**Naivný Bayesovský klasifikátor**

Je algoritmus strojového učenia využívajúci Bayesové pravidlo založené na teórii pravdepodobnosti pre klasifikovanie kategórie vzorky (kde napr. vzorka je textový dokument a kategóriou môže byť šport či umenie). Na reprezentáciu textových dát pri učení sa klasifikácie je používaný model bag-of-words, ktorý spočíva v spočítavaní frekvencií každého slova v trénovacích vzorkách pre kategóriu do ktorej spadajú. Majme napríklad 5 trénovacích vzoriek s dvoma rôznymi ohodnoteniami (kategóriami) produktu Good a Bad v nasledujúcej tabuľke.

|  |  |
| --- | --- |
| **Text vzorky** | **Kategória (review)** |
| excellent headset | Good |
| mic doesn’t work | Bad |
| love this headset | Good |
| bad quality mic | Bad |
| great mic | Good |

Reprezentácia textu vzoriek pre každú kategóriu sa cez model bag-of-words dá vyjadriť nasledujúcou tabuľkou, kde atribúty sú slová a ich hodnoty sú frekvencie v dvoch rôznych kategóriách.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Kategórie** | **excellent** | **headset** | **mic** | **doesn’t** | **work** | **love** | **this** | **bad** | **quality** | **great** |
| Good | 1 | 2 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| Bad | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |

Spočítavanie frekvencií slov je takmer celý proces učenia sa a prostredníctvom tejto tabuľky máme najväčší nástroj pre vypočítavanie pravdepodobností pre každú kategóriu do ktorej môže neznáma vzorka patriť. Na predikovanie kategórie neznámej vzorky jednoducho vybereme kategóriu s najväčšou vypočítanou pravdepodobnosťou. Vzorky použité na spočítavanie frekvencií slov sa nazývajú ako trénovacia množina

**Výber kategórií**

Na to, aby sme videli ako presne funguje celý Naivný Bayesovský klasifikátor, vo webovej aplikácii máme v sekcii “klasifikačné parametre a ich výsledky“ vybraté tieto 3 kategórie:

* ohsu-Bacterial\_Infections\_and\_Mycoses
* ohsu-Stomatognathic\_Diseases
* news-Medicine

**Výber vzorky**

Klikneme na Klasifikuj a nižšie v trojuholníkovom grafe, predstavujúci pravdepodobnosti niekoľkých klasifikovaných neznámych vzoriek je vidieť jednu zle klasifikovanú vzorku ktorá patrí do Bacterial\_infections. Keď na ňu klikneme, zobrazia sa nám ďalšie grafy prostredníctvom ktorých môžeme vidieť výpočtové procesy vzorky.

Trojuholníkový graf napravo obsahuje slová kliknutej vzorky a ukazuje ich vypočítané pravdepodobnosti spočívajúce ku ktorým kategóriám najviac korešpondujú.

Nižšie od trojuholníkových grafom je graf vyjadrujúci priebeh výpočtov pravdepodobností po slovách v neznámej vzorke. Čiarkované čiary predstavujú pravdepodobnosti samostatných slov, ktoré je možné vidieť v pravom trojuholníkovom grafe. Plná čiara vyjadruje sumarizujúcu pravdepodobnosť ktorá bere do úvahy všetky predošlé slová v jednom segmente slov. Segmenty sú oddelované tučným bodom, ktorý farbou vyjadruje kategóriu ku ktorej by bol príslušný segment slov klasifikovaný. Po každom tomto bode sa tieto sumarizujúce pravdepodobnosti resetujú (pre lepšiu vizualizáciu výpočtov cez celú vzorku). Dĺžku segmentov je možné zmeniť prostredníctvom slideru vyššie – “Resetovy bod sumarizacnych pravdepodobnosti”.

**Slová**

Nastavme ho na 35. Uvidíme na začiatku druhého segmentu slov ako pravdepodobnostne vedie kategória, ktorá spôsobila chybnú klasifikáciu (Stomatognathic\_Diseases) druhého segmentu slov. Najväčší chybný vplyv mali v tomto segmente slová tumors (39. slovo), bony (47), tumor (49/56) dissection (66). Tieto slová je možné vidieť aj v grafe uvedenom nižšie, ktorý hovorí o ich jedinečnosti cez výpočet percentuálneho rozdielu medzi najpravdepodobnejším a druhým najpravdepodobnejším slovom medzi kategóriami. Pozrime sa na slovo tumor (49) bližšie cez atribúty trénovacej množiny, ktorý je vyjadrený modelom bag-of-words.

**Atribúty trénovacieho datasetu**

Graf týchto atribútov je zobrazený vyššie (atributy trenovacieho datasetu). Pre zobrazenie najfrekventovanejších používaných slov kategórie spôsobujúcu chybu klikneme na dropdown pre výber tejto kategórie (Stomatognathic\_Diseases) a nájdeme slovo tumor napr. cez ctrl+f. Je vidieť, že frekvencia tohto slova je veľmi malá a pomerový rozdiel s ostatnými kategóriami nieje tak veľký, no napriek tomu toto slovo malo zásadný vplyv na pravdepodobnostný výpočet. (Tento príklad nieje veľmi šťastný, o niečo neskôr spomeniem slovo na ktorom je vidieť krajšie vidieť frekvencie spolu aj s pomerom)

V tomto grafe taktiež môžeme vidieť výsledné pravdepodobnosti slov cez možnosť “Pravdepodobnost slov”. Cez tieto dve možnosti je vidieť, ako pravdepodobnosť nezávisí len od samotných frekvencí slov pretože pomer hodnôt medzi kategóriami sa zásadne mení.

