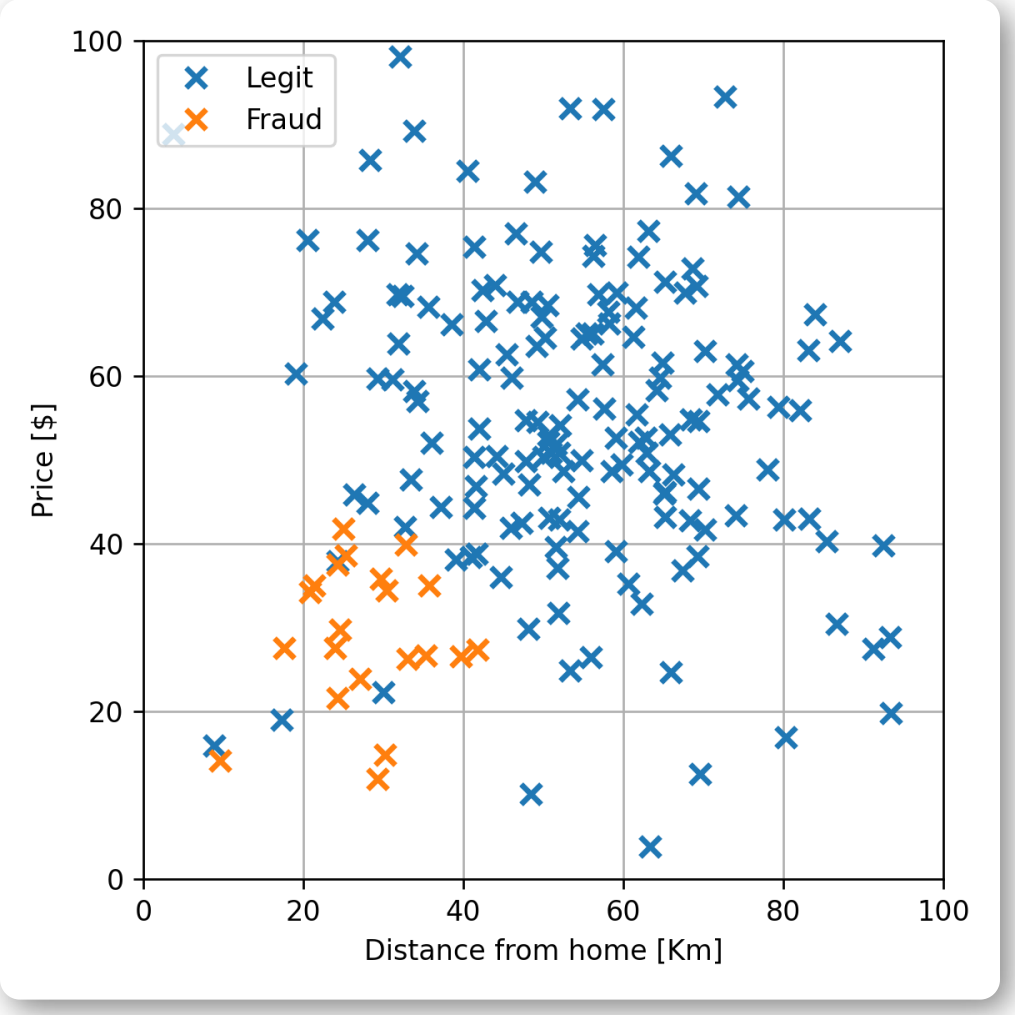


- היינו רוצים מודל אשר יכול לייצג בנפרד את שני האיזורים.
- לצערנו המבחר של המודלים בהם אנו יכולים לא גדול.
- המגבלה הזו נובעת מהצורך שהמודל ייצג פילוגים חוקיים.

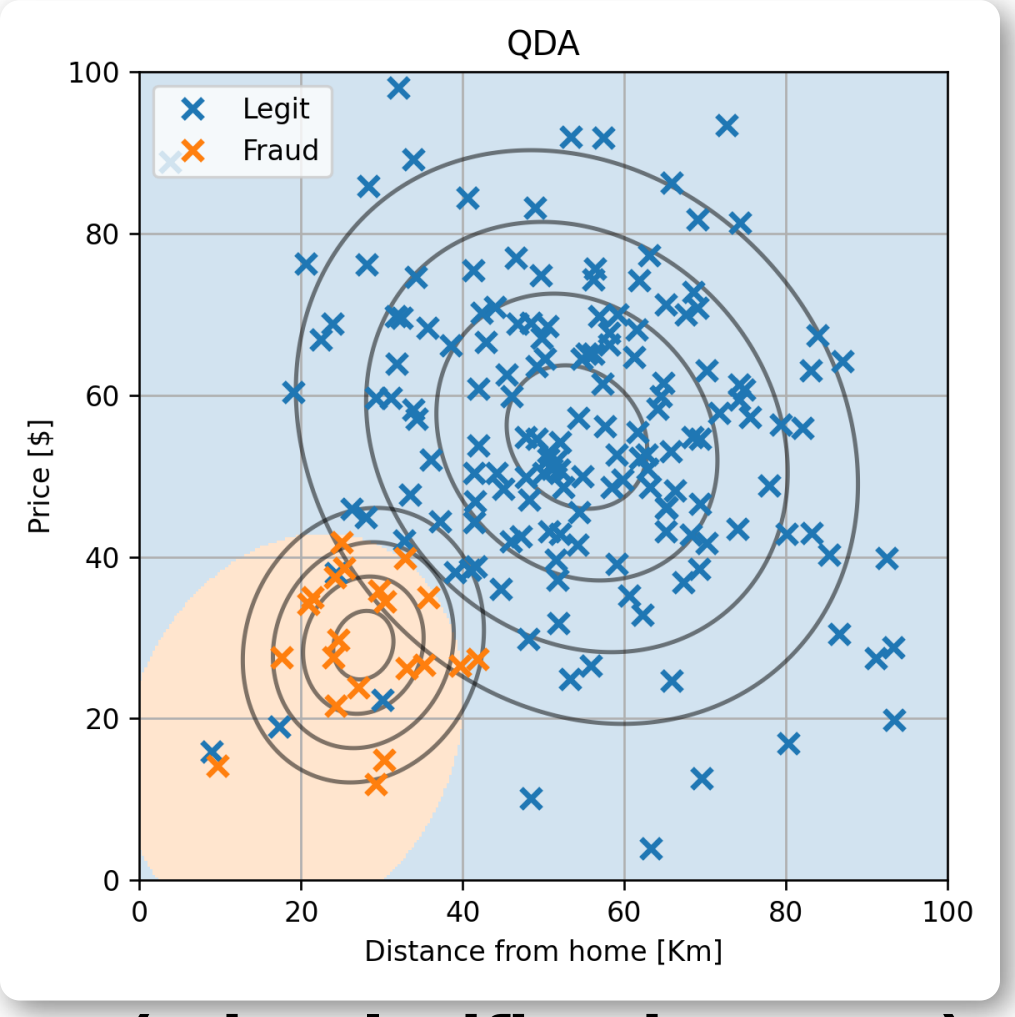
הערה: במקרה זה ניתן להשתמש ב GMM + EM.

