PLIN037 Sémantika a počítače

OP VK Mezi bohemistikou a informatikou
www.projekt-inova.cz

Zuzana Nevěřilová
xpopelk@fi.muni.cz

Centrum zpracování přirozeného jazyka, B203
Fakulta informatiky, Masarykova univerzita

7. dubna 2016

Klasifikace parafrází

- obtížný úkol, se kterým se každý vyrovná po svém ($\rightarrow$ nízká mezianotátorská shoda)
- anotační manuál (který buď nezachytí všechny případy, nebo ho nikdo nebude číst)
- řešení neshody (třetí anotátor)
- řešení náhodné shody? (výpočet Cohen κ nebo Fleiss κ)

Klasifikace parafrází

- obtížný úkol, se kterým se každý vyrovná po svém ($\rightarrow$ nízká mezianotátorská shoda)
- anotační manuál (který buď nezachytí všechny případy, nebo ho nikdo nebude číst)
- řešení neshody (třetí anotátor)
- řešení náhodné shody? (výpočet Cohen κ nebo Fleiss κ)

Lehký úvod do strojového učení

učení = změna stavu na základě vnějších podnětů

strojové učení:

1) máme program (který se už nebude měnit), 2) máme vstupní (trénovací) data

výsledek: 3) program se z trénovacích dat naučí klasifikovat libovolná další data

Aby bylo strojové učení užitečné, musí mít vysokou přesnost:

- velké množství trénovacích dat
- velká rozmanitost trénovacích dat
- vhodně definovaná klasifikační úloha
- vhodný algoritmus strojového učení (podle povahy dat)

Lehký úvod do strojového učení

Velmi důležité je znát data:

- velikost datového souboru
- způsob pořízení dat (kvůli možných chybám)
- původce dat
- distribuce sledovaného jevu

(tady se opravdu hodí statistika)

Lehký úvod do strojového učení

známe-li data, můžeme se pustit do strojového učení

Základní techniky:

- rozhodovací stromy, rozhodovací seznamy
- vzdálenosti
- naivní Bayesovský klasifikátor
- k-NN (nearest neighbors), klastrování
- neuronové sítě

Vzdálenost (distance) a metrika (metric)

Vzdálenost D je funkce definovaná na kartézském součinu $X \times X$ s nezápornými hodnotami.

$$D : X \times X \to \mathcal{R}$$

$$\forall x, y : D(x, y) \geq 0$$

Vzdálenost (distance) a metrika (metric)

Vzdálenost D je funkce definovaná na kartézském součinu $X \times X$ s nezápornými hodnotami.

$$D : X \times X \to \mathcal{R}$$

$$\forall x, y : D(x, y) \geq 0$$

Vzdálenost je metrika, pokud:

- $D(x, y) = 0 \Leftrightarrow x = y$ (identita)
- $D(x, y) + D(y, z) \geq D(x, y)$ (trojúhelníková nerovnost)
- $D(x, y) = D(y, x)$ (symetrie)

Vzdálenost (distance) a metrika (metric)

Vzdálenost D je funkce definovaná na kartézském součinu $X \times X$ s nezápornými hodnotami.

$$D : X \times X \to \mathcal{R}$$

$$\forall x, y : D(x, y) \geq 0$$

Vzdálenost je metrika, pokud:

- $D(x, y) = 0 \Leftrightarrow x = y$ (identita)
- $D(x, y) + D(y, z) \geq D(x, y)$ (trojúhelníková nerovnost)
- $D(x, y) = D(y, x)$ (symetrie)

Množinu X nazýváme metrický prostor.

Vzdálenost mezi body

Triviální diskrétní metrika:

- $D(x, y) = 0 \Leftrightarrow x = y$
- $D(x, y) = 1 \Leftrightarrow x \neq y$

Vzdálenost mezi body

Euklidovská vzdálenost:

- $D(x, y) = \sqrt{(x - y)^2}$ – jednorozměrná
- $D(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$,
  $p = (p_1, p_2), q = (q_1, q_2)$ – dvourozměrná
- $D(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$,
  $p = (p_1, p_2, p_3), q = (q_1, q_2, q_3)$ – trojrozměrná
- $D(p, q) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_n - q_n)^2}$,
  $p = (p_1, \ldots, p_n), q = (q_1, \ldots, q_n)$ – n-rozměrná

Čtverec euklidovské vzdálenosti: $D^2(x, y) = (x - y)^2$

Vzdálenost mezi body

Euklidovská vzdálenost:

- $D(x, y) = \sqrt{(x - y)^2}$ – jednorozměrná
- $D(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$,
  $p = (p_1, p_2), q = (q_1, q_2)$ – dvourozměrná
- $D(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$,
  $p = (p_1, p_2, p_3), q = (q_1, q_2, q_3)$ – trojrozměrná
- $D(p, q) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_n - q_n)^2}$,
  $p = (p_1, \ldots, p_n), q = (q_1, \ldots, q_n)$ – n-rozměrná

Čtverec euklidovské vzdálenosti: $D^2(x, y) = (x - y)^2$

Manhattanská vzdálenost

- $D(p, q) = |p_1 - q_1| + |p_2 - q_2|, p = (p_1, p_2), q = (q_1, q_2)$

Vzdálenost mezi vektory

Kosinová vzdálenost:

$$cos(\theta) = \frac{A \cdot B}{|A||B|}$$

Vzdálenost mezi vektory

Kosinová vzdálenost:

$$cos(\theta) = \frac{A \cdot B}{|A||B|}$$

$$A = (a_1, a_2), B = (b_1, b_2)$$

$$cos(\theta) = \frac{a_1 b_1 + a_2 b_2}{\sqrt{a_1^2 + a_2^2}\sqrt{b_1^2 + b_2^2}}$$

Vzdálenosti mezi řetězci

Levenshteinova vzdálenost a její varianty

(nejmenší) počet operací (transpozic, transposition), které řetězec s_1 převedou na řetězec s_2

Vzdálenosti mezi řetězci

Levenshteinova vzdálenost a její varianty

(nejmenší) počet operací (transpozic, transposition), které řetězec s_1 převedou na řetězec s_2

Hammingova vzdálenost

počet pozic, které jsou mezi dvěma řetězci rozdílné (non-matching characters)

Vzdálenosti mezi řetězci

Levenshteinova vzdálenost a její varianty

(nejmenší) počet operací (transpozic, transposition), které řetězec s_1 převedou na řetězec s_2

Hammingova vzdálenost

počet pozic, které jsou mezi dvěma řetězci rozdílné (non-matching characters)

Jaro-Winkler, Soundex (založen na hláskové podobnosti)

Vzdálenosti mezi množinami

Jaccardův koeficient: $Q = \frac{2|A \cap B|}{|A \cup B|}$

Sørensenův–Diceův koeficient: $Q = \frac{2|A \cap B|}{|A| + |B|}$

Editační vzdálenost stromů (tree edit distance)

operace (podobné jako u L. vzdálenosti)

- přidat uzel u
- smazat uzel u (a připojit jeho potomky k rodiči u)
- přejmenovat uzel

Editační vzdálenost stromů (tree edit distance)

operace (podobné jako u L. vzdálenosti)

- přidat uzel u
- smazat uzel u (a připojit jeho potomky k rodiči u)
- přejmenovat uzel

řádově těžší úloha (použití rekurze)

Podobnost

Podobnost je (nějakým způsobem) převrácená hodnota vzdálenosti.