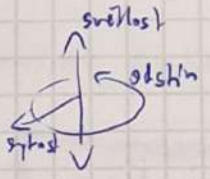
Data science

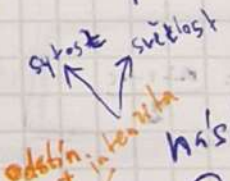
1) seženeme a připravíme data (ETL: extraction, transformation, loading do cílového systému)
   - data wrangling (munging)
2) na data se koukáme z různých úhlů, filtrujeme, těžíme
   ⇝ trendy, ...
3) grafický design, prezentace ⟶ efektivně, jasně
   - odhalíme strukturu
   - storytelling with data
4) interakce s daty – InfoVis, HCI (human computer interaction)

Exploratorní (EDA) ✗✗ explanatorní vizualizace

- mezi námi a daty
- mezi daty a publikem

70% 👁, 30% zbylé smysly

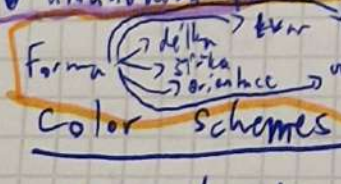




Sakadický pohyb očí – nespojitý

Obrazová paměť – předvědomé zpracování – nástroje, jak něco zvýraznit
Krátkodobá paměť – vědomé zpracování
Dlouhodobá paměť

- barva, velikost, orientace, lokace



Color schemes

- monochromatic – různé intenzity stejného odstínu
- analogous – vedle sebe
- complementary
- triadic – rovnostranný trojúhelník

vizuální chunky

= spojení objektů na základě předvědomých atributů

dobře se pamatuje max. 3–4 chunky (v krátkodobé paměti)
- při opakování se dostanou do dlouhodobé paměti

⟶ storytelling je o vztazích těchle

kvantitativní ✗✗ kategorické rozdíly

- např. velikost → měřitelné
- např. barva, ID, číslo účtu, číslo dveří → pozorovatelné, ale neměřitelné (nebo spíš že s nimi nejde- ... dělat aritm. op.)
  - nominální – bez lin. usp.
  - ordinální – usp., ale nekvantifikovat (např. velikost: small, medium, large)
  - hier. intervalové

Gestalt principy (organizující principy)

- blízkost ⇒ skupina
- podobnost – barva, tvar, orientace
- enclosure (ohrada) + co je ohraničené, tvoří jeden celek
- closure (spojení, uzávěr) – [] ◯
- continuity
- spojení ⇒ skupina

Efekty kontextu

– víc vnímáme rozdíly než abs. hodn.

Statistiky

- průměr
- medián
- modus
- midrange = $\frac{min+max}{2}$

average

- std. odch.
- spread = max – min

variation

↳ ovlivněno extrémy !!

korelace → -1 → neg. korelace
až
1 → pozit. korelace

Tabulka ✗✗ Graf

Tabulka:
- potřebujeme číst jednotlivé přesné hodnoty, porovnávat je, je víc úrovní agregace (Σ, avg, ...)

Graf:
- chceme vidět pattern, outliery
- jako celek, porovnáváme vůči celku

Kvantitativní hodnoty

- body => scatter plot
- úsečky => < spojení bodů => vzorce / best fits -> trendy

- bars
- boxes => distribuce - box plot
- plocha -> pie chart
  - -> bubble chart
- barva -> bubble plot s různými intenzitami nebo odstíny

Kategorické hodnoty

- pozice
- odstín
- tvar značky
- fill pattern -> pozor na Moiré effect (že to dělá iluzi takového vlnění)
- typ čáry

---

2 Distribuce

box ploty ~> **histogramy** ~> **density histogramy** ~> **kernel density Function (KDF)** ~>

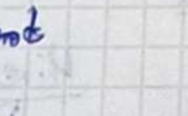
- box ploty -> medián, min, max, 1. a 3. kvartil
- histogramy -> chlívečky
- density histogramy -> hodnoty procenta z celku
- KDF -> místo chlívečků už jednotlivé hodnoty, na které se dají ⌒ a posčítají se

~> **kum. distrib. funkce** -> ≤ hodnot ~> **QQ plot** -> kvantily proti sobě

!

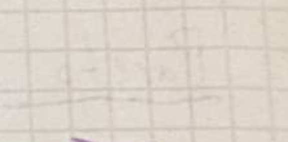
Ještě víc distribucí

- freq. polygon



- ridgeline plot
  - prostě víc jejich KDF

- stacked density chart



nebo třeba s histogramem

- candlestick chart
- violin plot
  - KDF · 90° + zrcadlemo

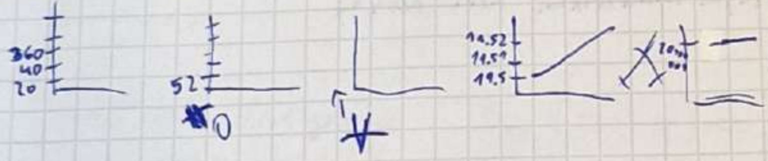
Nejistota

Lhaní

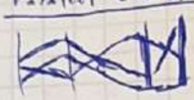
Lie Factor := velikost efektu ve vizualizaci / velikost efektu v datech

< 0.95 nebo > 1.05 => FUJ

Manipulace s Y



Parallel coords.



Polar chart

Icicle tree

Sunburst

-> hierarchie
- listy na obvodu

Circle packing

Sankey diagram

=> tok dat

Parallel sets

Waterfall chart

- bar začíná tam, kde předchozí skončil
=> změny

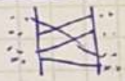
Gantt chart

-> "rozvrh"

Sparklines

-> datawords
-> "text s minigrafy"

Slopegraph



Tag/word cloud

Stem-and-leaf plot

"histogram", ale v chlívečcích jsou ty hodnoty
-> tabulka

Jointplot

scatter/hex plot + histogram(y)

Small multiple

prostě víc stejných grafů s různými daty

Treemap

=> hierarchie / část celku

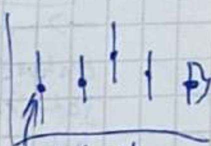
Arc diagram

-> 3D: arc map
- takové chobotničky

-> když spojíme konce do kruhu: radial chart

může být čistší hierarchical edge bundling

Binned scatterplot



chlívečky

prům. / medián / ..

konf. int.

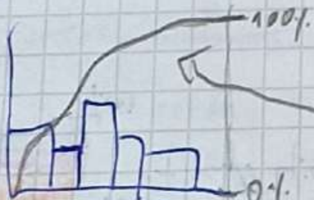
Heatmap

-> pomocí odstínů
- kvantitativní atributy
- korelace

- často s dendogramem
-> shlukuje podobné objekty

- correlogram = když uděláme heatmapu korelací víc atributů

Pareto chart



- prostě něco - třeba bar chart + kum. % nebo průběžný součet - třeba line chart

Hive plots

Bag plot

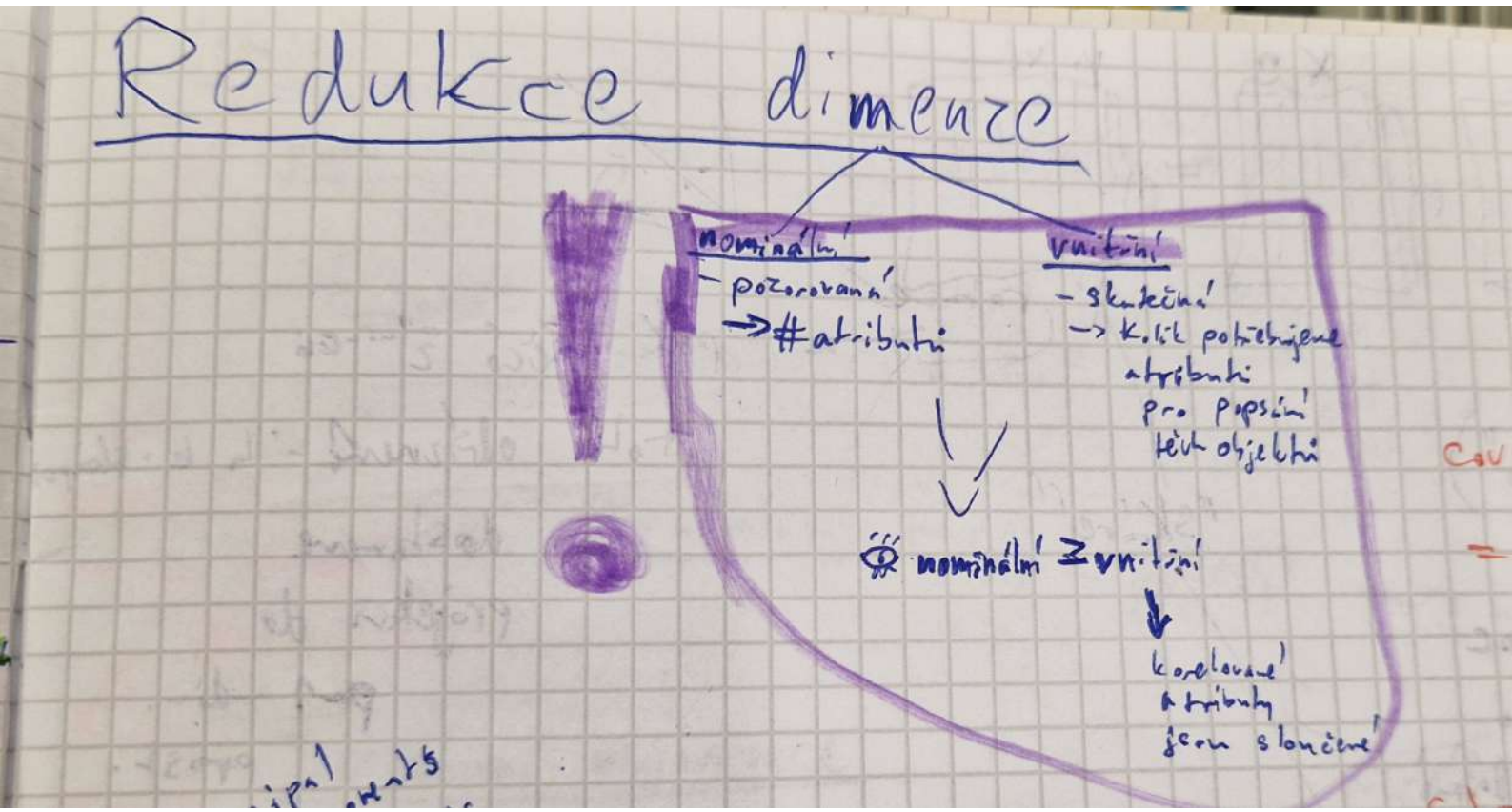
3x med.

Konv. obal v tom 3x med.

BioFabric

vrchol

Redukce dimenze



PCA

předp. už vycentrované

$$\mathrm{var}(X) := \frac{\sum_{i=1}^{n}(X_i - \mathbb{E}[X])^2}{n} = \frac{\sum_{i=1}^{n} X_i^2}{n} = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2]$$

$$\mathrm{cov}(X,Y) := \frac{\sum_{i=1}^{n}(X_i - \mathbb{E}[X])(Y_i - \mathbb{E}[Y])}{n} \rightsquigarrow \mathrm{cor}(X,Y) := \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{cov}(X,X)\,\mathrm{cov}(Y,Y)}}$$

kov. matice Σ = pro X je to XX^T

1. PC: $Y_1 = a_1^T X$ t.ž. $\mathrm{var}(Y_1)$ je max. při $a_1^T a_1 = 1$

$$\mathrm{var}(Y_1) = \mathrm{var}(a_1^T X) = \mathbb{E}\left[(a_1^T X)(a_1^T X)^T\right] =$$

$$= \mathbb{E}[a_1^T \underbrace{XX^T}_{\Sigma} a_1] = a_1^T \mathbb{E}[\Sigma]\, a_1 = a_1^T \Sigma a_1$$

Lagrangeovy multiplikátory – max. $f(x,y)$ při $g(x,y) = c$

$\Lambda(x,y,\lambda) = f(x,y) - \lambda(g(x,y) - c)$ ← chceme max.

↳ tedy $\Lambda(a_1, \lambda) = a_1^T \Sigma a_1 - \lambda(a_1^T a_1 - 1)$

chceme: $\frac{\partial \Lambda(a_1,\lambda)}{\partial a_1} = 0$

$2\Sigma a_1 - 2\lambda a_1$

$\Sigma a_1 = \lambda a_1 \Rightarrow a_1$ je vl. vektor Σ,

λ je k němu příslušné vl. číslo

$\mathrm{var}(Y_1) = a_1^T \Sigma a_1 = \lambda \underbrace{a_1^T a_1}_{1} = \lambda \Rightarrow \lambda := \lambda_1$

2. PC: $Y_2 = a_2^T X$, s max. $\text{var}(Y_2)$ při $a_2^T a_2 = 1$ a $\text{cov}(Y_2, Y_1) = 0$

$$\Lambda(a_2, \lambda, \kappa) = a_2^T \Sigma a_2 - \lambda(a_2^T a_2 - 1) - \kappa(a_2^T a_1)$$

$$\frac{d}{da_2} = 0$$

$$2\Sigma a_2 - 2\lambda a_2 - \kappa a_1 = 0 \quad / \cdot \frac{a_1^T}{2}$$

$$a_1^T \Sigma a_2 - \lambda a_1^T a_2 - \frac{\kappa}{2} a_1^T a_1 = 0$$

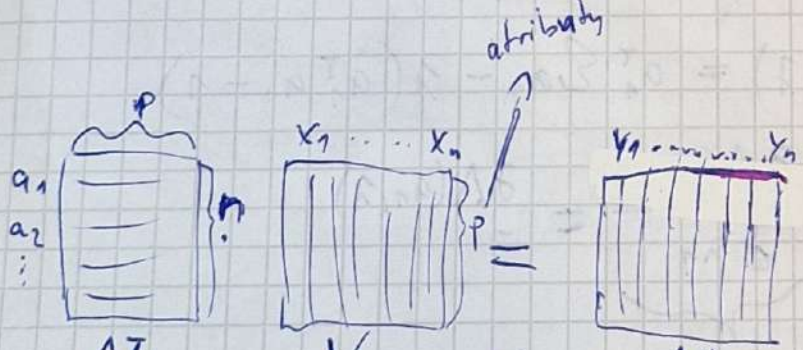
$a_1^T \Sigma a_2 = 0$, $\lambda a_1^T a_2 = 0$, $\frac{\kappa}{2} a_1^T a_1 = 1$

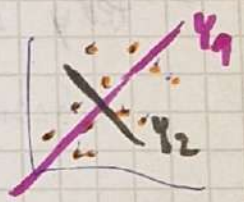
$$\Rightarrow \kappa = 0$$

$$\Rightarrow \Sigma a_2 = \lambda a_2 \rightarrow \lambda := \lambda_2$$

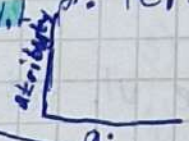
$0 = a_2^T \Sigma a_1 = a_2^T \lambda_1 a_1 = \lambda_1 a_2^T a_1 \rightarrow a_2^T a_1 = 0$

k. PC: $Y_k = a_k^T X$ s max. $\text{Var}(Y_k)$ při $a_k^T a_k = 1$ a $\forall l < k: \text{cov}(Y_k, Y_l) = 0$

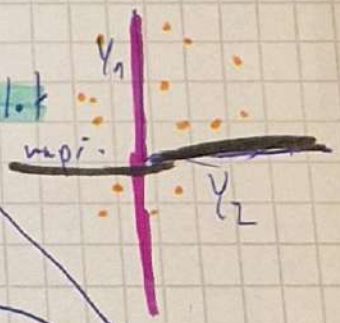




loadings
↳ loadings plot pro PC_i:



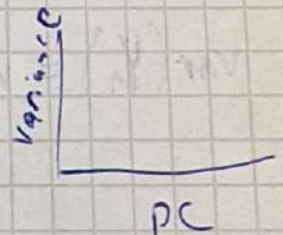
score plot

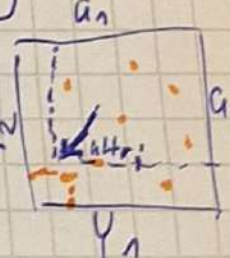


$\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$ ... kolik variance nese i-tá PC

↳ biplot

↳ Scree plot





MDS → vzdálenosti ⟿ souřadnice

Klasické – vzdálenosti jsou Euklidovské

matice vzdáleností D, souřadnice (neznámé) X ve sloupcích ↳ předp. centrované

$$B := X^T X$$

$$D^2_{ij} = (x_i - x_j)^T (x_i - x_j) = x_i^T x_i - x_j^T x_i - x_i^T x_j + x_j^T x_j =$$

$$= \langle x_i, x_i \rangle + \langle x_j, x_j \rangle - 2 \underbrace{x_i^T x_j}_{B_{ij}}$$

$$m := (\langle x_1, x_1 \rangle, \dots, \langle x_n, x_n \rangle)$$

$$D^2 = m \mathbb{1}_n^T + \mathbb{1}_n m^T - 2B$$

$C := I_n - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^T = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}$ .. centrovací matice

$$-\frac{1}{2} C D^2 C = -\frac{1}{2} C \left(m \mathbb{1}_n^T + \mathbb{1}_n m^T - 2B \right) C$$

$$= -\frac{1}{2} C m \underbrace{\mathbb{1}_n^T C}_{0} - \frac{1}{2} C \mathbb{1}_n \underbrace{m^T C}_{0} + CBC$$

$$\mathbb{1}_n^T C = \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \cdots \\ -\frac{1}{n} & \ddots & \\ \vdots & & \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots \end{pmatrix} = 0_n$$

$$\underbrace{1 - \frac{1}{n}}_{\frac{n-1}{n}} \quad \underbrace{-\frac{1}{n} - \frac{1}{n} \cdots}_{(n-1)\text{-krát}} = \frac{(n-1) - (n-1)}{n} = 0$$

$$= CBC = B$$ (předpokládali jsme už vycentrované)

$$X^T X = B = Q \Lambda Q^T = \underbrace{(Q \Lambda^{1/2})}_{X^T} \underbrace{(\Lambda^{1/2} Q^T)}_{X}$$

(vezmeme jen nenulové)

$$GOF = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

a z nich jen prvních m

⟹ **Alg:** 1) $D \to D^2$

2) $B := -\frac{1}{2} C D^2 C$

3) spektrální rozklad $B = Q \Lambda Q^T$

(ne)metrické – známe δ_{ij} ... dissimilarities

když už známe nějakou konfiguraci X – matici $n \times m$ (n = #bodů, m = dim. prostoru)

$$D_{ij}(X) := \sqrt{\sum_{a=1}^{m} (x_{ia} - x_{ja})^2}$$

↳ chyba je $\sigma^2(X) := \sum_{i<j} (D_{ij}(X) - \delta_{ij})^2$

lépe: raw stress $\sigma_r^2(X) := \sum_{i<j} w_{ij} (D_{ij}(X) - \delta_{ij})^2$

w_{ij} – váhy – kvůli chybějícím datům

normalizovaný stress $\sigma_n^2(X) := \frac{\sigma_r^2(X)}{\sum_{i<j} w_{ij} \delta_{ij}^2}$

Kruskalův stress $\sigma_1(X) := \sqrt{\sigma_n^2(X)}$

0 .. super

>0.2 .. BLE

Alg.: 1) X náhodně

2) Opakujeme, dokud se nám to nelíbí:

3) Spočítáme stress X

4) Upravíme X, abychom zmenšili stress

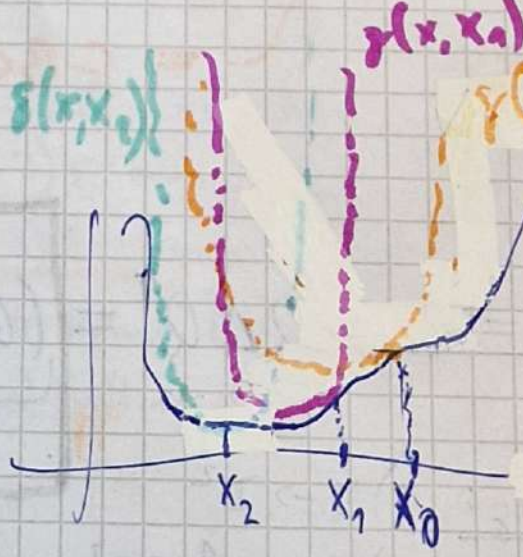
iterativní majorizace

$f(x) \rightsquigarrow g(x,z)$

majorizační funkce, dotýká se f v bodě z

$(f(z) = g(z,z))$,

$f(x) \leq g(x,z)$



1) $z \leftarrow z_0$

2) Najdeme x t.ž. $g(x,z) \leq g(z,z)$

3) P.K. $F(z) - F(x) < \varepsilon$ => konec

4) Jinak $z \leftarrow x$, a jdeme (2)

OP

$$\sigma(X) = \sum_{i<j} w_{ij} (D_{ij}(X) - \delta_{ij})^2 =$$

$$= \underbrace{\left(\sum_{i<j} w_{ij} \delta_{ij}^2\right)}_{\eta_\delta^2} + \underbrace{\left(\sum_{i<j} w_{ij} D_{ij}^2(X)\right)}_{\eta^2(X)} - 2 \underbrace{\left(\sum_{i<j} w_{ij} \delta_{ij} D_{ij}(X)\right)}_{\rho(X)} =: \tau(X, Z)$$

kvadr. majorizační funkce

$\eta^2(X)$ → $\mathrm{trace}(X^T V X)$ → $w_{ij} A_{ij}$ → $\begin{matrix} i & j \\ 1 & -1 \\ -1 & 1 \end{matrix}$ (i, j)

$\rho(X)$ → $\mathrm{trace}(X^T B(Z) Z)$ → $\frac{w_{ij} \delta_{ij}}{D_{ij}(Z)}$

SMACOF: Y ↘ Function convex

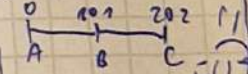
Může ale být potřeba transformace těch "vzdáleností" ($\delta_{ij}(p_{ij})$) vstupních dat (proximities) → disparities $\hat{D}_{ij}$

metrické: $\hat{D}_{ij} = f(p_{ij})$ → mapování: $b \cdot p_{ij}$ ... poměr., $a + b \cdot p_{ij}$ ... int., $a + b \cdot l_{ij}(p_{ij})$ ... int.

nemetrické: ordinální $p_{ij} < p_{jk} \Rightarrow \hat{D}_{ij} < \hat{D}_{jk}$

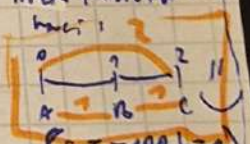
nominální: $p_{ij} = p_{jk} \Rightarrow \hat{D}_{ij} = \hat{D}_{jk}$

mapě: $|AB| = 101$, $|BC| = 101$, $|AC| = 102$ ⇒ udělá se



FUJ

ale když dovolíme int. transformaci:



(a = −100, b = 1)

disparity ($n \times (n-1)/2$)

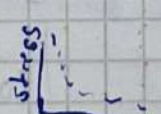
$$\sigma_r(\hat{D}, X) = \sum_{i<j} w_{ij} (D_{ij}(X) - \hat{D}_{ij})^2 = \eta_{\hat{D}}^2 + \eta^2(X) - 2\rho(\hat{D}, X)$$

musíme optimalizovat $\hat{D}$ i X

Configuration plot

- prostě ten výsledek – umístění bodů na ty souřadnice

Scree plot



Shepard plot

- u nemetrického MDS

Stress plot

Bubble plot

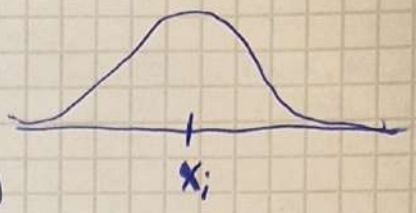
= configuration plot + ty body mají velikost podle toho, jaký přínos! stress

třeba stříbavě

t-SNE → pro vizualizace (ne redukci dim.)

nejdřív SNE:

$$p_{j|i} := \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$



Pravd., že si x_i vezme x_j za svého souseda

$$q_{i|j} := \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}$$

$\left(\sigma := \frac{1}{\sqrt{2}}\right)$

entropie distribuce p: $H(p) := -\sum_{i=1}^{N} p(x_i) \log p(x_i)$

↳ KL divergence → "vyjádření, jak moc se dvě distribuce liší"

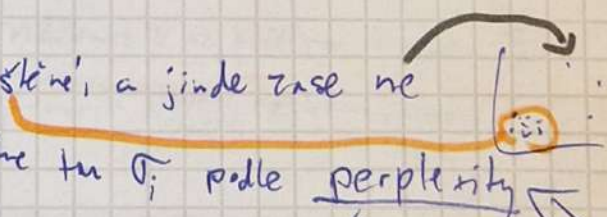
$$D_{KL}(p \| q) := \sum_{i=1}^{N} p(x_i)\left(\log p(x_i) - \log q(x_i)\right) = \sum_{i=1}^{N} p(x_i) \log \frac{p(x_i)}{q(x_i)}$$

↳ očekávané množství informace, kterou ztratíme, když použijeme q místo p

cost function: $C := \sum_a KL(P_a \| Q_a) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$

↳ pomocí gradient descent minimalizujeme – "pružiny"

ALE někde to může být malinkatý, a jinde zase ne



⇒ u každého bodu nastavíme ten σ_i podle perplexity

kolik přibližně má být bližší bod mít sousedů (~ 5–50)

PARAMETR

$$\underbrace{Perp(P_i)}_{\text{dáno}} = 2^{H(P_i)}$$

bin. vyhl.

hejbeme s σ_i, aby to myšlo

SNE → t-SNE

- zesymetrizováno: $q_{ij} = q_{ji} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq l} \exp\left(-\|y_k - y_l\|^2\right)}$

  → rychlejší konv., taktéž problémů s outliery

  ↖ ne jen vzdálenosti do i, ale všechny

- t-distrib. pro q ⇒ méně crowdingu + lépe se počítá

$$p_{ij} = p_{ji} = \frac{p_{j|i} + p_{i|j}}{2n}$$ (aby nebyl problém s outliery — široké)

UMAP

k-simplex ≡ k-dim. △ ⇝ simplexový komplex ≡ skupina simplexů, které dohýkají na nižší-dim. simplexy

Čechův
- ∀ bod → 0-simplex
- k+1 protínajících se koulí → k-simplex

Vietoris-Ripsův
- jednodušší –
z těch k+1 koulí se nemusí protínat všechny, ale stačí po párech

Pro ten poloměr koulí bude mít každý bod svou metriku, aby ta koule zasahovala ke k-tému nejbližšímu sousedovi (k-NN) — PARAMETR

→ ale NE - uděláme, že 1. NN tam bude patřit a ti ostatní postupně se snižující se pravděpod. + protože NN-vztah není symetrický, tak zesymetrizujeme (např. $P(A) \oplus P(B) := P(A) + P(B) - P(A) \cdot P(B)$)

- plně určen 0- a 1-simplexy
⇒ dobře reprezentovatelný jako graf

klika → spojení do jednoho simplexu

⇓

graf s hranami ohodnocenými pravděpodobnostmi

Alg: 1) ∀ bod si spočítáme k-NN, vzdálenost k nejbližšímu sousedovi ρ_i a σ_i tak aby

$$\sum_{j=1}^{k} \exp\left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right) = \log_2 k$$

aby to do 1. NN byl $e^0 = 1$

2) Sestrojíme graf $H = (X, E, w)$; E ... mezi vrcholem a jeho k-NN

$$w_h(x_i, x_j) := \exp\left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right)$$

3) Body rozmístíme do toho nízko-dim. prostoru (nějak přibližně)
⇒ na základě Euklid. vzdál. spočítáme váhy hran w_e (do PARAMETR min_dist — váha 1, pak klesá)

4) Minimalizujeme cross-entropy: → SGD + negative sampling

$$\sum_{e \in E} \underbrace{w_h(e) \log\left(\frac{w_h(e)}{w_e(e)}\right)}_{\text{přitažlivá síla pro body blízko ve vys. dim. prost.}} + \underbrace{(1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_e(e)}\right)}_{\text{odpudivá síla}}$$