דוגמא למדגם שמתאים למודל של QDA

לצורך הדגמה נסתכל על גירסא של המדגם שבה יש רק איזור אחד של ההונאות:



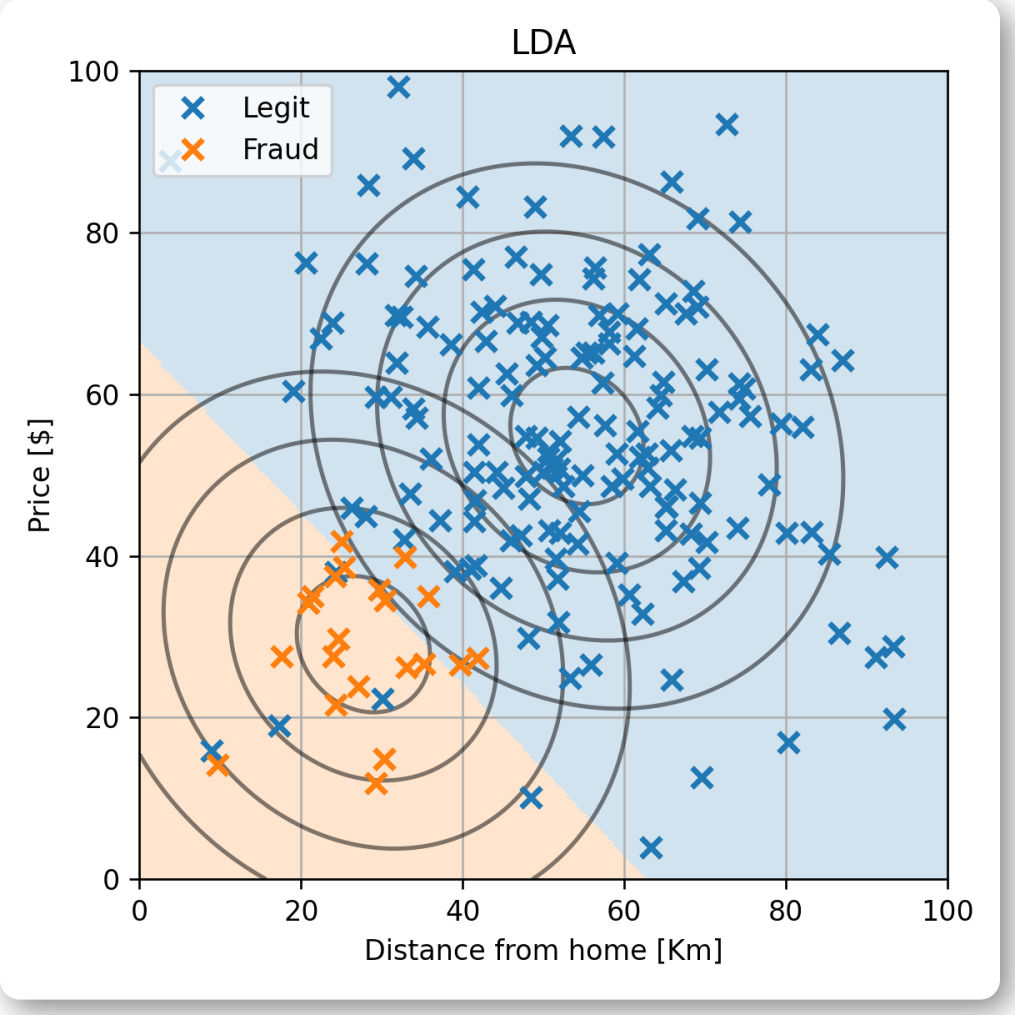
מודל QDA



שגיאת החיזוי (miscalssification rate) על ה test set במקרה הזה הינה 0.

מודל LDA

רק לשם השוואה, נציג גם את התוצאה המתקבלת ממודל ה LDA:



הגישות שראינו עד כה

הגישה הדיסקרימינטיבית

מדגם

▼

חזאי בעל ביצועים טובים על המדגם

הגישה הגנרטיבית

מדגם

▼

הפילוג המשותף של x ו y על סמך המדגם

▼

חזאי אופטימלי בהינתן הפילוג המשותף

הגישה הדיסקרימינטיבית הסתברותית

ברוב פונקציות המחיר החזאי האופטימאלי יהיה תלוי רק בפילוג המותנה של y בהינתן x.

הגישה הדיסקרימינטיבית הסתברותית

מדגם

▼

הפילוג המותנה של y בהינתן x על סמך המדגם

▼

חזאי אופטימלי בהינתן הפילוג המותנה

ההתייחסות לגישה זו במקרות אחרים

- **גישה זו מכוונת ישירות למציאת החזאי ולא מנסה ללמוד את התכונות של המדגם לכן נחשבת לגישה דיסקרימינטיבית.**

- **השם גישה דיסקרימינטיבית הסתברותית לא מופיע במקרים אחרים.**
- **במרבית המקרים מציינים שיש שתי גישות דיסקרימינטיבית אך לא נותנים להם שמות שונים.**

שימוש במודלים פרמטריים

- אנו נבחר מודל פרמטרי אשר יתאר את הפילוג המותנה, $p_{\mathrm{y}|\mathbf{x}}(y|\boldsymbol{x})$.
- נשערך את פרמטרים של המודל בשיטות דומות לגישה הגנרטיבית (MLE ו MAP).

MLE על הפילוג המותנה

$$
\begin{aligned}
\boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{N} \log\left(p_{\mathbf{x},\mathrm{y}}(\boldsymbol{x}^{(i)}, y^{(i)}; \boldsymbol{\theta})\right) \\
&= \arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{N} \log\left(p_{\mathrm{y}|\mathbf{x}}(y^{(i)}|\boldsymbol{x}^{(i)}; \boldsymbol{\theta}) p_{\mathbf{x}}(\boldsymbol{x}^{(i)})\right) \\
&= \arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{N} \log\left(p_{\mathrm{y}|\mathbf{x}}(y^{(i)}|\boldsymbol{x}^{(i)}; \boldsymbol{\theta})\right) - \sum_{i=1}^{N} \log\left(p_{\mathbf{x}}(\boldsymbol{x}^{(i)})\right) \\
&= \arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{N} \log\left(p_{\mathrm{y}|\mathbf{x}}(y^{(i)}|\boldsymbol{x}^{(i)}; \boldsymbol{\theta})\right)
\end{aligned}
$$

- **המשמעות היא ש אין צורך לדעת או לשערך את הפילוג של $\mathbf{x}$.**
- **ניתן להגיע לאותה תוצאה גם עבור משערך MAP.**
- **שימו לב שהפילוג השולי של $\mathbf{x}$ אינו משפיע על הסיווג.**

היתרון של הדיסקרימינטיבית הסתברותית

צריכה לקיים את התנאים הבאים: $p_{\mathbf{x},\mathrm{y}}(\boldsymbol{x},y)$

1. $p_{\mathbf{x},\mathrm{y}}(\boldsymbol{x},y;\boldsymbol{\theta}) \geq 0 \qquad \forall \boldsymbol{x},y,\boldsymbol{\theta}$
2. $\int\int p_{\mathbf{x},\mathrm{y}}(\boldsymbol{x},y;\boldsymbol{\theta})d\boldsymbol{x}dy = 1 \qquad \forall \boldsymbol{\theta}$

עבור בעיות סיווג צריכה לקיים את התנאים הבאים: $p_{\mathrm{y}|\mathbf{x}}(y|\boldsymbol{x})$

1. $p_{\mathrm{y}|\mathbf{x}}(y|\boldsymbol{x};\boldsymbol{\theta}) \geq 0 \qquad \forall \boldsymbol{x},y,\boldsymbol{\theta}$
2. $\sum_{y=1}^{C} p_{\mathrm{y}|\mathbf{x}}(y|\boldsymbol{x};\boldsymbol{\theta}) = 1 \qquad \forall \boldsymbol{x},\boldsymbol{\theta}$

האינטגרל על כל הערכים התחלף בסכום סופי של איברים.
נראה כעת כיצד ניתן לבנות מודלים המקיימים תנאים אלו.

