Jak se počítají bayesiánské modely

Aleš Vomáčka

Jak se počítají bayesiánské modely

Analytický přístup

- Klasický přístup jsme si už ukázali, skrze bayesův teorém.

$$P(Hypotza|Data) = \frac{P(Data|Hypotza) * P(Hypotza)}{P(Data)}$$

Analytický přístup

- Výhoda - pokud funguje, výpočetně velmi rychlé (srovnatelné s maximum likelihood.)
- Nevýhoda - Výpočet $P(Data)$ vyžaduje integraci (resp. je integrál)
- Důvodem je nutnost standardizovat $P(Data|Hypotza) * P(Hypotza)$, aby šlo o platnou distribuční funkci

Analytický přístup

- Integrál $P(Data)$ je v praxi spočitatelný za dvou podmínek:
  - Pokud používáme tzv. konjugální priory
  - pokud model samotný není moc složitý (tzn. nejde o multilevel nebo aditivní model apod.)

Konjugální priory

- Konjugální prior je takový prior, který má rozdělení ze stejné rodiny jako posterior
- Např. Pokud má náš prior i naše likelihood normální rozdělení, potom bude mít i posterior normální rozdělení.

Konjugální priory

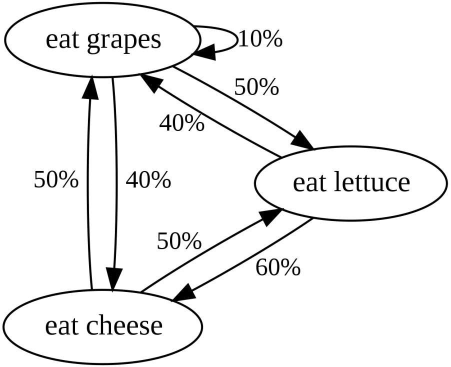
- V praxi jsme tedy omezení na specifické priory:
  - Normální prior pro klasickou lineární regresi
  - Beta prior pro logistickou regresi
- Často ale chceme vybírat priory na základě teorie bez omezení + konjugální priory neřeší problém s výpočtem složitých modelů.

Markov chain Monte Carlo

- Pokud se nedají bayesiánské modely počítat analyticky, dají se aproximovat simulačně
- Technika na výpočet bayesiánských modelů se nazývá Markov Chain Monte Carlo

Markov chain

- Sekvence stavů, kdy následující stav závisí pouze na tom stávajícím, ale ne na těch předchozích



Monte Carlo

- Pouze fancy název pro techniku simulace dat.
- Monte Carlo původně krycí název pro simulační techniky využité pro konstrokci atomové bomby.
- Dnes běžně využíváme např. u funkci `rnorm()`, `rbinom()` apod.

Markov Chain Monte Carlo

- Kombinací těchto dvou technik můžeme tahat vzorky z postoriorní dostribuce, aniž bychom museli počítat $P(Data)$.
- Likelihood parametrů, které odhadujeme je vlastně prostor, kterým “cvrnkáme” fyzikální částici.
- Částici pravidelně zastavíme, zapíšeme si její pozici a vypustíme znovu.
- Místa, kde se částice zastaví jsou vzorky z posteriu, s dostatkem vzorků dostaneme dobrou představu o tvaru posterioru

Markov Chain Monte Carlo

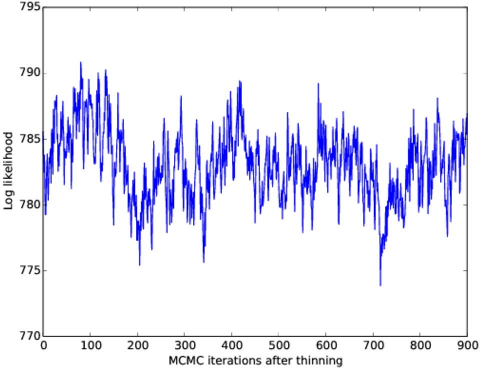
- Stan (jazyk, který používáme pro výpočet modelů) využívá Hamiltonské MCMC
- Grafické vysvětlení zde

Markov Chain Monte Carlo

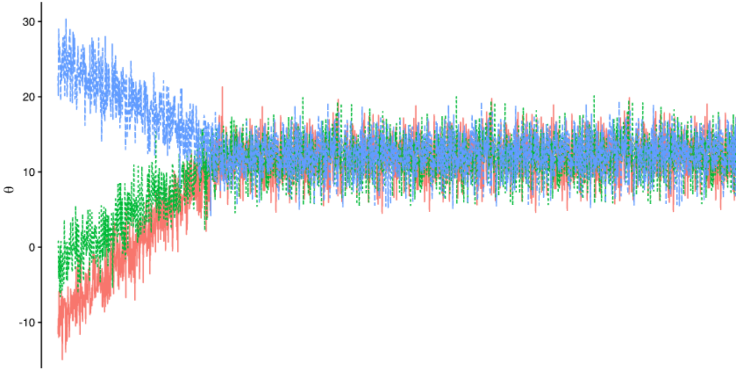
- Konceptuální znalost MCMC je důležitá:
  - Abyste chápali, proč je výpočet bayesiánských modelů tak pomalý (ve srovnání s MLE)
  - Abyste věděli, jak interpretovat diagnostické grafy.

R InteRmezzo

Špatná konvergence



Špatná konvergence



Divergentní tranzice