Gépi tanulás – III.

Szepesvári Csaba

MTA SZTAKI

2005 ápr. 11

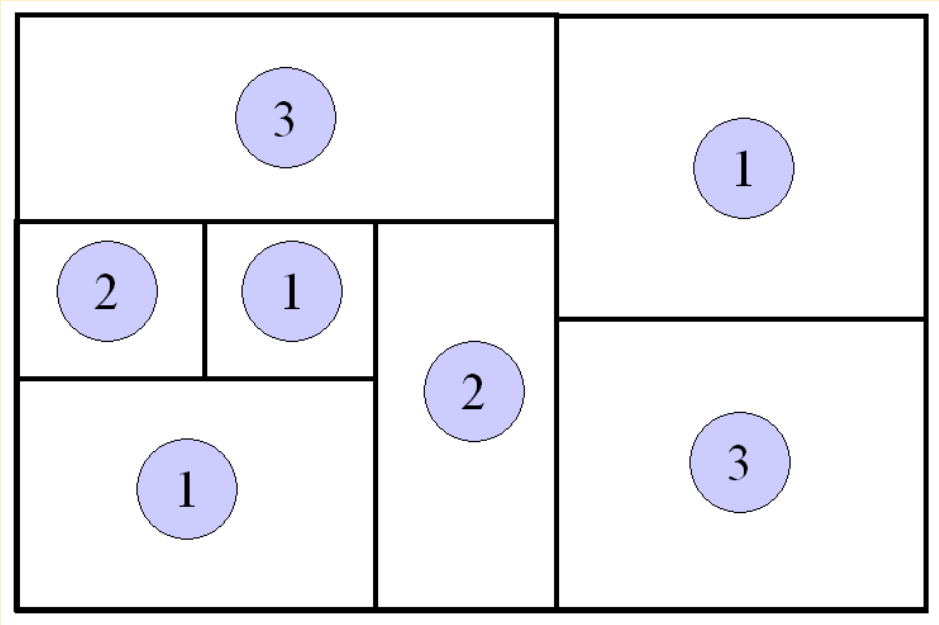
1. Döntési fák

2. Felügyelet nélküli tanulás
   - Klaszter-analízis
   - EM algoritmus
   - Gauss Mixture Models
   - Főkomponens analízis

3. Összefoglalás

Döntési fák

- Ötlet: input tér partícionálása



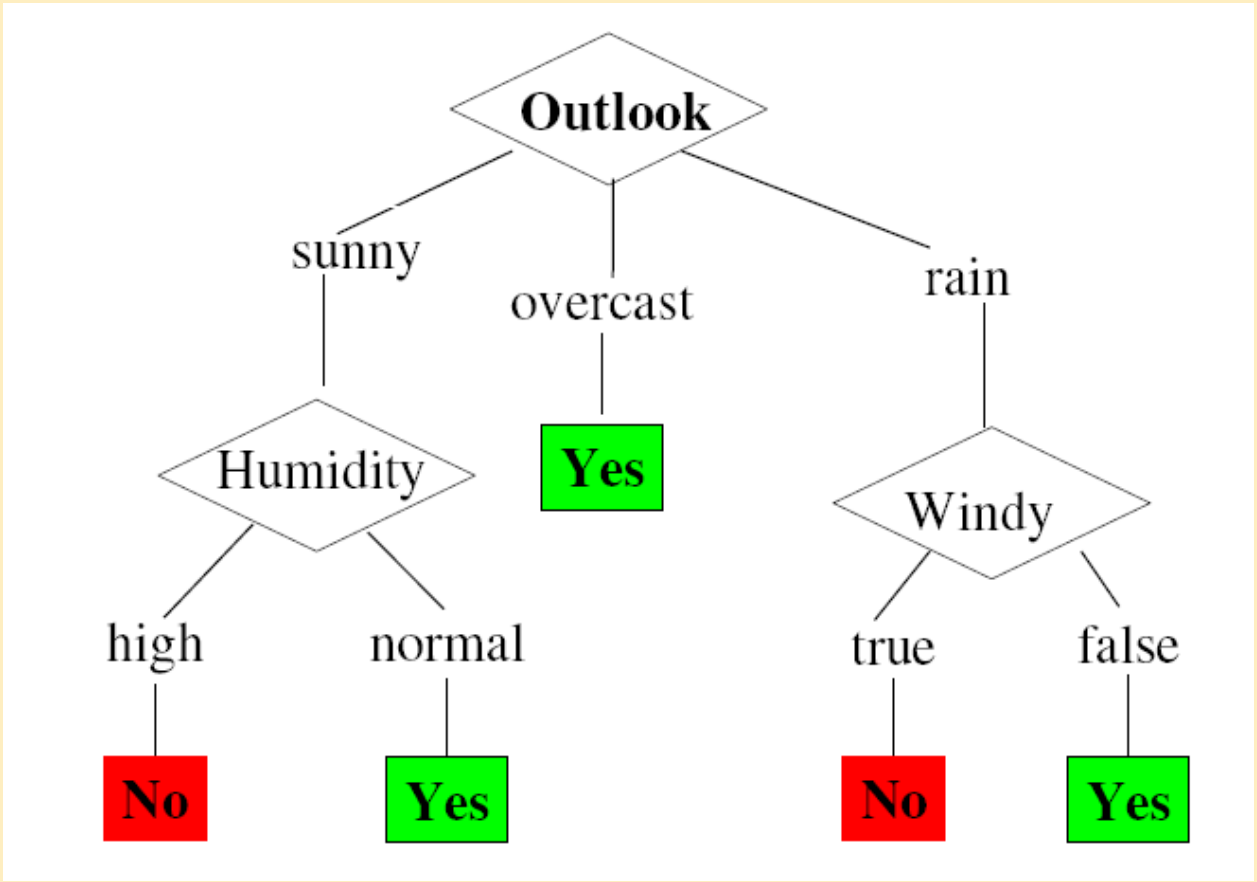
- Oszd meg és uralkodj (divide and conquer)!
- „Top-down” partícionálás; lehetőségek:
  - Egyszerre egy attibútum (ID3, C4.5) vs. általános helyzetű egyenesek mentén (CART)
  - Melyik attribútum?
  - Hol vágjunk? Hány részre?

Döntési fák

Fogalmak

- Csúcspontok, ágak, levelek
- Nem levél csúcspontok: egy-egy attribútumra vonatkozó teszt
- Ág: a teszt egy kimenetele (pl. Szín=piros)
- Levél: Címke, vagy eloszlás a címkéken
- Klasszifikáció: A fa gyökerétől egy levélig egy út bejárása

Példa: Teniszezzünk?



Döntési fák építése (Quinlan, '93)

- Top-down építkezés
  - Induláskor minden példa a gyökérhez tartozik
  - Lépésenként egy attribútumot választva rekurzívan partícionáljuk a példákat
- Bottom-up nyírás (pruning)
  - Részfákat/ágakat teszteljünk és dobjuk ki őket, ha a CV-vel becsült hibaarányon ez javít

Melyiket válasszam..?

- Melyik a legalkalmasabb attribútum?
  - Amelyik a legkisebb fához vezet
  - Heurisztika: Válasszunk olyan attribútumot, amelyik a „legegyszerűbb" alfeladathoz (nódushoz) vezet
- Jóság: amelyik attribútumnál a legnagyobb, azt választjuk ki
- Legismertebb mértékek:
  - Information gain (ID3/C4.5)
  - Information gain ratio
  - Gini index

Információ nyereség I.

- X, Y val.változók
- Y entrópiája: *Várható kódhossz, optimális kódolás mellett*:

$$H(Y) = -\mathbb{E}[\log_2 p(Y)] \left(= -\sum_i \mathbb{P}(Y = y_i) \log_2 \mathbb{P}(Y = y_i) \right).$$

- Individuális feltételes entrópia: Y entrópiája azokban az esetekben, amikor $X = x$:

$$H(Y|X = x) = \sum_i \mathbb{P}(Y = y_i|X = x) \log_2 \mathbb{P}(Y = y_i|X = x)$$

  azaz Y kódhossza, feltéve, hogy $X = x$.
- (Megj.: $H(Y|X = x) \neq -\mathbb{E}[\log_2 \mathbb{P}(Y)|X = x]$!)
- Feltételes entrópia:
  $\mathcal{H}(Y|X) = \mathbb{E}[H(Y|X)] = \sum_x \mathbb{P}(X = x)H(Y|X = x)$; Y várható kódhossza, feltéve, hogy X értéke felhasználható a kódoláskor.
  - (Megj.: $H(Y, X) = H(X) + \mathcal{H}(Y|X) = (H(Y) + \mathcal{H}(X|Y)))$

Információ nyereség II.

- X miatti információ nyereség Y-ra nézve: várhatólag mennyivel csökken Y entrópiája, azáltal, hogy X-et felhasználhatjuk a kódoláskor:
$$I(Y,X) = H(Y) - \mathcal{H}(Y|X).$$
- (Megj. $I(Y,X) = I(X,Y)$, ezért nem $I(Y|X)$-et használunk a jelölésben)
- *Köv:* Ha X hasznos információt hordoz Y-ra nézve, akkor $I(Y,X)$ nagy!

Példa biztosításmatematikából

Feladat

Adott ügyfél megéri-e a 80-at?

$Y = 1$ – hosszú életű, $Y = 0$ – nem hosszú életű.
Múltbeli adatok alapján a következőt találjuk:

- $I(Y|\text{Hajszín}) = 0.01$
- $I(Y|\text{Dohányos}) = 0.2$
- $I(Y|\text{Nem}) = 0.25$
- $I(Y|\text{SzSzUtolsóSzJegye}) = 0.0001$

Melyik attribútumokat érdemes figyelembe venni?

IG: mennyire érdekes egy 2D-s kontingencia tábla.