סיווג בינארי

עבור התנאי השני יהיה:

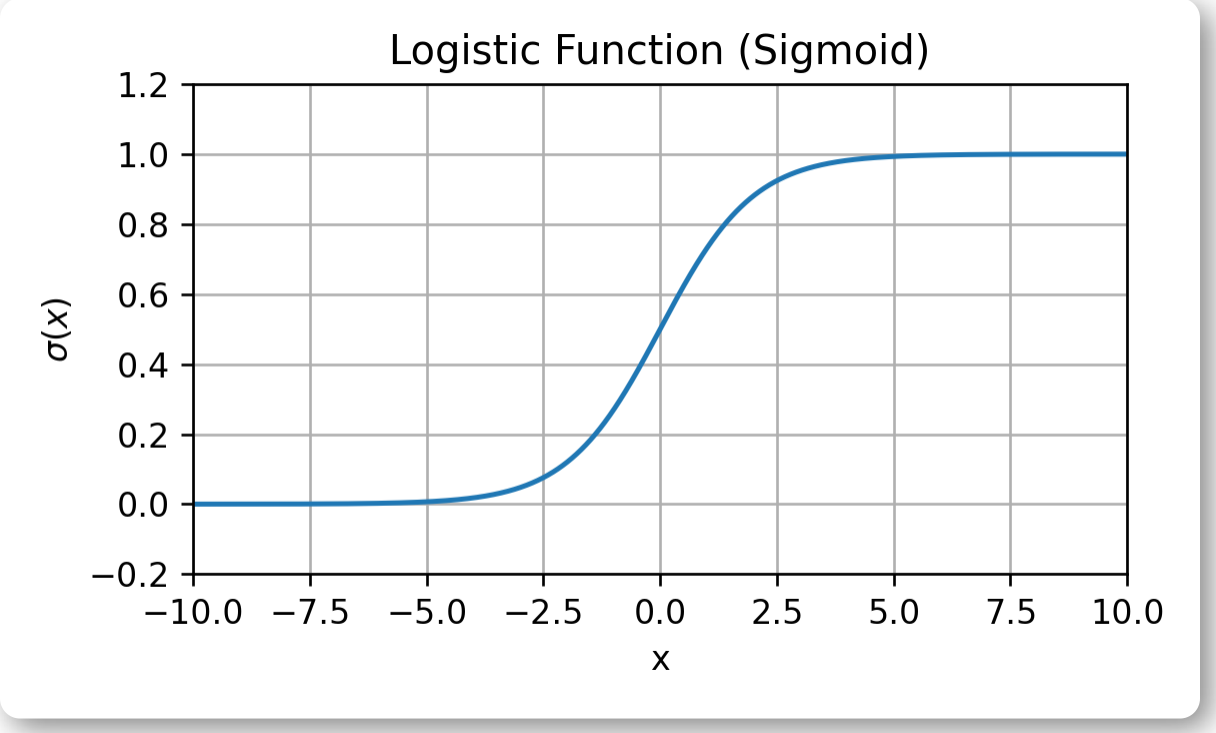
$$p_{\mathrm{y}|\mathbf{x}}(0|\boldsymbol{x};\boldsymbol{\theta}) + p_{\mathrm{y}|\mathbf{x}}(1|\boldsymbol{x};\boldsymbol{\theta}) = 1 \qquad \forall \boldsymbol{x}, \boldsymbol{\theta}$$

דרך פשוטה לקיים תנאי זה הינה למצוא פונקציה $f(x;\theta)$ אשר מחזירה ערכים בין 0 ל 1 ולהגדיר את המודל באופן הבא:

$$p_{\mathrm{y}|\mathbf{x}}(1|\boldsymbol{x};\boldsymbol{\theta}) = f(\boldsymbol{x};\boldsymbol{\theta})$$
$$p_{\mathrm{y}|\mathbf{x}}(0|\boldsymbol{x};\boldsymbol{\theta}) = 1 - f(\boldsymbol{x};\boldsymbol{\theta})$$

הפונקציה הלוגיסטית

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



הערה: מקובל לכנות את הפונקציה הזו סיגמואיד (sigmoid).

הפונקציה הלוגיסטית

כל מודל פרמטרי מהצורה:

$$p_{\mathrm{y}|\mathbf{x}}(1|\boldsymbol{x};\boldsymbol{\theta}) = \sigma(f(\boldsymbol{x};\boldsymbol{\theta}))$$
$$p_{\mathrm{y}|\mathbf{x}}(0|\boldsymbol{x};\boldsymbol{\theta}) = 1 - \sigma(f(\boldsymbol{x};\boldsymbol{\theta}))$$

יהיה מודל פרמטרי חוקי עבור f שמקבלת ערכים חיוביים ושליליים.

תכונות

- **רציפה**
- **מונוטונית עולה**
- $1 - \sigma(z) = \sigma(-z)$
- $\frac{d}{dz} \log(\sigma(z)) = 1 - \sigma(z)$

Binary Logistic Regression

ב Binary Logistic Regression (כלומר, $y \in \{0,1\}$) נשתמש במודל שהצגנו קודם:

$$p_{\mathrm{y}|\mathbf{x}}(1|\boldsymbol{x};\boldsymbol{\theta}) = \sigma(f(\boldsymbol{x};\boldsymbol{\theta}))$$
$$p_{\mathrm{y}|\mathbf{x}}(0|\boldsymbol{x};\boldsymbol{\theta}) = 1 - \sigma(f(\boldsymbol{x};\boldsymbol{\theta}))$$

נמצא את הפרמטרים של המודל בעזרת MLE:

$$\begin{aligned}
\boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta}} -\sum_{i=1}^{N} \log\left(p_{\mathrm{y}|\mathbf{x}}(y^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{\theta})\right) \\
&= \arg\min_{\boldsymbol{\theta}} -\sum_{i=1}^{N} I\{y^{(i)}=1\}\log(\sigma(f(\boldsymbol{x}^{(i)};\boldsymbol{\theta}))) \\
&\qquad + I\{y^{(i)}=0\}\log(1-\sigma(f(\boldsymbol{x}^{(i)};\boldsymbol{\theta}))) \\
&= \arg\min_{\boldsymbol{\theta}} -\sum_{i=1}^{N} y^{(i)}\log(\sigma(f(\boldsymbol{x}^{(i)};\boldsymbol{\theta}))) + (1-y^{(i)})\log(1-\sigma(f(\boldsymbol{x}^{(i)};\boldsymbol{\theta})))
\end{aligned}$$

Binary Logistic Regression

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} -\sum_{i=1}^{N} y^{(i)} \log(\sigma(f(\boldsymbol{x}^{(i)};\boldsymbol{\theta}))) + (1-y^{(i)}) \log(1-\sigma(f(\boldsymbol{x}^{(i)};\boldsymbol{\theta}))$$

במרבית המקרים לא ניתן יהיה לפתור באופן אנליטי ונחפש את הפתרון בשיטות נומריות כגון אלגוריתם ה gradient descent עליו נרחיב בהמשך ההרצאה.

הערה: בגישה הגנרטיבית שתארנו גם לא ניתן בד"כ לחשב אנליטית את פתרון הסבירות המרבית. במקרה הגאוסי זה ניתן כמו שראינו.

Binary Logistic Regression

$$p_{\mathrm{y}|\mathbf{x}}(1|\boldsymbol{x};\boldsymbol{\theta}) = \sigma(f(\boldsymbol{x};\boldsymbol{\theta}))$$
$$p_{\mathrm{y}|\mathbf{x}}(0|\boldsymbol{x};\boldsymbol{\theta}) = 1 - \sigma(f(\boldsymbol{x};\boldsymbol{\theta}))$$

עבור misclassification rate החזאי האופטימאלי יהיה:

$$h(\boldsymbol{x}) = \arg\max_{y} p_{\mathrm{y}|\mathbf{x}}(y|\boldsymbol{x};\boldsymbol{\theta}) = \begin{cases} 1 & \sigma(f(\boldsymbol{x};\boldsymbol{\theta})) > 0.5 \\ 0 & \text{else} \end{cases} = \begin{cases} 1 & f(\boldsymbol{x};\boldsymbol{\theta}) > 0 \\ 0 & \text{else} \end{cases}$$

סיווג לא בינארי

ניתן להרחיב את השיטה לבניית מודלים באמצעות פונקציית ה softmax.

