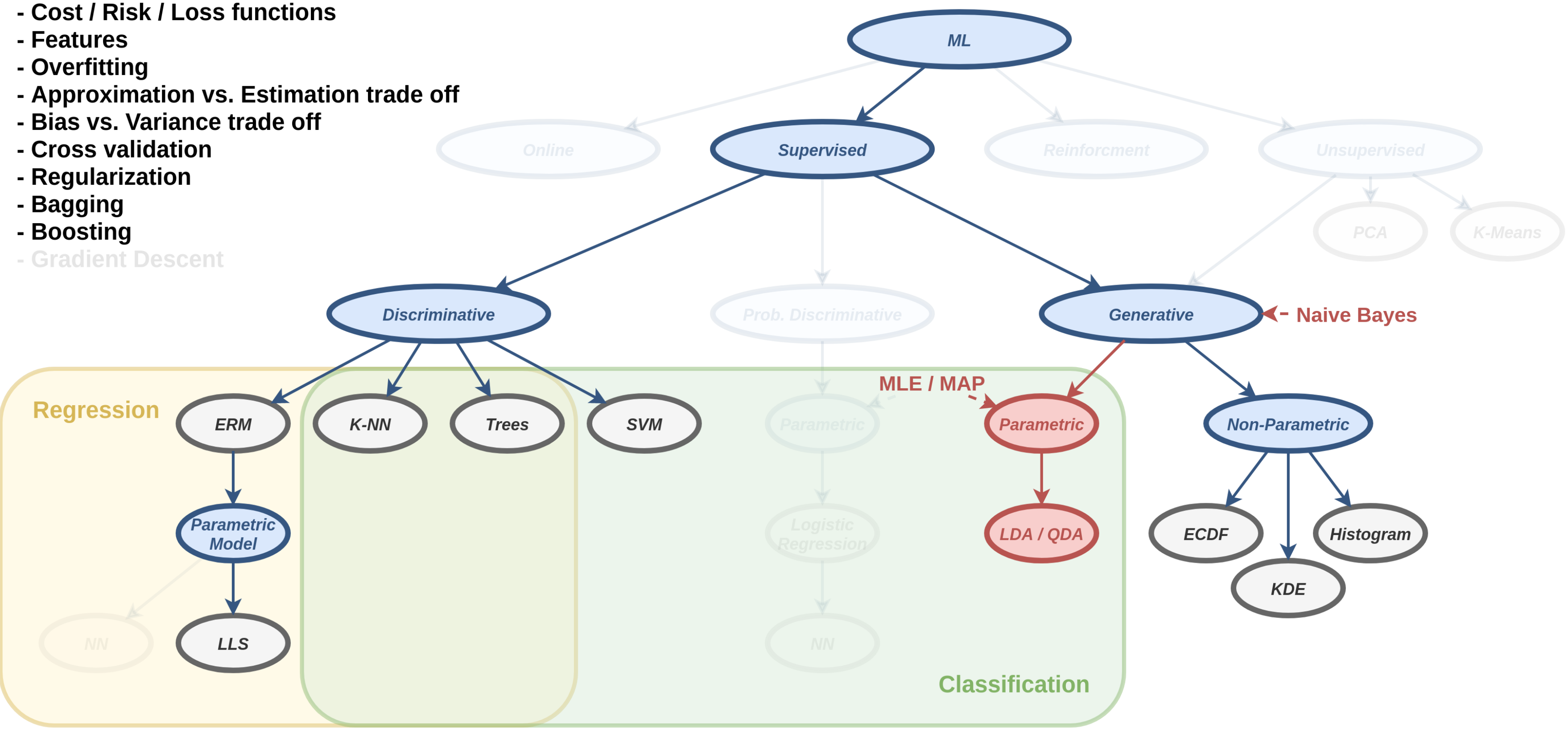
הרצאה 8 - שיערוך פילוג בשיטות פרמטריות וסיווג גנרטיבי

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

נניח שאנו רוצים לחזות האם אדם מסויים פרק את הכתף על פי הסימפטומים שלו. לשם כך נסתכל על המדגם הבא

	פריקה	כאב בכתף	נפיחות	סימנים כחולים	נימול ביד	נזלת
1	+	+	+	+	+	-
2	+	+	+	+	-	-
3	+	+	+	-	+	-
4	+	+	+	+	-	+
5	-	-	-	-	-	-
6	-	+	-	-	-	+
7	-	-	-	+	-	-
8	-	+	-	-	-	-

האם לאדם עם כל הסימפטומים יש פריקה של הכתף?

שיערוך נאיבי - חוסר תלות בין המשתנים

נניח שאנו רוצים לשערך צפיפות הסתברות של משתנה D ממדי. שיטה מאד נאיבית (לא מתוחכמת) לפתור את הבעיה היא להתעלם מהקשר בין המשתנים השונים ולהניח שהם בלתי תלויים סטטיסטית. זאת אומרת ש:

$$p_{\mathrm{x}}(\boldsymbol{x}) = p_{\mathrm{x}_1}(x_1)p_{\mathrm{x}_2}(x_2)\dots p_{\mathrm{x}_D}(x_D) = \prod_{d=1}^{D} p_{\mathrm{x}_d}(x_d)$$

- חיסרון: שהיא מגבילה מאד את הפילוגים שניתן ללמוד.

מסווג בייס נאיבי - Naïve Bayes Classification

תזכורת במשימות סיווג קיים $\mathrm{x} \in R^D$ ו-$\mathrm{y} \in \{1, 2, ..., C\}$. בנוסף מוגדרת הסתברות משותפת $P(\mathrm{x}, \mathrm{y})$.
נוכל להשתמש בשערוך הנאיבי לפתרון בעיות סיווג.
נניח כי:

$$p_{\mathrm{x}|\mathrm{y}}(\boldsymbol{x}|y) = p_{\mathrm{x}_1|\mathrm{y}}(x_1|y)p_{\mathrm{x}_2|\mathrm{y}}(x_2|y)\ldots p_{\mathrm{x}_D|\mathrm{y}}(x_D|y) = \prod_{d=1}^{D} p_{\mathrm{x}_d|\mathrm{y}}(x_d|y)$$

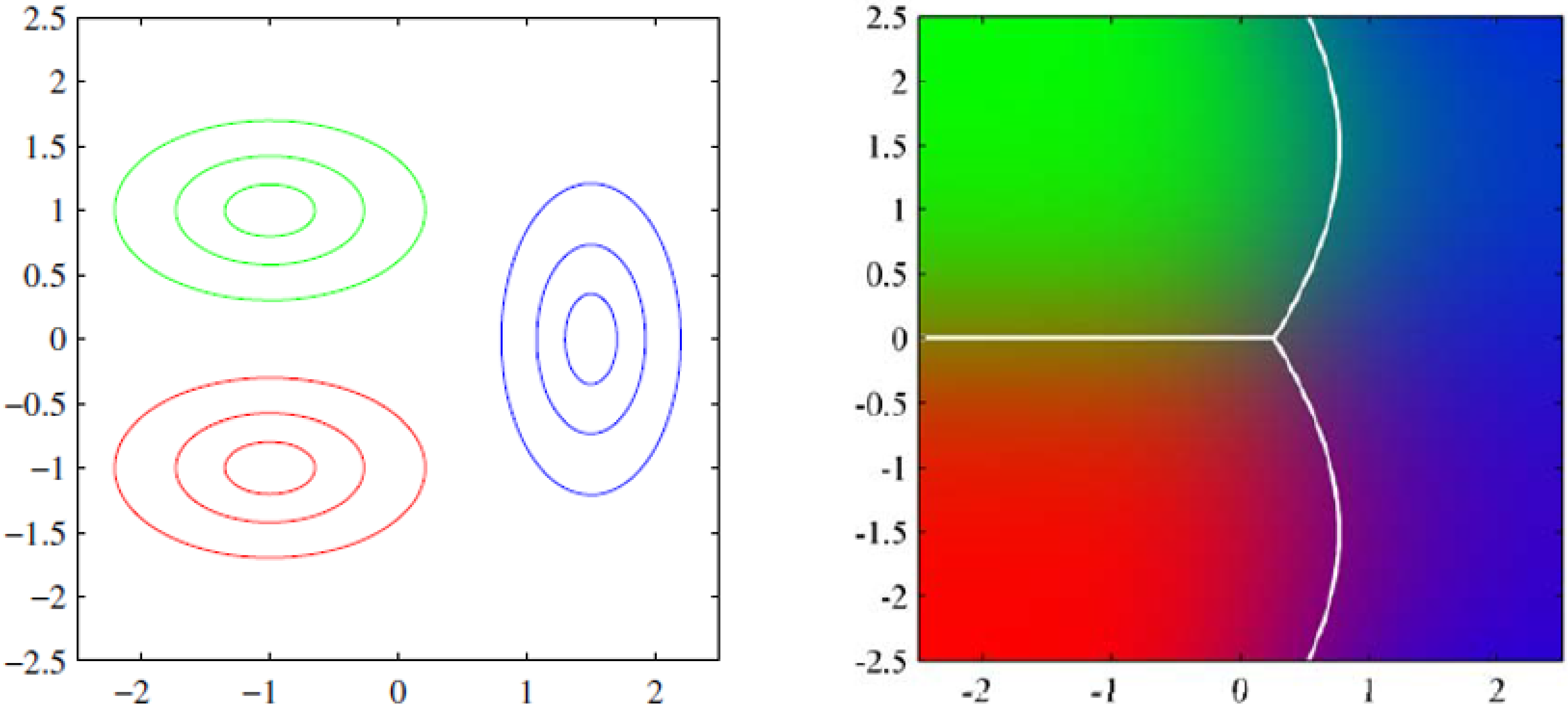
- **זו כמובן הנחה מאוד פשטנית שאינה מתקיימת במציאות.**
- **אנו לא נרצה להניח חוסר תלות בין x ל y.**

מסווג בייס נאיבי - Naïve Bayes Classification

החזאי אשר ממזער את הסיכוי לטעות יהיה:

$$\begin{aligned}\hat{y}=h(\boldsymbol{x})&=\underset{y}{\arg\max}\, p_{\mathrm{y}|\boldsymbol{x}}(y|\boldsymbol{x})\\&=\underset{y}{\arg\max}\, p_{\boldsymbol{x}|\mathrm{y}}(\boldsymbol{x}|y)p_{\mathrm{y}}(y)\\&=\underset{y}{\arg\max}\, p_{\mathrm{y}}(y)\prod_{d=1}^{D}p_{\mathrm{x}_d|\mathrm{y}}(x_d|y)\end{aligned}$$

מסווג בייס נאיבי - Naïve Bayes Classification



$p(x|y),\ y = 1, 2, 3$ $\qquad$ $p(y|x),\ y = 1, 2, 3$

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

דוגמא 1 - זיהוי פריקה של הכתף

תחת הנחת חוסר התלות נשערך בנפרד את ההסתברות המותנית של כל אחד מהרכיבים $p_{\mathrm{x}_d|\mathrm{y}}(x_d|y), x_d \in \{0,1\}$. לדוגמא:

$$p_{\mathrm{x}_{\mathrm{pain}}|\mathrm{y}}(x_{\mathrm{pain}}|1) = \begin{cases} \frac{4}{4} = 1 & 1 \\ \frac{0}{4} = 0 & 0 \end{cases}$$

$$p_{\mathrm{x}_{\mathrm{pain}}|\mathrm{y}}(x_{\mathrm{pain}}|0) = \begin{cases} \frac{2}{4} = 0.5 & 1 \\ \frac{2}{4} = 0.5 & 0 \end{cases}$$

$$p_{\mathrm{x}_{\mathrm{swelling}}|\mathrm{y}}(x_{\mathrm{swelling}}|1) = \begin{cases} \frac{4}{4} = 1 & 1 \\ \frac{0}{4} = 0 & 0 \end{cases}$$

$$p_{\mathrm{x}_{\mathrm{swelling}}|\mathrm{y}}(x_{\mathrm{swelling}}|0) = \begin{cases} \frac{0}{4} = 0 & 1 \\ \frac{1}{4} = 1 & 0 \end{cases}$$

ובאופן דומה לשאר הרכיבים.

דוגמא 1 - זיהוי פריקה של הכתף

החיזוי עבור המקרה בו מופיעים כל הסימפטומים יהיה

$$\hat{y}=\arg\max_{y} p_{\mathrm{y}}(y)\prod_{d=1}^{D}p_{\mathrm{x}_d|\mathrm{y}}(1|y)$$

זאת אומרת שעלינו לבדוק האם:

$$p_{\mathrm{y}}(1)\prod_{d=1}^{D}p_{\mathrm{x}_d|\mathrm{y}}(1|1)\overset{?}{>}p_{\mathrm{y}}(0)\prod_{d=1}^{D}p_{\mathrm{x}_d|\mathrm{y}}(1|0)$$

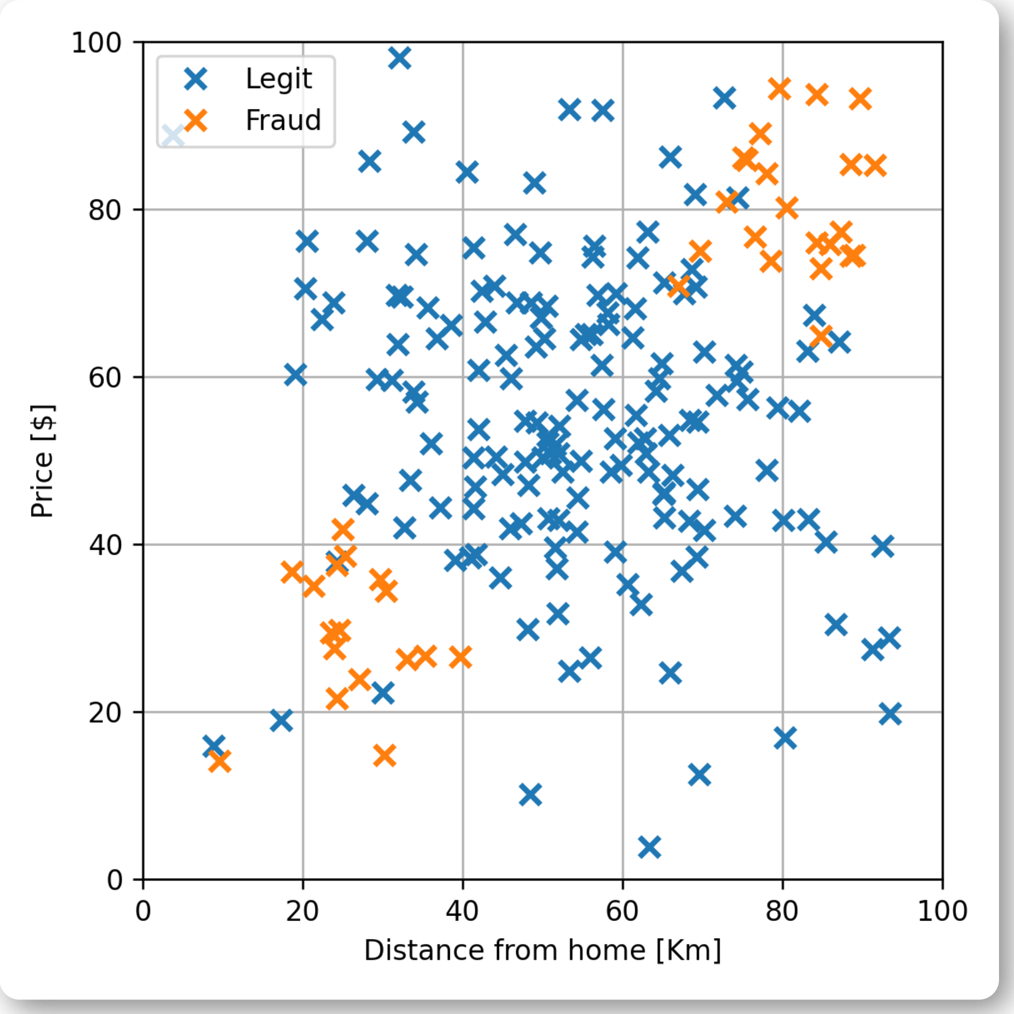
$$0.5\times 1\times 1\times 0.75\times 0.5\times 0.25\overset{?}{>}0.5\times 0.5\times 0\times 0.25\times 0\times 0.25$$

$$0.09375\overset{?}{>}0$$

מכיוון שתנאי זה מתקיים, החיזוי יהיה שישנה פריקה של כתף.

דוגמא 2 - זיהוי הונאות אשראי

ננסה להשתמש בשיטה זו לבעיית חיזוי הונאות האשראי



שיטות פרמטריות

- דומה לשימוש שעשינו במודלים פרמטריים בגישה הדטרמיניסטית.
- נגביל את הצורה של הפונקציה שאותה אנו רוצים לשערך למודל פרמטרי.
- נסמן את וקטור הפרמטרים ב θ.
- כאן המודל חייב לייצר פילוג חוקי עבור כל בחירה של פרמטרים.
- מגבלה קשה אשר מצמצמת מאד את המודלים הפרמטריים שאיתם ניתן לעבוד.

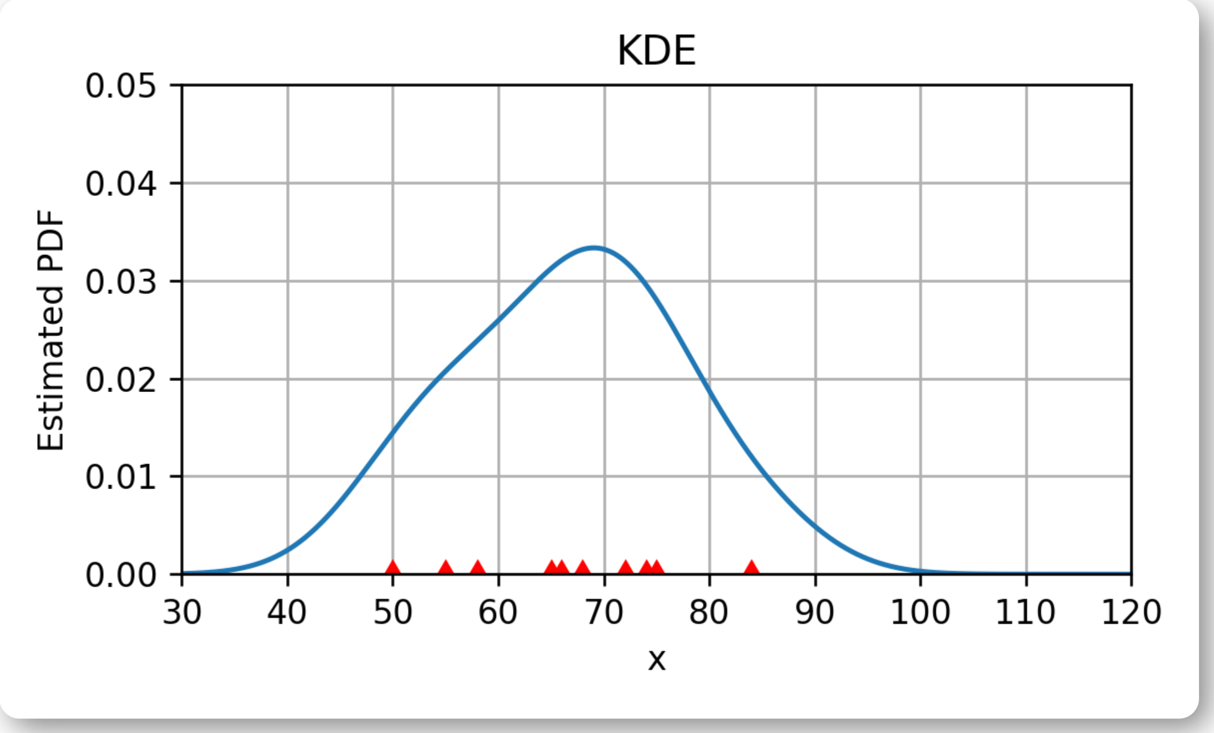
בחירת הפרמטרים

- נרצה למצוא דרך לתת "ציון" לכל בחירה של פרמטרים ולחפש את הפרמטרים אשר מניבים את הציון הטוב ביותר.
- נציג שתי גישות שונות להתייחס לפרמטרים של המודל.
- כל גישה מובילה לדרך מעט שונה לבחירה של הפרמטרים.

דוגמא: שיערוך הפילוג של זמן הנסיעה

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

משערך ה KDE של הפילוג (לא דנו הסמסטר) הינו:



נרצה לשערך פרמטרים של פילוג נורמלי שיתאר בצורה טובה את הדגימות במדגם.

הגישה הלא-בייסיאנית
(קלאסית או תדירותית (Frequentist))

$$p_{\mathrm{x}}(\boldsymbol{x};\boldsymbol{\theta})$$

- נתייחס לפרמטרים כאל מספרים שאותם יש לקבוע על מנת שהמודל יתאר בצורה טובה את המדגם.
- ההנחה היא כי יש ערך לא ידוע של הפרמטר שהוא ה"טוב" ביותר.

Maximum Likelyhood Estimator ((MLE

- נסמן ב $p_{\mathcal{D}}(\mathcal{D};\boldsymbol{\theta})$ את ההסתברות לקבלת המדגם $\mathcal{D}=\{\boldsymbol{x}^{(i)}\}$.
- גודל זה מכונה ה**סבירות** (**likelihood**) של המדגם.
- אנו מעוניינים למצוא את הפרמטרים θ אשר מניבים את הסבירות הכי גבוהה.
- מקובל לסמן את פונקציית ה likelihood באופן הבא:

$$\mathcal{L}(\boldsymbol{\theta};\mathcal{D}) \triangleq p_{\mathcal{D}}(D;\boldsymbol{\theta})$$

משערך ה MLE של θ הוא הערך אשר ממקסם את פונקציית ה likelihood:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \underset{\boldsymbol{\theta}}{\arg\max}\, \mathcal{L}(\boldsymbol{\theta};\mathcal{D})$$

Maximum Likelyhood Estimator ((MLE

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$$

- נרשום את בעיית האופטימיזציה כבעיית מינימיזציה:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\min_{\boldsymbol{\theta}} \; -\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$$

- כאשר הדגימות במדגם הם i.i.d:

$$p_{\mathcal{D}}(\mathcal{D}; \boldsymbol{\theta}) = \prod_{i} p_{\mathbf{x}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

ולכן:

Maximum Likelyhood Estimator ((MLE

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\min_{\boldsymbol{\theta}} -\mathcal{L}(\boldsymbol{\theta};\mathcal{D}) = \arg\min_{\boldsymbol{\theta}} -\prod_i p_{\mathbf{x}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})$$

במקרים רבים נוכל להחליף את המכפלה על כל הדגימות בסכום, על ידי החלפת פונקציית ה likelihood ב log-likelihood:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\min_{\boldsymbol{\theta}} -\log\mathcal{L}(\boldsymbol{\theta};\mathcal{D}) = \arg\min_{\boldsymbol{\theta}} -\sum_i \log\left(p_{\mathbf{x}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})\right)$$

הערה: בקורסים "עיבוד אותות אקראיים" ו-"הסקה סטטיסטית" מרחיבים הרבה בנושא תכונות משערך זה ואחרים.

דוגמא - זמן הנסיעה

ננסה להתאים למדגם מודל של פילוג נורמלי:

$$p_{\mathrm{x}}(x;\boldsymbol{\theta})=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

וקטור הפרמטרים הינו $\boldsymbol{\theta}=[\mu,\sigma]^\top$.

דוגמא - זמן הנסיעה

$$p_{\mathrm{x}}(x;\boldsymbol{\theta})=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

נרשום את בעיית האופטימיזציה של מציאת משערך ה MLE:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}&=\underset{\boldsymbol{\theta}}{\arg\min}\ -\sum_i\log\left(p_{\mathrm{x}}(x^{(i)};\boldsymbol{\theta})\right)\\&=\underset{\boldsymbol{\theta}}{\arg\min}\ -\sum_i\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}\right)\right)\\&=\underset{\boldsymbol{\theta}}{\arg\min}\ \sum_i\log(\sigma)+\tfrac{1}{2}\log(2\pi)+\frac{(x^{(i)}-\mu)^2}{2\sigma^2}\\&=\underset{\boldsymbol{\theta}}{\arg\min}\ N\log(\sigma)+\frac{1}{2\sigma^2}\sum_i(x^{(i)}-\mu)^2\end{aligned}$$

נפתור על ידי גזירה והשוואה ל-0.

דוגמא - זמן הנסיעה

נסמן את ה objective ב f:

$$f(\boldsymbol{\theta};x) = N\log(\sigma) + \frac{1}{2\sigma^2}\sum_i (x^{(i)}-\mu)^2$$

$$\begin{cases} \frac{\partial f(\boldsymbol{\theta};x)}{\partial \mu} = 0 \\ \frac{\partial f(\boldsymbol{\theta};x)}{\partial \sigma} = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} -\frac{1}{\sigma^2}\sum_i (x^{(i)}-\mu) = 0 \\ \frac{N}{\sigma} - \frac{1}{\sigma^3}\sum_i (x^{(i)}-\mu)^2 = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} N\mu - \sum_i x^{(i)} = 0 \\ N\sigma^2 - \sum_i (x^{(i)}-\mu)^2 = 0 \end{cases}$$

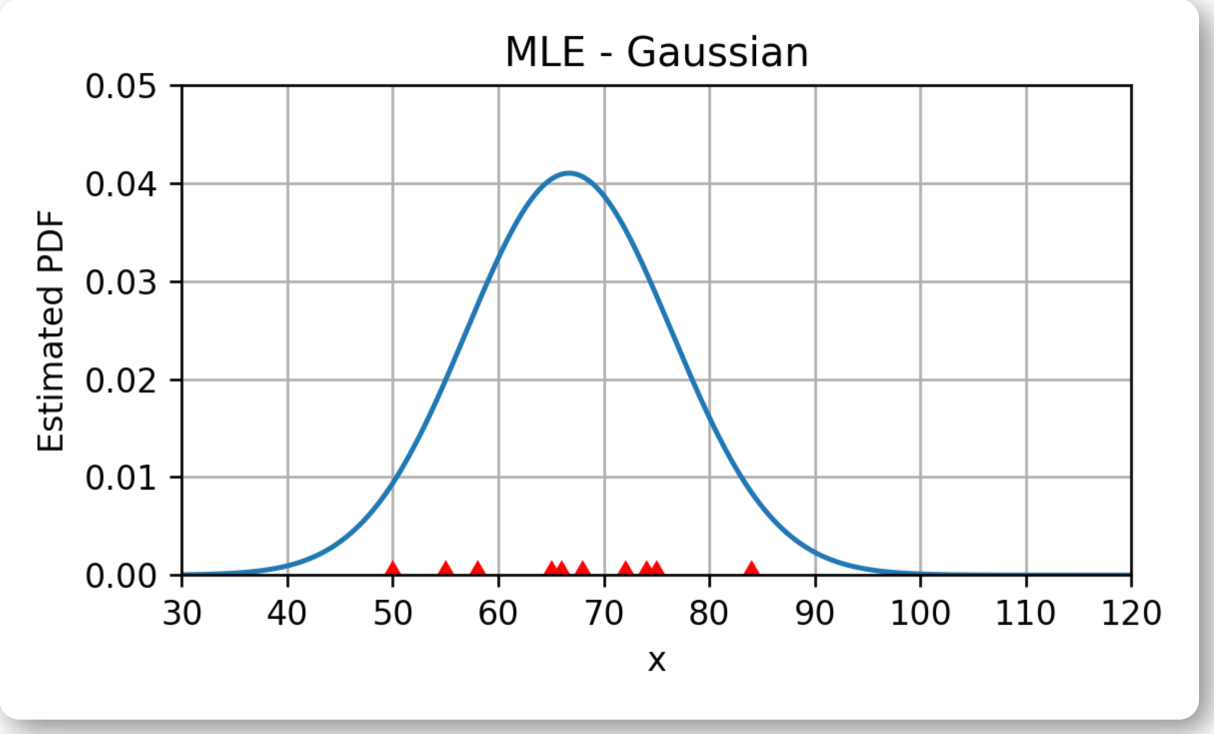
$$\Leftrightarrow \begin{cases} \mu = \frac{1}{N}\sum_i x^{(i)} \\ \sigma^2 = \frac{1}{N}\sum_i (x^{(i)}-\mu)^2 \end{cases}$$

דוגמא - זמן הנסיעה

במקרה של הנסיעות בכביש החוף נקבל:

$$\mu = 66.7 \ [\mathrm{min}]$$

$$\sigma = 9.7 \ [\mathrm{min}]$$



הגישה הבייסיאנית

- הפרמטרים של המודל הם ריאליזציות (הגרלות) של משתנה אקראי.
- גישה זו מניחה שיש בידינו מודל לפילוג המשותף של הפרמטרים והמדגם.

$$p_{\mathcal{D},\boldsymbol{\theta}}(\mathcal{D},\boldsymbol{\theta})=p_{\mathcal{D}|\boldsymbol{\theta}}(\mathcal{D}|\boldsymbol{\theta})p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$$

תחת ההנחה שבהינתן הפרמטרים הדגימות במדגם הם i.i.d:

$$p_{\mathcal{D},\boldsymbol{\theta}}(\mathcal{D},\boldsymbol{\theta})=p_{\boldsymbol{\theta}}(\boldsymbol{\theta})\prod_i p_{\mathbf{x}|\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$$

- עלינו לקבוע את $p_{\mathbf{x}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta})$ ואת $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$.
- זכרו בשבגישה הקודמת הנחנו כי θ קבוע אבל לא ידוע.

הגישה הבייסיאנית

A Prioiri Distribution

הפילוג השולי של הפרמטרים $p_\theta(\theta)$, מכונה לרוב **הפילוג הפריורי** (**prior distribution**) או **הא-פריורי** (**a priori distribution**), זאת אומרת הפילוג של θ לפני שראינו את המדגם.

A Posteriori Distribution

הפילוג של הפרמטרים בהינתן המדגם $p_{,\theta|\mathcal{D}}(\theta|\mathcal{D})$ מכונה **הפילוג הפוסטריורי** (**posterior distribution**) או **א-פוסטריורי** (**a posteriori distribution**) (או הפילוג בדיעבד), זאת אומרת, הפילוג אחרי שראינו את המדגם.

Maximum A-posteriori Probaility ((MAP

MAP משערך את הערך אשר ממקסם את הפילוג האפוסטריורי (הערך הכי סביר של θ בהינתן המדגם $p_{\theta|\mathcal{D}}(\theta|\mathcal{D})$):

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}|\mathcal{D}}(\boldsymbol{\theta}|\mathcal{D}) = \arg\min_{\boldsymbol{\theta}} \ -p_{\boldsymbol{\theta}|\mathcal{D}}(\boldsymbol{\theta}|\mathcal{D})$$

על פי חוק בייס, נוכל לכתוב זאת כ:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \ -\frac{p_{\mathcal{D}|\boldsymbol{\theta}}(\mathcal{D}|\boldsymbol{\theta})p_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{p_{\mathcal{D}}(\mathcal{D})} = \arg\min_{\boldsymbol{\theta}} \ -p_{\mathcal{D}|\boldsymbol{\theta}}(\mathcal{D}|\boldsymbol{\theta})p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$$

כאשר הדגימות במדגם **בהינתן** θ הן i.i.d, מתקיים:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \ -p_{\boldsymbol{\theta}}(\boldsymbol{\theta})\prod_i p_{\mathbf{x}|\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$$

Maximum A-posteriori Probaility ((MAP

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \ -p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_i p_{\mathbf{x}|\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$$

גם כאן נוכל להפוך את המכפלה לסכום על ידי מזעור מינוס הלוג של הפונקציה:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \ -\log\left(p_{\boldsymbol{\theta}}(\boldsymbol{\theta})\right) - \sum_i \log\left(p_{\mathbf{x}|\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})\right)$$

ההבדל בין MLE ל MAP

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\min_{\boldsymbol{\theta}} \; -\sum_i \log\left(p_{\mathbf{x}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})\right)$$

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \; -\log\left(p_{\boldsymbol{\theta}}(\boldsymbol{\theta})\right) - \sum_i \log\left(p_{\mathbf{x}|\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})\right)$$

- האיבר $-\log\left(p_{\boldsymbol{\theta}}(\boldsymbol{\theta})\right)$ מוסיף את הידע שיש לנו לגבי איזה ערכים של θ יותר סבירים.
- ראינו תוספת שכזו כאשר דיברנו על רגולריזציה.
- ניתן לחשוב על בעיית ה MAP כעל בעיית MLE עם רגולריזציה.
- בתרגיל הבית אתם תראו את השקילות שבין בעיית MAP לבין לבעיית MLE עם רגולריזציה.

בגישה בייסיאנית השערוך הוא בעיית חיזוי

- אנו מתייחסים גם למדגם וגם לפרמטרים בריאליזציות של משתנים אקראיים.
- אנו מניחים שאנו יודעים את הפילוג המשותף שלהם.
- ואנו מנסים למצוא את הערך של הפרמטרים בהינתן המדגם.
- זוהי בדיוק בעיית חיזוי קלאסית של משתנה אקראי אחד בהינתן משתנה אקראי אחר על סמך הפילוג המשותף.

דוגמא - הוספת prior

- נחזור לדוגמא של התאמת מודל של פילוג נורמלי לפילוג של זמן הנסיעה בכביש החוף.
- לשם הפשטות נקבע את סטיית התקן של המודל ל $\sigma = 10$.
- הפרמטר היחיד של המודל יהיה μ:

$$p_{\mathrm{x}|\mu}(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

דוגמא - הוספת prior

$$p_{\mathrm{x}|\mu}(x|\mu)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- נניח שיש לנו ידע קודם על פילוג הצפוי של μ.
- נניח שהפילוג הא-פריורי של μ הוא גם פילוג נורמלי עם תוחלת $\mu_\mu=60$ וסטיית תקן של $\sigma_\mu=5$:

$$p_\mu(\mu)=\frac{1}{\sqrt{2\pi}\sigma_\mu}\exp\left(-\frac{(\mu-\mu_\mu)^2}{2\cdot\sigma_\mu^2}\right)$$

נרשום את משערך ה MAP של μ:

$$\hat{\mu}_{\text{MAP}}=\arg\min_{\mu}\ -\log\left(p_\mu(\mu)\right)-\sum_i\log\left(p_{\mathrm{x}|\mu}(\boldsymbol{x}^{(i)}|\mu)\right)$$

דוגמא - הוספת prior

$$\hat{\mu}_{\text{MAP}} = \arg\min_{\mu} \; -\log\left(p_{\mu}(\mu)\right) - \sum_i \log\left(p_{\mathbf{x}|\mu}(\boldsymbol{x}^{(i)}|\mu)\right)$$

גזירה והשוואה ל-0 נותנת את התוצאה הבאה:

$$\frac{\partial f(\boldsymbol{\theta}; x)}{\partial \mu} = 0$$

$$\Leftrightarrow \frac{1}{\sigma_{\mu}^2}(\mu - \mu_{\mu}) - \frac{1}{\sigma^2}\sum_i (x^{(i)} - \mu) = 0$$

$$\Leftrightarrow \left(\frac{1}{\sigma_{\mu}^2} + \frac{N}{\sigma^2}\right)\mu = \frac{\mu_{\mu}}{\sigma_{\mu}^2} + \frac{1}{\sigma^2}\sum_i x^{(i)}$$

$$\Leftrightarrow \mu = \frac{\frac{\sigma^2}{N\sigma_{\mu}^2}\mu_{\mu} + \frac{1}{N}\sum_i x^{(i)}}{\frac{\sigma^2}{N\sigma_{\mu}^2} + 1}$$

זו למעשה ממוצע ממושקל בין הממוצע של x במדגם לבין μ_{μ}.

דוגמא - הוספת prior

עבור הדוגמא שלנו נקבל:

$$\mu = 64.8 \ [\text{min}]$$

ערך זה מעט יותר קרוב ל60 מאשר התוצאה שקיבלנו בשיערוך ה MLE. זאת משום ה prior ש"מושך" את הפרמטרים לאיזורים הסבירים יותר ולכן הוא מקרב אותו ל $\mu_\mu = 60$.

שימוש בשיערוך פרמטרי לפתרון בעיות supervised learning

נראה עכשיו איך להשתמש בשערוך הצפיפות שתארנו צורך פתרון בעיות סיווג בלמידה מפוקחת. נציג שיטה אשר משתמשת במודל של פילוג נורמלי וב MLE לפתרון בעיות סיווג.

Quadratic Discriminant Analysis (QDA)

- נשתמש בפילוג נורמלי וב MLE על מנת לשערך את $p_{\mathrm{x}|\mathrm{y}}(\boldsymbol{x}|y)$
- אנו צריכים לשערך מודל עבור כל אחת מ C המחלקות של y:
  - וקטור תוחלת μ_c עבור כל אחד מהמחלקות.
  - מטריצת קווריאנס Σ_c עבור כל אחד מהמחלקות.

$$p_{\mathrm{x}|\mathrm{y}}(\boldsymbol{x}|c;\boldsymbol{\mu}_c,\Sigma_c)=\frac{1}{\sqrt{(2\pi)^D|\Sigma_c|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^\top\Sigma_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)\right)$$

הפילוג המשותף של x ו y יהיה:

Quadratic Discriminant Analysis (QDA)

בעיית האופטימיזציה של MLE תהיה:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\text{MLE}} &= \arg\min_{\boldsymbol{\theta}} \; -\log \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) \\
&= \arg\min_{\boldsymbol{\theta}} \; -\sum_i \log\left(p_{\mathrm{x}|\mathrm{y}}(\boldsymbol{x}^{(i)}|y^{(i)}; \boldsymbol{\mu}_y, \Sigma_y) p_{\mathrm{y}}(y^{(i)})\right) \\
&= \arg\min_{\boldsymbol{\theta}} \; -\sum_i \log\left(p_{\mathrm{x}|\mathrm{y}}(\boldsymbol{x}^{(i)}|y^{(i)}; \boldsymbol{\mu}_y, \Sigma_y)\right) + \log\left(p_{\mathrm{y}}(y^{(i)})\right) \\
&= \arg\min_{\boldsymbol{\theta}} \; -\sum_i \log\left(p_{\mathrm{x}|\mathrm{y}}(\boldsymbol{x}^{(i)}|y^{(i)}; \boldsymbol{\mu}_y, \Sigma_y)\right)
\end{aligned}$$

בהינתן ש-p_{y} ידוע.

Quadratic Discriminant Analysis ((QDA

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \underset{\boldsymbol{\theta}}{\arg\min} \ -\sum_i \log\left(p_{\mathbf{x}|\mathrm{y}}(\boldsymbol{x}^{(i)}|y^{(i)};\boldsymbol{\mu}_y,\Sigma_y)\right)$$

- נחלק את הסכימה לסכימות נפרדות על כל אחת מהמחלקות.

Quadratic Discriminant Analysis (QDA)

- נגדיר לשם כך את הסימונים הבאים:
  - $\mathcal{I}_c = \{i : \ y^{(i)} = c\}$ - זאת אומרת, אוסף האינדקסים של הדגמים במדגם שמקיימים $y^{(i)} = c$.
  - $|\mathcal{I}_c|$ - מספר האינדקסים ב $\mathcal{I}_c$

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \underset{\boldsymbol{\theta}}{\arg\min} \ -\sum_{i\in\mathcal{I}_1} \log\left(p_{\mathbf{x}|\mathrm{y}}(\boldsymbol{x}^{(i)}|1;\boldsymbol{\mu}_1,\Sigma_1)\right) - \sum_{i\in\mathcal{I}_2} \log\left(p_{\mathbf{x}|\mathrm{y}}(\boldsymbol{x}^{(i)}|2;\boldsymbol{\mu}_2,\Sigma_2)\right) - \dots$$

Quadratic Discriminant Analysis ((QDA

עבור המחלקה ה c נקבל את בעיית האופטימיזציה הבאה:

$$\hat{\boldsymbol{\mu}}_{c,\text{MLE}}, \hat{\Sigma}_{c,\text{MLE}} = \underset{\boldsymbol{\mu}_c, \Sigma_c}{\arg\min} \; -\sum_{i \in \mathcal{I}_c} \log\left(p_{\mathbf{x}|\mathrm{y}}(\boldsymbol{x}^{(i)} | c; \boldsymbol{\mu}_c, \Sigma_c)\right)$$

$$= \underset{\boldsymbol{\mu}_c, \Sigma_c}{\arg\min} \sum_{i \in \mathcal{I}_c} \log\left(\sqrt{|\Sigma_c|}\right) + \frac{1}{2}(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_c)^\top \Sigma_c^{-1}(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_c$$

- ניתן לפתור את הבעיה הזו על ידי גזירה והשוואה ל-0.
- הפיתוח עבור Σ_c הוא מעט מורכב ואנו לא נראה אותו בקורס זה ונקפוץ ישר לפתרון. הפיתוח מודגם בקורס "עיבוד וניתוח מידע".

Quadratic Discriminant Analysis ((QDA

החישוב של μ_c

$$f(\boldsymbol{\theta};x) = \sum_{i\in\mathcal{I}_c} \log\left(\sqrt{|\Sigma_c|}\right) + \frac{1}{2}(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_c)^\top \Sigma_c^{-1}(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_c)$$

$$\frac{\partial f}{\partial \boldsymbol{\mu}_c} = 0$$

$$\Leftrightarrow -\sum_{i\in\mathcal{I}_c} \Sigma_c^{-1}(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_c) = 0$$

$$\Leftrightarrow |\mathcal{I}_c|\boldsymbol{\mu}_c - \sum_{i\in\mathcal{I}_c} \boldsymbol{x}^{(i)} = 0$$

$$\Leftrightarrow \boldsymbol{\mu}_c = \frac{1}{|\mathcal{I}_c|}\sum_{i\in\mathcal{I}_c} \boldsymbol{x}^{(i)}$$

Quadratic Discriminant Analysis (QDA)

הפרמטרים של המודל יהיו:

$$p_{\mathrm{y}}(c) = \frac{|\mathcal{I}_c|}{N}$$

$$\boldsymbol{\mu}_c = \frac{1}{|\mathcal{I}_c|}\sum_{i\in\mathcal{I}_c}\boldsymbol{x}^{(i)}$$

$$\Sigma_c = \frac{1}{|\mathcal{I}_c|}\sum_{i\in\mathcal{I}_c}\left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_c\right)\left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_c\right)^T$$

הצגנו כאן משערך אינטואיטבי להסתברות השיוך לכל מחלקה. אפשר להראות שזו התוצאה המתקבלת גם מהנחת מודל פילוג מולטינומיאלי למשתנה הקטגורי של המחלקה.

Quadratic Discriminant Analysis ((QDA

עבור פונקציית מחיר של סיכוי הטעות, החזאי האופטימאלי יהיה:

$$\begin{aligned}\hat{y} = h(\boldsymbol{x}) &= \arg\max_y p_{\mathrm{x}|\mathrm{y}}(\boldsymbol{x}|y;\boldsymbol{\mu}_y,\Sigma_y)p_{\mathrm{y}}(y)\\ &= \arg\max_y \ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_y)^\top\Sigma_y^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_y)+\log\left(\frac{p_{\mathrm{y}}(y))}{\sqrt{|\Sigma_y|}}\right)\end{aligned}$$

המקרה הבינארי - משטח הפרדה ריבועי

עבור המקרה של סיווג בינארי (סיווג לשתי מחלקות) מתקבל החזאי הבא:

$$h\left(x\right)=\begin{cases}1 & \boldsymbol{x}^T C\boldsymbol{x}+\boldsymbol{a}^T\boldsymbol{x}+b>0\\ 0 & \text{otherwise}\end{cases}$$

כאשר:

$$C=\frac{1}{2}(\Sigma_0^{-1}-\Sigma_1^{-1})$$

$$\boldsymbol{a}=\Sigma_1^{-1}\boldsymbol{\mu}_1-\Sigma_0^{-1}\boldsymbol{\mu}_0$$

$$b=\tfrac{1}{2}\left(\boldsymbol{\mu}_0^T\Sigma_0^{-1}\boldsymbol{\mu}_0-\boldsymbol{\mu}_1^T\Sigma_1^{-1}\boldsymbol{\mu}_1\right)+\log\left(\frac{\sqrt{|\Sigma_0|}p_{\text{y}}(1)}{\sqrt{|\Sigma_1|}p_{\text{y}}(0)}\right)$$

התנאי שקיבלנו $\boldsymbol{x}^T C\boldsymbol{x}+\boldsymbol{a}^T\boldsymbol{x}+b>0$ הוא ריבועי ב $\boldsymbol{x}$ ומכאן מקבל האלגוריתם את שמו.

(Linear Discriminant Analysis (LDA

- מניח שלפונקציות הפילוג של המחלקות השונות יש את אותה מטריצת הקווריאנס.
- שהפרמטרים של המודל יהיו כעת:
  - וקטור תוחלת μ_c עבור כל אחד מהמחלקות.
  - מטריצת covariance אחת Σ משותפת לכל המחלקות.

$$p_{\mathrm{x}|\mathrm{y}}(\boldsymbol{x}|c;\boldsymbol{\mu}_c,\Sigma)=\frac{1}{\sqrt{(2\pi)^D|\Sigma|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)\right)$$

(Linear Discriminant Analysis (LDA

פתרון בעיית ה MLE:

$$p_{\mathrm{y}}(c) = \frac{|\mathcal{I}_c|}{N}$$

$$\boldsymbol{\mu}_c = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \boldsymbol{x}^{(i)}$$

$$\Sigma = \frac{1}{N} \sum_i \left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}} \right) \left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}} \right)^T$$

(Linear Discriminant Analysis (LDA

עבור פונקציית מחיר של miscalssification rate, החזאי האופטימאלי המתקבל ממודל זה הינו:

$$\begin{aligned}\hat{y}=h(\boldsymbol{x}) &= \underset{y}{\arg\max}\, p_{\mathrm{x}|\mathrm{y}}(\boldsymbol{x}|y;\boldsymbol{\mu}_c,\Sigma)p_{\mathrm{y}}(y) \\ &= \underset{y}{\arg\max}\, -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_y)^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_y)+\log(p_{\mathrm{y}}(y)) \\ &= \underset{y}{\arg\min}\, \boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{\mu}_y-\frac{1}{2}\boldsymbol{\mu}_y^\top\Sigma^{-1}\boldsymbol{\mu}_y-\log(p_{\mathrm{y}}(y))\end{aligned}$$

המקרה הבינארי

עבור המקרה של סיווג בינארי (סיווג לשני מחלקות) מתקבל החזאי הבא:

$$h\left(x\right)=\begin{cases}1 & \boldsymbol{a}^{T}\boldsymbol{x}+b>0\\0 & \text{otherwise}\end{cases}$$

כאשר:

$$\boldsymbol{a}=\Sigma^{-1}\left(\boldsymbol{\mu}_{1}-\boldsymbol{\mu}_{0}\right)$$

$$b=\tfrac{1}{2}\left(\boldsymbol{\mu}_{0}^{T}\Sigma^{-1}\boldsymbol{\mu}_{0}-\boldsymbol{\mu}_{1}^{T}\Sigma^{-1}\boldsymbol{\mu}_{1}\right)+\log\left(\frac{p_{\text{y}}\left(1\right)}{p_{\text{y}}\left(0\right)}\right)$$

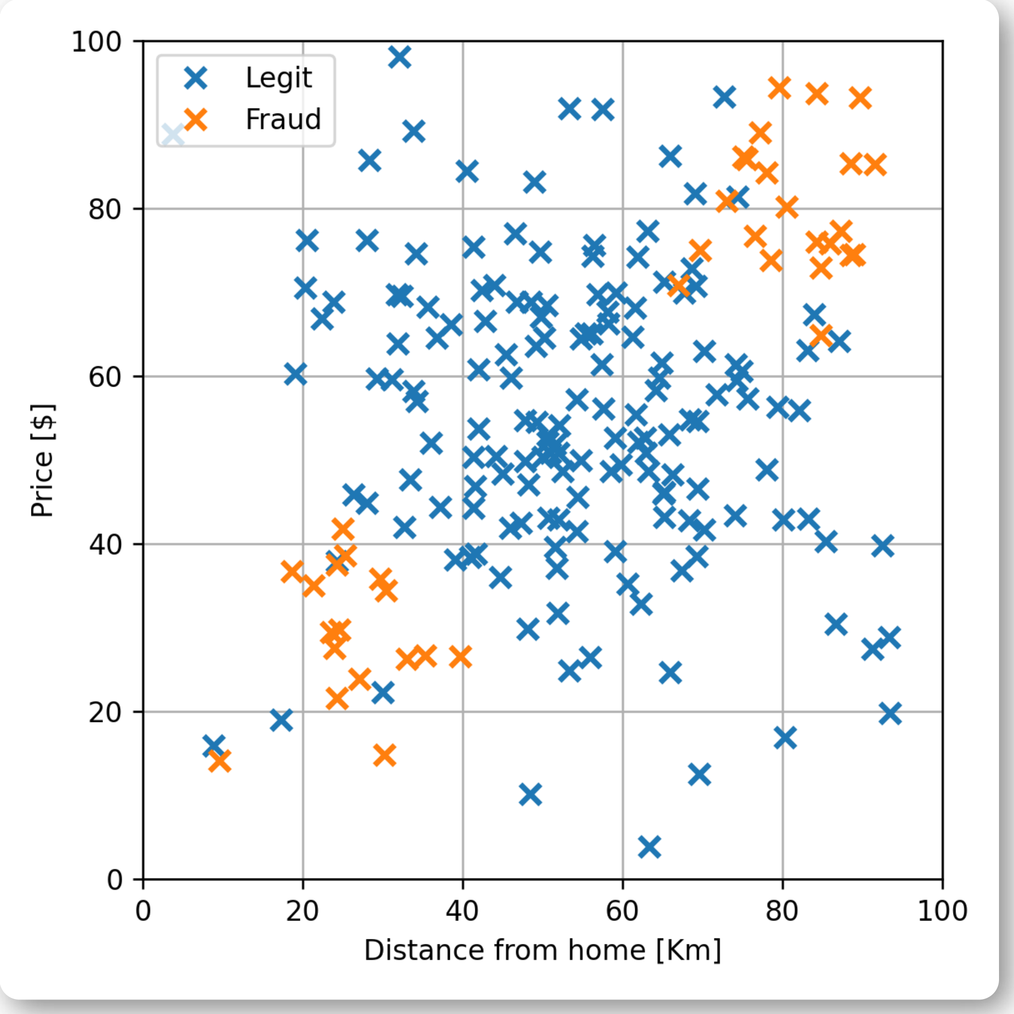
התנאי שקיבלנו $a^{T}\boldsymbol{x}+b>0$ הוא לינארי ב $\boldsymbol{x}$ ומכאן מקבל האלגוריתם את שמו.

המקרה הכללי (לא בינארי)

במקרה הכללי המרחב יהיה מחולק ל C איזורים שהשפות שלהם יהיו מורכבות מהמישורים המתקבלים מהשפות שבין כל זוג מחלקות. דוגמא למקרה עם 3 מחלקות תופיע בתרגול.

דוגמא

נסתכל שוב על הבעיה של חיזוי עסקאות שחשודות כהונאות:



התאמה של מודל QDA

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

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

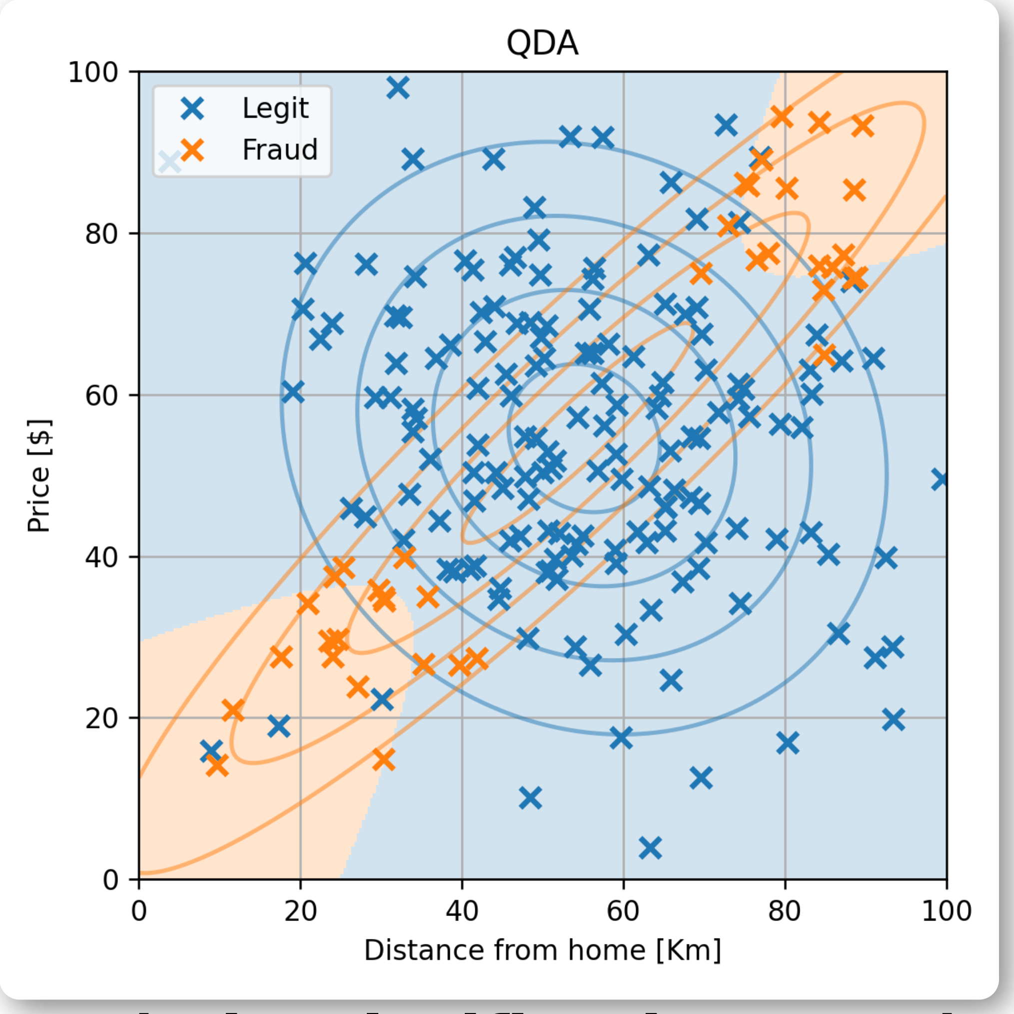
$$\boldsymbol{\mu}_0 = \frac{1}{|\mathcal{I}_0|} \sum_{i \in \mathcal{I}_0} \boldsymbol{x}^{(i)} = [55.1, 54.6]^\top$$

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

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

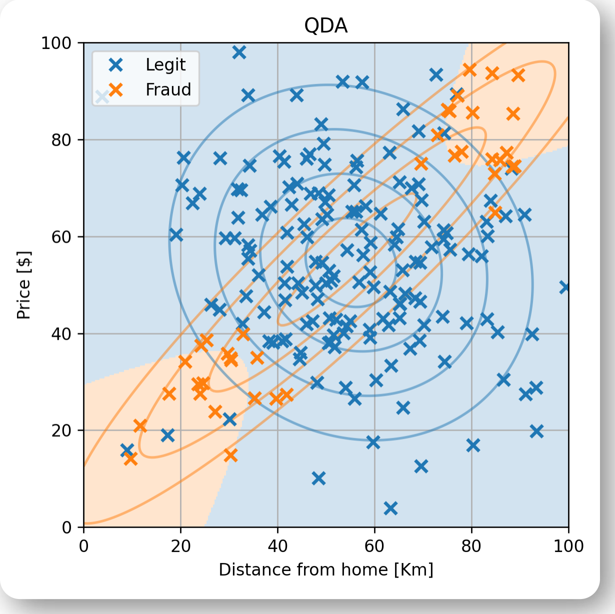
$$\Sigma_1 = \frac{1}{|\mathcal{I}_1|} \sum_i \left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}}\right) \left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}}\right)^T = \begin{bmatrix} 817.9 & 730.5 \\ 730.5 & 741.7 \end{bmatrix}$$

