Haladó Gépi Tanulás 4

Kornai András

HGT 2022/9/29

Óraterv

- Projektterv-megbeszélés
- Markov és Rejtett Markov (HMM) modellek
- Gaussian Mixtures (GMM)

„M” csoport: Matúz Máté, Mészáros Bálint, Medgyes Csaba, Oroszki Norbert

A terv kezdetleges, a rendelkezésre áll1o adatok, illetve Benchmarkok függvényében változhat. Alapjáraton SentimentAnalysis-t szeretnénk vizsgálni legfőképp Netflix (vagy hasonló szolgáltató/filmértékelő oldal) kommentjei alapján

M csoport menetrend (hetek)

4-6	irodalom és a rendelkezésre álló adatok értelmezése, Benchmarkok átnézése
7	Adatértelmezés, adattisztítás és feldolgozás
8-11	Modellezés
12-13	Eredmények összegzése, modell kiértékelése és tesztelése, dokumentáció elkészítése

„Tomi” csoport

Keywords: OLS Bootstrap, Resampling, Pairs, Residual, Wild Bootstrap, MacKinnon-White errors, White-test, Bootstrapped Sampling Distribution of Regression Coefficients, Bootstrapped Confidence Interval, Hungarian Balance Sheet Data, Cross-Sectional Data, Observational Unit Bootstrapping

Plan: My project is built on the ideas of MacKinnon and Efron to bootstrap linear regressions, whose standard errors are not constant, that is, heteroskedastic. The aim of the bootstrap in our case is to get an accurate standard errors and be able to build a CI on regression coefficients. I am gonna demonstrate it on a Hungarian Balance Sheet data of firms choosing a year. I have been building a Python package for several months and am gonna be deploy on this dataset, where the target variable is the log(sales). The regressors will be later determined. The main topic would be how accurately we can estimate the coefficients and the how accurate our CIs on those coefficients. These stuff will be shown with different sample sizes.

„Tomi” csoport kritika

- Nyelvi: csak formális nyelvezetet lehet használni, semmi I'm gonna. Nekem mindegy, de ez megakadályozza a szöveg újrafelhasználását (cikkben, disszertációban), mert ilyen helyeken garantált, hogy valamelyik bíráló beleköt.
- Nem értem rendesen. Se az adatok nincsenek leírva kellő alapossággal, se a javasolt módszer. Mi a CI?
- Ha már régóta építesz egy programcsomagot, tervezz be egy előadást ennek ismertetéséről
- Általában is kérek menetrendet
- Nincs specifikálva, hogy az eredmény hogy lesz kiértékelve. Nálunk (NLP) nem lehet elfogad(tat)ni olyan cikket, amiben nincs evaluation

„B” csoport

TDK-zni is lehet (mail, preview)

Klasszikus Markov Modellek

- A kiindulás eleve nyelvészeti volt: Andrej Andrejevics Markov azt vette észre, hogy a következő hang (nála: betű) sokkal jobban megjósolható ha tudjuk mi volt az előző
- Véges automata, állapotai a betűk, élek x-ből y-ba valószínűségi súlyokkal (empirikus munka: megadni ilyen súlyokat pl. magyarra)
- Lehetnek az állapotok betüpárok (bigram), sőt betű-n-esek is (ngram) (empirikus munka: kimérni mennyivel jobb bigramra mint unigramra, trigramra mint bigramra, stb)
- Adatok: átmeneti mátrix ami megmondja hogy az i-edik állapotból a j-edikbe milyen valószínűséggel megyünk át
- Tétel: ebből a mátrixból egyszerűen kiszámolható a határeloszlás (ha sokat futtatjuk a folyamatot melyik állapotnak mi lesz a valószínűsége)

Fontos trükkök

- Nem kell n-állapotot visszakövetni, elég állapot-n-esekkel dolgozni *unigram modellben*
- Fel szoktunk venni (csendes vagy hangos) kezdő- és végállapotokat
- A hangos explicit kezdő- ill végjelet jelent (regexp ^, $)
- Simítunk! Nem hagyunk 0 valószínűségü élet akkor se, ha nem találkoztunk olyan adattal
- Matematikai oka: kikerüljük a szingularitásokat (nem fog soha kelleni zéróval osztani)
- Filozófiai oka: nem tudhatjuk milyen adatot fogunk még látni

Rejtett Markov Modellek

- Egy állapothoz többféle kimenő jel is tartozhat (diszkrét modell)
- Illetve jelek egy eloszlása is (folytonos modell)
- Példa diszkrét (szintaxis) modellre: az állapotok szófajok, a kimenő jelek szavak
- Mit tud ez? Szavak sorozatából visszafejteni a szófajokat
- Mennyire jól? Kb. 96%