**Vlastnosti trénovacej množiny**

Na to, aby sme mohli počítať pravdepodobnosť slova je treba zohľadniť, koľko slov sa vyskytuje vo všetkých vzorkách jednej kategórie. Ak by sme mali napr. 10 slov v jednej kategórií a 3 z nich by boli ‚zdravie‘, je omnoho pravdepodobnejšie, že slovo zdravie bude viac korešpondovať s touto kategóriou ako s kategóriou ktorá obsahuje 7 takýchto slov s celkovým počtom slov až 300.

V našom prípade rozdiel v počtoch slov medzi kategóriami (Stomatognathic a Bacterial) je cca 7-násobný. Toto je možné spozorovať pri najvyššie zobrazenom grafe (Vlastnosti trenovacej mnoziny). Defaultne sú nastavené počty trénovacích vzoriek pre každú kategóriu, pomocou interaktívnej legendy však môžeme zakliknúť “Slova“ pre zobrazenie počtov slov v každej kategórii.

Ak sa napr. pozreme na 4. alebo 5. slovo (treatment a disease), v grafe „atributy trenovacieho datasetu“, môžeme vidieť, že kategória Bacterial ich má najväčší počet, ak sa však pozreme na ich výsledné pravdepodobnosti, kategória Stomatognathic má prekvapivo pravdepodobnejší výskyt tohto slova.

Pravdepodobnosť slova neznámej vzorky sa teda počíta počtom výskytov (z trénovacích vzoriek) určitého slova v kategórii vydeleným celkovým počtom slov (z trénovacích vzoriek) v kategórii. Tieto vypočítané pravdepodobnosti sú násobené každým ďalším slovom vo vzorke a tvoria výslednú pravdepodobnosť - **Likelihood**.

Klasifikačný proces vzorky obsahuje ešte jedno vynásobenie s tzv. **Priorom**. Tým, že sa jedná len o jedno vynásobenie navyše, pri vzorkách s dlhým textom mávajú tieto priory veľmi slabý vplyv.

Tak ako sme brali do úvahy celkový počet slov v kategóriách, môžeme brať do úvahy aj počty vzoriek v každej kategórii. Ak máme kategóriu so zriedkavejšími výskytmi vzoriek, chceli by sme, aby bola o niečo menej pravdepodobná vo výslednej klasifikácii.

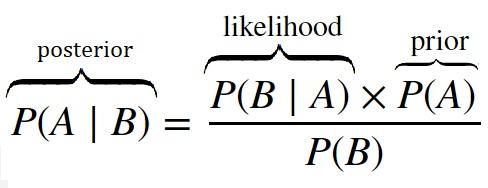
Pravdepodobnosť Prioru dosiahneme počtom vzoriek v kategórii vydelenú celkovým počtom vzoriek vo všetkých kategóriách. Čím majú kategórie podobnejšie počty vzoriek, tým budú mať menší vplyv na výsledok. Ak však tam aj je veľký rozdiel, pri klasifikácii dlhých textov to bude mať tak či tak slabý vplyv na výsledok.

**Vizualizácia vplyvu priorov**

Vplyv je možné vidieť v grafe pravdepodobností slov. Kedže máme zakliknutú vzorku s veľkým počtom slov, tento vplyv priorov je veľmi slabý a nieje vidieť odlišnosť vo výsledných pravdepodobnostiach, ktoré sú označené oranžovou a fialovou farbou. Na to, aby sme videli lepšie vplyv priorov, zredukujme text označenej vzorky prostredníctvom slideru “Selekcia slov vzorky” nižšie. Stiahnime bod konca na začiatok pre výber napr. len prvých 2 slov. Uvidíme, že výsledné pravdepodobnosti bez prioru a s priorom sa odlišujú omnoho viac. Každým pridaným slovom môžeme sledovať, ako sa tieto výsledné pravdepodobnosti približujú k sebe.

**Bayesové pravidlo**

Proces výpočtu neznámej vzorky pre určenú kategóriu je inšpirovaný už spomínaným Bayesovým pravidlom, kde B je neznáma vzorka a A je kategória:



* Presnejšie vyjadrenie výpočtov pravdepodobností je vyjadrený vo webovej aplikácii.
* Tento algoritmus je možné modifikovať tak, aby vedel pracovať aj s reálnymi číslami v atribútoch (Gaussian Naive Bayes).
* Hlavná nevýhoda algoritmu je predpoklad nezávislosti slov - kedže počítame pravdepodobnosti každého slova zvlášť, vytvára sa predpoklad, že slová niesú medzi sebou v žiadnom vzťahu.
* Výhodou algoritmu je rýchlosť trénovania a klasifikácie, vystačí si s malým množstvom trénovacích dát a poradí si aj s veľkým počtom atribútov tak, ako sme to aj videli pri našej textovej klasifikácii.
* Pri vypočítavaní likelihoodu sa môže stať, že niektorá kategória neobsahuje v trénovacej množine žiadne slovo ktoré sa vypočítava v neznámej vzorke a spôsobí nulovú celkovú pravdepodobnosť likelihoodu. Preto sa používa Laplace smoothing – pridanie jedného výskytu slova do všetkých atribútov kategórie.