Sok-értékű attribútumok

- Probléma: Ha az attribútum extrém sok értéket vesz fel.. (pl. sz.sz.)
- $Y|X = x$ nagyobb eséllyel „egyszerű”, így:
  - IG jó eséllyel sok értéket felvevő attribútumot választ
  - .. ez túltanításhoz vezethet

Megoldás?

Túltanítás elkerülése

- .. pl. relativizáljuk az információ nyereséget:

$$IR(Y|X) \stackrel{\text{def}}{=} I(Y,X)/H(Y|X)$$

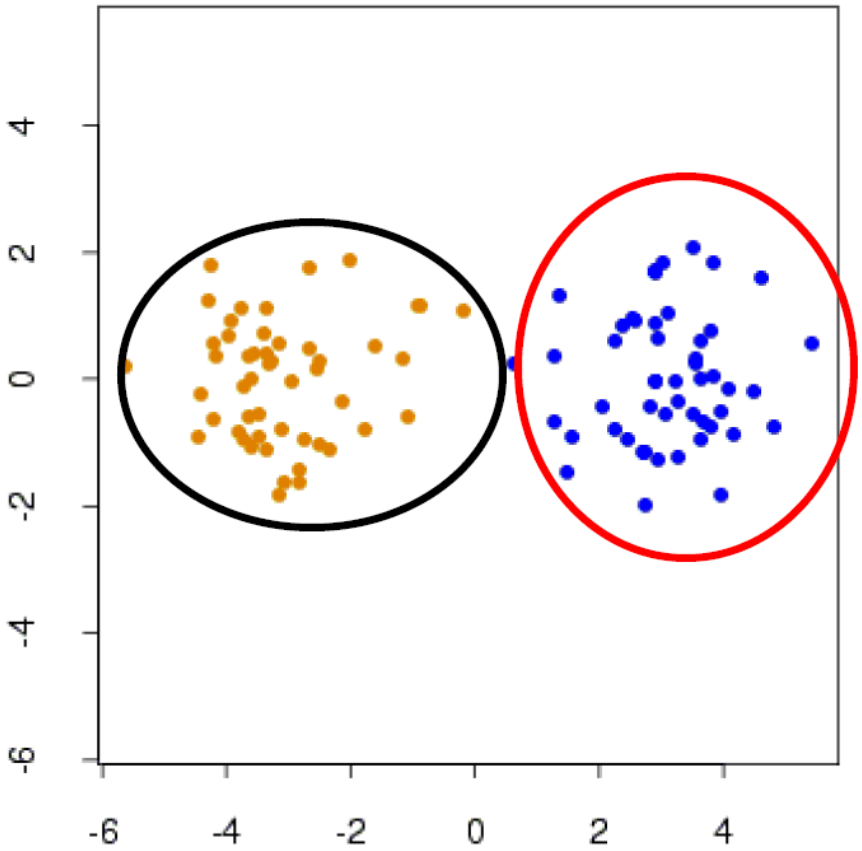
  $\Longrightarrow$Arányosított információ nyereség
- Más jóság mértékek; pl. Gini index: $G(Y) = 1 - \sum_i P(Y = y_i)^2$; $G(Y|X = x) = 1 - \sum_i P(Y = y_i|X = x)^2$; $\mathcal{G}(Y|X) = E[G(Y|X)]$
- Early stopping: Álljunk le a fa növesztésével (pl. ha a tanítási minták száma egy adott ágban már kicsi)
- Tanítsunk túl, majd vagdossuk le a részfákat/ágakat független validációs adatokon mért teljesítmény segítségével (+++)
- A leveleknél használjunk Laplace korrekciót
- ..

Egyéb praktikus kérdések

- Hiányzó attribútum értékek
- Klasszifikációs költségek
- Számításigény

$\Longrightarrow$ Quinlan's C4.5 (1993), majd C5.0

Felügyelet nélküli (unsupervised) tanulás



- Adatok: $X_1, \ldots, X_n$; $X_i \sim p(\cdot)$, p nem ismert
- Kérdések:
  - $p = ?$
    – sűrűségfüggvény becslés
  - Milyen természetes csoportok különíthetőek el a adatban?
    – klaszterezés

Klaszter-analízis

- Klaszterező függvény: $C : \mathbb{R}^d \to \{1, 2, \ldots, k\}$, $C =?$, $k =?$
- Kvantáló függvény: $Q : \mathbb{R}^d \to \{x_1, x_2, \ldots, x_k\}$, $Q =?$, $x_1, \ldots, x_k =?$, $k =?$
- Objektumok csoportosítása úgy, hogy
  - .. a hasonlóbbak kerüljenek azonos csoportba
  - .. kvantálási hiba minimális legyen
- Alapprobléma: Hogy definiáljuk a hasonlóságot/kvantálási hibát?

