1

Bias je

- * libovolná omezující podmínka (jiná než konsistence s trénovacími daty), která je užita pro výběr hypotézy.
- konsistence s trénovacími daty.
- libovolná omezující podmínka, která je užita pro výběr hypotézy.

Data obsahují atributy Barva:{red, blue, green}, Tvar:{circle,triangle,rectangle}, Velikost:{small, large} a Obsah (reálné kladné číslo).

Pak pro dvě hypotézy H1: red & circle a H2:(red & circle) or (Obsah < 10) platí, že

- * H2 je obecnější než H1
- H1 a H2 nejsou v relaci "býti obecnější"
- H1 je obecnější než H2

Pro dvě hypotézy h1 a h2 platí, že h1 je obecnější nebo stejně obecná jako h2, tedy (h1 >= h2)

- právě když každá instance, která splňuje h1 také splňuje h2.
- právě když existuje instance, která splňuje h2 a nesplňuje h1.
- *právě když každá instance, která splňuje h2 také splňuje h1.

Jedna z hypotéz, na níž je postaveno učení s učitelem (supervised learning) říká následující (vyberte variantu, které je nejblíž pravdě):

- * libovolná funkce, která dobře aproximuje cílový koncept (hypotézu) na dostatečně velké množině trénovacích příkladů bude také dobře aproximovat cílovou funkci na testovací množině za předpokladu stejného pravděpodobnostního rozložení.
- libovolná funkce, která dobře aproximuje cílový koncept (hypotézu) na dostatečně velké množině trénovacích příkladů bude také dobře aproximovat cílovou funkci na datech z jiného praděpodobnostního rozložení..
- libovolná funkce, která dobře aproximuje cílový koncept (hypotézu) na malé množině trénovacích příkladů bude též dobře aproximovat cílovou funkci na větší množině příkladů..

V které situaci je celková správnost (accuracy) nejméně vhodná?

- Datová sada se skládá z mnoha tříd s přibližně stejným počtem příkladů v každé třídě

- Datová sada se skládá ze dvou tříd.
- * Počet příkladů v jednotlivých třídách se výrazně liší

2

Supervised anomaly detection: pro testování těchto metod potřebujeme, příklady, u nichž víme, zda jsou anomální nebo normální. Můžeme použít

- * generátor umělých dat (tzv. ground truth)
- Asociační pravidla
- Náhodně vygenerovaná data i třídy

Semi-supervised anomaly detection: pro testování těchto metod potřebujeme příklady, u nichž víme, zda jsou anomální nebo normální. Můžeme použít

- náhodně vygenerovaná data i třídy
- * dvoutřídní data např. z kaggle
- časté vzory

Mějme čtyři dokumenty

1. She sat down.
2. She drank coffee.
3. She spent much time in learning text mining.
4. She invested significant efforts in learning text mining.

a převrácenou hodnotu eukleidovské vzdálenosti jako míru podobnosti. Po převedení na bag-of-words (reprezentace TF, term frequency) bude větě "She drank coffee." nejpodobnější

- * She sat down
- She invested significant efforts in learning text mining.
- She spent much time in learning text mining.

Klasifikace dokumentů: Pro úpravu textového dokumentu jsme použili stemming a vytvořili jsme dokument-term matici s hodnotami TF-IDF. Bez stemmingu

- * můžeme získat lepší výsledek (např. accuracy)
- nemůžeme získat lepší výsledek (např. accuracy)
- získáme vždy horší výsledek (např. accuracy)

3

Který způsob prohledávání prostoru hypotéz používají algoritmy pro učení rozhodovacích stromů, např. C4.5 nebo CART?

- * hill-climbing (greedy search, heuristické prohledávání)
- beam search
- depth-first search (prohledávání do hloubky)

Učení rozhodovacího stromu: Mějme data obsahující dvě třídy. Kdy bude informační zisk nejmenší?

- *Když poměr pozitivních a negativních příkladů na jednotlivých cestách (tj. po použití atributu) je podobný jako před použitím.
- Když po použití atributu jedna cesta klasifikuje jen negativní příklady.
- Když po použití atributu jedna cesta klasifikuje jen pozitivní příklady.

Mějme syntaktický strom vytvořený mělkým syntaktickým analyzátorem. Je-li v kořeni stromu (usel hloubky 0) věta, která neobsahuje předložkové fráze, potom mělký syntaktický strom má hloubku

- * 1
- 2
- 0

Pro daný dataset a daný support>0 a confidence>0 asociační pravidla pokrývají (tj. pro příklad z datasetu existuje pravidlo, které pro daný příklad platí)

- * jen část datasetu
- Všechny příklady z datasetu
- N/support příkladů z datasetu, kde N je počet příkladů v datasetu

Vyberte variantu, která je nejblíž pravdě.
Častý vzor (frequent pattern, large itemset)

- se dá použít pro klasifikaci, pokud pravá strana má určitý tvar.
- se nikdy nedá použít pro klasifikaci.
- *se vždycky dá použít pro klasifikaci.

Zipfův zákon říká, že pokud nejčastější slovo má rank 1, frekvence F libovolného slova je

- 1/e^F
- log(F)
- *1/F

4

Tokenizujte (převeďte na tokeny) následující větu

after reading for 3 h, he decided to read for another two.

Předpokládejte, že členy, osobní zájmena, číslovky a předložky jsou stop slova a odstraňte je.

Proveďte stemming a napište výsledek.

- read h , decid read another

Předpokládejte kolekci 1000 dokumentů, kde se každý tvar slovesa a slova after a another vyskytují vždy ve 100 dokumentech.

Všechna ostatní slova se vyskytují v 10 dokumentech.

Napište tf-idf reprezentaci vaší předchozí odpovědi. Logaritmus nemusíte počítat. Chcete-li však, použijte se základem 10.

- read h , decid read another
- ~ read decid another h ,
- [0.02, 0.01, 0.1, 0.1, 0.1]

Napište větu, bez ohledu na sémantiku, avšak syntakticky správně utvořenou

1. která bude obsahovat jiná slovesa a předložky než věta z první otázky
2. a bude mít stejnou kombinaci (tj. nezáleží na pořadí) nenulových hodnot tf-idf jako vaše odpověď na Otázku 2.

- Before swimming within 20 km, she took another swim with boys 2

5

Uveďte aspoň dva příklady desambiguačních úloh ve zpracování přirozeného jazyka.

- Word sense desambiguation:
- Například v NER, ke slovo Apple může představovat několik různých entit — název firmy, ovoce, jméno osoby.
- Druhak v sentiment analysis, kdy slovo s negativním sentimentem je ovlivněno (i opozitně) kvůli slovům v okolí.
- Výběr klasifikace textového objektu z několika možností podle kontextu textového objektu.

Vyberte odpověď, která se nejvíc blíží pravdě.

Při desambiguačních úlohách generujeme učící příklad

- * z pravého a levého kontextu desambiguavaného slova
- z levého kontexu desambiguavaného slova

- z celého dokumentu

Při analýze sentimentu definujeme sentiment/opinion jako pětici.

Uveďte aspoň tři prvky této pětice.

- Nech jsou věci označeny vzestupnými indexy:
  - (o_i, f_ik, s_iklq, h_l, t_q), kde
  - o_i je target,
  - f_ik k-tá featura i,
  - s_iklq je sentiment fatury k o objektu i v čase q od protagonisty l,
  - h_l je l-tý protagonista
  - t_q je q-tý čas, vyjadřující, kdy byl názor vyjádřen

6

Extrakce informace z textu. Uveďte dva příklady obvyklých a možných extrahovaných relací (Relation extraction) a ke každé příklad,

např. Relace = býti králem(), příklad = býti králem(Lávra).

- Relace = být podřízený, příklad = být podřízený(rektor, docent);
- Relace = interaguje, příklad = interaguje(peptin, škrob);
  - být podřízený(rektor, docent): raději být podřízený(Bareš, Bouda),
  - rektor a docent vidím spíš jako typy pojmenovaných entit

Extrakce informace z textu: Uveďte dva typy obvyklých pojmenovaných entit a pro každý jednu konkrétní entitu (příklad),

např. Entita=král Příklad Král Lávra.

- entita: místo; příklad: Brno
- entita: jméno; příklad: Lubomír
- entita: čas; příklad: odpoledne
- entita: organizace; příklad: Apple
- entita: stát; příklad: Česká Republika

Učící příklady v úloze extrakce pojmenovaných entit (Name Entity Recognition) z textu mají nejčastěji tvar

- * levý kontext, pojmenovaná entita, pravý kontext, typ pojmenované entity
- matice dokument-term
- pojmenovaná entita, typ pojmenované entity

Extrakce informace z textu tak, jak jsme o ní mluvili, je z pohledu strojového učení

- * supervised learning

- unsupervised learning
- first order clustering

7

Pro redukci sloupců v matici document-term se používají metody feature selection,

které používají pro ohodnocení atributů rankovací funkci. Uveďte jednu takovou (stručně, názvem nebo jednou větou)

- Jako rankovací fci můžeme použít tyto metriky:
  - F1 míra (pro nevyrnané třídy i s jinou konstantou),
  - Chi metrika, či
  - Information gain.

Pro výběr rozumného počtu sloupců v matici document-term se hodí učící křivka,

kde na ose X je počet atributů a na ose Y accuracy. Která z odpovědí nejlépe popisuje její tvar? Pro redukci atributů jsme použili stemming.

