Kategorizácia textu

Márius Šajgalík

Úloha kategorizácie

- Učenie s učiteľom

- Máme objekty zaradené do niekoľkých tried
- Všímame si **črty**, ktoré ich diskriminujú

- Samotná kategorizácia nových objektov na základe čŕt

Klasifikátory

- Naivný Bayes
  - na základe čŕt, učíme sa pravdepodobnosť priradenia objektu do každej z tried
- LDA (Linear discriminant analysis)
  - učíme sa váhy čŕt - lineárne kombinácie, ktoré charakterizujú jednotlivé triedy
- SVM (Podporný vektorový stroj)
  - hľadáme najlepšiu oddeľujúcu priamku/rovinu
  - viaceré typy, lineárny SVM je veľmi rýchly

Kategorizácia textu

- Tradične - slová sú črty
- Viaceré štatistické metriky, väčšinou odvodené od TF-IDF
- Zameriavajú sa na výpočet diskriminačných váh

Frekvenčná tabuľka

	Frekvencia slova W	Frekvencia ostatných slov
Frekvencia v kategórií C	A	B
Frekvencia v ostatných kategóriách	C	D

$$N = A + B + C + D$$

Štatistické metriky

IDF = log(N/(A+C))

RF = log(2+A/max(C,1))

TDS = A/(A+B) / ((A+C)/N)

IG = N *

A / N * log(A*N) / ((A+C)*(A+B)) +

B / N * log(B*N) / ((B+D)*(A+B)) +

C / N * log(C*N) / ((A+C)*(C+D)) +

D / N * log(D*N) / ((B+D)*(C+D))

Nevýhody

- Zbytočne veľa čŕt - celý slovník
- Nemáme vzťahy medzi črtami

Riešenie

- Zapojenie ontológií
  - prepojíme príbuzné slová
  - Problém: ktoré sú to tie príbuzné slová?

Riešenie

- Zapojenie ontológií
  - prepojíme príbuzné slová
  - Problém: ktoré sú to tie príbuzné slová?
- Využijeme vektory čŕt
  - Podobnosť na základe používaného kontextu
  - Podobné slová sú blízko seba vo vektorov priestore čŕt
  - Zároveň máme oveľa menej čŕt na trénovanie

Využitie kategorizácie

- Veľa metód sa snaží vypočítať charakteristické vlastnosti objektov
- Vieme využiť diskriminačnosť kategórií na vyhodnotenie danej metódy

Príklad č. 1

- Moja dizertačka
- Výpočet kľúčových slov
  - tradičný problém - čo sú to kľúčové slová?

If the implementation is hard to explain, it's a bad idea.

PEP 20, Verse 17

Príklad č. 1

- Moja dizertačka
- Výpočet kľúčových slov
  - kľúčové slová **diskriminujú** dokument, resp. kategóriu, do ktorej patrí

Príklad č. 1

- Moja dizertačka
- Výpočet kľúčových slov
  - kľúčové slová diskriminujú dokument, resp. kategóriu, do ktorej patrí
- Modelovanie záujmov používateľa
  - snažím sa charakterizovať záujmy používateľa slovami
  - záujmy charakterizujú - odlišujú daného používateľa

Príklad č. 2

- Hlboké neurónové siete
- Máme veľa vrstiev - môžeme mať aj viac výstupov
- Snažíme sa, aby niektoré vrstvy boli diskriminačné
  - na výstup pripojíme napr. SVM