Klaszterező algoritmusok – alaptípusok

- Kombinatorikus módszerek
  - Közvetlenül az adatokkal dolgoznak
- Keverék-eloszlások modellezése (Mixture Modeling)
  - Valamilyen keverék-eloszlást feltételez
- Módusz-keresés
  - Sűrűségfüggvények móduszának keresése

- Klaszteren belüli különbözőség (dissimilarity) minimalizálása
  ekvivalens: klaszterek közötti különbözőség maximalizálása

Miért?

- Klaszteren belüli különbözőség:

$$W(C) = \sum_p \sum_q \mathbb{I}(C(X_p) = C(X_q))\, d(X_p, X_q)$$

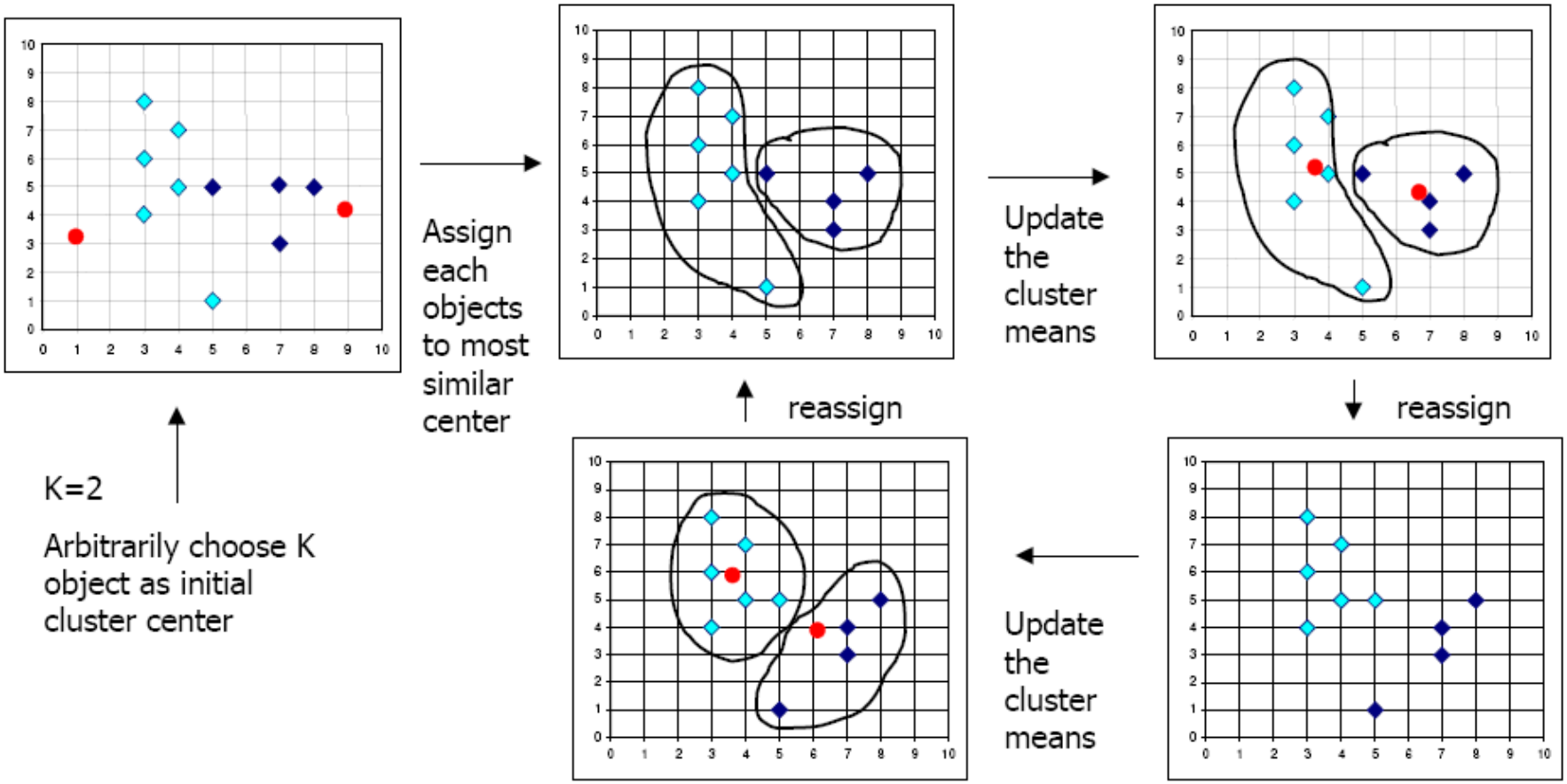
- Klaszterek közötti különbözőség:

$$B(C) = \sum_p \sum_q \mathbb{I}(C(X_p) \neq C(X_q))\, d(X_p, X_q)$$

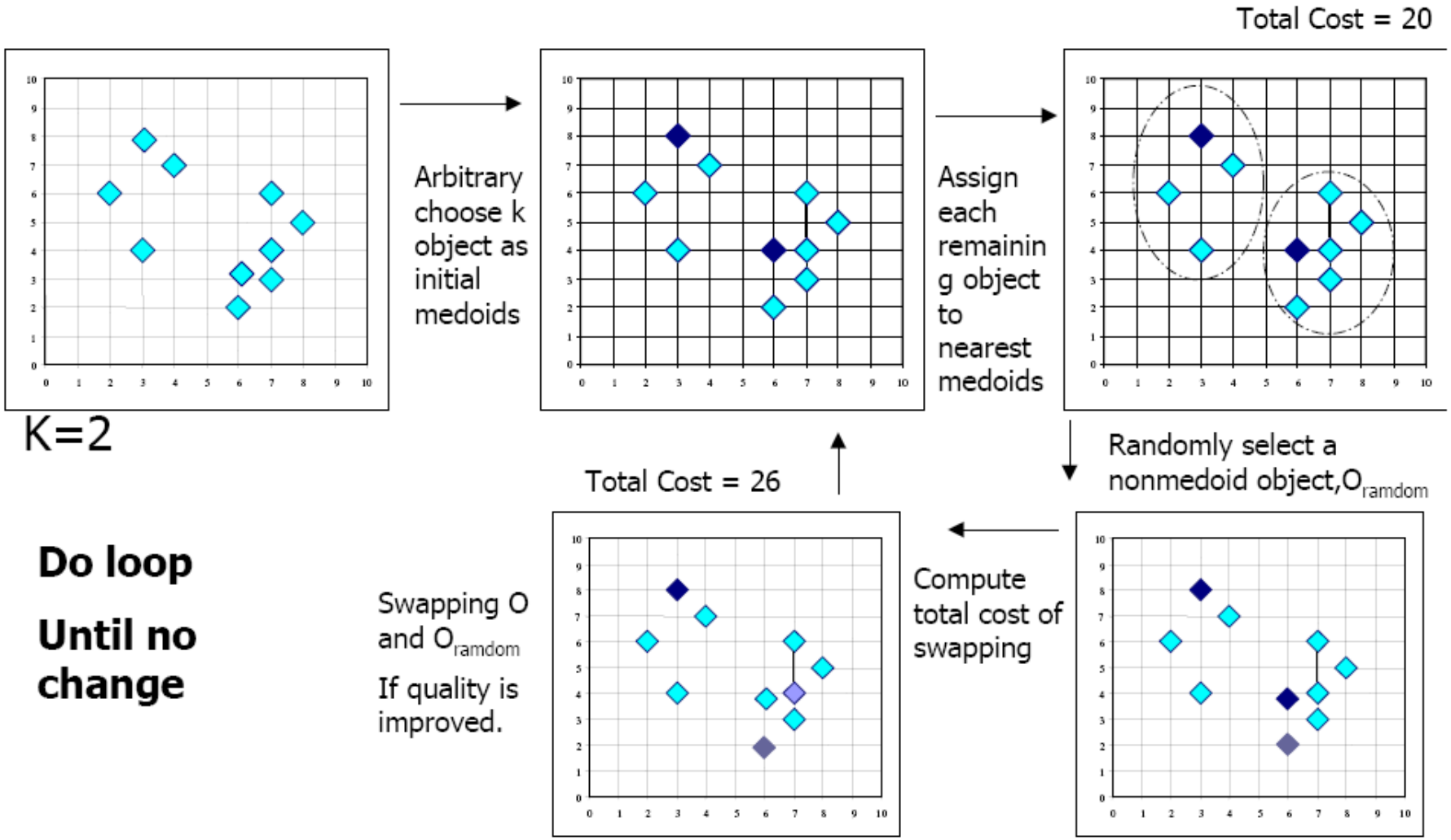
- Teljes különbözőség:

$$T(C) = W(C) + B(C) \equiv \text{const.}$$

k-Means


K=2
Arbitrarily choose K object as initial cluster center
Assign each objects to most similar center
reassign
Update the cluster means
reassign
Update the cluster means

k-Medoids


Total Cost = 20
Arbitrary choose k object as initial medoids
Assign each remainin g object to nearest medoids
Randomly select a nonmedoid object,O ramdom
Compute total cost of swapping
Total Cost = 26
Swapping O and O ramdom
If quality is improved.
Do loop
Until no change
K=2

Mennyi legyen k értéke?

- Néha a problémával jön k értéke is!:)
- Ha k nem adott, akkor $W_k(C_k^*)$-ot megvizsgáljuk..
- Ha k nő, $W_k(C_k^*)$ csökken..
  - Büntessük a nagyobb k értéket (MDL, MAP/BIC,..)
  - Validációs adat: cross-validation
- pl. „X-means” (D. Pelleg, A. Moore, 2000)

EM algoritmus – sűrűségfüggvény becslés

- $D_n = (X_1, X_2, \ldots, X_n)$; $X_i \sim p$
- $p = ?$
- Maximum likelihood:
  - $p = p(\cdot; \theta)$;
  - $L(\theta) = \log p(D_n; \theta) = \sum_i \log p(X_i; \theta)$ – függetlenség
  - $$\theta_{ML} = \text{argmax}_\theta \sum_i \log p(X_i; \theta)$$
- Rejtett változók: $(X_i, Z_i) \sim p$

Miért?

- Keverék eloszlások
- Hidden-Markov Models (rejtett Markov modellek)

Modellezés keverék eloszlásokkal

Példa: Gaussok keveréke

Keverék:

$$p(x) = \sum_{j=1}^{r} w_j \mathcal{N}(x; \mu_j, \Sigma_j); \quad w_j \geq 0, \quad \sum_{j=1}^{r} w_j = 1$$

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Generatív szemlélet:

$$Z_i \sim (w_1, \ldots, w_r), \quad i = 1, \ldots, n$$
$$X_i \sim \mathcal{N}(x; \mu_{Z_i}, \Sigma_{Z_i})$$

Paraméterek:

$$\theta = (w_1, \mu_1, \Sigma_1, \ldots, w_r, \mu_r, \Sigma_r) = ?$$

Keverékek - II.

- $D_n = (X_1, X_2, \ldots, X_n)$; $X_i \sim p$
- $p = p(x, z; \theta) = ?$
- Marginalizálás: $p(x; \theta) = \mathbb{E}[p(x, Z; \theta)]$
- Maximum likelihood:

$$\theta_{ML} = \mathrm{argmax}_\theta \sum_i \log p(X_i; \theta) = ?$$

- Rejtett változók értékének dekódolása: $\mathrm{argmax}_z p(z|x; \theta) = ?$ (pl. HMM)

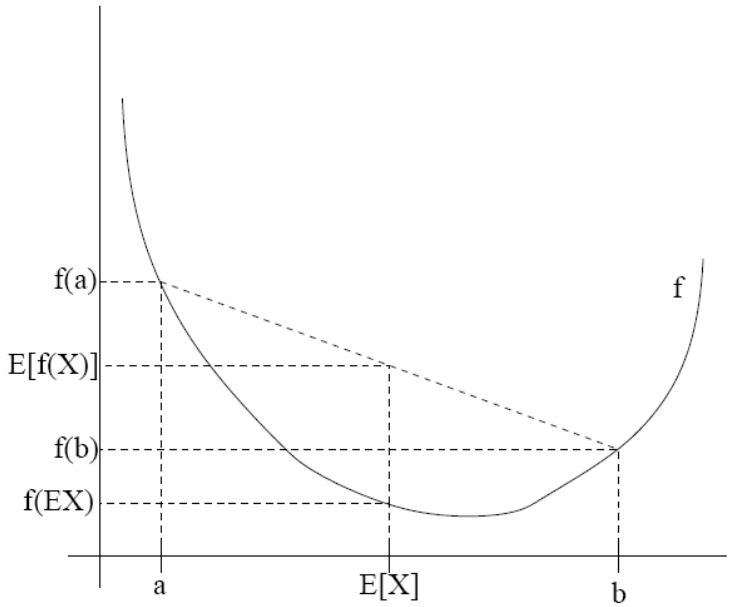
Jensen egyenlőtlensége

Tétel

Ha $f : D \to \mathbb{R}$ konvex, és X D-értékű val.változó, akkor

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}X)$$

(f konkáv, akkor $\mathbb{E}[f(X)] \leq f(\mathbb{E}X)$)



EM algoritmus: első lépések

- Ha Z_i-t megfigyeltük volna θ_{ML} számolható lenne
- De Z_i-t nem figyeltük meg..
- Log-likelihood-ra alsó becslés:

$$\begin{aligned} L(\theta) = \sum_i \log p(X_i;\theta) &= \sum_i \log \sum_z p(X_i, z;\theta) \\ &= \sum_i \log \sum_z Q_i(z) \frac{p(X_i, z;\theta)}{Q_i(z)} \text{ (Jensen :)} \\ &\geq \sum_i \sum_z Q_i(z) \log \frac{p(X_i, z;\theta)}{Q_i(z)} \end{aligned}$$

  – aholis $Q_i(\cdot)$ egy tetszőleges eloszlás.
- Jensen: $Y_i(x) = \frac{p(x,Z_i;\theta)}{Q_i(Z_i)}$-re, $Z_i \sim Q_i(\cdot)$, $f(w) = \log(w)$ mellett.
- Hogy válasszuk Q_i-t?

