

Üzleti Elemzések Módszertana

8. Előadás: Generatív modellezés

Kuknyó Dániel
Budapesti Gazdasági Egyetem

2023/24
2.félév

1 Bevezetés

2 Naïv Bayes

3 Gauss-i keverékek

1 Bevezetés

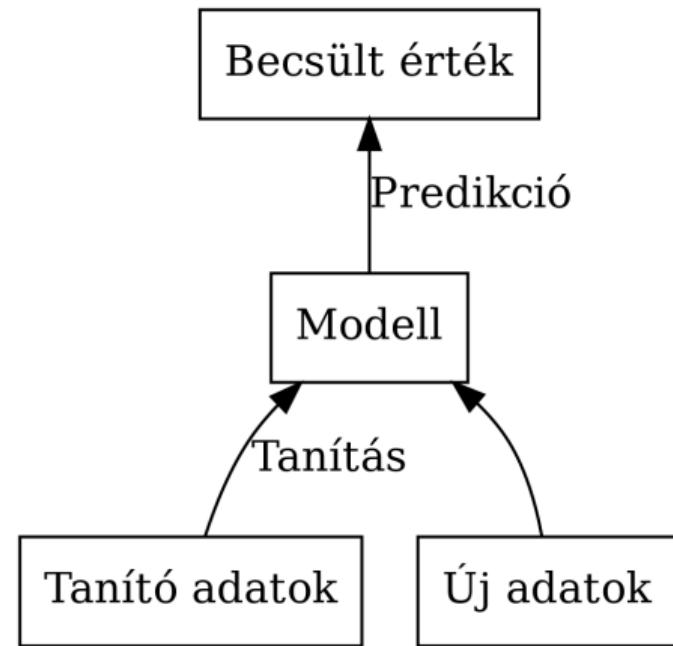
2 Naív Bayes

3 Gauss-i keverékek

Diszkriminatív modellezés

Diszkriminatív modellezés esetén a modellezés eljárasa:

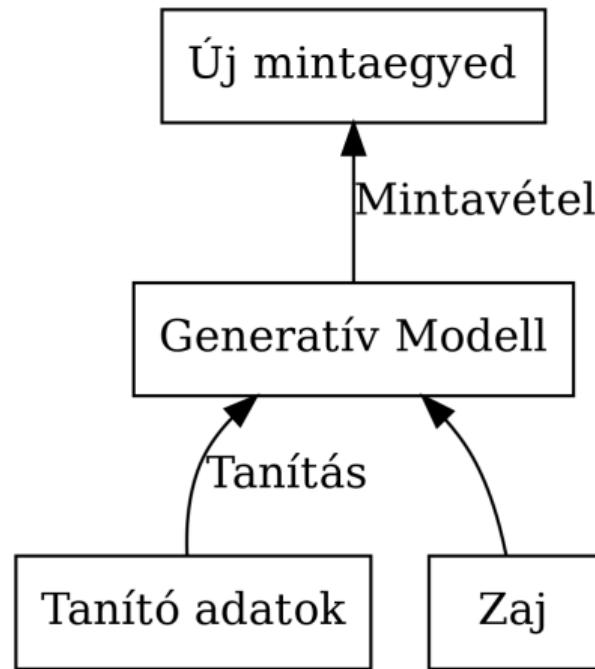
- ① Tanító adatok gyűjtése
- ② Modell tanítása a tanító adatokon
- ③ Nem látott mintaegyedeken predikció elvégzése



Generatív modellezés

Generatív modellezés esetén a folyamat a következőképpen módosul:

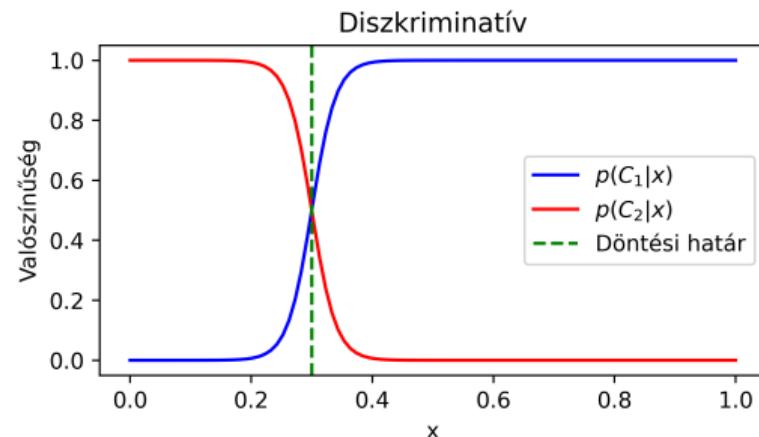
- ① Tanító adatok gyűjtése
- ② Modell tanítása a tanító adatokon
- ③ Zaj beengedése a rendszerbe
- ④ Új mintaegyed létrehozása



Diszkriminatív modellezés

Diszkriminatív modell

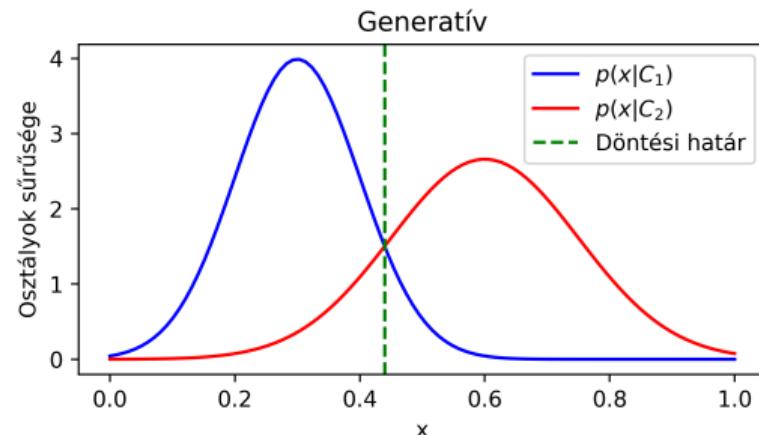
A tanítás célja, hogy megkülönböztesse a különböző adatkategóriákat. Ezt használja fel arra, hogy megbecsülje az adatpontok osztályba tartozását.



Generatív modellezés

Generatív modell

A célja, hogy megtanulja az adatok eloszlását. Megtanulja, hogyan generálódnak az adatok, és ezt felhasználva képesek új adatokat generálni.



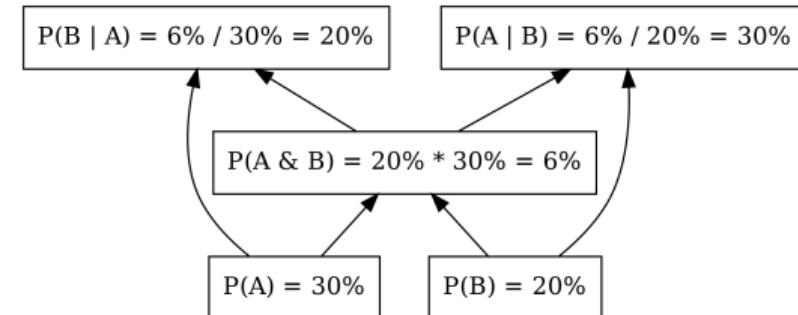
Feltételes valószínűségek

Valamely A esemény feltételes valószínűsége azt jelenti, mekkora az esély A esemény bekövetkezésére feltéve, hogy B esemény már megtörtént. Ennek jelölése:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Ennek megfelelően például az a valószínűség, hogy egy rendelés csalóktól érkezik feltéve, hogy kupont használtak:

$$P(\text{Csaló}|\text{Kupon}) = \frac{P(\text{Csaló} \cap \text{Kupon})}{P(\text{Kupon})}$$



Inverz feltételes valószínűségek

Az inverz feltételes valószínűség kiszámítható a **Bayes-tételnek megfelelően** a feltételes valószínűség és a nem feltételes valószínűségek segítségével:

$$P(B|A) = \frac{P(A|B) P(B)}{P(A)}$$

A probabilisztikus modellek ezt veszik alapul. A probabilisztikus osztályozás célja valamely \mathcal{L} címkehalmaz valószínűségét megbecsülni adott x változóhalmaz alapján:

$$P(\mathcal{L}|x) = \frac{P(x|\mathcal{L}) P(\mathcal{L})}{P(x)}$$

Ebben az esetben az a valószínűség, hogy egy vásárlás csalóktól érkezik feltéve, hogy kupont használtak:

$$P(\text{Csaló}|\text{Kupon}) = \frac{P(\text{Kupon}|\text{Csaló}) P(\text{Csaló})}{P(\text{Kupon})}$$

1 Bevezetés

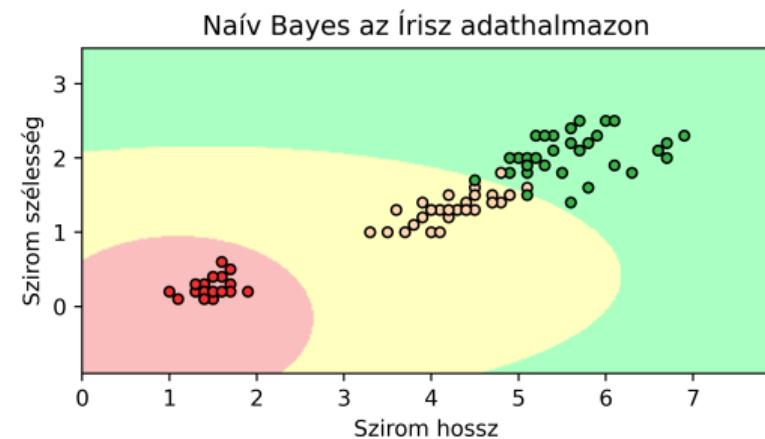
2 Naív Bayes

3 Gauss-i keverékek

Naív Bayes

Probabilisztikus osztályozók egy családja,
amely a Bayes-tételt alkalmazza jellemzőkre
úgy, hogy közben erősen függetlennek
tekinti a jellemzőket.

A naív Bayes modell naivitása abból a
feltételezésből ered, hogy egy jellemző
jelenléte független bármely másik jellemző
jelenlététtől, ha adott egy osztály változó.
Ez jelentősen csökkenti az algoritmus
számítási igényét.



A Gauss-eloszlás

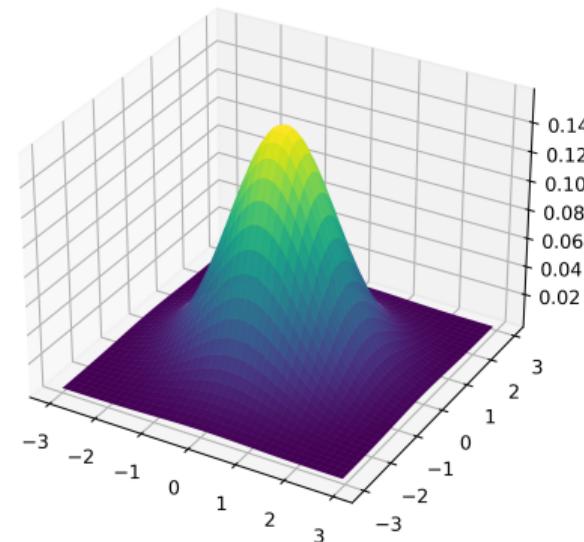
Egy X véletlen változó Gauss-i eloszlást követ, ha a valószínűség eloszlása a következő függvényt követi:

Gauss-eloszlás

$$f(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu))}$$

Ahol:

- μ : Az eloszlás várható értéke, vagyis a középre húzás tendenciája
- Σ : A kovarianciamátrix, amely megadja az adatok szóródásának mértékét minden dimenzióban

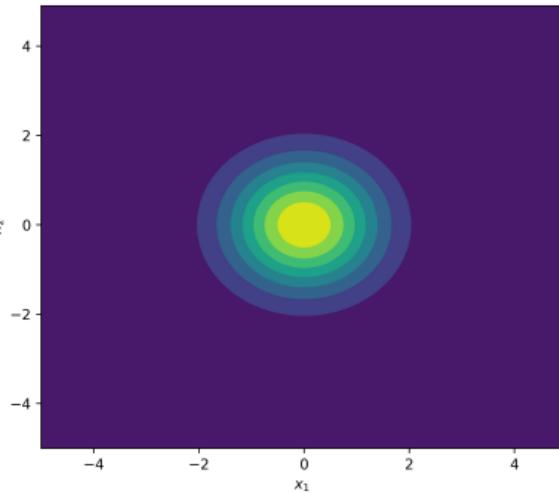
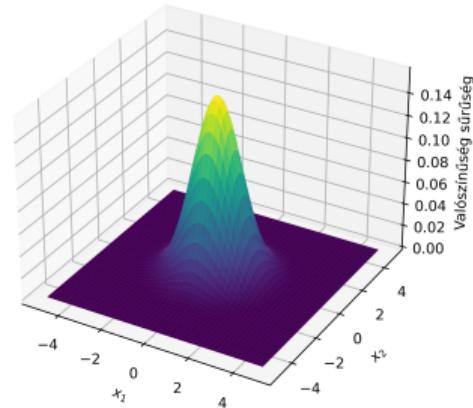


Példák Gauss-eloszlásra

Az ábrán a következő paraméterekkel definiált többváltozós Gauss-eloszlások láthatók:

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

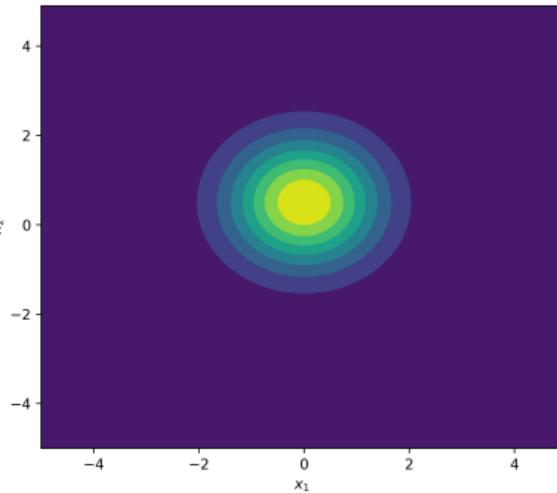
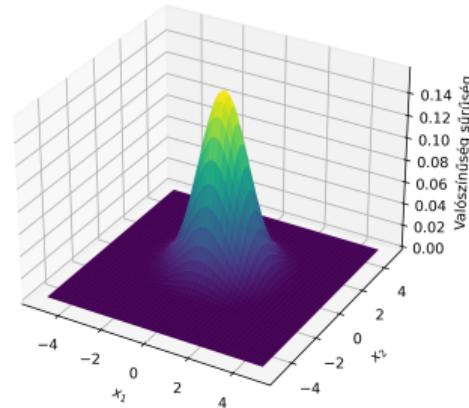


Példák Gauss-eloszlásra

Az ábrán a következő paraméterekkel definiált többváltozós Gauss-eloszlások láthatók:

$$\mu = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

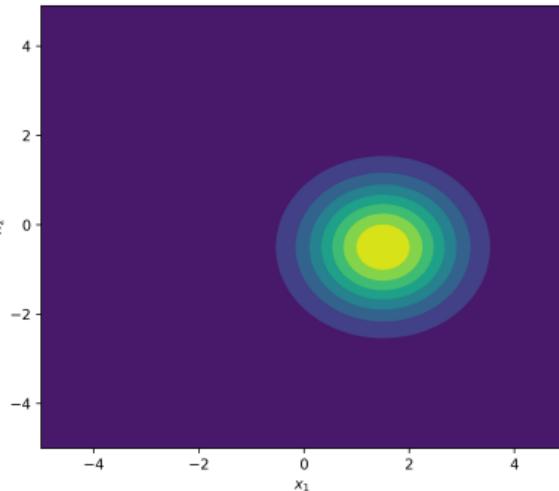
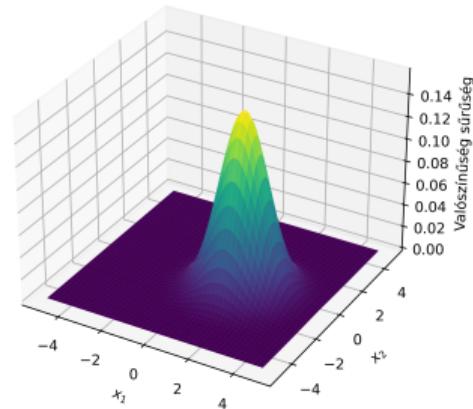


Példák Gauss-eloszlásra

Az ábrán a következő paraméterekkel definiált többváltozós Gauss-eloszlások láthatók:

$$\mu = \begin{pmatrix} 1.5 \\ -0.5 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

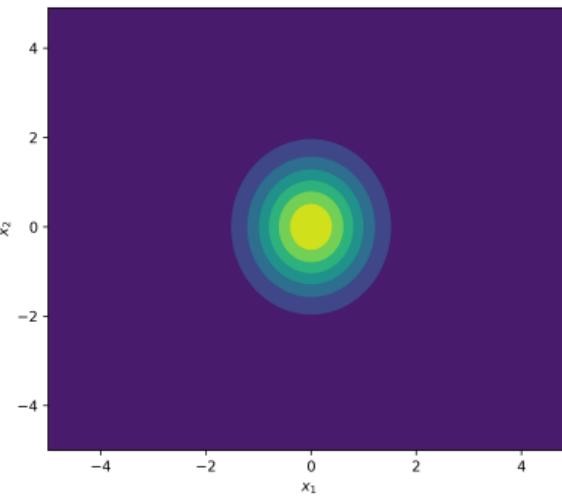
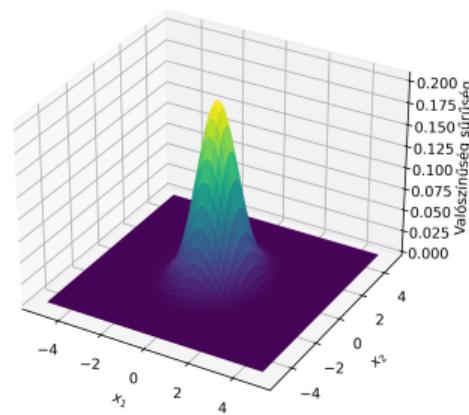


Példák Gauss-eloszlásra

Az ábrán a következő paraméterekkel definiált többváltozós Gauss-eloszlások láthatók:

$$\mu = \begin{pmatrix} 1.5 \\ -0.5 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 0.6 & 0 \\ 0 & 1 \end{pmatrix}$$

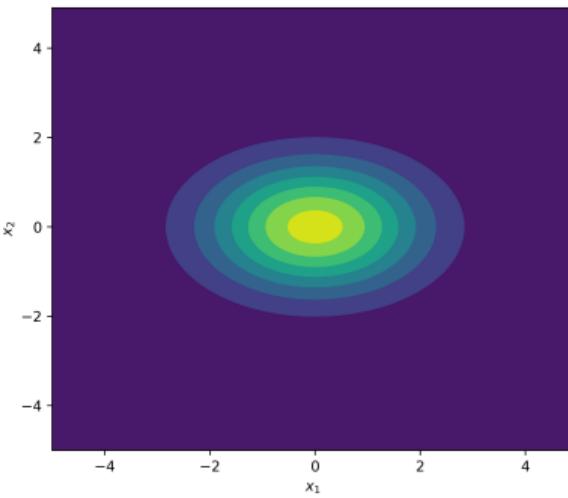
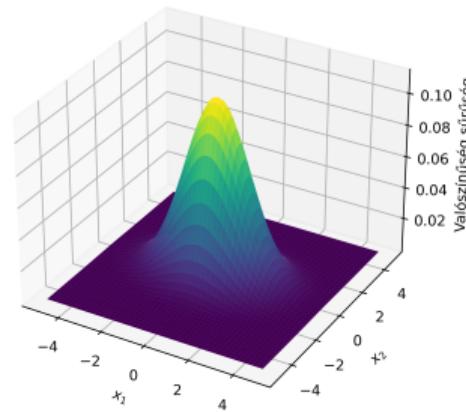