פונקציית ה Softmax

לוקחים וקטור z באורך C ומייצרים ממנו וקטור אשר יכול לייצג פילוג דיסקרטי חוקי.

$$\text{softmax}(\boldsymbol{z}) = \frac{1}{\sum_{c=1}^{C} e^{z_c}} [e^{z_1}, e^{z_2}, \ldots, e^{z_C}]^\top$$

או פונקציה עם טווח רב-ממדי:

$$\text{softmax}(\boldsymbol{z})_i = \frac{e^{z_i}}{\sum_{c=1}^{C} e^{z_c}}$$

פונקציית ה Softmax

$$\text{softmax}(\boldsymbol{z})_i = \frac{e^{z_i}}{\sum_{c=1}^{C} e^{z_c}}$$

תכונות

- $\text{softmax}(\boldsymbol{z} + a)_i = \text{softmax}(\boldsymbol{z})_i \ \forall i$.
- $\frac{\partial}{\partial z_j} \log(\text{softmax}(\boldsymbol{z}))_i = \delta_{i,j} - \text{softmax}(\boldsymbol{z})_j$

הפונקציה הלוגיסטית כמקרה פרטי

עבור וקטור באורך 2: $z = [a, b]$, נקבל:

$$\text{softmax}(\boldsymbol{z})_1 = \frac{e^a}{e^a + e^b} = \frac{1}{1 + e^{b-a}} = \sigma(a - b)$$

$$\text{softmax}(\boldsymbol{z})_2 = \frac{e^b}{e^a + e^b} = 1 - \sigma(a - b)$$

Non-Binary) Logistic Regression)

עבור C פונקציות פרמטריות כלשהן, $f_c(\boldsymbol{x};\boldsymbol{\theta}_c)$, ניתן לבנות מודל פרמטרי חוקי באופן הבא:

$$p_{\mathrm{y}|\mathbf{x}}(y|\boldsymbol{x};\boldsymbol{\theta}) = \frac{e^{f_y(\boldsymbol{x};\boldsymbol{\theta}_y)}}{\sum_{c=1}^{C} e^{f_c(\boldsymbol{x};\boldsymbol{\theta}_c)}}$$

לשם נוחות נסמן:

- $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \ldots, \boldsymbol{\theta}_C^\top]^\top$.
- $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}) = [f_1(\boldsymbol{x};\boldsymbol{\theta}_1), f_2(\boldsymbol{x};\boldsymbol{\theta}_2), \ldots, f_C(\boldsymbol{x};\boldsymbol{\theta}_C)]^\top$

נוכל לרשום את המודל הפרמטרי באופן הבא:

$$p_{\mathrm{y}|\mathbf{x}}(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathrm{softmax}(\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}))_y$$

Non-Binary) Logistic Regression)

$$p_{\mathrm{y}|\mathbf{x}}(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathrm{softmax}(\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}))_y$$

משערך ה MLE של מודל זה יהיה נתון על ידי:

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta}} -\sum_{i=1}^{N} \log\left(p_{\mathrm{y}|\mathbf{x}}(y^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{\theta})\right) \\ &= \arg\min_{\boldsymbol{\theta}} -\sum_{i=1}^{N} \log(\mathrm{softmax}(\boldsymbol{f}(\boldsymbol{x}^{(i)};\boldsymbol{\theta}))_{y^{(i)}})\end{aligned}$$

היתירות בייצוג של מודל ה logistic regression

- במקרה הבינארי לא היינו צריכים להגדיר 2 פונקציות פרמטריות.
- במקרה הכללי מספיק להגדיר $C-1$ פונקציות פרמטריות.
- הסתברות של $C-1$ מחלקות תקבע באופן מוחלט את המחלקה האחרונה כך שהיא תשלים את ההסתברות ל-1.
- כל שינוי מהצורה של $f_c(\boldsymbol{x};\boldsymbol{\theta}_c) \rightarrow f_c(\boldsymbol{x};\boldsymbol{\theta}_c) + g(\boldsymbol{x})$ לא ישנה את הפילוג המותנה

במקרים מסויימים נרצה לבטל יתירות זו. ניתן לעשות זאת על ידי קיבוע של $f_1(\boldsymbol{x};\boldsymbol{\theta}_1) = 0$

Linear Logistic Regression

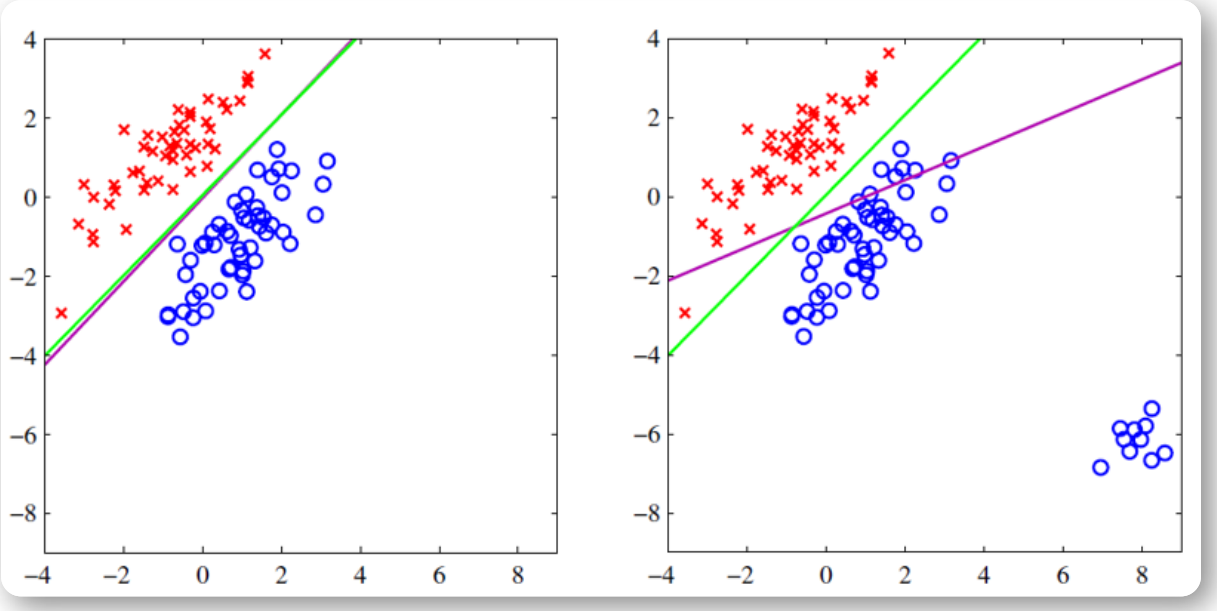
המקרה שבו הפונקציות הפרמטריות הם לינאריות:

$$f_c(\boldsymbol{x};\boldsymbol{\theta}_c) = \boldsymbol{\theta}_c^\top \boldsymbol{x}$$

אפשר כמובן להוסיף סף ע"י תוספת קבוע לאגף ימין.

- **במקרה זה פונקציית ה objective היא קמורה (convex) ומובטח ש gradient descent, במידה והוא מתכנס, יתכנס למינימום גלובלי.**
- **כאשר הפונקציה f אינה לינארית, פונקציית המחיר בד"כ אינה קמורה ואז ייתכנו הרבה מינימות מקומיות.**

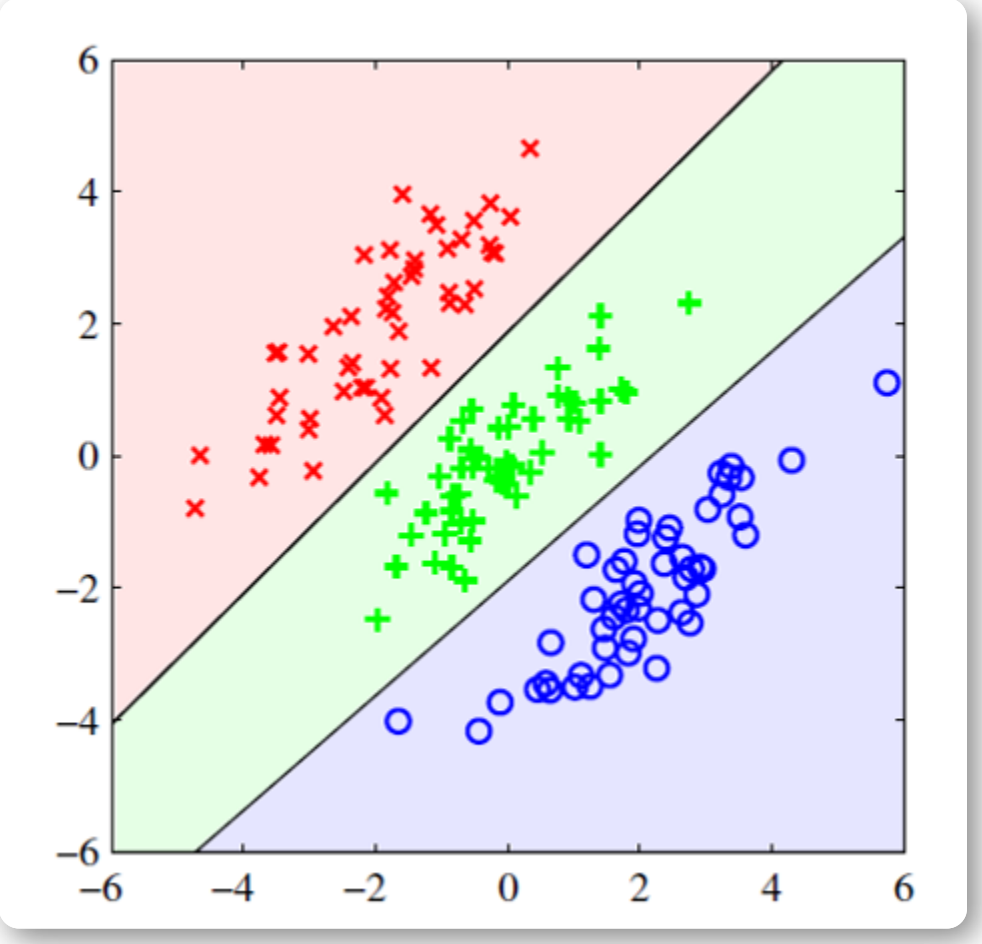
Linear Logistic Regression



ירוק - סיווג בינארי לוגיסטי. סגול - קריטריון אחר.

האיור מתוך, C.M. Bishop, Pattern Recognition and Machine Learning

Linear Logistic Regression



סיווג לוגיסטי לשלוש מחלקות.

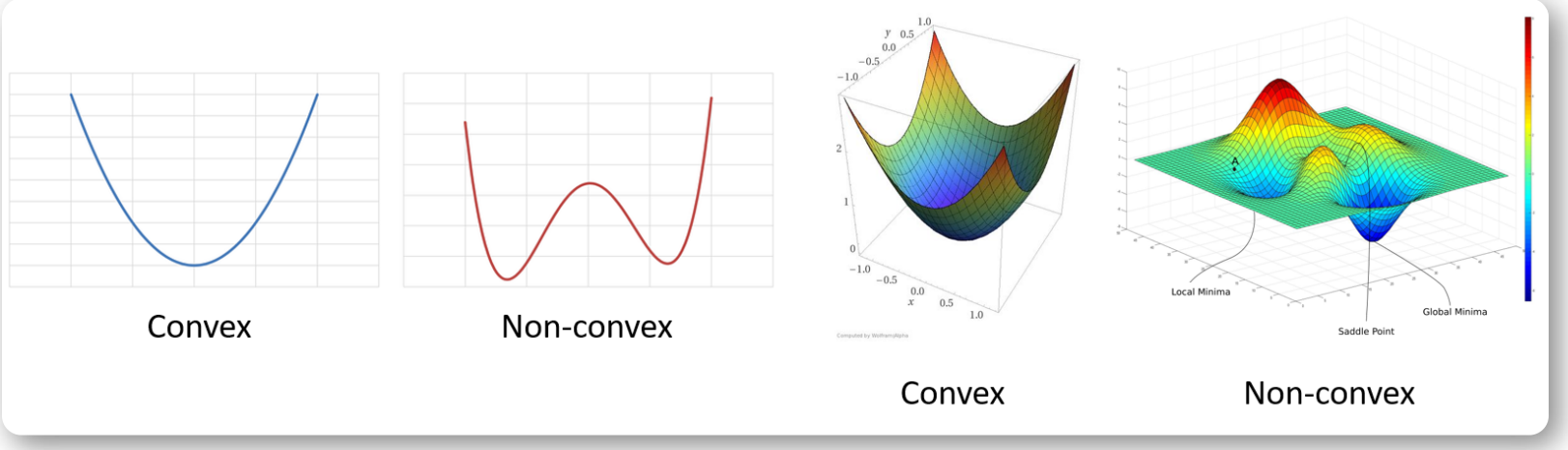
האיור מתוך, C.M. Bishop, Pattern Recognition and Machine Learning

Gradient descent (שיטת הגרדיאנט)

האלגוריתם מנסה למצוא מינימום מקומי על ידי התקדמות בצעדים קטנים בכיוון שבו הפונקציה יורדת הכי מהר.

Gradient descent (שיטת הגרדיאנט)

- אלגוריתם חמדן (greedy): מנסה בכל איטרציה לשפר את מצבו לעומת המצב הנוכחי
- יתכנס למינימום מקומי.
- הדרישה היחידה הינה היכולת לחשב את הנגזרת של פונקציית המטרה.



Gradient descent (שיטת הגרדיאנט)

עבור בעיית המינמיזציה:

$$\arg\min_{\boldsymbol{\theta}} \quad g(\boldsymbol{\theta})$$

- **מאתחלים את $\theta^{(0)}$ לנקודה אקראית כל שהיא.**
- **חוזרים על צעד העדכון הבא עד שמתקיים תנאי עצירה:**

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}^{(t)})$$

בתרגול תראו שעבור גודל צעד קטן כיוון הגרדיאנט השלילי הוא זה המבטיח ירידה מרבית בערך הפונקציה.

את הפרמטר η יש לקבוע מראש, והוא יקבע את גודל הצעדים שהאלגוריתם יעשה.

תנאי עצירה

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}^{(t)})$$

- **מספר צעדי עדכון שנקבע מראש:** $t > \text{max-iter}$.
- **הנורמה של הגרדיאנט קטנה מערך סף:** $\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta})\|_2 < \epsilon$
- **השיפור בפונקציית המטרה קטן מערך סף:** $g(\boldsymbol{\theta}^{(t-1)}) - g(\boldsymbol{\theta}^{(t)}) < \epsilon$
- **שימוש בעצירה מוקדמת על מנת להתמודד עם התאמת יתר (נרחיב על כך בהרצאה הבאה)**

הבעיות של האלגוריתם

- התכנסות למינימום מקומי ותלות באיתחול
- לא ניתן לקבוע בוודאות האם האלגוריתם התכנס
- בעיית הבחירה של גודל הצעד

שתי הבעיות הראשונות מונעות הגעה לאופטימום אך עדיין לא מפריעות לאלגוריתם להניב תוצאות טובות.
הבעיה של בחירת גודל צעד עלולה למנוע מהאלגוריתם להניב תוצאות רלוונטיות תוך מספר סביר של צעדים.

דוגמא גודל צעד קטן

דוגמא גודל צעד גדול

דוגמא גודל צעד גדול מידי

בעיית הבחירה של גודל הצעד

- כאשר יהיו בבעיה כיוונים שונים בהם ישנו הבדל גדול בקצב השינוי של הפונקציה לרוב לא יהיה גודל צעד אשר יגרום לפונקציה להתכנס במספר סביר של צעדים. ראו למשל איור בשקפים הקודמים.
- gradient descent בצורתו הפשוטה אינו מאד שימושי.
- למזלנו ישנם מספר שיפורים שניתן לעשות על מנת להתמודד עם בעיה זו.
- לצערנו בקורס זה לא נספיק לכסות שיפורים אלו.
- האלגוריתם נידון בהרחבה בקורס "אופטימזציה" ובקורס "למידה עמוקה".

שיפורים נפוצים

1. הוספה של רכיב תנע (מומנטום) לאלגוריתם
2. שימוש בגודל צעד אשר משתנה במהלך הריצה

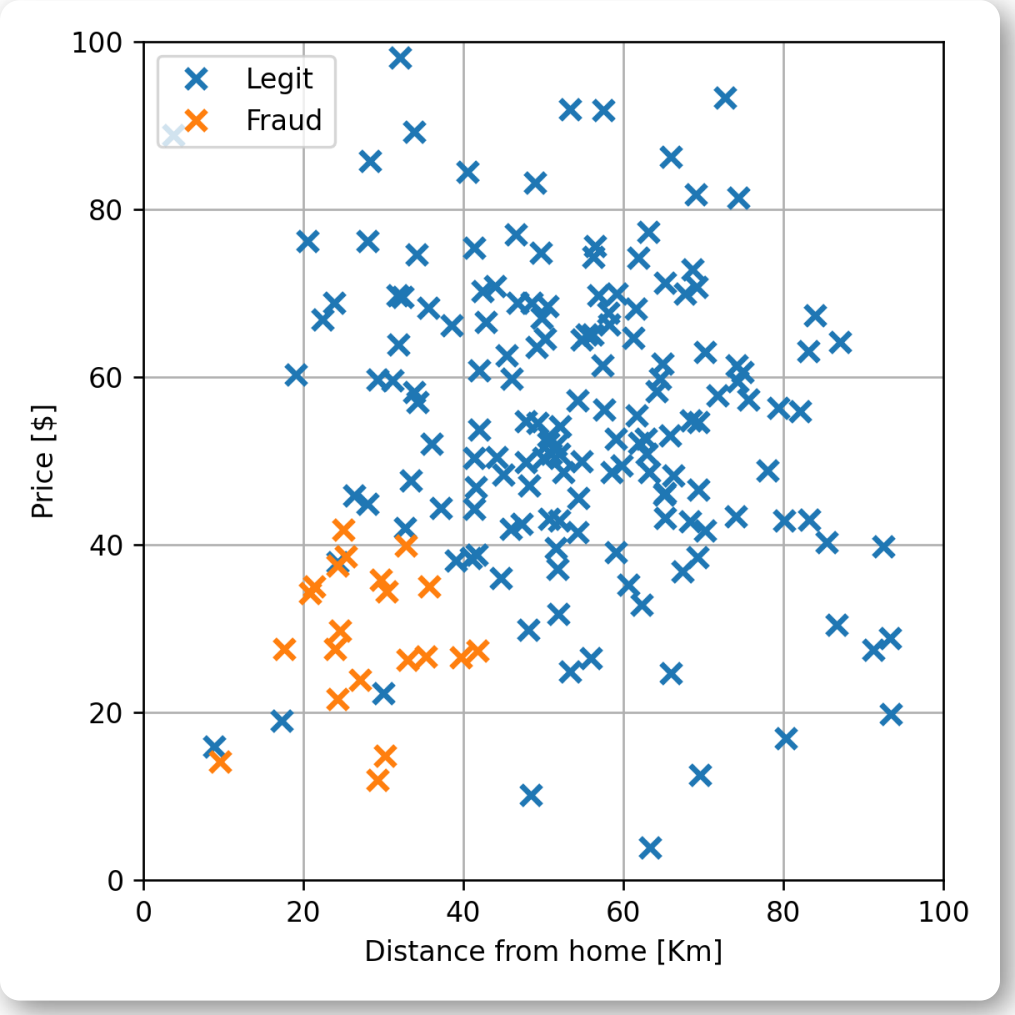
לקריאה על נושא:

1. An overview of gradient descent optimization algorithms
2. Why Momentum Really Works

בתרגיל הרטוב תשתמשו במימוש קיים ADAM.

דוגמא: Linear Logistic Regression

נחזור לבעיה של חיזוי עסקאות החשודות כהונאות אשראי.



דוגמא: Linear Logistic Regression

נשתמש במודל של linear logistic regression:

$$p_{\mathrm{y}|\mathbf{x}}(y|\boldsymbol{x};\boldsymbol{\theta}) = \begin{cases} \sigma(\boldsymbol{x}^\top \boldsymbol{\theta}) & y = 1 \\ 1 - \sigma(\boldsymbol{x}^\top \boldsymbol{\theta}) & y = 0 \end{cases}$$

נמצא את הפרמטרים של המודל בעזרת MLE:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} -\sum_{i=1}^{N} y^{(i)} \log(\sigma(\boldsymbol{x}^{(i)\top}\boldsymbol{\theta})) + (1 - y^{(i)}) \log(1 - \sigma(\boldsymbol{x}^{(i)\top}\boldsymbol{\theta}))$$

כלל העדכון של האלגוריתם יהיה:

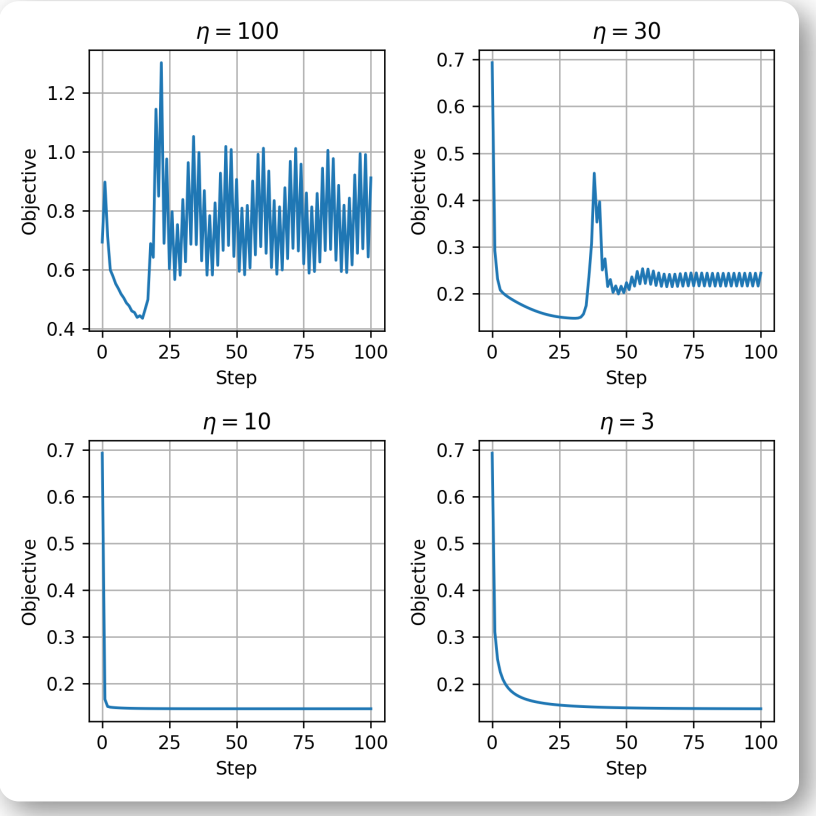
$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \eta \sum_{i=1}^{N} \left(y^{(i)}(1 - \sigma(\boldsymbol{x}^{(i)\top}\boldsymbol{\theta})) - (1 - y^{(i)})\sigma(\boldsymbol{x}^{(i)\top}\boldsymbol{\theta}) \right) \boldsymbol{x}^{(i)}$$

נזכור כי עבור בעיה זו, פונקציית המחיר קמורה.

