

# 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 7

# Obsah přednášky

- 1 Entropie
- 2 Perplexita
- 3 Vyhledávání kolokací
- 4 Vyhodnocování úspěšnosti
- 5 Lingvistická anotace

# Entropie náhodné veličiny

## ■ Míra informace náhodné veličiny

- kolik informace získáme, když se dozvíme hodnotu náhodné veličiny
- „hodnota informace“, kterou nám veličina dává
- měří se v bitech
- nulová entropie = jsme schopni určit hodnotu veličiny se 100% jistotou

## ■ Počátky

- 40. léta (Shannon)
- potřeba přenést informaci co nejmenší možnou zprávou

# Entropie

## ■ Vzorec

- $H(p) = H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$
- $X$  = množina možných hodnot
- $p$  = pravděpodobnostní rozložení

## ■ Příklad – hod dvěma mincemi, počítáme panny

- $p(0) = 1/4$ ,  $p(1) = 1/2$ ,  $p(2) = 1/4$
- $H(p) = -(1/4 \log_2(1/4) + 1/2 \log_2(1/2)) + 1/4 \log_2(1/4) = -(-2/4 - 1/2 - 2/4) = 1.5$  bitu

## ■ Pokud budou na obou mincích padat pouze panny

- $p(0) = 0$ ,  $p(1) = 0$ ,  $p(2) = 1$
- $H(p) = -(\log_2(1)) = -(0) = 0$
- $\rightarrow$  nemusíme předávat žádnou informaci, abychom zjistili, že padly dvě panny

# Podmíněná entropie

## ■ Podobně jako podmíněná pravděpodobnost

- $H(X|Y)$  – entropie veličiny  $X$  za předpokladu, že známe hodnoty veličiny  $Y$

- $H(p) = H(X|Y) = \sum_{x \in X} p(x)H(Y|X = x)$

## ■ Řetízkové pravidlo (chain rule)

- $H(X, Y) = H(X) + H(Y|X)$

# Perplexita

- $2^H$ 
  - stejná(podobná) informace, jiné číslo
  - „počet možností, které mohou nastat“, pokud by jejich rozložení bylo rovnoměrné
- Měřitko kvality jazykových modelů
  - menší perplexita  $\Rightarrow$  lepší model
  - (ale vždy to tak nemusí být)

# Vyhledávání kolokací

## ■ Kolokace

- různé definice
- fráze, jejíž význam se neskládá z významů jejích částí
- nějakým způsobem „významné“ spojení dvou slov
- např. idiomy, ale nejen
- základní škola, silný čaj, ...

## ■ Jakým způsobem vyhledat v korpusu kolokace?

- případně statisticky určit „sílu“ libovolné kolokace na základě dat?
- odlišit „strong tea“ od „powerful tea“

# Jakým způsobem vyhledat v korpusu kolokace?

- Prosté frekvence sekvencí slov v korpusu?
  - → „of the”, „in the”, ...
- Frekvence filtrovaných sekvencí slov?
  - na základě slovních druhů jednotlivých slov
  - → „New York”, „United States”, ...
  - ale třeba i „last week”
- T-test
  - aplikace testování hypotéz
  - předpokládáme, že se slova chovají standardně (nulová hypotéza) = podle svých obvyklých pravděpodobnostních rozložení
  - vyvrácení nulové hypotézy = kolokace
  - problém: při dostatečně velkých datech je nulová hypotéza vyvrácena téměř vždy



# Mutual information (vzájemná informace)

- Míra informace, kterou jedna náhodná proměnná říká o jiné
  - vzorec:  $MI(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
  - 0, pokud jsou veličiny nezávislé
  - čím vyšší, tím více hodnoty jedné vlastnosti určují hodnoty druhé vlastnosti
- Příklad použití – kolokace
  - X: výskyt slova  $a$  (např. „základní“) v textu
  - Y: výskyt slova  $b$  (např. „škola“) v textu
  - MI je měřítkem „síly“ kolokace těchto dvou slov
  - je tím vyšší, čím vyšší je počet souvýskytů slov a tím nižší, čím jsou slova častější

# LogDice

## ■ Dice

- $2f_{AB}/(f_A + f_B)$
- jednoduché, nezávislé na velikosti korpusu
- ale čísla, která z toho lezou, nejsou hezká

## ■ logDice

- $14 + \log_2 Dice$
- většinou mezi 0 a 10
- +1 znamená dvojnásobný počet souvýskytů
- 0 je méně než 1 souvýskyt na 16 tisíc výskytů

# Vyhodnocování úspěšnosti NLP nástrojů

- Gold standard (grand truth)
  - správně označovaná data, většinou manuálně
  - považujeme je za správná (ale i lidé dělají chyby)
  - porovnáváme výstup nástroje s gold standardem
- Příklad: klasifikace diskusních příspěvků
  - 3 třídy: pozitivní, negativní, neutrální

## Confusion matrix (matice záměn)

Příklad: V gold standard datech je 100 příspěvků z každé třídy. Řádky říkají, jak dokumenty z těchto tříd klasifikoval testovaný systém.

	pozitivní	neutrální	negativní
pozitivní	65	5	5
neutrální	30	90	5
negativní	5	5	90

- větší čísla na diagonále = lepší
- dobře znázorňuje nejčastější typy chyb
- (systém z příkladu je poměrně úspěšný, ale často označuje pozitivní příspěvky jako neutrální)

# Klasifikace do dvou tříd

	pozitivní	negativní
pozitivní	<i>true positives (TP)</i>	<i>false positives (FP)</i>
negativní	<i>false negatives (FN)</i>	<i>true negatives (TN)</i>

- **precision/přesnost:**  $TP / (TP + FP)$
- **recall/pokrytí:**  $TP / (TP + FN)$
- **F-score:**  $2 * precision * recall / (precision + recall)$
- **accuracy/přesnost:**  $(TP + TN) / (TP + FP + TN + FN)$

v diagnostických testech (medicína apod.):

- **sensitivity:**  $TP / (TP + FN)$  (totéž, co recall)
- **specificity:**  $TN / (FP + TN)$

# Lingvistická anotace dat

- Např. při vytváření gold standardu pro vyhodnocování
- Základní parametr: mezinotátorská shoda
  - inter-annotator agreement
  - podobné: intra-annotator agreement
  - protože člověku se zdá, že je to jasné, ale není
  - anotátoři se také mohou shodnout náhodou – tzv. chance agreement
- Často existuje velká šedá zóna
  - kde lidé neví, nebo se neshodnou z principiálních důvodů
  - můžeme se rozhodnout šedou zónu do gold standardu nezahrnout

# Mezianotátorská shoda

## ■ Základní verze

- počet shod / celková velikost dat
- např. 100 komentářů, u 80 z nich se anotátoři shodli  
→ 80 %
- dále  $A_o$  (observed agreement)

## ■ Lépe: nechceme počítat náhodné shody

- $A_e$  (expected agreement)
- $\frac{A_o - A_e}{1 - A_e}$
- Cohen's kappa (a další)