Példák Gauss-eloszlásra

Az ábrán a következő paraméterekkel definiált többváltozós Gauss-eloszlások láthatók:

$$\mu = \begin{pmatrix} 1.5 \\ -0.5 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

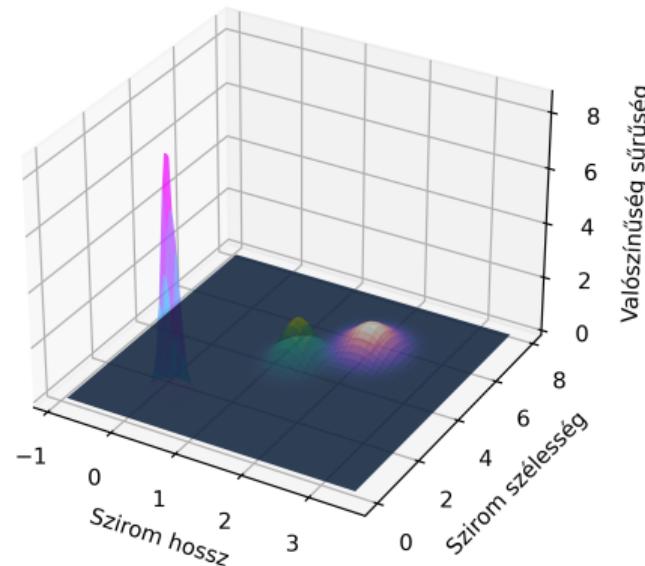


Becsült eloszlások az Írisz adathalmaz esetén

Az ábrán az előző képen látható modell által becsült valószínűség eloszlások láthatók.

Ebben az esetben az illesztett modell egy **Gauss-i naív Bayes osztályozó**, tehát az előfeltételezése, hogy az adatok egy Gauss-i eloszlásból származhattak.

Az algoritmus sikeresen megtalálta az összes osztályhoz tartozó eloszlást, pedig ez nem hiperparamétere az algoritmusnak.



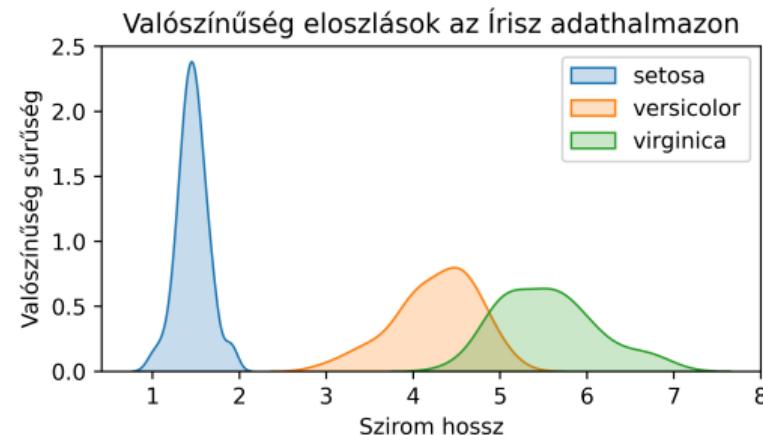
Gauss-i naív Bayes

A paraméterek becslése a tanító adathalmazra:

- A várható érték $\mu_{i,j}$ az i jellemző átlaga minden olyan egyedre, amely j osztályba tartozik
- A kovariancia mátrix Σ a várható értéktől való négyzetes eltéréseket tartalmazza minden i, j esetén

A predikció minden C_j osztály utólagos valószínűsége a Bayes tételek megfelelően:

$$P(C_j|x_1 \dots n) = \frac{P(C_j) \prod_{i=1}^n P(x_i|C_j)}{P(x_1, \dots, x_n)}$$



1 Bevezetés

2 Naív Bayes

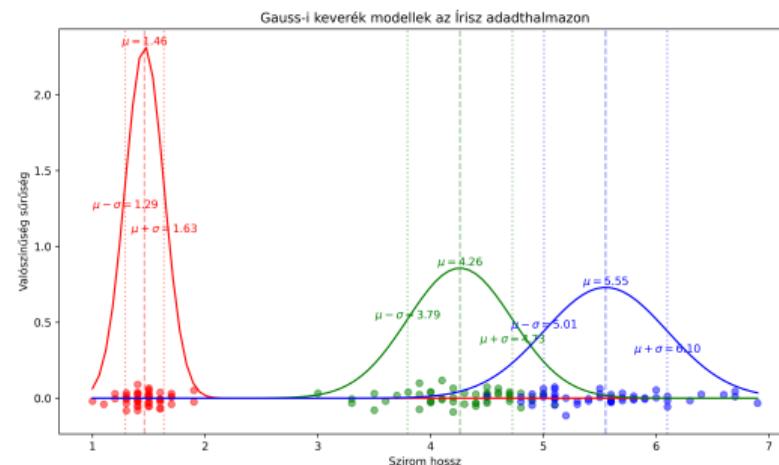
3 Gauss-i keverékek

Gauss-i keverékek

Probabilisztikus modellezési eljárás melynek alap feltételezése, hogy **az adatok ismeretlen paraméterű Gauss-i eloszlások által lettek generálva.**

Az eljárás megadja, hogy milyen komplex eloszlásból származhatna az adatkészlet, ha véletlen minta lenne.

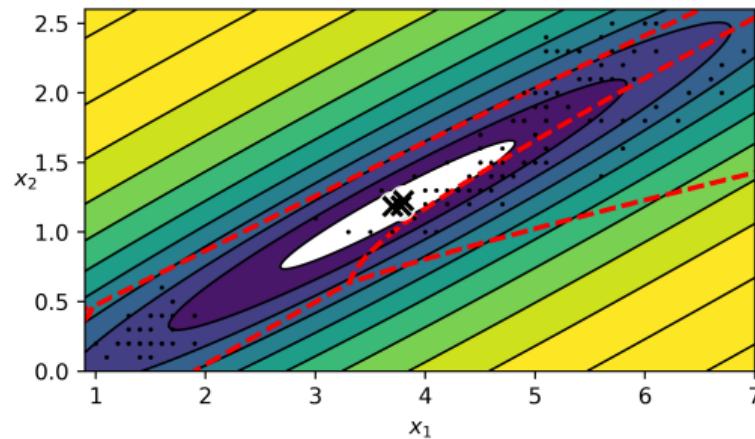
A modell célja megkeresni a Gauss-eloszlások μ és Σ paramétereit.



A modellezési eljárás

A Gauss-i keverékek modellje az EM (elvárás-maximalizálás) algoritmus alapján keresi meg az eloszlások paramétereit:

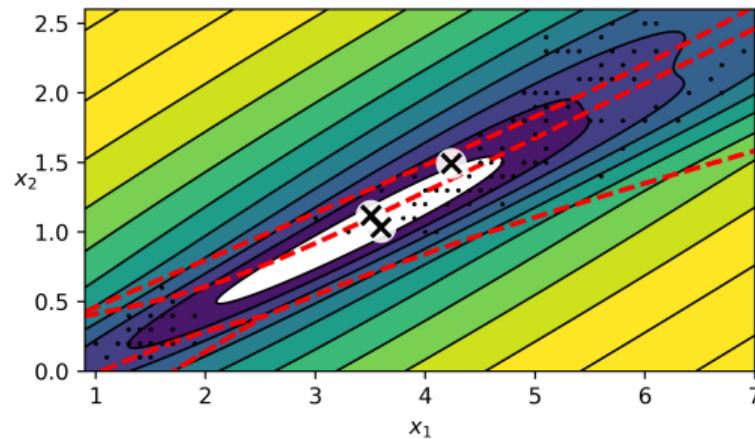
- ① **Inicializáció:** Gauss komponensek számának kiválasztása és véletlenszerű paraméter inicializáció
- ② **Elvárás:** Felelősségek kiszámítása, amely az a valószínűség minden adatpontra, hogy melyik Gauss komponensből származnak
- ③ **Maximalizálás:** Gauss paraméterek frissítése a valószínűségek maximalizálására, új várható értékek és kovarianciamátrixok kiszámítása



A modellezési eljárás

A Gauss-i keverékek modellje az EM (elvárás-maximalizálás) algoritmus alapján keresi meg az eloszlások paramétereit:

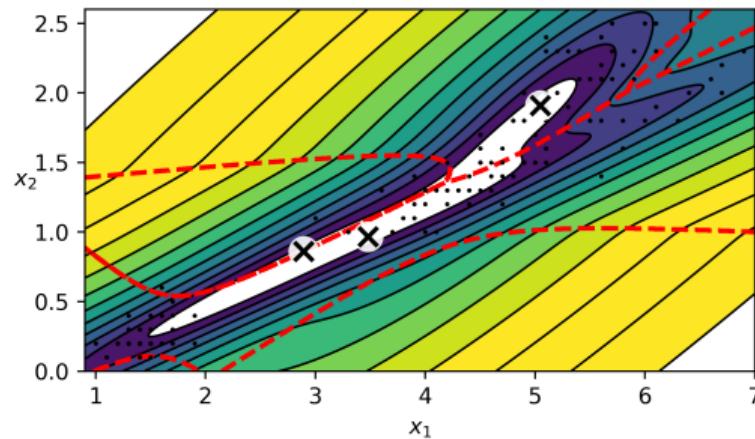
- ① **Inicializáció:** Gauss komponensek számának kiválasztása és véletlenszerű paraméter inicializáció
- ② **Elvárás:** Felelősségek kiszámítása, amely az a valószínűség minden adatpontra, hogy melyik Gauss komponensből származnak
- ③ **Maximalizálás:** Gauss paraméterek frissítése a valószínűségek maximalizálására, új várható értékek és kovarianciamátrixok kiszámítása



A modellezési eljárás

A Gauss-i keverékek modellje az EM (elvárás-maximalizálás) algoritmus alapján keresi meg az eloszlások paramétereit:

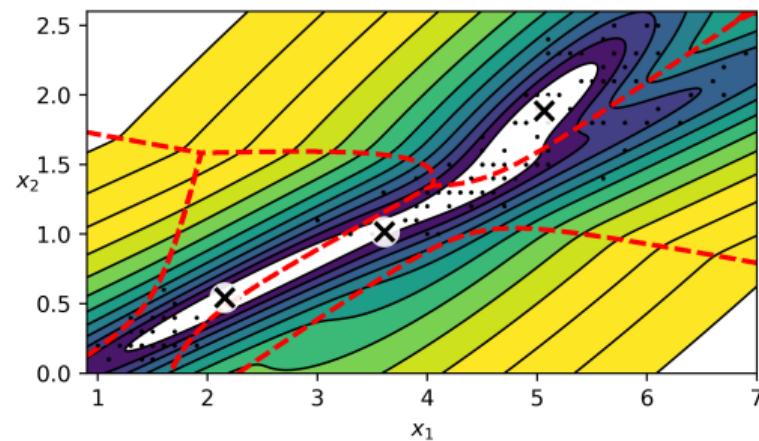
- ① **Inicializáció:** Gauss komponensek számának kiválasztása és véletlenszerű paraméter inicializáció
- ② **Elvárás:** Felelősségek kiszámítása, amely az a valószínűség minden adatpontra, hogy melyik Gauss komponensből származnak
- ③ **Maximalizálás:** Gauss paraméterek frissítése a valószínűségek maximalizálására, új várható értékek és kovarianciamátrixok kiszámítása



A modellezési eljárás

A Gauss-i keverékek modellje az EM (elvárás-maximalizálás) algoritmus alapján keresi meg az eloszlások paramétereit:

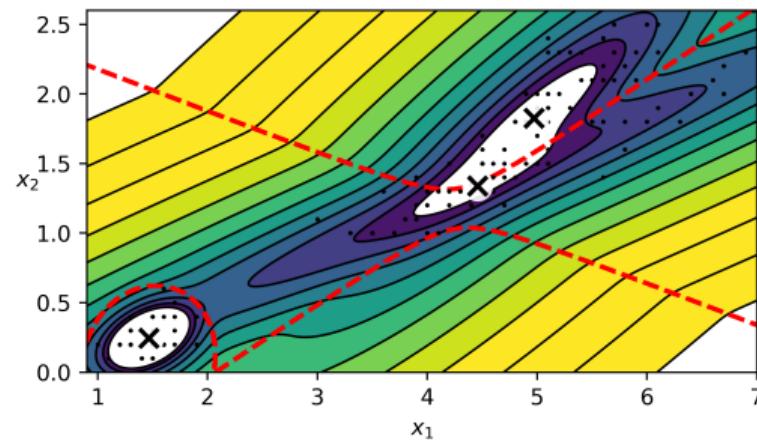
- ① **Inicializáció:** Gauss komponensek számának kiválasztása és véletlenszerű paraméter inicializáció
- ② **Elvárás:** Felelősségek kiszámítása, amely az a valószínűség minden adatpontra, hogy melyik Gauss komponensből származnak
- ③ **Maximalizálás:** Gauss paraméterek frissítése a valószínűségek maximalizálására, új várható értékek és kovarianciamátrixok kiszámítása



A modellezési eljárás

A Gauss-i keverékek modellje az EM (elvárás-maximalizálás) algoritmus alapján keresi meg az eloszlások paramétereit:

- ① **Inicializáció:** Gauss komponensek számának kiválasztása és véletlenszerű paraméter inicializáció
- ② **Elvárás:** Felelősségek kiszámítása, amely az a valószínűség minden adatpontra, hogy melyik Gauss komponensből származnak
- ③ **Maximalizálás:** Gauss paraméterek frissítése a valószínűségek maximalizálására, új várható értékek és kovarianciamátrixok kiszámítása

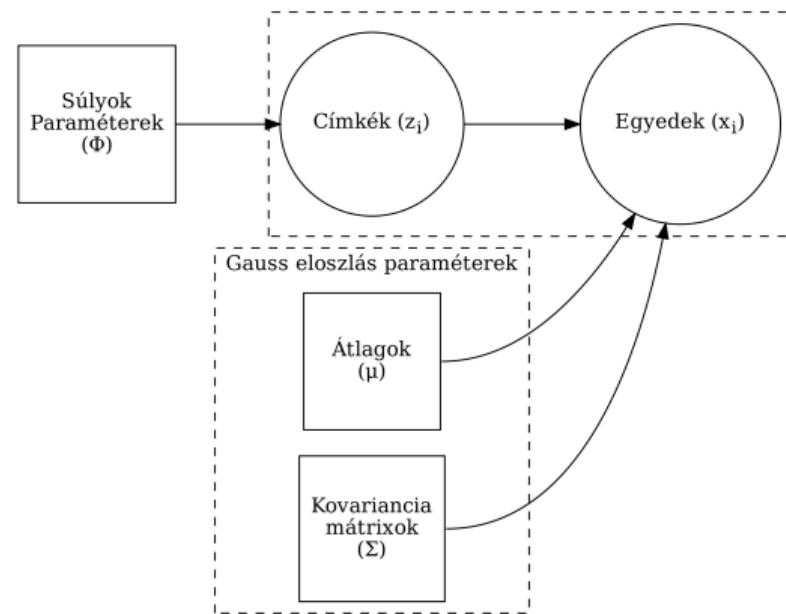


Az eljárás modellje

Minden adatponthoz véletlenszerűen rendelődik klaszter k klaszterből. A j klaszter választásának valószínűsége ϕ . Az i mintaegyedhez rendelt klaszter z_i .

Ha $z_i = j$ az i mintaegyed pozíciója szerint x_i véletlen minta választódik ki abból a Gauss eloszlásból μ, Σ paraméterekekkel:
 $x_i \sim N(\mu_j, \Sigma_j)$, azaz a
 $P(x_i|z_i = k, \mu_k, \Sigma_k)$ valószínűség.

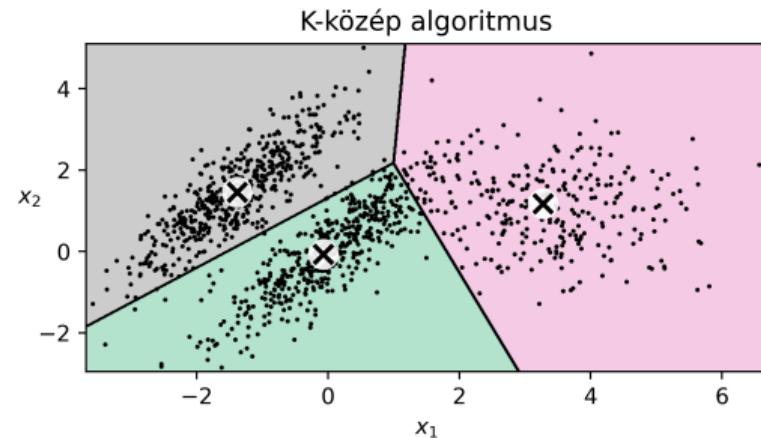
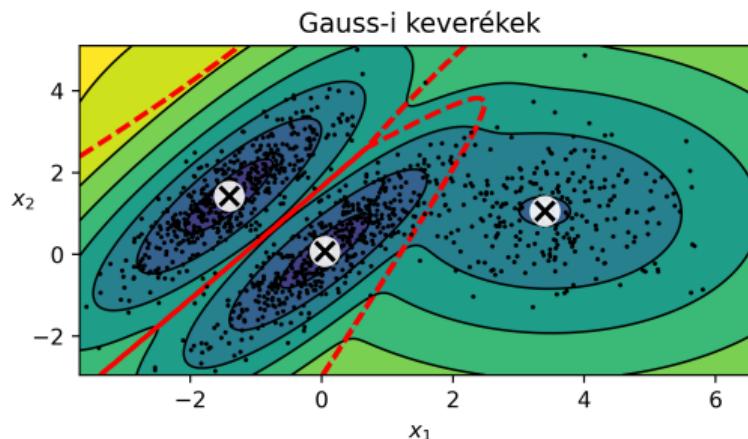
A körök véletlen változók és a téglalapok fix értékek.



A Gauss-i keverékek és K -közép kapcsolata

Az EM algoritmus a K -közép generalizált változata, ami nem csak a centroidokat keresi meg, hanem a klaszterek méretét, orientációját és a relatív súlyaikat is.

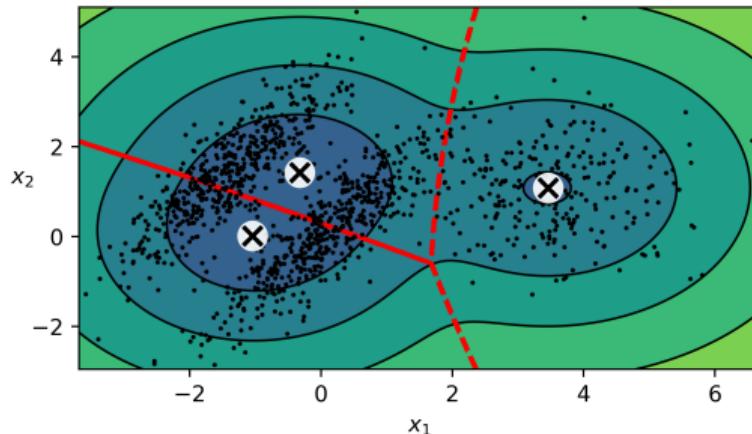
Ezzel sikerül javítania a K -közép problémáin amik miatt nem teljesít jól irregularis sűrűségű, szóródású és formájú klaszterek esetén.



