

Základy matematiky a statistiky pro humanitní obory

II

Pavel Rychlý Vojtěch Kovář

Fakulta informatiky, Masarykova univerzita
Botanická 68a, 602 00 Brno, Czech Republic

{pary, xkovar3}@fi.muni.cz

30. 3. 2011

Obsah přednášky

- 1 Podmíněná pravděpodobnost
- 2 Nezávislé jevy
- 3 Bayesův vzorec
- 4 „Paradoxy“: Aplikace Bayesova vzorce

Podmíněná pravděpodobnost

- Pravděpodobnost jevu A , za předpokladu, že nastal jev B
 - značíme $P(A|B)$
 - např. pravděpodobnost deště zítra v poledne za předpokladu deště dnes v poledne
 - např. pravděpodobnost, že součet dvou hodů kostkou bude 8, pokud první výsledek byl 3
 - např. pravděpodobnost deště zítra v poledne za předpokladu, že dnes skončíme o 10 minut dřív
 - např. pravděpodobnost, že člověk je bezdomovec, pokud má vousy delší než 5 cm
 - → jevy A a B mohou, ale nemusí mít kauzální souvislost



Podmíněná pravděpodobnost

- Výpočet podmíněné pravděpodobnosti
 - $P(A|B) = P(A, B)/P(B)$
 - kde $P(A, B)$ je pravděpodobnost, že jevy A a B nastanou současně

Nezávislé jevy

- Jevy A a B jsou nezávislé, pokud
 - to, jestli nastal jev B, neovlivní pravděpodobnost jevu A
 - $P(A|B) = P(A)$
- Pro nezávislé jevy platí
 - $P(A, B) = P(A) * P(B)$
 - pozor: platí **pouze** pro nezávislé jevy
- Reálné jevy nebývají téměř nikdy dokonale nezávislé

Bayesův vzorec

■ Převod mezi podmíněnými pravděpodobnostmi

- $$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

■ Důkaz

- $$P(A|B) = P(A, B) / P(B)$$

- $$P(B|A) = P(B, A) / P(A)$$

- $$P(A|B) * P(B) = P(A, B) = P(B|A) * P(A)$$

„Paradoxy“: Aplikace Bayesova vzorce – I

- Mějme následující soutěž
 - troje dveře, za jedněmi z nich je výhra
 - moderátor, který ví, kde je výhra
 - vybereme si dveře 1
 - moderátor soutěže otevře dveře 3
 - za nimi výhra není
 - nyní máme možnost svou volbu změnit
- Vyplatí se změnit volbu a vybrat dveře 2?



„Paradoxy“: Aplikace Bayesova vzorce – I (2)

- Označme si události následovně
 - V_1, V_2, V_3 : výhra je za dveřmi 1, 2 nebo 3
 - X : moderátor otevřel dveře 3
 - (předpokládáme, že v případě, že výhra je za dveřmi, které jsme si vybrali, se moderátor rozhoduje náhodně)
- Vyjádřeme pravděpodobnosti
 - $P(V_1) = P(V_2) = P(V_3) = 1/3$
 - $P(X|V_1) = 1/2$
 - (vybrali jsme správně, moderátor rozhoduje náhodně)
 - $P(X|V_2) = 1$
 - (vybrali jsme špatně, moderátor má jedinou možnost)
 - $P(X|V_3) = 0$
 - (moderátor nevybere dveře s cenou)



Příklad 1

„Paradoxy“: Aplikace Bayesova vzorce – I (3)

- Spočteme podmíněné pravděpodobnosti pro událost X

- $P(V_1|X) = \frac{P(X|V_1)*P(V_1)}{P(X)} = \frac{\frac{1}{2}*\frac{1}{3}}{\frac{1}{2}} = 1/3$

- $P(V_2|X) = \frac{P(X|V_2)*P(V_2)}{P(X)} = \frac{1*\frac{1}{3}}{\frac{1}{2}} = 2/3$

- $P(V_3|X) = \frac{P(X|V_3)*P(V_3)}{P(X)} = \frac{0*\frac{1}{3}}{\frac{1}{2}} = 0$

- Jak to?

- otevření dveří moderátorem ve 2/3 případů určí správné dveře
 - (ve 2/3 případů si vybereme na začátku špatně)
 - představme si variantu hry, kdy máme 1000 dveří a moderátor otevírá 998

Zajímavosti

- Pouze 13 % lidí změní svou původní volbu
 - a při opakování pokusu se chovají stále stejně
- Obdobný pokus s holuby
 - holubi se během 30 dní naučili téměř vždy změnit původní volbu
- (zdroj a více informací viz Wikipedia: Monty Hall problem)

„Paradoxy“: Aplikace Bayesova vzorce – II

- Testování drog mezi zaměstnanci
- Mějme k dispozici test, který odhalí požití drogy na 99 %
 - je pozitivní v 99 % případů, kdy zkoumaný požil drogu
 - je negativní v 99 % případů, kdy zkoumaný nepožil drogu
- Dále dejme tomu, že 0,5 % zaměstnanců skutečně požilo drogu
- Záměr vedení firmy
 - otestovat všechny zaměstnance
 - propustit ty, kteří budou mít pozitivní test
- Je tento záměr správný?
- Kolik procent propuštěných bude propuštěno neoprávněně?

„Paradoxy“: Aplikace Bayesova vzorce – II (2)

■ Označme události

- D : testovaný zaměstnanec požil drogu
- N : testovaný zaměstnanec nepožil drogu
- pos : test zaměstnance je pozitivní
- neg : test zaměstnance je negativní

■ Vyjádřeme známé pravděpodobnosti

- $P(D) = 0,005$
- $P(N) = 0,995$
- $P(pos|D) = 0,99$ („true positive“)
- $P(pos|N) = 0,01$ („false positive“)
- $P(pos) = P(truepositive) + P(falsepositive) =$
 $0,99 * 0,005 + 0,01 * 0,995 = 0,0149$

„Paradoxy“: Aplikace Bayesova vzorce – II (3)

- Chceme zjistit $P(D|pos)$
 - pravděpodobnost, že zaměstnanec požil drogu za předpokladu, že má pozitivní test
 - $P(D|pos) = \frac{P(pos|D)*P(D)}{P(pos)}$
 - $= \frac{0,99*0,005}{0,0149}$

„Paradoxy“: Aplikace Bayesova vzorce – II (3)

- Chceme zjistit $P(D|pos)$
 - pravděpodobnost, že zaměstnanec požil drogu za předpokladu, že má pozitivní test
 - $P(D|pos) = \frac{P(pos|D)*P(D)}{P(pos)}$
 - $= \frac{0,99*0,005}{0,0149}$
- $P(D|pos) = 0,3322$
- Kde je problém?
 - testy s úspěšností 99 % jsou relativně časté



„Paradoxy“: Aplikace Bayesova vzorce – III

■ Morfologické značkování nejednoznačných slov

- např. „jak“
- 80 % výskytů v textu je spojka
- 20 % výskytů v textu je podstatné jméno

■ Cíl

- chceme maximalizovat podíl správně označkových výskytů
- bez dalších informací (např. o kontextu)

■ Otázky

- jaký je optimální postup?
- jaké úspěšnosti značkování lze takto dosáhnout?

Závěry

- Přemýšlejme nad čísly a nad tím, co znamenají
 - i 99 % může být hodně málo
- V jednoduchosti je síla
 - i zdánlivě hloupý postup může být optimální
 - je třeba domýšlet věci do důsledků