Bayesův vzorec:
jak nová data mění pravděpodobnost jevu

J. Valenta[1], D. Krátký[2], O. Říha[3], M. Moravec[4]

[1]Gymnázium Jateční, [2]Gymnázium Tanvald,

[3]Gymnázium Velké Meziříčí, [4]Církevní gymnázium Plzeň

[1]Valenta.j@gymjat.cz, [2]dany.kratky.dk@gmail.com,

[3]rihaond@gmail.com, [4]moravecmatej@post.cz

Abstrakt:

Bayesův vzorec umožňuje výpočet aposteriorní pravděpodobnosti náhodných jevů pomocí předem známých informací (apriorních pravděpodobností), lze ho aplikovat v řadě různých odvětví. V případě našeho výzkumu jsme vypočítali pravděpodobnost deště na základě dalších vnější pozorování. Využívali jsme k tomu statistik meteorologických institucí a otevřených zdrojů z let 2011-2022. Tento výzkum si nečiní nárok na to, aby byl přesným a spolehlivým ve smyslu předpovědi počasí, ale je spíše ukázkou možné aplikace Bayesova vzorce k určení, jak použiti nových dat může změnit odhad pravděpodobnosti některého jevu.

1 Teoretický úvod

Bayesův vzorec je rovnicí znázorňující vztah mezi podmíněnou pravděpodobností a opačnou podmíněnou pravděpodobností. Vzorec byl poprvé zveřejněn v roce 1763, 2 roky po autorově smrti, ale povědomí o něm stagnovalo až do dob 2. poloviny 20. století, kdy začal být široce využíván a dal vzniknout novému odvětví statistiky – bayesovské.

Teorie pravděpodobnosti, do které tento vzorec spadá, se zabývá popisováním vlastností náhodných jevů. Ty jevy, které se dají ovlivnit nějakými jinými nezávislými jevy, můžeme popisovat právě pomocí zmíněného vzorce.

Podmíněná pravděpodobnost je důležitou součástí tohoto vzorce:

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

Popisuje pravděpodobnost jevu $P(A)$ za předpokladu, že mu jistě předcházel jev B. Pravděpodobnost samotného jevu $P(A)$ jsme schopni vypočítat, za předpokladu, že známe pravděpodobnosti úplného systému vzájemně neslučitelných jevů $B_1, B_2, \ldots, B_n$ a jejich vzájemné vztahy. Proto:

$$P(A) = \sum_{i}^{n} P(A|B_i) \cdot P(B_i)$$

Bayesův vzorec nám pomáhá s vyčíslením pravděpodobnosti jednoho prvku z daného systému vzájemně neslučitelných jevů, popsaným výše [1].

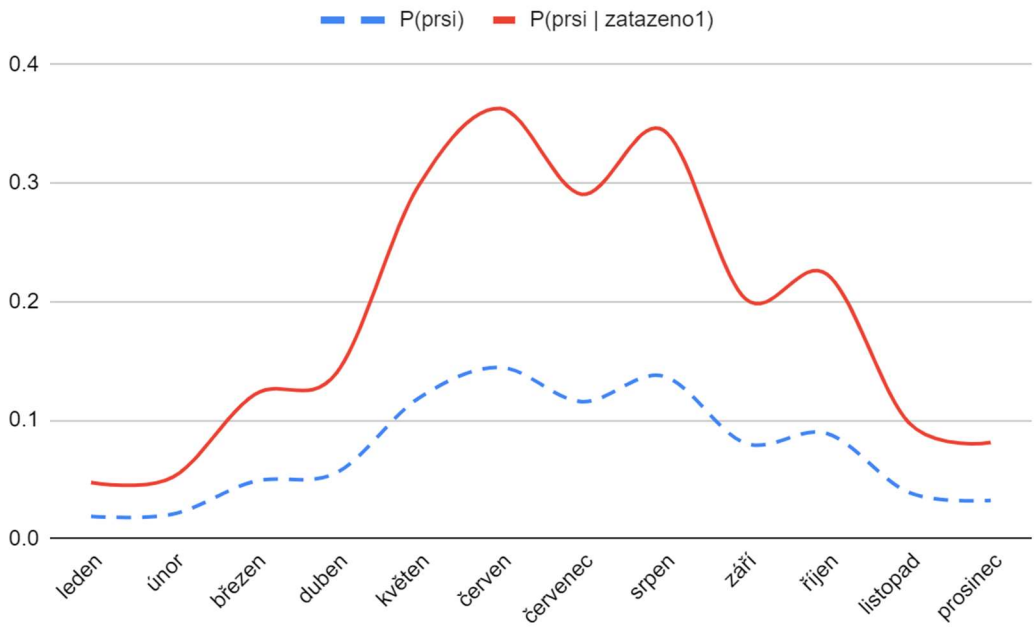
$$P(B_k|A) = \frac{P(A|B_k) \cdot P(B_k)}{P(A)}$$

