Tradiční vs Bayesiánská statistika

Když se na škole vyučuje statistika, jedná se obvykle o teoretický výklad, ze kterého je poměrně těžké si odnést představu, jak ji lze využít v praxi. Zkusíme si spolu spočítat jeden příklad, který se může nejen hodit, ale také si ukážeme důvod, proč je zpočátku tak těžké tradiční statistickou inferenci pochopit.

Zadání

Zítra se chystáme na výlet do Prahy a rádi bychom věděli, jestli si s sebou máme vzít deštník. Zavoláme proto třem svým kamarádům a nezávisle na sobě se jich zeptáme, jestli v Praze prší. Všichni shodně odpověděli, že prší. Kamarádi jsou ovšem vtipálci a s pravděpodobností 1/3 nám zalhali. Máme deštník nechat doma?

Tradiční přístup

S populací budeme zacházet jako s uspořádanou trojicí (XYZ) , kde jednotlivé prvky nabývají na hodnotách D (déšť) nebo S (slunce). Celá populace se skládá z osmi různých hodnot: (DDD), (DDS), (DSD), (SDD), (DSS), (SDS), (SSD), (SSS).

Protože můžeme předpokládat, že jsou odpovědi nezávislé, můžeme s nimi zacházet jako s binomickou distribucí a použít vzorec

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

V našem případě je n=3, p=2/3 a k odpovídá počtu kamarádů, kteří řekli pravdu. Teď už snadno spočítáme pravděpodobnosti jednotlivých jevů P(k) pro počet odpovědí, které odpovídají dešti.

P(3) = P({DDD}) = 8/27

P(2) = P({DDS}) = 12/27

P(1) = P({DSS}) = 6/27

P(0) = P({SSS}) = 1/27

Jak budeme odpovědi interpretovat?

Pokud nejsme z cukru, pak si deštník vezmeme pouze v případě, že všude prší, což odpovídá pravděpodobnosti P({DDD}) = 8/27.

Na druhou stranu pokud chceme jistotu, že jsme připraveni na déšť, pak chceme vědět, že nikde neprší, což odpovídá pravděpodobnosti 1 - P({SSS}) = 1 - 1/27 = 26/27.

Nejen, že jsme získali dvě možné odpovědi, ale obě interpretace také popírají intuitivní chápání odpovědí. Co když všichni kamarádi bydlí na stejném místě? Počítání pravděpodobnosti jevů {DDS} nebo {DSS} sice dává smysl v matematickém duchu, ale nedává smysl v tom, jaké počasí vlastně v Praze je?

Bayesiánská statistika

Pojďme to zkusit lépe a spočítat si pravděpodobnost, s jakou prší, aby byl matematický výpočet v souladu s intuitivním chápáním dat.

K tomu využijeme Bayesovu větu o podmíněné pravděpodobnosti

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

V našem příkladě jev A odpovídá reálnému dešti, zatímco jev B jsou odpovědi našich kamarádů. Počítáme tedy podmíněnou pravděpodobnost deště za předpokladu, že všichni kamarádi odpověděli kladně, p(d|{DDD}).

Dosazením do Bayesovy věty získáme

$$P(d|\{DDD\}) = \frac{P(\{DDD\}|d) * P(d)}{P(\{DDD\})}$$

Tedy pravděpodobnost, že prší při třech kladných odpovědích, p(d|{DDD}), odpovídá pravděpodobnosti tří kladných odpovědí za deště, p({DDD}|d), násobené [nepodmíněnou] pravděpodobností deště, p(d), a dělené [nepodmíněnou] pravděpodobností tři kladných odpovědí, p({DDD}).

Výrazu p(d|{DDD}) říkáme a posteriori pravděpodobnost, zatímco výraz p(d) nazýváme a priori pravděpodobnost. Ta odpovídá naší víře nebo předpokladům za stavu, kdy nemáme žádné další informace.

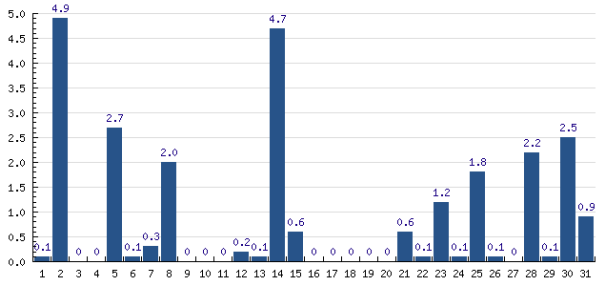
Pravděpodobnost ve jmenovateli spočítáme pomocí věty o úplné pravděpodobnosti, p({DDD}) = p({DDD}|d} * p(d) + p({DDD}|s) * p(s).

A protože pravděpodobnosti odpovědí jsme spočítali už v první části pomocí binomického rozdělení, získáme snadno finální vzorec.

$$P(d|\{DDD\}) = \frac{8/27 * P(d)}{8/27 * P(d) + 1/27 * (1 - P(d))}$$

$$P(d|\{DDD\}) = \frac{8 * P(d)}{1 + 7 * P(d))}$$

K vyřešení úlohy ovšem potřebujeme ještě zjistit, jaká je a priori pravděpodobnost deště v Praze bez dalších předpokladů? Jedna možnost je najít si informace na internetu.



V březnu v Praze pršelo 20 z celkových 31 dnů, tedy můžeme postavit $p(d) = 20/31$. Dosazením dostáváme odpověď na naši původní úlohu, že šance na déšť na základě odpovědí našich kamarádů je 0.936, tedy 93.6%.

Ještě zajímavější odpověď ale můžeme získat, pokud a priori pravděpodobnost deště neznáme? Řekněme, že pokud je alespoň poloviční šance na déšť, deštník si raději vezmeme.

V tom případě položíme podmínku $p(d|\{DDD\}) >= 0.5$ a dosazením našeho vzorce získáme $p(d) >= 1/9$. Takže pokud alespoň jednou za posledních devět dní pršelo, nebo pokud pršelo alespoň čtyřikrát za poslední měsíc, bude jistější si deštník vzít.