בחירת η

נריץ את האלגוריתם מספר קטן של צעדים עבור ערכי η שונים:

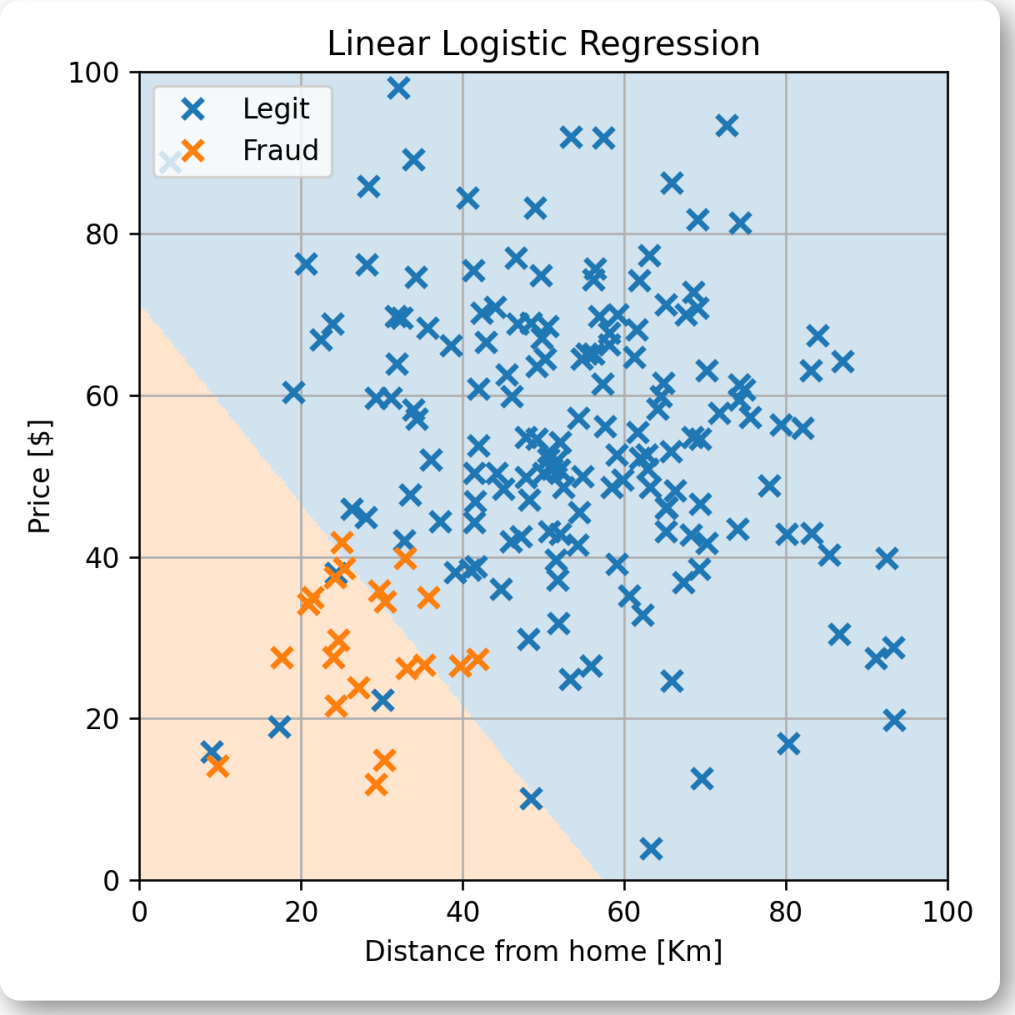
בחירת η



- $\eta = 30$ **ו** $\eta = 100$ **מתאימים למקרה של** η **גדול מידי.**
- **נבחר את** $\eta = 10$**.**

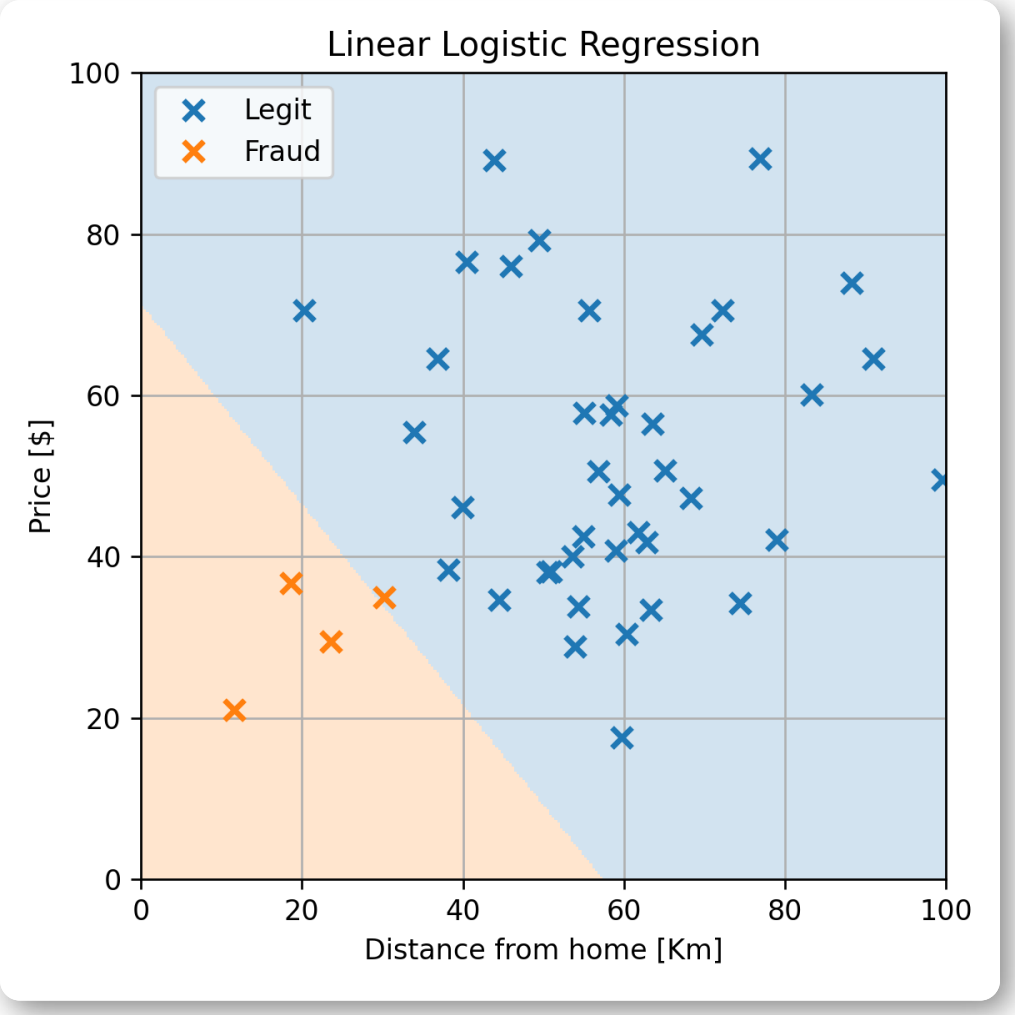
דוגמא: Linear Logistic Regression

נריץ את האלגוריתם עם $\eta = 10$ ונקבל את החזאי הבא:



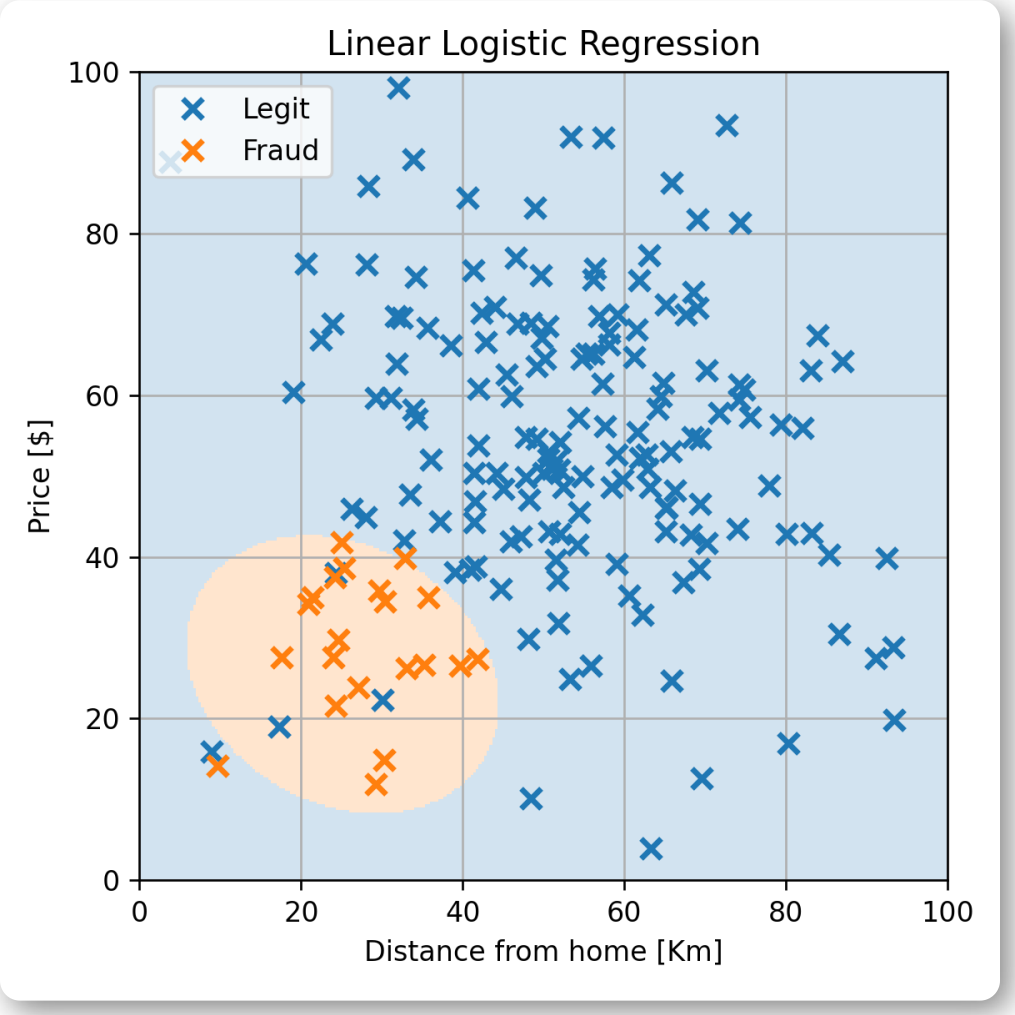
דוגמא: Linear Logistic Regression

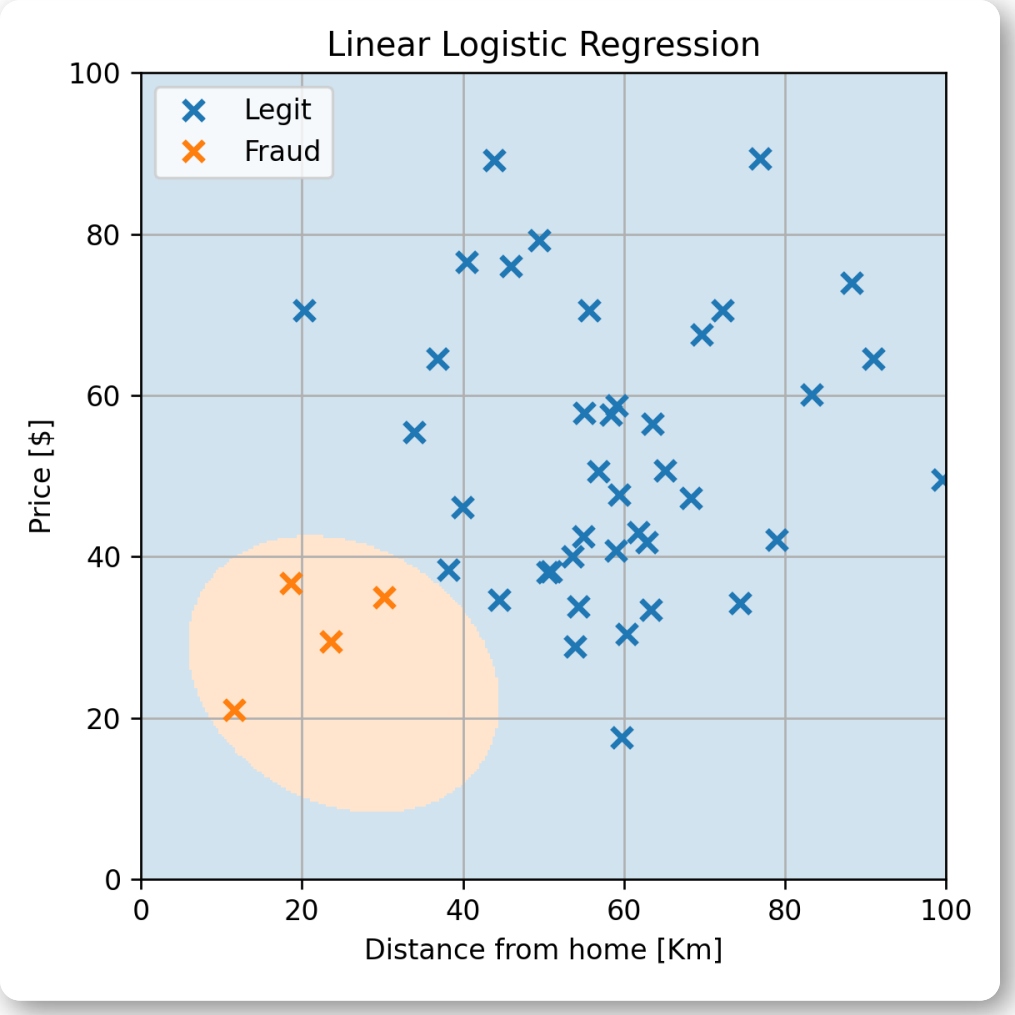
נקבל misclassification rate של 0.02 על ה test set:



שימוש במודל מסדר גבוה יותר

נוכל להשתמש בכל מודל שנרצה.
נחליף את $f(x;\theta)$ בפולינום מסדר שני ונקבל:



שימוש במודל מסדר גבוה יותר

נקבל misclassification rate של 0 על ה test set: