Dimenziócsökkentő eljárások

Dimenziócsökkentés

- Egy adatállománynak számos jellemzője lehet.
- Tekintsük dokumentumok egy halmazát, amelyben minden dokumentumot egy olyan vektor reprezentál, melynek elemei az egyes szavak előfordulási gyakoriságai az adott dokumentumban. Az ilyen esetekben általában több ezer vagy több tízezer attribútum (elem) van, a szótár minden szavához egy.
- Másik példaként tekintsük idősorok egy halmazát, amely különböző részvények egy 30 éves időintervallum folyamán feljegyzett napi záróértékeiből áll. Az attribútumokból, amelyek itt a konkrét napokhoz tartozó árak, ebben az esetben is több ezer van.

Dimenziócsökkentés előnyei

- A dimenziócsökkentés egyik legfontosabb haszna, hogy számos adatbányászati algoritmus jobban működik, ha a dimenziószám - az adatok attribútumszáma - kisebb. Ennek oka részben az, hogy a dimenzió csökkentésével kiküszöbölhetőek a lényegtelen jellemzők és csökkenthető a zaj, részben pedig a dimenzió probléma.
- A dimenzió csökkenésével az adatbányászati algoritmus(ok) számára szükséges idő és memóriamennyiség is csökken.

Dimenziócsökkentés előnyei

- Egy másik előnye, hogy a dimenzió csökkentése egy érthetőbb modellhez vezethet, mert a modellben kevesebb attribútum fog szerepelni.
- Emellett a dimenziócsökkentés adatok könnyebb ábrázolását teszi lehetővé. Még ha a dimenziócsökkentés nem is redukálja az adatokat két- vagy háromdimenzióssá, az adatokat gyakran ábrázoljuk attribútumpárjaik vagy attribútum-hármasaik alapján, és az ilyen kombinációk száma így jelentősen csökken.

A dimenzió probléma

- A dimenzió probléma azt a jelenséget jelenti, hogy számos adatelemzés lényegesen nehezebbé válik az adatok dimenziójának növekedésével. Speciálisan, a dimenzió növekedésével az adatok egyre ritkábban helyezkednek el az általuk kitöltött térben.
- Osztályozásnál ez azt is jelentheti, hogy nem lesz elég adatobjektum ahhoz, hogy létrehozzunk egy olyan modellt, amely minden lehetséges objektumot megbízhatóan besorol egy osztályba.

A dimenzió probléma

- Klaszterezésnél a sűrűség és a pontok közötti távolság definíciói, amelyek ennél a módszernél kritikus fontosságúak, veszítenek jelentőségükből.
- Ennek eredményeként sok klaszterező és osztályozó algoritmus (és más adatelemző algoritmusok) számára problémát jelentenek a magas dimenziójú adatok -- csökken az osztályozás pontossága és gyenge minőségű klaszterek jönnek létre.

Dimenziócsökkentés

- Transzformáljuk úgy alacsonyabb dimenzióba a több/sokdimenziós adatunkat, hogy a benne rejlő variancia minél kisebb részét veszítsük csak el
- Feltevés: az eredeti m-dimenziós adatpontok egy bizonyos m’-dimenziós altéren (vagy legalábbis annak közelében) helyezkednek el → az adatpontokat az altér tengelyeire transzformálva jól reprezentálható az eredeti adat
- Mi lehet ez a m’-dimenziós altér?

A dimenziócsökkentés lineáris algebrai módszerei

- A legáltalánosabb dimenziócsökkentési megközelítések között is van néhány, főként folytonos adatok esetén, mely a lineáris algebra módszereit alkalmazva képezi le a magas dimenziójú térben lévő adatokat egy alacsonyabb dimenziójú térbe.
- A főkomponens analízis (PCA -- Principal Component Analysis), illetve a szinguláris felbontás (SVD -- Singular Value Decomposition) ilyen, lineáris algebrai módszerek, amelyek kapcsolódnak is egymáshoz és gyakran használják őket dimenziócsökkentésre.

A főkomponens analízis (PCA)

- **A főkomponens analízis (PCA -- Principal Component Analysis)** egy ilyen, lineáris algebrai módszer, amely olyan új attribútumokat (főkomponenseket) tár fel, amelyek:
- (1) az eredeti attribútumok lineáris kombinációi,
- (2) ortogonálisak (merőlegesek) egymásra, és
- (3) az adatokban fellelhető ingadozást maximálisan kifejezik.
- Az első két főkomponens például az adatok ingadozását maximálisan kifejezi két olyan ortogonális attribútummal, melyek az eredeti attribútumok lineáris kombinációi.