התאמה של מודל QDA



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

הבעיה של הגישה הגנרטיבית הפרמטרית

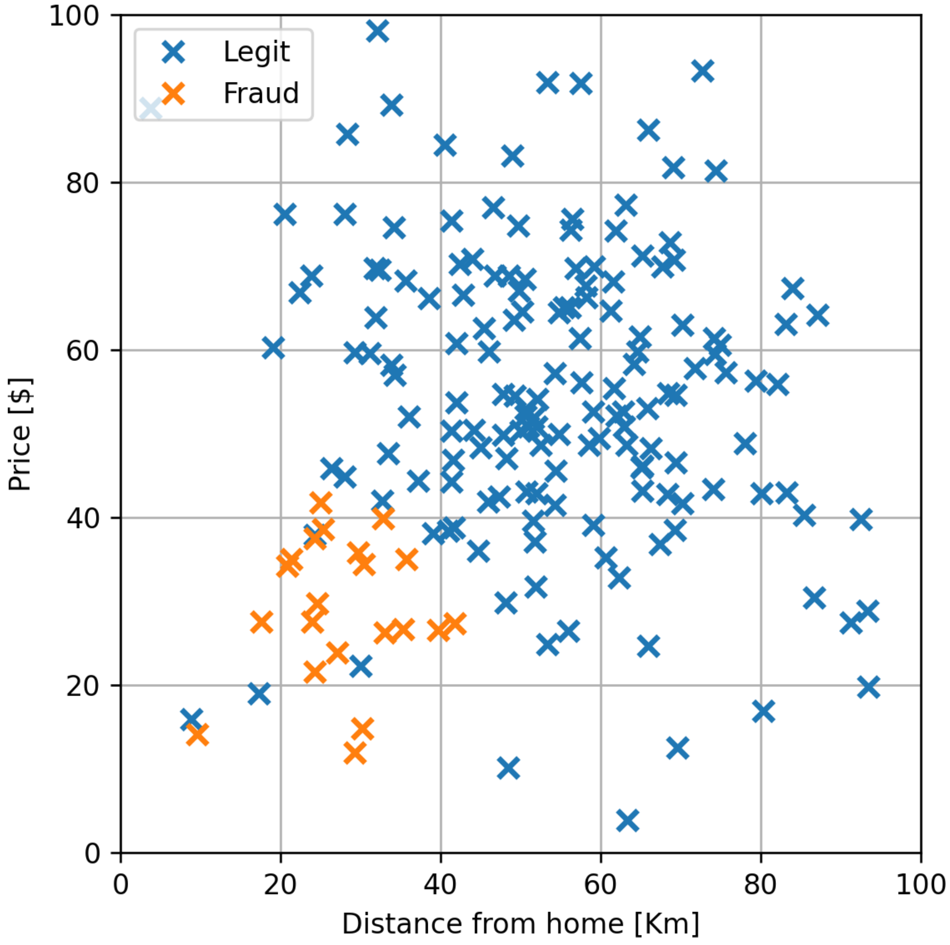


- הינו רוצים מודל אשר יכול לייצג בנפרד את שני האיזורים.
- לצערינו המבחר של המודלים בהם אנו יכולים לא גדול.
- המגבלה הזו נובעת מהצורך שהמודל ייצג פילוג חוקיים.

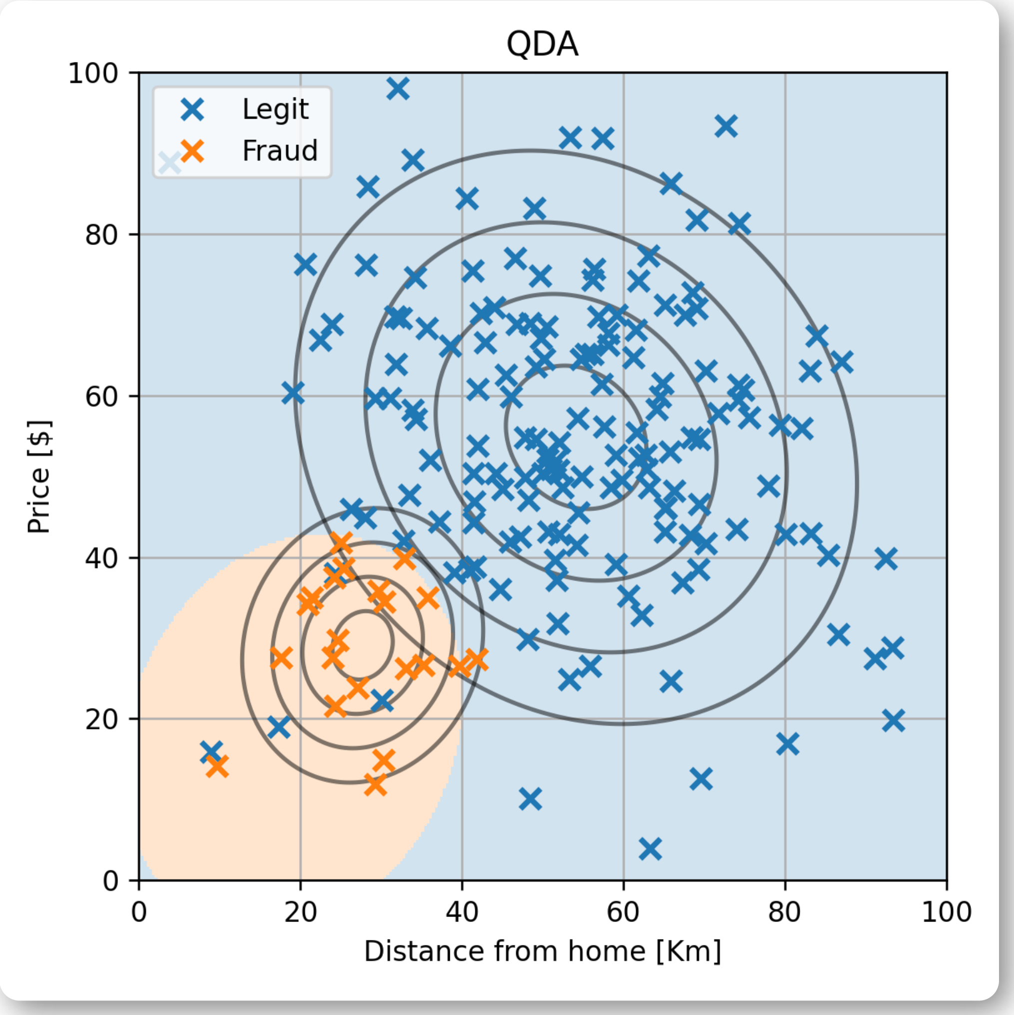
הערה: קיימות הרבה הרחבות שנידונות במקורסים מתקדמים יותר, למשל, תערובת של גאוסיאניים.

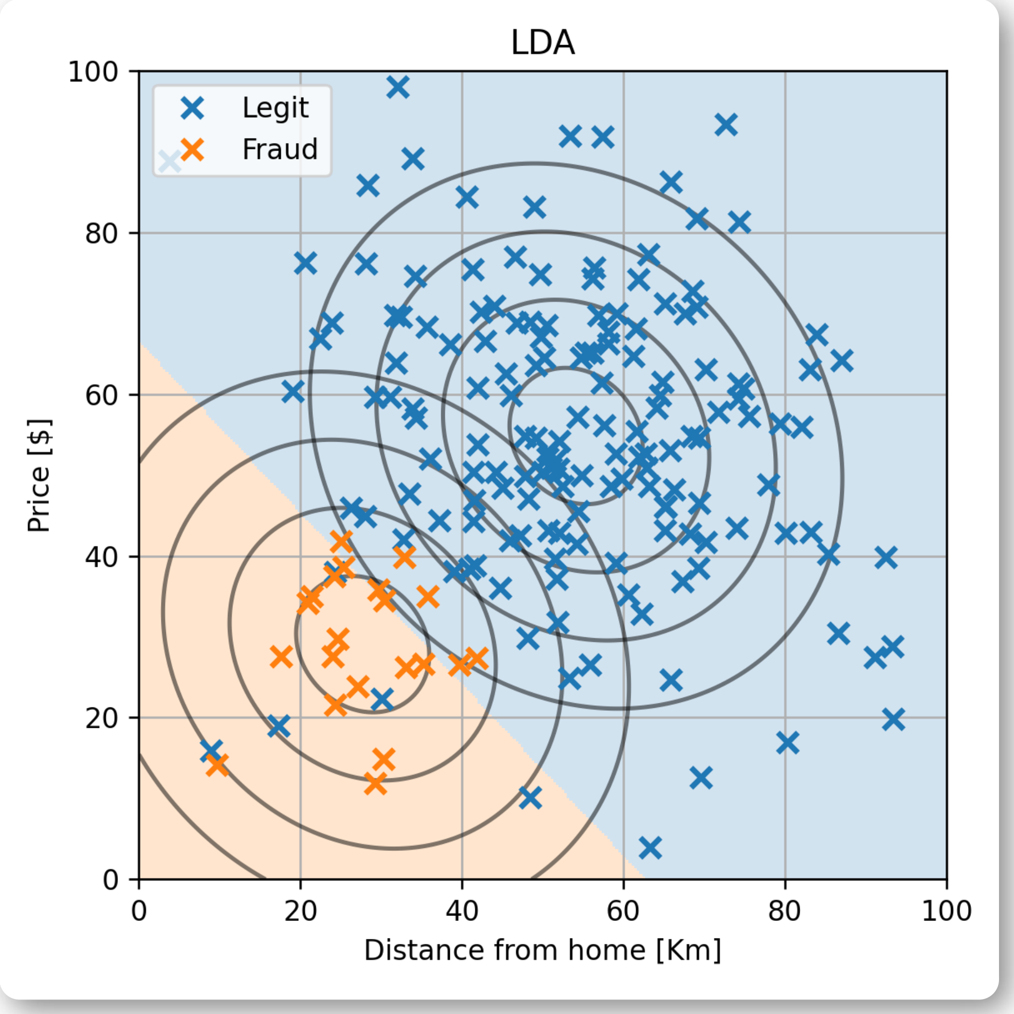
דוגמא למדגם שמתאים למודל של QDA

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



מודל QDA



מודל LDA

נציג גם את התוצאה המתקבלת ממודל ה LDA: