

Zadání druhého projektu z MSP

Projekt se skládá ze dvou úloh.

- využití bayesovských odhadů a simulace rozdělení z naměřených dat.
- regresní analýza.

Odevzdání projektu je do konce 13. výukového týdne - tedy do 17.12. 23:59:59.

Forma odevzdání je jediný soubor PDF. Vhodně strukturovaný podle úloh, kde bude okomentovaný kód a okomentované požadované výstupy.

Název souboru: *Příjmení_MSP_2*-projekt. Za *Příjmení* si každý dosadí to svoje!!
Soubor vložíte do příslušného adresáře v E-learningu (bude zřízen).

ÚLOHA 1 – Bayesovské odhady – 4. body

a) Konjugované apriorní a aposteriorní rozdělení, prediktivní rozdělení [2 body]

Předpokládáme, že počet **připojení na internetovou síť za 1 ms** je popsán náhodnou veličinou s Poissonovým rozdělením s parametrem λ , t.j. $X \sim Po(\lambda)$.

O parametru λ máme následující expertní odhad: **každých 5 ms by mělo nastat 10 připojení**.
Pozorovali jsme připojení po **dobu 100 ms**. Pozorování o počtu připojení za každou 1 ms jsou uvedené v souboru `measurements.csv` ve sloupci „úloha_1 a“.

Vášim zadáním je z této expertní informace určit konjugované apriorní rozdělení k parametru Poissonova rozdělení a na základě pozorování určit aposteriorní rozdělení. Dále určete apriorní a aposteriorní prediktivní rozdělení pozorování.

Požadovaný výstup:

- 1) Do jednoho obrázku vykreslíte **apriorní a aposteriorní hustotu** parametru Poissonova rozdělení λ .
- 2) Do **jednoho** obrázku vykreslíte apriorní a aposteriorní **prediktivní hustotu pozorování x za jeden časový interval**.
- 3) Sestrojte **95% interval spolehlivosti** pro parametr λ z apriorního a aposteriorního rozdělení a **porovnejte** je.
- 4) Vyberte si **dva aposteriorní bodové odhady** parametru λ , **porovnejte** je a **okomentujte** jejich výběr.
- 5) Vyberte si **jeden apriorní** a **jeden aposteriorní** bodový odhad počtu pozorování a porovnejte je.

Nápověda.

Pro určení apriorní, aposteriorní a prediktivních hustot využijte tabulky konjugovaných rozdělení, např. https://en.wikipedia.org/wiki/Conjugate_prior

b) Aproximace diskrétním rozdělením [2 body]

Integrál ve jmenovateli Bayesově větě je ve většině praktických aplikací důvodem, proč nejsme schopni odvodit aposteriorní hustotu analyticky. Jeden ze způsobů, jak překonat tento problém a odhadnout parametru (ne vektor parametrů) je, že zvolíme diskrétní aproximaci a neřešitelný integrál přejde na sumu.

Poznámka:

Nyní řešíme odhad aposteriorní hustoty a parametrů v případě, že apriorní informace (hustota) je ve formě naměřených hodnot (sloupec „uloha_1 b)_prior“) a rozdělení procesu, který sledujete, je také ve tvaru naměřených hodnot (sloupec „uloha_1 b)_pozorovania“). Tedy místo zadání dvou hustot máme naměřené hodnoty a s pomocí tříděného statistického souboru odhadneme hustoty. Pak se plocha pod hustotou spočítá součtem četností (obdobu numerického počítání integrálu obdélníkovou metodou).

Víme, že délka zpracování procesu v milisekundách ms má **odseknuté normální rozdělení** (truncated normal distribution)

viz.: https://en.wikipedia.org/wiki/Truncated_normal_distribution

s parametry

$$\mu = 3, \sigma^2 = 1, a = 1$$

Naší úlohou je odhadnout parametr b , t.j. **maximální dobu trvání procesu**. Máme historické záznamy o jeho délce trvání (sloupec „uloha 1 a)_prior“) na počítačích podobné výkonové řady. Provedli jsme sérii pozorování po 10, číslo série pozorování v tabulce v sloupci „skupina“. Z těchto záznamů vyjádříte apriorní informaci o parametru b .

Ve sloupci „uloha_1 b)_pozorovania“ jsou naše pozorování délky trvání procesu Vyjádřete **funkci věrohodnosti** (sloupec „uloha_1 b)_pozorovania“) (v tomto případě také jen její diskrétní aproximace) a následně diskrétní aposteriorní hustotu.

Požadovaný výstup:

- 1) Do jednoho grafu vykreslíte apriorní, aposteriorní hustotu a funkci věrohodnosti. Funkci věrohodnosti **normujte tak, aby jej součet byl 1** kvůli porovnatelnosti v obrázku.
- 2) **Z aposteriorní hustoty určete 95% interval spolehlivosti** (konfidenční interval) **pro parametr b** .
- 3) Vyberte **dva bodové odhady parametru b** a **spočítejte je**.

ÚLOHA 2 – Regrese – 8. bodů

Disclaimer: data (včetně „příběhu“) jsou vygenerovaná a nemusí mít dobrý obraz v realitě. Berte proto prosím výsledky z regrese s „rezervou“. Díky.

Podařilo se Vám pomocí stroje času vrátit do doby „zlatého věku“ sociálních sítí a rozhodli jste se konkurovat Facebooku a Twitteru. V souboru Data_v1.0.csv máte k dispozici záznamy od více než 500 uživatelů o rychlosti odezvy (sloupec ping [ms]) během používání Vaší aplikace. Ke každému zápisu máte navíc k dispozici o počtu uživatelů (sloupec ActiveUsers) v daném okamžiku, o procentu uživatelů, kteří momentálně interagují s prezentovaným obsahem (sloupec InteractingPct), o procentu uživatelů, kteří jen tupě scrollují po Vaší obdobě timeline/twitterfeedu (sloupec ScrollingPct) a o operačním systému zařízení ze kterého se uživatel připojil (OSType).

Úkoly a požadované výstupy:

- 1) Pomocí zpětné eliminace určete vhodný regresní model. Za výchozí „plný“ model považujte plný kvadratický model (všechny interakce druhého řádu a všechny druhé mocniny, které dávají smysl).
 - Zapište rovnici Vašeho finálního modelu.
 - Diskutujte splnění předpokladů lineární regrese a základní regresní diagnostiky.
 - Pokud (až během regresního modelování) identifikujete některé „extrémně odlehle hodnoty“ můžete ty „nejodlehlejší“ hodnoty, po alespoň krátkém zdůvodnění, vyřadit.

[4. body]

- 2) Pomocí Vašeho výsledného modelu identifikujte, pro které nastavení parametrů má odezva nejproblematictější hodnotu.

[1. bod]

- 3) Odhadněte hodnotu odezvy uživatele s Windows, při průměrném nastavení ostatních parametrů a vypočtěte konfidenční interval a predikční interval pro toto nastavení.

[1. bod]

- 4) Na základě jakýchkoli vypočtených charakteristik argumentujte, zdali je Váš model „vhodný“ pro další použití.

[2. body]