A főkomponens analízis (PCA)

- A főkomponens analízis célja dimenziók (attribútumok) olyan új halmazának a keresése, mely jobban tükrözi az adatok variabilitását.
- Az első dimenziót úgy választjuk meg, hogy a maximálisan lehetséges mértékű variabilitást hordozza.
- A második dimenzió merőleges az elsőre és ezen kényszerfeltétel mellett a fennmaradó variabilitásból a lehető legtöbbet hordozza, és így tovább.

A főkomponens analízis (PCA) - előnyök

- A PCA segítségével könnyebb megtalálni az adatokat legjobban jellemző mintázatokat. Ezért a PCA mintázat-kereső módszerként is használható.
- Gyakori, hogy az adatok nagyon nagy mértékben tömöríthetők az információtartalom lényeges csökkenése nélkül. Utóbbi miatt a PCA-n alapuló dimenziócsökkentés relatíve alacsony dimenziót eredményez, mely már olyan eszközökkel is kezelhető, ami a kiinduló adatokon nem használható.
- Mivel az adatokban található zaj (reményeink szerint) gyengébb, mint a mintázatok, a dimenziócsökkentés a zajok nagy részét képes kiküszöbölni. (adatbányászat, adatelemző algoritmusok)

A főkomponens analízis (PCA)

Egy többváltozós (több, folytonos attribútummal rendelkező) adatsor variabilitását többek között az **S** kovarianciamátrixszal jellemzik.

- Ha adott egy m-szer n-es **D** adatmátrix, melynek m sora az adatokat azonosítja, n oszlopa pedig a jellemzőket, akkor kovarianciamátrixa az az **S** mátrix, melynek s_{ij} eleme az i. és j. attribútum (oszlop) kovarianciája.
- Két attribútum kovarianciája azt méri, hogy a két mennyiség mennyire erősen függ egymástól.

A főkomponens analízis (PCA)

- Ha a **D** adatmátrix előzetes feldolgozásával azt olyan alakúra hoztuk, hogy az egyes attribútumok középértéke 0, akkor $\mathbf{S}=\mathbf{D}^T\mathbf{D}$.
- A PCA célja olyan transzformáció megkeresése:
- Minden új (különböző elemekből álló) attribútumpár kovarianciája 0.
- Az attribútumok aszerint vannak rendezve, hogy milyen mértékben járulnak hozzá a szóráshoz: az első attribútum járul hozzá a szóráshoz a legnagyobb mértékben. Az ortogonalitási feltétel miatt minden egyes bevont attribútum olyan mértékben járul hozzá a szóráshoz, amennyire csak lehetséges.

A főkomponens analízis (PCA)

- Az adatoknak az a transzformációja, mely ezeket a tulajdonságokat eredményezi, megkapható a kovarianciamátrix sajátértékeinek elemzésével.
- Legyenek e célból $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ az **S** mátrix sajátértékei. A sajátértékek mindegyike nemnegatív.
- Legyen $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$ az **S** sajátvektoraiból álló mátrix.
- Tegyük fel végül, hogy a **D** adatmátrix az előfeldolgozás során úgy lett átalakítva, hogy minden attribútum (oszlop) középértéke 0.

A főkomponens analízis (PCA)