- * křivka je rostoucí, 1. derivace klesající
- křivka je rostoucí, 1. derivace rostoucí
- křivka je klesající

Po převodu kolekce dokumentů na document-term matrix a po odstranění stop-slov bývá počet sloupců

nejčastěji

- * v řádu tisíců
- v řádu desítek
- v řádu stovek

8

Unsupervised anomaly detection: pro testování těchto metod potřebujeme příklady, u nichž víme, zda jsou anomální nebo normální. Můžeme použít

(vyberte nejlepší variantu)

- * dvoutřídní data např. z kaggle
- náhodně vygenerovaná data i třídy
- časté vzory

Unsupervised anomaly detection: pro testování těchto metod potřebujeme příklady, u nichž víme, zda jsou anomální nebo normální. Můžeme použít

(vyberte nejlepší variantu)

- asociační pravidla

- *generátor umělých dat (tzv. ground truth)
- náhodně vygenerovaná data i třídy

Popište jednu z metod detekce anomálií použitých v našem projektu, ale ne LOF.

- DSMetric fitne strom a počítá occurences na listech.
- DCPMetric fitne strom a počítá "jiné" sousedy pro každý element => Toto seřazení nám umožní odstranit elementry v nejméně stejnorodém sousedství.
- KDNMetric, fitne NearestNeighbors a počítá "jiné" sousedy pro každý element => Toto seřazení nám umožní odstranit elementry v nejméně stejnorodém sousedství.
- k-Disagreeing Neighbors (kDN) - jedná se o procento sousedních instancí, které náleží jiné třídě než daná instance (kterou počítáme). Jinak řečeno, jde o procentuální vyjádření nesouhlasných sousedů.

LOF (Local Outlier Factor) je metoda

- * unsupervised
- semi-supervised
- supervised

9

Jedna z klasických metod detekce anomálií - kNN - počítá outlier faktor jako vzdálenost bodu k jeho k nejbližších sousedů.

Jiná - KDN - počítá z k nejbližších sousedů odlehost příkladu jako relativní počet sousedů, kteří nesdílí s příkladem stejnou třídu.

Z obou vychází CODB. Jak se od nich liší?

- Bere v potaz vzdálenost i hustotu v závislosti na třídách =>
    - class outlier factor se počítá z podobnosti ke K nejbližším sousedům (tzn. KDN) +
    - + kNN
    - + 1/celková vzdálenost od ostatních své třídy (== novinka) +

Pro detekci anomálních textů se používá řada metrik, jedna z následujících však ne. Označte ji.

- * Information-theoretic meta-features (e.g. Attribute Entropy)
- Rank Features (e.g. Distribution of POS tri-grams list)
- Part of Speech and Syntax Features (e.g. Percentage of words that are adjectives)

Pro detekci anomálních textů se používá řada metrik, jedna z následujících však ne. Označte ji.

- * Simple meta-features (e.g. Number of examples)
- Simple Surface Features (e.g. Average sentence length)

- Readability Measures

Pro detekci anomálních textů se používá řada metrik, jedna z následujících však ne. Označte ji.

- Rank Features (e.g. Distribution of Prepositions list)
- Obscurity of Vocabulary Features (e.g. Top 10 000 words in Gigawords)
- *Model-based features (e.g. Proportion of leaves to the class)

10

Klíčové slovo (keyword) dokumentu D je podle Mika Scotta takové slovo,

(Vyberte variantu, která je nejblíž pravdě.)

- * které se v D vyskytuje významně častěji než v referenční korpusu.
- které se v D vyskytuje častěji než v referenční korpusu.
- které se vyskytuje v D a nevyskytuje se v referenčním korpusu.

Pro detekci anomálního segmentu v textu se používají specifické míry vzdálenosti.

(Vyberte tu nejsmysluplnější.)

- * Vzdálenost založená na shlukování (např. average linkage clustering)
- Počet klíčových slov delších než tři znaky
- Rozdíl počtu určitých a neurčitých členů v textu

Pro detekci anomálního segmentu v textu se používají specifické míry vzdálenosti.

Vyberte tu nejsmysluplnější.

- *Vzdálenost segmentu textu od jeho doplňku (tj. zbytku textu)
- Počet klíčových slov delších než tři znaky
- Rozdíl počtu určitých a neurčitých členů v textu

Klíčová slova dokumentu charakterizují jednak aboutness tohoto dokumentu, jednak jeho styl.

Která z variant častěji vypovídá o stylu než o aboutness?

- * spojky
- substantiva
- Adjektiva

Vyberte odpověď, která je podle vás pravdě nejblíž.
Klíčová slova dokumentu charakterizují jednak aboutness tohoto dokumentu, jednak jeho styl.

Která z variant nejčastěji vypovídá o aboutness?

- *substantiva
- číslovky
- předložky