Regularizáció a keverék modellek esetén

A keverékek esetén a regularizáció a kovariancia mátrixokra tett megkötésekkel érhető el. Ez a `covariance_type` paraméter. A lehetséges értékei a következők:

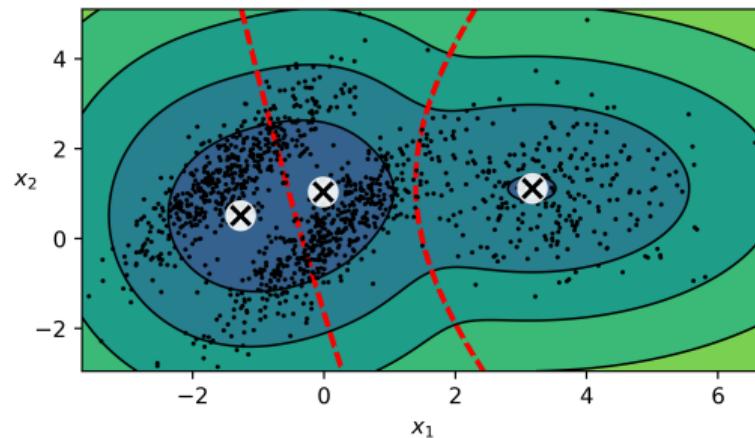
`spherical`: Kör alakú klaszterek, különböző átmérőkkel (szórással).



Regularizáció a keverék modellek esetén

A keverékek esetén a regularizáció a kovariancia mátrixokra tett megkötésekkel érhető el. Ez a `covariance_type` paraméter. A lehetséges értékei a következők:

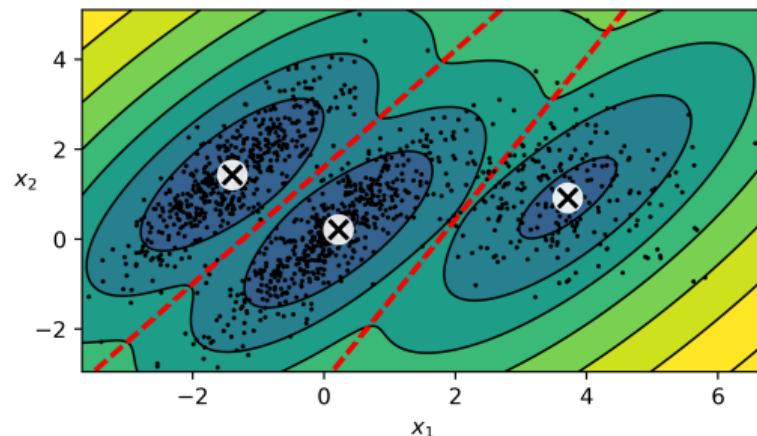
`diag`: Csak ellipszoid alakú lehet a klaszter, és a tengelyeinek a koordináta-rendszer tengelyeivel párhuzamosnak kell lennie.



Regularizáció a keverék modellek esetén

A keverékek esetén a regularizáció a kovariancia mátrixokra tett megkötésekkel érhető el. Ez a covariance_type paraméter. A lehetséges értékei a következők:

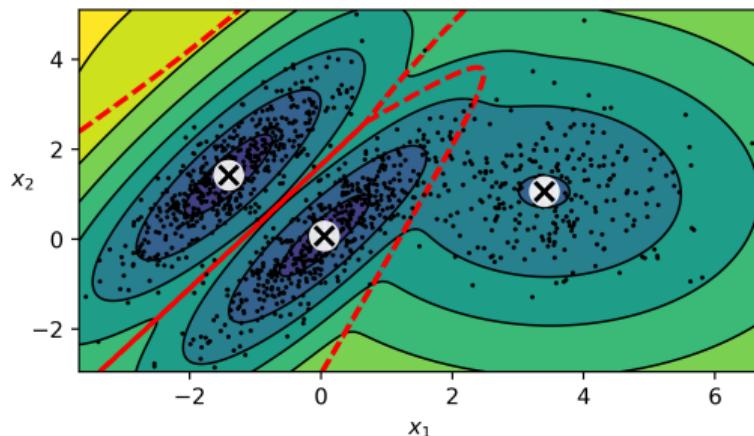
tied: minden létrejövő klaszternek ugyanolyan formájúnak, méretűnek, és orientációjúnak kell lennie.



Regularizáció a keverék modellek esetén

A keverékek esetén a regularizáció a kovariancia mátrixokra tett megkötésekkel érhető el. Ez a `covariance_type` paraméter. A lehetséges értékei a következők:

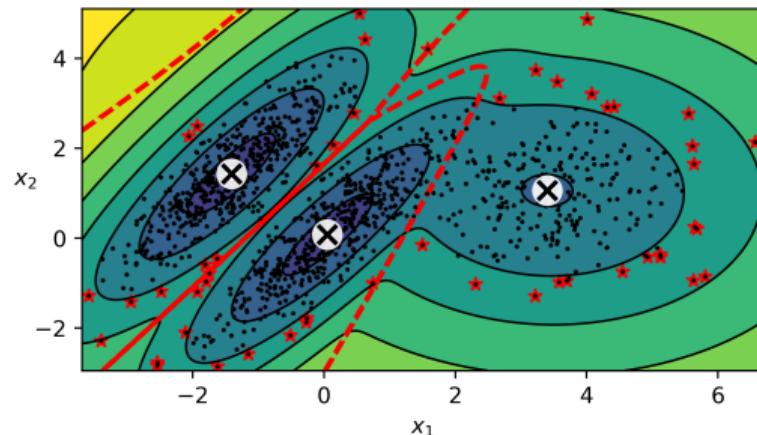
`full`: Nincs regularizáció.



Anomália detekció Gauss-i keverékekkel

A keverék modellek esetén minden olyan mintaegyed, amelyhez **adott küszöbnél alacsonyabb valószínűség sűrűség tartozik, anomáliának számít.**

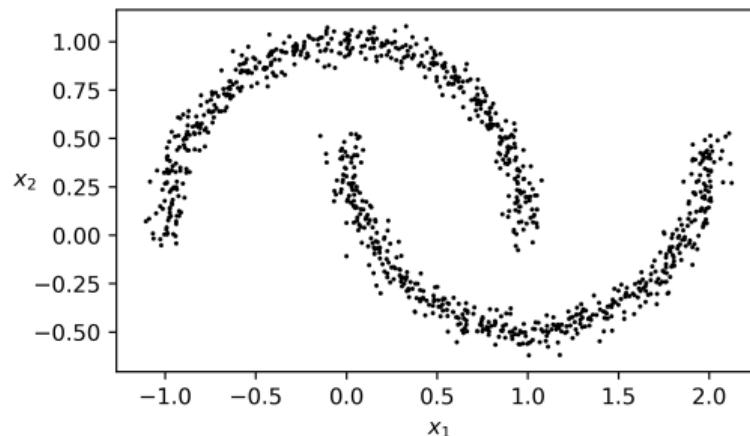
Ha túlságosan magas az anomáliák aránya, a küszöbértéket csökkenteni kell. Az ábrán a küszöbérték a negyedik percentilis.



GMM klaszterezésre

A következő példában a GMM feladata klaszterezni a `make-moons` által létrehozott irreguláris formájú klasztereiket.

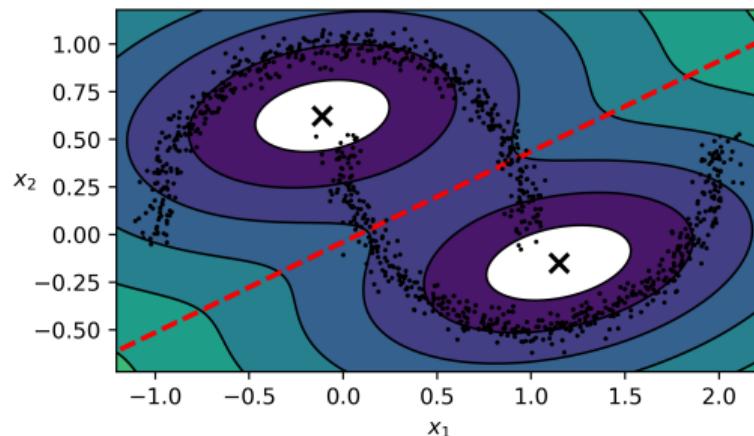
Mivel a GMM nem egy klaszterező algoritmus, a két hold alakú klasztert nem képes megfelelően definiálni.



GMM klaszterezésre

A következő példában a GMM feladata klaszterezni a `make-moons` által létrehozott irreguláris formájú klasztereket.

Mivel a GMM nem egy klaszterező algoritmus, a két hold alakú klasztert nem képes megfelelően definiálni.

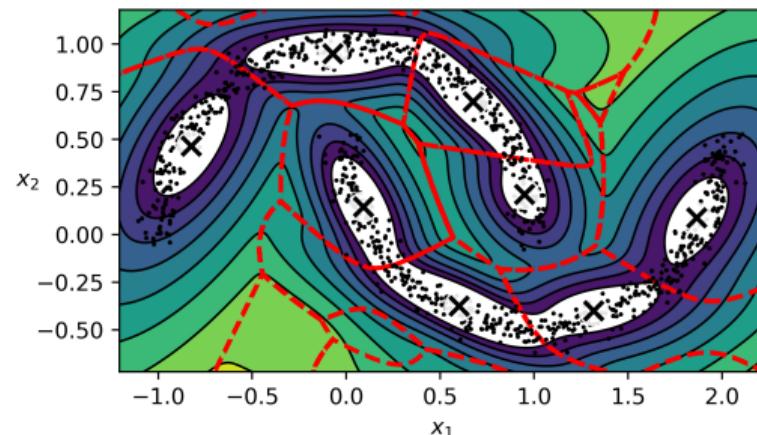


GMM sűrűségek becslésére

A 2 generátor által definiált klaszter konfiguráció szuboptimális, mert nem képesek megfelelően szeparálni a két holdhoz tartozó klasztert.

Az ábrán látható, mi történik, ha 2 helyett 10 generátor illeszkedik az adatpontokra.

Ez továbbra sem lesz képes megfelelő módon elvégezni a klaszterezés feladatát, de az adatpontok sűrűségét már jóval pontosabban becsüli meg.



Optimális generátorszám megtalálása

A keverék modellek esetén **nem megbízható a könyök módszer és a sziluett sem**, mert ezek izotróp (azonos formájú) klaszterekkel képesek jól működni.

Ehelyett rendelkezésre áll a **BIC** és **AIC** **információs kritériumok**, amelyek a modell komplexitását és jóságát veszik alapul.

BIC

$$BIC = \log(m) - 2 \cdot \log(\hat{L})$$

Ahol:

- m : Az egyedek száma
- p : A modell paramétereinek száma
- \hat{L} : A modell likelihood függvényének maximuma

Optimális generátorszám megtalálása

A keverék modellek esetén **nem megbízható a könyök módszer és a sziluett sem**, mert ezek izotróp (azonos formájú) klaszterekkel képesek jól működni.

Ehelyett rendelkezésre áll a **BIC** és **AIC információs kritériumok**, amelyek a modell komplexitását és jóságát veszik alapul.

AIC

$$AIC = 2 \cdot p - 2 \cdot \log(\hat{L})$$

Ahol:

- m : Az egyedek száma
- p : A modell paramétereinek száma
- \hat{L} : A modell likelihood függvényének maximuma

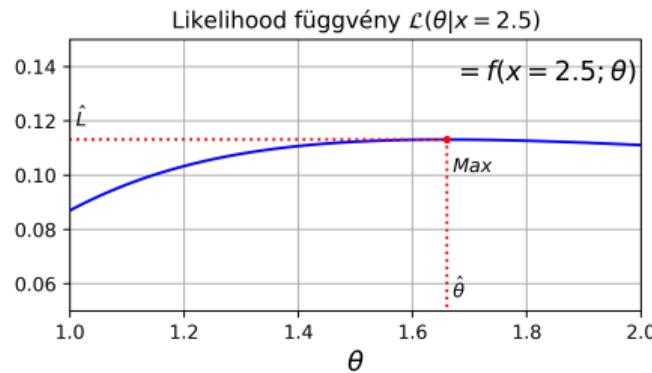
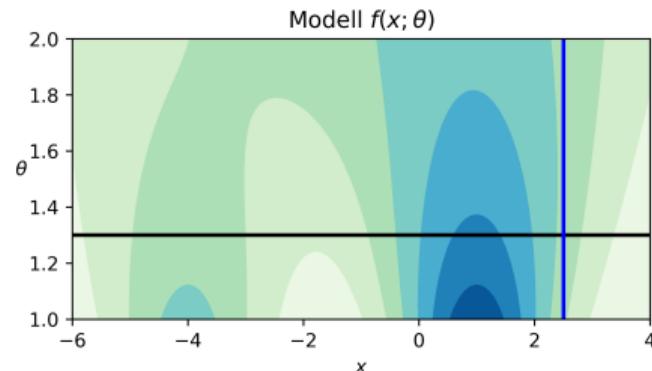
Valószínűség és likelihood

Valószínűség

A valószínűség azt adja meg, mekkora esélye van egy x kimenetnek, ha adott θ paraméterhalmaz.

Likelihood

A likelihood azt jelenti, hogy mennyire valószínű egy θ paraméterhalmaz, miután ismert x kimenet.



BIC, AIC diagram

Az optimális generátorszám ott található, ahol a minden k generátorszámra kiszámolt BIC és AIC információs kritériumoknak a legalacsonyabb az értéke:

