Základy matematiky a statistiky pro humanitní obory II

Vojtěch Kovář

Fakulta informatiky, Masarykova univerzita
Botanická 68a, 602 00 Brno, Czech Republic
xkovar3@fi.muni.cz

část 8

Obsah přednášky

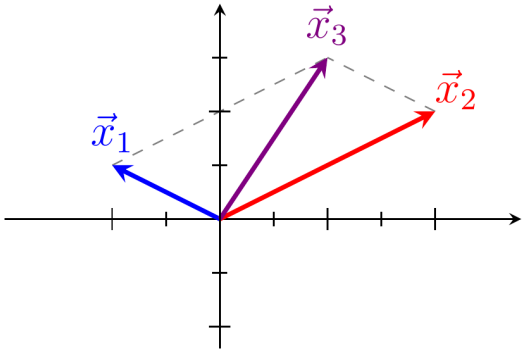
Vektorový prostor

- Základní pojem lineární algebry
- Abstrakce pro „šipky" nebo polynomy
  - „šipky" = veličiny, které mají velikost a směr
  - axiomatická definice
- Vektor
  - uspořádaná n-tice (reálných) čísel
  - n = dimenze vektorového prostoru
  - skalár = číslo
  - šipky → geometrická reprezentace (ve 2D, 3D)
- Operace nad vektory
  - sčítání (po složkách)
  - násobení vektoru skalárem
  - skalární součin dvou vektorů

Operace nad vektory

- Sčítání
  - $(a_1, a_2, a_3) + (b_1, b_2, b_3) = (a_1 + b_1, a_2 + b_2, a_3 + b_3)$
- Násobení vektoru skalárem
  - $c * (a_1, a_2, a_3) = (c * a_1, c * a_2, c * a_3)$
- Velikost vektoru
  - $|(a_1, a_2, a_3)| = \sqrt{{a_1}^2 + {a_2}^2 + {a_3}^2}$
- Skalární součin (dot product, inner product)
  - dvě ekvivalentní definice
  - $(a_1, a_2, a_3) * (b_1, b_2, b_3) = a_1 b_1 + a_2 b_2 + a_3 b_3$
  - $u * v = |u| * |v| * cos\Phi$
- Definovány pro libovolný počet rozměrů

Sčítání vektorů – geometrická reprezentace



zdroj: https://commons.wikimedia.org/wiki/File:Two_noncolinear_vectors_plus_addition_dotted.png

Matice

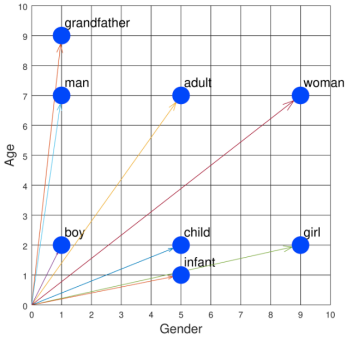
- Lineární zobrazení mezi vektorovými prostory
  - typ funkce nad vektory
  - vektor * matice = jiný vektor
  - „měníme obrázek vykreslený vektory"
- Uplatnění např. v kvantové mechanice

Cosine similarity (kosinová podobnost)

- Dva vektory spolu svírají úhel
  - kosinus tohoto úhlu je měřítko podobnosti vektorů
  - vektory směřující stejným směrem: 1
  - pravoúhlé (ortogonální) vektory: 0
  - vektory směřující opačným směrem: -1
- Výpočet (z definice skalárního součinu)
  - $u * v = |u| * |v| * cos\Phi$
  - $S_C = cos\Phi = \frac{u * v}{|u| * |v|}$

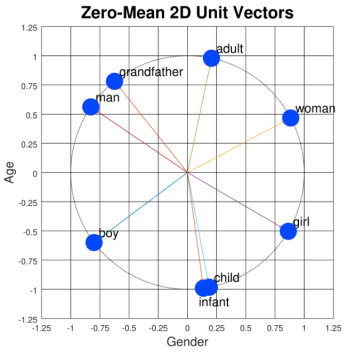
Word embeddings

Slova reprezentovaná jako vektory



Original Word Vectors			
grandfather	[	1,	9]
man	[	1,	7]
adult	[	5,	7]
woman	[	9,	7]
boy	[	1,	2]
child	[	5,	2]
girl	[	9,	2]
infant	[	5,	1]

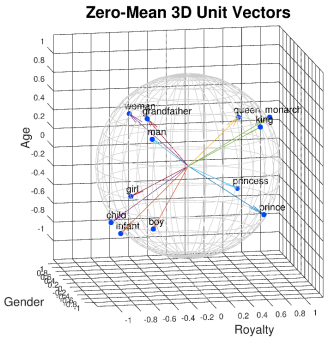
Word embeddings



Zero-Mean 2D Unit Vectors			
grandfather	[	-0.6247,	0.7809
man	[	-0.8275,	0.5615
adult	[	0.2060,	0.9785
woman	[	0.8844,	0.4668
boy	[	-0.8000,	-0.6000
child	[	0.1871,	-0.9823
girl	[	0.8638,	-0.5039
infant	[	0.1366,	-0.9906

Podobnost mezi slovy: cosine similarity

Word embeddings



Zero-Mean 3D Unit Vectors				
		Gender	Age	Royalty
grandfather	[	-0.5426,	0.6412,	-0.5426]
man	[	-0.7191,	0.4576,	-0.5230]
woman	[	0.7741,	0.4168,	-0.4764]
boy	[	-0.6971,	-0.5070,	-0.5070]
girl	[	0.7543,	-0.4642,	-0.4642]
king	[	-0.5570,	0.5064,	0.6583]
monarch	[	0.0676,	0.4730,	0.8785]
queen	[	0.6608,	0.3558,	0.6608]
prince	[	-0.5846,	-0.4252,	0.6909]
princess	[	0.6484,	-0.3990,	0.6484]
child	[	0.0733,	-0.5866,	-0.8066]
infant	[	0.0642,	-0.7057,	-0.7057]

Word embeddings – ale:

- Ne 3 dimenze, ale např. 300
  - přesto: představa slov/frází jako míst na povrchu zeměkoule je celkem blízko realitě
- Dimenze neodpovídají „hezkým" vlastnostem slov
  - jako např. věk, slovní druh, ...
  - nejsme schopni dimenze pojmenovat
  - ale jsou určeny kontexty
  - většina „hezkých" vlastností slov je v nich nějakým způsobem zakódována
- Jsme schopni s takto reprezentovanými slovy pěkně počítat
  - „king" - „man" + „woman" = nějaký vektor
  - jehož nejbližší soused je „queen"
- Počátek: word2vec, Tomáš Mikolov (bývalý student VUT Brno)