EM algoritmus: Q_i választása

$$
\begin{aligned}
L(\theta) = \sum_i \log p(X_i;\theta) &= \sum_i \log \sum_z p(X_i, z;\theta) \\
&= \sum_i \log \sum_z Q_i(z) \frac{p(X_i, z;\theta)}{Q_i(z)} \\
&\geq \sum_i \sum_z Q_i(z) \log \frac{p(X_i, z;\theta)}{Q_i(z)}
\end{aligned}
$$

- Hogy válasszuk Q_i-t?
- Ötlet: Úgy, hogy fent egyenlőség legyen.
- Jensen: $Y_i \equiv const \Longrightarrow$ egyenlőség
- $Q_i(z) \propto p(X_i, z;\theta)$, spec. $\sum_z Q_i(z) = 1$ miatt

$$
Q_i(z) = \frac{p(X_i, z;\theta)}{\sum_z p(X_i, z;\theta)} \equiv p(z|x;\theta)
$$

EM algoritmus

Algoritmus

Input: $(X_1, \ldots, X_n)$ – adatok; $p(x, z; \theta)$ modell

$t = 0$

Konvergenciáig ismételni:

- E-step: Minden i-re, $Q_i(z) := p(z|X_i; \theta_t)$
- M-step: $\theta_{t+1} := \text{argmax}_\theta \sum_i \sum_z Q_i(z) \log \frac{p(X_i, z; \theta)}{Q_i(z)}$
- $t := t + 1$

EM algoritmus: konvergencia

Állítás

Az EM algoritmus konvergál $L(\theta)$ egy lokális maximumhelyéhez

Most csak: $L(\theta)$ sohasem csökken

- *Def.:*

$$A(\theta;\theta') = \sum_i \sum_z Q_i(z;\theta) \log \frac{p(X_i, z;\theta')}{Q_i(z;\theta)}$$

- Volt: $L(\theta_t) = A(\theta_t;\theta_t)$
- $A(\theta_t;\theta_{t+1}) = \max_{\theta'} A(\theta_t;\theta') \geq A(\theta_t;\theta_t) = L(\theta_t)$
- $L(\theta_{t+1}) \geq A(\theta_t;\theta_{t+1})$? – igen; volt:

$$L(\theta) \geq \sum_i \sum_z Q_i(z) \log \frac{p(X_i, z;\theta)}{Q_i(z)}$$

spec. $\theta = \theta_{t+1}, Q_i(z) = Q_i(z;\theta_t)$-ra is áll.

EM algoritmus: egy másik interpretáció..

- EM $L(\theta)$-t nem csökkenti
- Leállási feltétel: $L(\theta)$ nem sokat változik
- Koordinátánkénti optimalizálás:
  - $Q = (Q_1, \ldots, Q_n)$;
  - $J(Q, \theta) = \sum_i \sum_z Q_i(z) \log \frac{p(X_i, z; \theta)}{Q_i(z)}$
  - Volt: $L(\theta) \geq J(Q, \theta)$
  - Interpretáció:
    - E-step: Q-ban maximalizálás
    - M-step: θ-ban maximalizálás

Inicializálás

- EM iteratív; nagyon érzékeny a kezdeti értékekre
- Rossz pontból indítva $\Longrightarrow$ rossz pontba konvergál
- Gyors és minőségi inicializálásra van szükség
- Gyakran: k-means (GMM-hez)
- Alternatívák: Gaussian splitting, hierarchikus k-means

Gauss Mixture Models

Modell:

$$p(x;\theta) = \sum_{j=1}^{r} w_j \mathcal{N}(x;\mu_j,\Sigma_j); \quad w_j \geq 0, \quad \sum_{j=1}^{r} w_j = 1$$

$$\mathcal{N}(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

- $p(x,z;\theta) = w_z \mathcal{N}(x;\mu_z,\Sigma_z)$
- $p(x) = \sum_z p(x,z;\theta)$
- Behelyettesítés..
  - $Q_i(z;\theta) = p(Z=z|X_i;\theta) = \frac{p(X_i|Z=z;\theta)p(Z=z;\theta)}{p(X_i;\theta)}$
  - ..

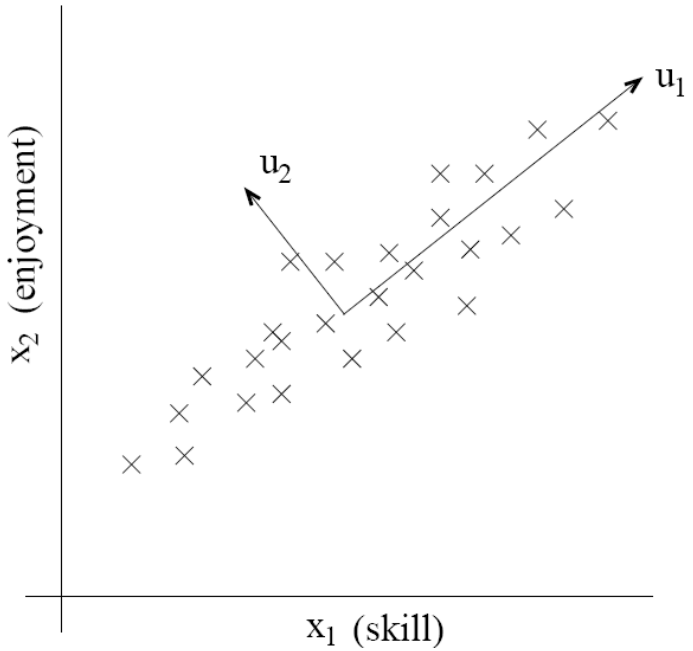
Miért érdekesek a GMM-ek

- Univerzális approximátorok
- Diagonális GMM-ek is univerzális approximátorok

Főkomponens analízis

Főkomponens analízis = Principal Component Analysis (PCA)

