brmiversity: Umělá inteligence a teoretická informatika

Přednáška č. 6

Petr Baudiš ⟨pasky@ucw.cz⟩

brmlab 2011

brmlab hackerspace prague

Outline

Pravděpodobnost

Pravděpodobnost: Stane nebo nestane se nějaká náhodná událost?

Dvě interpretace

- **Subjektivisti:** Stav mysli, stupeň víry.
- **Frekventisti:** Konvergence série experimentů.

(**Fuzzy logika** pracuje se stupni pravdivosti, to je něco jiného!)

- Provádíme sérii *pokusů*, ty nám dávají *výsledky*, množiny výsledků jsou *náhodné jevy*.

Matematická pravděpodobnost

- Náhodný jev A má pravděpodobnost $P(A) \in [0,1]$
- Součet pravděpodobností všech základních jevů (jendotlivých výsledků) je 1; negace jevu je $1-p$
- Jednoduchý případ — rovnoměrně náhodný jev A má po n pokusech, z toho m úspěšných, $P(A) \simeq m/n$

- Stane se A nebo B? $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Nezávislé jevy** A, B — pravděpodobnost jednoho nezávisí na výskytu druhého
- Stane se A a B zároveň? $P(A \cap B) = P(A) \cdot P(B)$, jsou-li nezávislé

Matematická pravděpodobnost

- **Podmíněná pravděpodobnost** — jevy A, B nejsou nezávislé
- Stane se A za předpokladu B? $P(A|B)$
- $P(A \cap B) = P(A|B) \cdot P(B)$

Bayesova věta

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

- Diskrétní "booleovský" jev A — $P(A)$ že (ne)nastane
- Spojitý jev X — interval čísel, očekávaná hodnota $\mathbb{E}[X]$

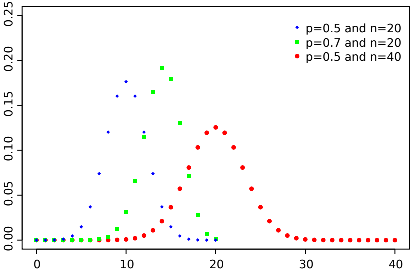
brmlab hackerspace prague

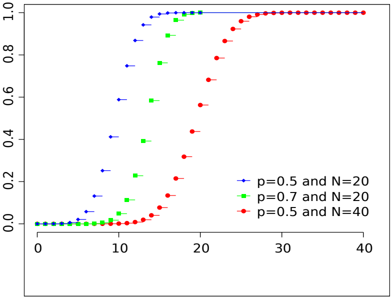
Statistika

- Teorie pravděpodobnosti se zabývá abstraktními pravděpodobnostmi jevů
- Statistika se zabývá pravděpodobnostmi, které jsme naměřili
- **Zákon velkých čísel:** Průměr (střední hodnota) naměřených hodnot konverguje k očekávané hodnotě

- Jak daleko jsou naměřené hodnoty od průměru?
- **Rozptyl** je střední kvadratická odchylka
- **Směrodatná odchylka** je odmocnina rozptylu

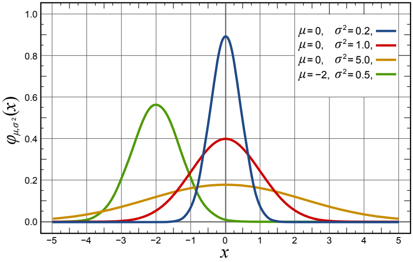
Pravděpodobnostní rozdělení

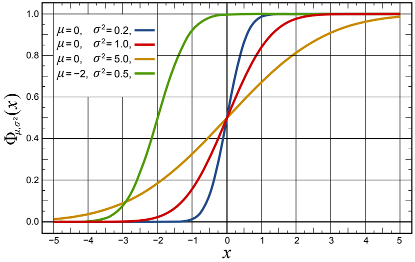




- Pravděpodobnostní rozdělení popisuje pravděpodobnosti různých hodnot při určitém typu pokusu
- Dává nám střední hodnotu a rozptyl pro dané parametry pokusu
- Bernoulliho rozdělení (p): Hod mincí
- Binomiální rozdělení (p, n): Série Bernoulliho trialů
- Poissonovo rozdělení (λ, k): Počet výskytů události za čas

Normální rozdělení





Máme-li mnoho měření, konvergují k normálnímu rozdělení:

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- **Interval spolehlivosti:** S pravděpodob. p bude $\mathbb{E}[X] = \mu \pm \epsilon$
- "ODS by volilo 20% $\pm$ 3% lidí" — s nějakou pravděpodobností (třeba 5%) by to bylo ještě více nebo méně
- 95% interval je 1.96σ