Naším cílem bylo prokázat aplikovatelnost Bayesova vzorce do reality. Vzorec jsme aplikovali k výpočtu pravděpodobnosti výskytu srážek za různých okolností, naměřených Českým hydrometeorologickým ústavem v letech 2011-2022, které mohou tento náhodný jev ovlivňovat.

2 Zpracování dat

Předpokládáme, že A je nahodný jev, že v určitý den bude pršet. $P(A)$ značí pravděpodobnost tohoto jevu bez jakéhokoliv kontextu. Tato hodnota byla spočítána zprůměrováním hodnot z hydrometeorologického ústavu ČR z let 2011-2022, z hodnot naměřených konkrétně stanicí v Praze, Karlov [2]. Jev B předpokladame za nahodný jev, že bude zatažená obloha. Jeho pravděpodobnost byla spočítána zprůměrováním hodnot z meteorologické stanice Košíky z let 2011-2022 [3]. $P(A|B)$ jedná se o podmíněnou pravděpodobnost toho, že bude pršet, pokud je zataženo. Tuto hodnotu jsme získali z webu [4]. Vztah mezi zataženou oblohou v ranních hodinách a následným deštěm jsme vypočítali pomocí zprůměrovaných dat z meteorologické stanice Košíkov, naměřených ve stejných letech.

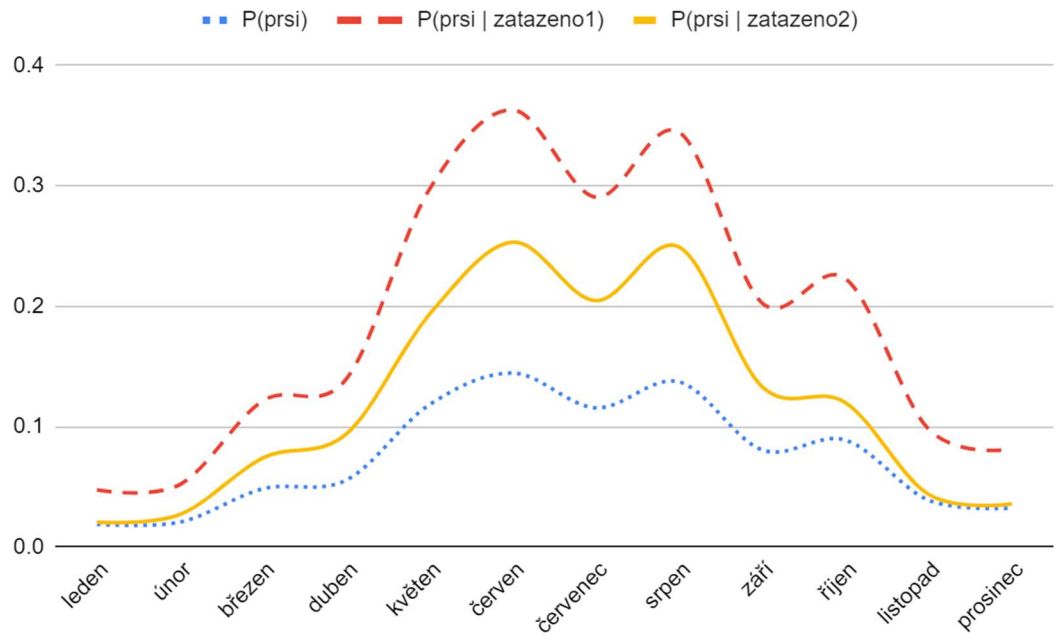
Z dosazených hodnot nám vyšly hodnoty pravděpodobnosti pro každý měsíc zvlášť. Pro lepší představu o těchto pravděpodobnostech jsme je vložili do tohoto grafu na obrázku 1. Šrafovaná čára značí pravděpodobnost, že bude pršet bez jakéhokoliv kontextu, pokud víme který je měsíc. Naproti tomu, plná čára značí pravděpodobnost deště, pokud víme, že je oblačno.



