Komprese dat neuronovou sítí

Aplikace I

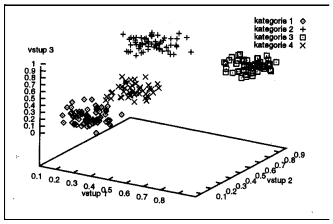
Komprese dat

- Objasňovat proč komprimovat by bylo mlácením prázdné slámy. Každý ví …
- Typy komprese
  - ztrátová,
  - bezeztrátová.
- Otázka: jaký typ komprese očekáváte u ANN?

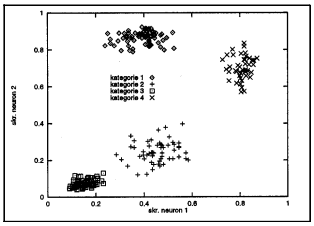
Neuronovou sítí?

- Jedna z aplikací:
  - kompresí k odstranění nepodstatné, irelevantní informace.
- Náš názor: právě zde jsou vhodné umělé neuronové sítě.

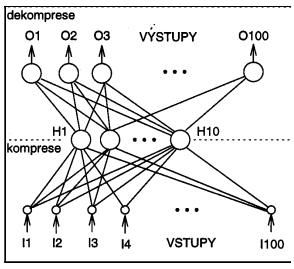
Příklad komprese dat v našem smyslu



Po kompresi:



Jak jsme toho dosáhli? Třeba takhle:



Diskuse

- Umíte si představit i jiný typ komprese za použití jiných umělých neuronových sítí?
  - Zde redukce dimenzionality,
  - jinde náhrada shluku vzorem,
  - jinde …?
- nn

Dílčí závěry diskuse:

- Tedy nejen BP, ale i
- Kohonenovu síť i
- ART

- můžeme použít pro kompresi dat.

Metody komprese založené na statistice

- Standardní situace:
  - počet sledovaných proměnných systému je velký, interpretaci jevů znepřehledňuje.
  - Nelze jejich počet redukovat, ale přitom zachovat nezkreslený obraz o původním systému?
- Metody mnoharozměrné statistické analýzy
  - komponentní analýza,
  - faktorová analýza.

- Jejich cílem je zhuštěný popis, redukce dimenze.
- Umožnila to **vzájemná korelace** původních proměnných.
- Nové proměnné jsou **syntetické**, často **hypotetické** a jsou zobecněním proměnných původních.

Budeme potřebovat pojmy:

Kovariance

$$\sigma_{xy} = \mathrm{Cov}(x, y) = E((x - E(x))\,(y - E(y))) = E(xy) - E(x)E(y)\,.$$

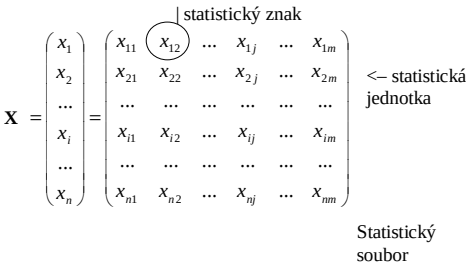
Kde $E(.)$ je střední hodnota proměnné $E(.) = \frac{1}{n}\sum_{i=1}^{n} (.)_i$

Korelace

$$\rho_{xy} = \sigma_{xy} / (D(x)D(y))^{1/2}\,,$$

Kde $D(.)$ je jejím rozptylem $D(.) = \frac{1}{n}\sum ((.)_i - E(.))^2$

Statistický soubor, znak?



Kovarianční matice I

$$\mathbf{C_i} = \begin{pmatrix} (x_{i1}-E(x_1))(x_{i1}-E(x_1)) & (x_{i2}-E(x_2))(x_{i1}-E(x_1)) & \dots & (x_{im}-E(x_m))(x_{i1}-E(x_1)) \\ (x_{i1}-E(x_1))(x_{i2}-E(x_2)) & (x_{i2}-E(x_2))(x_{i2}-E(x_2)) & \dots & (x_{im}-E(x_m))(x_{i2}-E(x_2)) \\ \dots & \dots & \dots & \dots \\ (x_{i1}-E(x_1))(x_{im}-E(x_m)) & (x_{i2}-E(x_2))(x_{im}-E(x_m)) & \dots & (x_{im}-E(x_m))(x_{im}-E(x_m)) \end{pmatrix}$$

Kovarianční matice II

- Čtvercová, symetrická matice,
  - na hlavní diagonále jsou rozptyly jednotlivých statistických znaků,
  - mimo jsou kovariance statistických znaků.
- Její význam:
  - kovarianční matice je vyjádřením míry závislosti mezi jednotlivými statistickými znaky.

Co je to PCA?

- PCA - Principal Component Analysis,
- Hotteling (1933), navázal na Pearsona (1901).
- Statistická metoda vyhledávající ve vícerozměrných datech skryté závislosti,
- umožňuje redukci dat z dimenze *m* do *n* (*m*>*n*).

Jak souvisí s kovariancí?

- Jak uvidíte dále (a přečtete si ve skriptech), matematický výpočet je založen na použití kovarianční matice.
- Charakteristickým polynomem matice **C** nazýváme determinant matice $\Lambda\mathbf{E} - \mathbf{C}$.
- Jeho kořeny jsou vlastní čísla matice **C** a
- vlastní čísla jsou **hlavními komponentami**.

Geometrická interpretace hlavních komponent

- Hledá se takové natočení os nového souřadného systému, kdy
  - nová 1. Osa (souřadnice z_1) prochází směrem maximálního rozptylu shluku původních vektorů a
  - promítá se tedy na ni maximum informace.
  - Druhá je na ní kolmá a platí totéž.
- Analogicky dále.

Motivační příklad

- Pro lepší pochopení definic a postupů
  - příklad „ze života“.
- Znáte všichni pohádku Mrazík a její postavy
  - Nastěnku, Ivánka, Marfušu,
  - macechu, babu Jagu, loupežníky a
  - dědečka Mrazíka?
- Tak si tedy hezky sedněte a poslouchejte ...

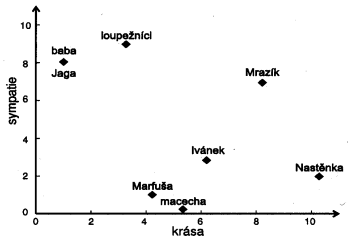
Hodnocení *krásy* a *sympatií*

Postava	Krása	Sympatie
baba Jaga	1	8
loupežníci	3	9
Marfuša	4	1
macecha	5	1
Ivánek	6	3
Mrazík	8	7
Nastěnka	10	2

$$\vec{X} = \begin{pmatrix} baba \\ loup \\ Marf \\ mace \\ Ivan \\ Mraz \\ Nast \end{pmatrix} = \begin{pmatrix} 1 & 8 \\ 3 & 9 \\ 4 & 1 \\ 5 & 1 \\ 6 & 3 \\ 8 & 7 \\ 10 & 2 \end{pmatrix}$$

Autorem příkladu je F. Daďa, hodnocení provedli jeho respondenti.

Rozložení postav v rovině



Připomenutí -
střední hodnota náhodného vektoru

$$\vec{\mu} = (E(x_1)\ E(x_2)\ \dots\ E(x_m)) = \left(\frac{\sum_{i=1}^{n} x_{i1}}{n} \quad \frac{\sum_{i=1}^{n} x_{i2}}{n} \quad \dots \quad \frac{\sum_{i=1}^{n} x_{im}}{n} \right).$$

V našem případě

$$\vec{\mu} = (E(kr) \quad E(sy)) = \left(\frac{\sum krasa}{n} \quad \frac{\sum symp}{n} \right) = (5{,}286 \quad 4{,}428).$$

Odchylky od střední hodnoty

Postava	? krása	? sympatie
baba Jaga	-4,286	3,571
loupežníci	-2,286	4,571
Marfuša	-1,286	-3,429
macecha	-0,286	-3,429
Ivánek	0,714	-1,429
Mrazík	2,714	2,571
Nastěnka	4,714	-2,429

Rozuměj: ? = Δ

Připomenutí - kovarianční matice

$$\mathbf{C_i} = \begin{pmatrix} (x_{i1} - E(x_1))(x_{i1} - E(x_1)) & (x_{i2} - E(x_2))(x_{i1} - E(x_1)) & \dots & (x_{im} - E(x_m))(x_{i1} - E(x_1)) \\ (x_{i1} - E(x_1))(x_{i2} - E(x_2)) & (x_{i2} - E(x_2))(x_{i2} - E(x_2)) & \dots & (x_{im} - E(x_m))(x_{i2} - E(x_2)) \\ \dots & \dots & \dots & \dots \\ (x_{i1} - E(x_1))(x_{im} - E(x_m)) & (x_{i2} - E(x_2))(x_{im} - E(x_m)) & \dots & (x_{im} - E(x_m))(x_{im} - E(x_m)) \end{pmatrix}$$

Tedy: Kovarianční matice pro babu

$$\mathbf{C_{baba}} = \begin{pmatrix} baba_{kr} - E(kr) & baba_{sy} - E(sy) \end{pmatrix} \begin{pmatrix} baba_{kr} - E(kr) \\ baba_{sy} - E(sy) \end{pmatrix} =$$

$$= (-4{,}286 \quad 3{,}572) \begin{pmatrix} -4{,}286 \\ 3{,}572 \end{pmatrix} = \begin{pmatrix} 18{,}367 & -15{,}306 \\ -15{,}306 & 12{,}755 \end{pmatrix}.$$

Podobně

- Vypočítáme
  - kovarianční matici ostatních postav a
  - kovarianční matici celkovou:

$$\mathbf{C} = \begin{pmatrix} 7{,}918 & -3{,}694 \\ -3{,}694 & 10{,}245 \end{pmatrix}.$$

Připomenutí - co to jsou hlavní komponenty?

- Hlavní komponenty jsou
  - vlastními čísly charakteristického polynomu kovarianční matice.
- Charakteristická matice kovarianční matice **C** je

$$\Lambda\mathbf{E} - \mathbf{C} = \begin{pmatrix} \lambda - 7{,}918 & -3{,}694 \\ -3{,}694 & \lambda - 10{,}245 \end{pmatrix}.$$

Výpočet

- Charakteristická čísla najdeme tak,
  - položíme rovný nule determinant této matice:
- $\det(\Lambda\mathbf{E} - \mathbf{C}) = (\lambda - 7{,}918)(\lambda - 10{,}245) - (-3{,}694)^2 = \lambda^2 - 18{,}163\,\lambda + 67{,}478$.

- Řešením kvadratické rovnice určíme, že

$\lambda_1 = 12{,}954$ a $\lambda_2 = 5{,}209$.

Výpočet II

Nyní vypočteme oba charakteristické vektory. Nejdříve pro $\lambda_1 = 12{,}954$:

$$(\mathbf{C} - \lambda_1\mathbf{E})\boldsymbol{\alpha}_1 = \mathbf{0}$$

$$\begin{pmatrix} 7{,}918 & -3{,}694 \\ -3{,}694 & 10{,}245 \end{pmatrix} - \begin{pmatrix} 12{,}954 & 0 \\ 0 & 12{,}954 \end{pmatrix}(\alpha_1 \quad \alpha_2) = (0 \quad 0),$$

Z čehož vyjde

$$\boldsymbol{\alpha}_1 = (0{,}591 \quad -0{,}806).$$

Výpočet III

Podobným výpočtem 2. charakteristický vektor

$\boldsymbol{\alpha}_2 = (0{,}806 \quad 0{,}591)$

a hledaná transformační matice matice nových souřadnic má tvar

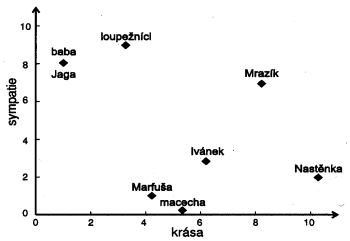
$$\alpha = \begin{pmatrix} 0{,}591 & 0{,}806 \\ -0{,}806 & 0{,}591 \end{pmatrix}$$

Výsledek

$\mathbf{z_i} = \boldsymbol{\alpha}\mathbf{x_i}$.

Postava	z_1	z_2
Baba Jaga	-5,81	5,53
Loupežníci	-5,43	7,72
Mrazík	-0,87	10,54
Iváček	1,15	6,57
Marfuša	1,57	3,79
macecha	2,16	4,59
Nastěnka	4,31	9,18

… a jeho interpretace. Srovnejte



a ...

… interpretace?

- Co znamenají syntetické proměnné **z**?
- Daďa navrhuje
  - z_1 **vtipnost** nebo **schopnost vyvolávat smích**,
  - z_2 **kladnost**.
- Souhlasíte?

Neuronový způsob výpočtu hlavních komponent

- **E. Oja** (doktorand T. Kohonena) si všiml,
  - že McCulloch-Pittsův perceptron s Hebbovs-kým učicím algoriotmem počítají
  - 1. hlavní komponentu.
- Za rozšíření mohou K.I. Diamantras a S.Y. Kung. To se označuje jako
  - APEX (Adaptive Principal Component Extraction) algoritmus.

APEX