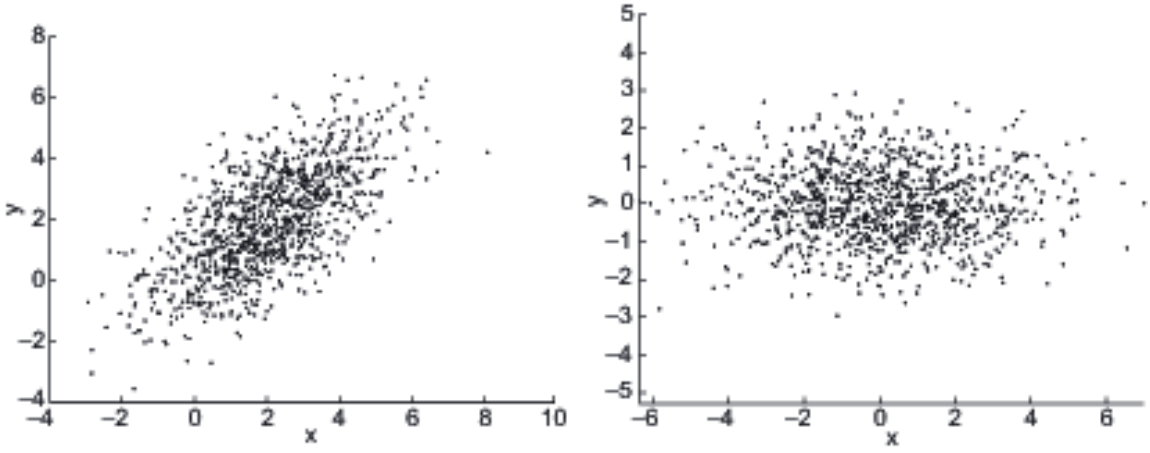
- Ekkor a következőket állíthatjuk.
- A **D'**= **DU** adatmátrix a transzformált adatok olyan halmaza, mely már teljesíti a fentebbi feltételeket.
- Minden új attribútum az eredetiek lineáris kombinációja. Konkrétabban, az i. attribútumot előállító lineáris kombináció súlyai az i. sajátvektor komponensei.
- Az i. új attribútum varianciája λ_i .
- A régi és az új attribútumok varianciáinak összege megegyezik.
- Az új jellemzőket **főkomponensek**nek nevezzük, így például az első új attribútum az első főkomponens.

A főkomponens analízis (PCA)

A legnagyobb sajátértékhez tartozó sajátvektor jelzi azt az irányt, amely irányban a legnagyobb az adatok varianciája. Ezt szemléletesebben úgy is kifejezhetjük, hogy ha az összes adatvektort rávetítenénk ezen sajátvektor által megadott egyenesre, akkor a keletkező értékeknek így lenne a legnagyobb a varianciájuk minden lehetséges irány közül. A második legnagyobb sajátértékhez tartozó sajátvektor azt az irányt adja meg, mely egyfelől merőleges az előzőre, és melyre nézve az adatoknak a legnagyobb a megmaradó varianciája.

A főkomponens analízis (PCA)

Ekkor **S** sajátvektorai új tengelyeket definiálnak. A PCA tehát úgyis tekinthető, mint a koordinátatengelyek olyan irányú forgatása, melyet az adatok variabilitása határoz meg. A teljes szóródás a transzformáció során megmarad, de az új attribútumok már korrelálatlanok lesznek.



(a) Eredeti pontok.

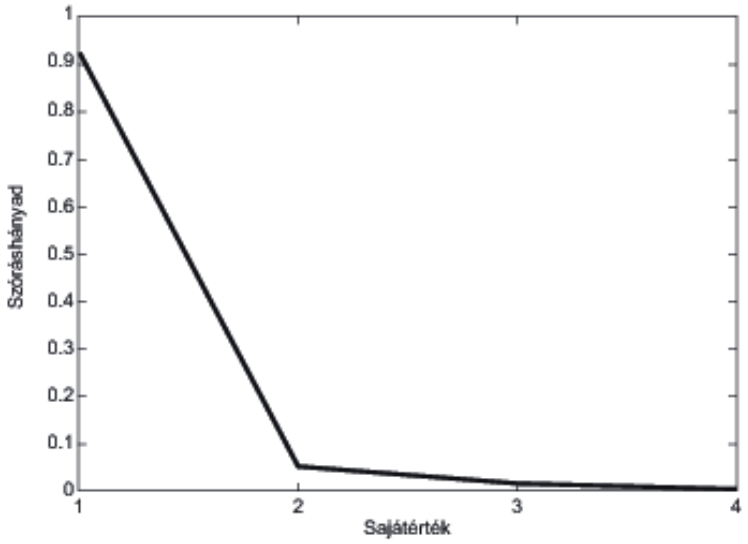
(b) Transzformált pontok.

A főkomponens analízis (PCA)

- **Írisz adatok**
- Ez a példa az íriszek (nőszirom) adatsorán mutatja be a dimenziócsökkentést. Ez az adatsor 150 adat objektumot (virágokat) tartalmaz. 50-50 virág van, három különböző fajból: nőszirom (Setosa), foltos nőszirom (Versicolor) és virginiai nőszirom (Virginica). Minden növény négy jellemzőjét tároljuk: csészelevél hossza és szélessége, sziromlevél hossza és szélessége.

A főkomponens analízis (PCA)

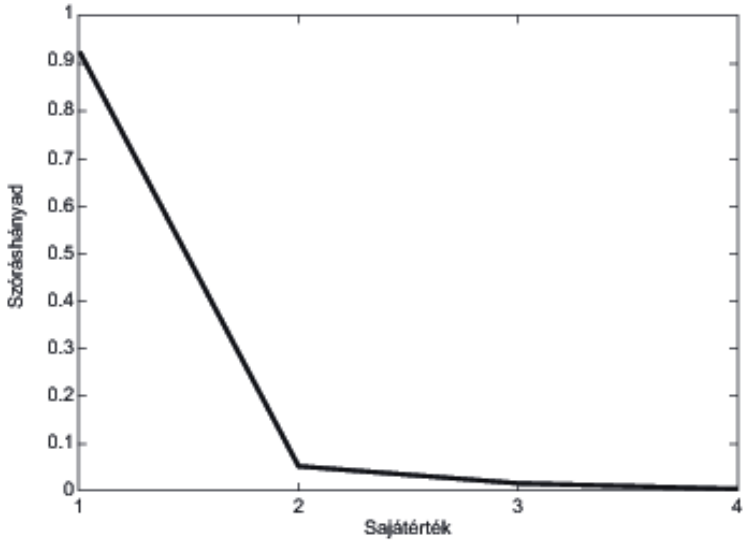
- **kőtörmelék ábra (scree plot):** a kovarianciamátrix egyes sajátértékei (a főkomponensek) milyen mértékben magyarázzák a teljes szórást. A megtartandó főkomponensek számának meghatározására használjuk azért, hogy megőrizzük az adatok ingadozásának a legnagyobb részét.



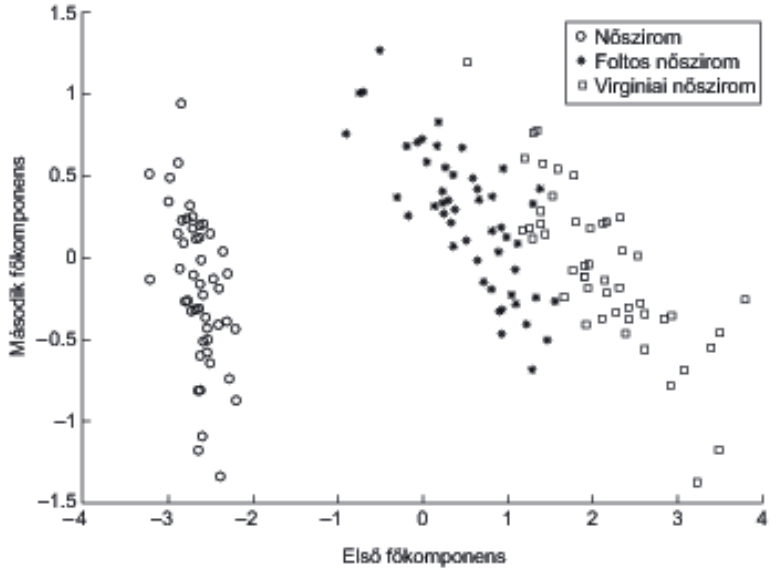
(a) Az egyes főkomponensek által képviselt szórások.

A főkomponens analízis (PCA)

- A nőszirom adatai esetében az első főkomponens magyarázza a szórás legnagyobb részét (92,5%), a második csak 5,3%-ot magyaráz, az utolsó kettő pedig együtt csak 2,2%-ot. Így ha csak az első két főkomponenst tartjuk meg, akkor az adatsor variabilitásának nagy részét megőrizzük.



(a) Az egyes főkomponensek által képviselt szórások.



(b) PCA a nőszirom adathalmazára alkalmazva.

A főkomponens analízis (PCA)

- **Összegzés**
- Centralizáljuk és normalizáljuk a **D** adatmátrixot
- Számoljuk ki a (centralizált és normalizált) **D** adatmátrixot jellemző kovariancia/szóródási-mátrixot
- Számoljuk ki a mátrix sajátértékeit
- Az m’ legnagyobb sajátértékhez tartozó sajátvektorból képezzük **U** projekciós mátrixot
- **D’=DU** adja meg a transzformált adatmátrixot
- $\mathbf{D'U^T}$ alakban kaphatjuk vissza az eredeti adatpontok egy közelítését

Bevezetés az adatbányászatba (Pang-Ning Tan, Michael Steinbach, Vipin Kumar) című tananyaga alapján készült (részben)