- Példa
  - Helikopter-vezetés távirányítással
  - Képesség
  - Élvezet
  - „karma” - rejtett változó
- Feladat: keressük azt az irányt (alteret), amelyikben az adat legjobban változik



Főkomponens analízis - Modellek

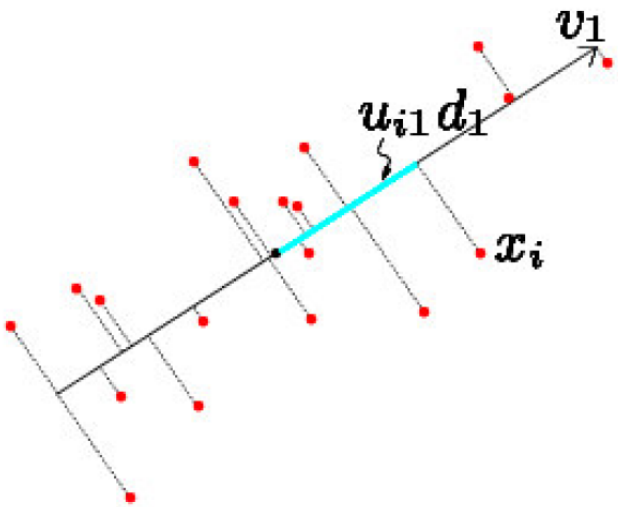
- Modell-1 (nem-parametrikus):
  - Projekció után az adatoknak a lehető legnagyobb legyen a varienciája
- Modell-2 (nem-parametrikus)
  - $P : \mathbb{R}^d \to \mathbb{R}^d$ projekció egy m dimenziós altérre
  - $Px = \sum_{j=1}^{m} v_j (v_j^T x)$, $v_i^T v_j = \delta_{ij}$.
  - $\operatorname{argmin}_{P \text{ proj.}} \mathbb{E}[\|PX - X\|_2^2] = ?$
- Modell-3 (parametrikus)
  - $X = AU + W$
  - $U \in \mathbb{R}^m$, $W \in \mathbb{R}^d$ v.v.
  - $A \in \mathbb{R}^{d \times m}$ mátrix
  - $(X_1, \ldots, X_n)$-t figyeljük meg, $A = ?$
  - Feltétel: U, W Gauss („ellipszisek")

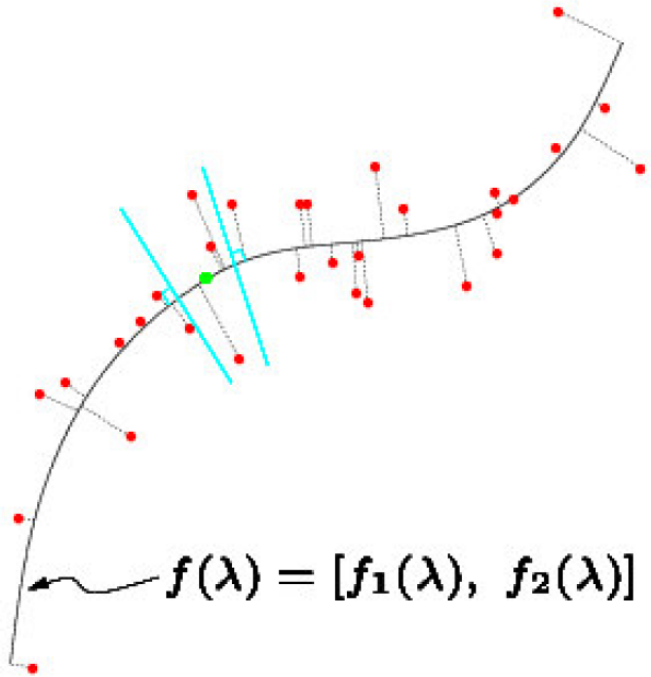
Algoritmus

Input: $(X_1, \ldots, X_n)$

- Centrálunk: $X_i' := X_i - m$, $m = (1/n)\sum_i X_i$.
- Empirikus kovariencia mátrix: $C = (1/n)\sum_i X_i'(X_i')^T$
- C sajátértékfelbontása: $C = \sum_{i=1}^{d} \lambda_i v_i v_i^T$, $\lambda_1 > \lambda_2 > \ldots \lambda_d$
- Output: $P = \sum_{j=1}^{m} v_j v_j^T$

Principal curves and surfaces





Összefoglalás

- Döntési fák: top-down, mohó (information gain); ID3/C4.5
- Felügyelet nélküli tanulás
  - Klaszterezés (k-means, k-medoid)
  - EM-algoritmus
  - Gauss-keverékek
  - Főkomponens analízis
- Legközelebb:
  - Független komponens analízis
  - Lineáris diszkriminánsok
  - NMF
  - Feature-selection
  - Feature weighting; boosting