Obrázek 1. Graf pravděpodobnosti deště za podmínky, že je oblačno

Z grafu lze vyčíst, že s novými daty se pravděpodobnost značně mění, ovšem tyto výsledky nemusí reflektovat realitu, jelikož v zimě je zatažená obloha pořád. Což nás přivedlo k přehodnocení dat, která využíváme a zvolení lepší techniky.

Z ČHMÚ jsme si vzali data ze stanice Praha, Karlov o denním úhrnu doby trvání slunečního svitu v Praze. Tyto hodnoty jsme porovnali s průměrnou délkou dne v tyto roční doby a výsledky použili jako podklad pro regulaci důležitosti oblačnosti. Tento přístup nám umožnil mnohem přesnější zpracování pravděpodobnosti vzhledem k typickému charakteru temné zimní oblohy, což je zobrazeno na obrázku 2. Tečkovaná čára značí pravděpodobnost bez kontextu. Šrafovaná čára značí náš první výsledek. Plná čára v tomto grafu označuje nejpřesnější výsledek.



Obrázek 2. Graf pravděpodobnosti deště za podmínky, že je oblačno

Tento graf je přesnější v zobrazení reality, neboť zahrnuje více faktorů, což ho dělá přesnějším.

3 Shrnutí

Jak jsme prokázali, tak Bayesův vzorec lze aplikovat na meteorologické jevy a získat lepší výsledky, než bychom získali bez jeho použití. Avšak je třeba do vzorce doplňovat co nejpřesnější údaje a vzít v úvahu všechen kontext. Také jsme ukázali důležitost zahrnutí, co nejvíce dat.

Poděkování

Toto poděkování bychom chtěli věnovat Mgr. Maksymu Drevalovi za jeho snahu a trpělivost učit nás základy kombinatoriky a pomoc s tímto projektem. Dále děkujeme celému týmu pořadatelů Týdnu vědy na Jaderce za možnost zúčastnit se.

Reference

[1] Wikipedie: Otevřená encyklopedie: Bayesova věta. Wikipedie: Otevřená encyklopedie [online]. [cit. 2023-06-20]. Dostupné z: https://cs.wikipedia.org/wiki/Bayesova_v%C4%9Bta

[2] Portál ČHMÚ : Historická data : Počasí : Denní data : Denní data dle z. 123/1998 Sb. Portál ČHMÚ [online]. Praha: Český hydrometeorologický ústav, 2023 [cit. 2023-06-20]. Dostupné z: https://www.chmi.cz/historicka-data/pocasi/denni-data/Denni-data-dle-z.-123-1998-Sb#

[3] Meteorologická stanice Košíky - Počasí Košíky. Meteorologická stanice Košíky [online]. Košíky [cit. 2023-06-20]. Dostupné z: http://www.pocasi-kosiky.cz/

[4] Bayesova věta - Elements of AI. Elements of AI [online]. Helsinki, 2022 [cit. 2023-06-20]. Dostupné z: https://course.elementsofai.com/cs/3/2