Štatistické rozdelenia dát

V pravdepodobnostných a štatistických modeloch reprezentujeme dáta ako realizácie náhodných premenných.

(Vo všeobecných metódach dátovej vedy však dáta nie vždy reprezentujeme ako realizácie náhodných premenných.)

„Teoretické“ rozdelenia

1. rozdelenia kategorických náhodných premenných
2. diskrétne typy rozdelení náh. premenných a náh. vektorov
3. spojité typy rozdelení náh. premenných a náh. vektorov
4. iné rozdelení náh. premenných a náh. vektorov
5. rozdelenia pravdepodobnosti na iných množinách

1.

Kategorické náhodné premenné:

Rovnomerné kategorické rozdelenie, kategorické rozdelenie dané vymenovaním množiny pravdepodobností jednotlivých kategórií

2.

Diskrétne jednorozmerné:

rovnomerné diskrétne, alternatívne, binomické, geometrické, Poissonovo, hypergeometrické, negatívne binomické, negatívne hypergeometrické, beta-binomické, Benfordovo, ...

Diskrétne mnohorozmerné:

multinomické, negatívne multinomické, ...

3.

Spojité jednorozmerné:

beta (špeciálny prípad rovnomerné rozdelenie na intervale (0,1)), gamma (špeciálny prípad exponenciálne rozdelenie), Paretovo, Normálne, Lognormálne, t, F, chíkvadrát, ...

Spojité mnohorozmerné:

mnohorozmerné normálne, rovnomerné na rôznych objektoch (najmä kocka alebo guľa), ...

4.

Rozdelenia, ktoré nie sú ani diskrétne, ani spojité:

singulárne mnohorozmerné normálne, Dirichletovo, rovnomerné na povrchu mnohorozmernej gule, von Misesovo-Fisherovo rozdelenie, ...

5.

Rozdelenia na iných množinách.

Rozdelenia na množine (alebo nejakej podmnožine) matíc, napr. Wishartovo rozdelenie, alebo maticové normálne rozdelenie;

Rozdelenia na množine nekonečných postupností vyplývajúce z modelov generujúcich náhodné postupnosti, predovšetkým z takzvaných Markovovských modelov;

Rozdelenia na množine funkcií vyplývajúce z modelov generujúcich takzvané náhodné procesy, napríklad Wienerov proces a Ornsteinov-Uhlenbeckov proces;

Rozdelenia na grafoch a sieťach vyplývajúce z teoretických modelov náhodných grafov a sietí, ako Erdosov-Polyov model náhodných grafov, alebo Barabasiho-Albertov model náhodných sietí.

„Praktické“ rozdelenia

- Niekedy sú teoretické rozdelenia v podstate dokonalé aj pre praktické aplikácie (mnohé fyzikálne aplikácie, niektoré informatické aplikácie, lotérie a hazardné hry, ...);

- Často sú teoretické rozdelenia jasne nedokonalé, ale aj tak používané (napríklad vo finančnej matematike, v prírodných a spoločenských vedách, v medicíne...). Predpoklad rozdelenia, hoci nie celkom presný, často vedie k tomu najlepšiemu, čo sme schopní reálne urobiť („*All models are wrong but some are useful.*“);

- Niekedy používame empirické rozdelenia pravdepodobnosti (napríklad úmrtnostné tabuľky v poisťovníctve, prípadne empirické rozdelenia založené na dátach v strojovom učení);

- Niekedy tiež používame zmesi teoretických rozdelení pravdepodobností na priblíženie sa realite (napríklad zmes normálnych rozdelení v partičnej analýze zhlukov).

FAQ k typom a rozdeleniam dát :)

Q: Sú všetky dáta normálne (gaussovské)?

A: Nie. Avšak mnohé dáta približne normálne sú, najmä menej-rozmerné spojité, intervalové dáta. Apriori nemáme dôvod sa domnievať, že dáta sú normálne. Avšak často povaha dát zaručuje, že ide o súčet malých, podobných, nezávislých náhodných vplyvov, vtedy je predpoklad normality dát oprávnený.

Navyše, aj ak dáta nie sú normálne, tak štatistiky, ktoré z dát vypočítame, môžu mať rozdelenie veľmi podobné normálnemu. Takže v niektorých situáciách môžeme použiť aj na nie-normálne dáta také metódy analýzy dát, ktoré sú vyvinuté za teoretického predpokladu normality.

Q: Ak použijem nesprávne rozdelenie na dáta, je zle?

A: Nie nutne. Veľmi stručne rozoberme dva typické ciele použitia pravdepodobnostných a štatistických modelov na dáta: klasické testovanie štatistických hypotéz (výpočet p-hodnôt) a klasifikácia objektov v strojovom učení (lineárny Bayesovský klasifikátor).

Q: Ako zistím, z akého rozdelenia pochádzajú dáta?

A: Môžeme mať exaktný model pozorovaní, ktorý jednoznačne určuje o aké konkrétne rozdelenie ide, prípadne o akú triedu rozdelení ide. Ak by som mal jasne určenú parametrickú triedu rozdelení (napríklad z literatúry ohľadom daného typu dát), použijem štatistické metódy odhadov parametrov, najčastejšie metódu maximálnej vierohodnosti. Ak neviem ani z akej triedy rozdelení sú moje dáta (alebo akou by mohli byť približne aproximované), skúšam rôzne triedy rozdelení a modelov, pričom ich primeranosť vyhodnocujem pomocou špeciálnych štatistických testov a kritérií, napríklad pomocou takzvaných informačných kritérií (*information criteria*, AIC, BIC, ...).

(Niekedy však nepotrebujem vedieť aké je štatistické rozdelenie mojich dát.)