Outline

1. Pravděpodobnost
2. Umělá inteligence
3. Složitost
4. Datové struktury

Zpracování neurčité informace

- Data o světě jsou neurčitá
- Úkony ve světě jsou neurčitá
- ...takže reálný svět je neurčitý
- Neurčitost: Usuzování, rozhodování, modelování, učení

Bayesovské sítě

- Chceme modelovat svět provázaných náhodných jevů
- Bayesovská síť: Graf (DAG), uzly jsou jevy, hrany jsou podmíněné vazby

- Co se stane, když vidím tohle?
- Co bych měl zjistit, abych si co nejvíce upřesnil obraz o světě?
- Na čem nejvíce závisí, že se tohle stane?
- Kvůli čemu se asi stalo tohle?

Pravděpodobnostní rozhodovací stromy

- Chceme posoudit dopady svých rozhodnutí
  Posloupnost rozhodnutí a neurčitých jevů
  vede k různému užitku
- Rozhodovací stromy: Větvení na rozhodnutích a jevech, list je užitek
- (Pozor, "rozhodovací stromy" se říká i něčemu jinému!)
- Kterou cestou má jet robot?

Influenční diagramy

- Podstromy rozhodovacích stromů se často opakují, nezáleží na předchozích rozhodnutích a jevech
- Influenční diagramy (rozhodovací grafy):
  Obecný graf, užitkové uzly udávají změnu užitku, když skrze ně projdeme

Otázky?

Příště UI: Modelování — Markovské modely, Kalmanův filtr.

Outline

Pravděpodobnostní algoritmy

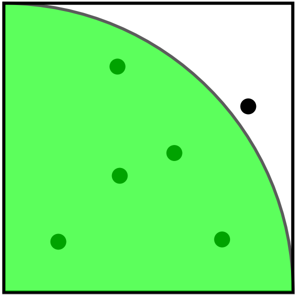
- Po algoritmu obvykle chceme, aby nám vrátil přesný výsledek za přesný čas
- Co když akceptujeme určitou *malou chybu*?
- Co když akceptujeme, že *jen asi* doběhneme včas?

- Model: Probabilistický Turingův stroj

Monte Carlo

- Monte Carlo algoritmus:
  Čím déle běžíme, tím
  přesnější výsledek dostaneme.
- Třída složitosti BPP: Problém
  řešitelný na probabilistickém TS
  v polynomiálním čase
  s pravděpodobností chyby $\leq 1/3$

- Obsah průniku kružnic
- Určitý integrál (plocha pod křivkou)
- Prvočíselné testy
- Monte Carlo Tree Search



brmlab hackerspace prague

Las Vegas

- Las Vegas algoritmus: *Očekávaná* doba běhu je jiná než nejhorší
- Třída složitosti ZPP: Problém řešitelný na probabilistickém TS v očekávaném polynomiálním čase

- Quicksort s náhodnou volbou pivota

- Je Las Vegas a Monte Carlo ekvivalentní?

Otázky?

Příště: Míry složitosti, Savitchova věta, konstruovatelné funkce.

Outline

Hashování

- Hash tabulka (v paměti — "interní")
- Musíme řešit kolize $\Rightarrow$ různé metody ukládání do tabulky
- Zajímá nás očekávaná délka řetězců l, počet testů při úspěšném (t^+) a neúspěšném (t^-) lookupu

Druhy hashování

- Separované řetězce: klasický hash se spojáky
- Uspořádané řetězce: trochu lepší t^- (a co skiplisty?)
- S přemisťováním: spoják přímo v tabulce, při kolizi přemístění
- Se dvěma ukazateli: ukazatel na začátek řetězce
- Srůstající hashování: řetězec po nejbližší volné políčko, triviální implementace, vkládáme na různá místa, příp. i do pomocné oblasti
- Dvojité hashování: jako srůstající, ale skáču chaoticky

Perfektní a univerzální hashování

Perfektní hashování: Chceme vyrobit read-only hash tabulku bez kolizí.

Univerzální hashování: Chceme vyrobit hashovací funkci odolnou k nerovnoměrnému rozdělení vstupu.

Otázky?

Příště: Univerzální a perfektní hashování.
A někdy doděláme ty haldy a externí hashování.

Děkuji vám

pasky@ucw.cz

Příště: Umělá inteligence.
Neuronové sítě (statistické zpracování dat).
Adaptivní agenti (komunikace a znalosti). Datové struktury.