Gépi Tanulás 1

Kornai András

KálmánCL 4/13

Van-e aki tud

- Programozni? (Python)
- Vektorokkal, mátrixokkal számolni? (Lin alg)
- Valószinűségekkel, statisztikával dolgozni?
- Csoportban dolgozni?

Kurzus webszájt:
https://nessie.ilab.sztaki.hu/~kornai/2021/KalmanCL

Mikről lesz szó (nem biztos, hogy ebben a sorrendben)

1. A fő feladatok: osztályozás, regresszió, generálás
2. A fő területek: beszéd- és írásfelismerés, információ-kinyerés, információ-visszakeresés, rangsorolás/ajánlás, biometrikus azonosítás, NLP feladatok
3. Leíró statisztika, lin alg, optimalizálás, információelmélet, adatredukció, PCA, LDA, jegymérnöklés
4. A fő gépi tanulók: lineáris osztályozás, maximum entrópia, rejtett markov (HMM), szomszédossági, maximális határ, genetikus/evolúciós, növelős (boosting), döntési fák, Bayes, NN
5. NLP feladatok: sorozat-cimkézés (POS, NER), darabolás (chunking), parszolás, anafora-feloldás, egyértelműsítés, nyelvazonosítás, szerep-cimkézés, jelentés-hasonlóság, parafrázis, szótárépítés, gépi fordítás

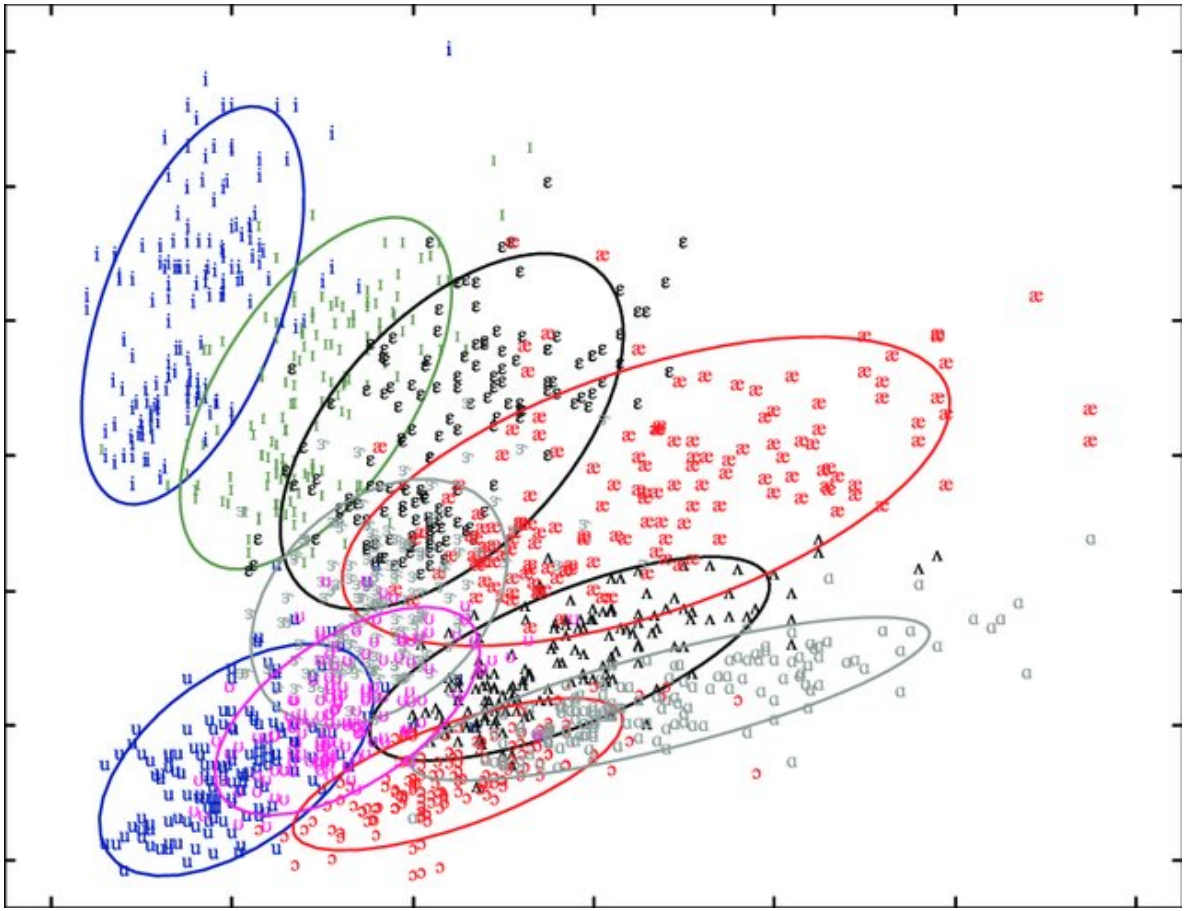
Csapjunk a lovak közé

```
@Article{   Peterson:1952,
author = {Gordon E. Peterson and Harold L. Barney},
title = {Control methods used in the study of vowels},
journal = {Journal of the Acoustical Society of America
volume = {24},
pages = {175--184},
year = {1952}}
```

Nagyon befolyásos cikk

Google Scholar 4,200 hivatkozás, Watrous, 1991 rekonstruálta az adatokat

Így néz ki



https://nessie.ilab.sztaki.hu/~kornai/2021/KalmanCL/PetersonBarney.t

Mi az adat?

33 Férfi (1); 28 Nő (2); 15 Gyerek (3) 10 mgh 2x (1520 felvétel)
Az adat 8 oszlopban: 1 nem; 2 beszélő; 3 fonéma; 4 ascii-fonéma; F0; F1; F2; F3

1	IY	[i]
2	IH	[I]
3	EH	[e]
4	AE	[ae]
5	AH	[^]
6	AA	[a]
7	AO	[o]
8	UH	[U]
9	UW	[u]
10	ER	[3]

Asterisk in ARPABET phoneme field means utterance failed of unanimous identification in listening test (26 listeners)

Mi az adat?

1	1	1	IY	160.	240.	2280.	2850.
1	1	1	IY	186.	280.	2400.	2790.
1	1	2	IH	203.	390.	2030.	2640.
1	1	2	IH	192.	310.	1980.	2550.
1	1	3	EH	161.	490.	1870.	2420.
1	1	3	*EH	155.	570.	1700.	2600.
1	1	4	*AE	140.	560.	1820.	2660.
1	1	4	AE	180.	630.	1700.	2550.
1	1	5	AH	144.	590.	1250.	2620.
1	1	5	AH	148.	620.	1300.	2530.

Mi az adat?

- Az adatok vektorok + cimkék (megfejtés)
- Lesz még olyan, hogy nem vektorok, hanem vektorok sorozatai, de ez steady-state
- Mi az első feladat?
- Szét kell vágni tanító, fejlesztő, és mérő-adathalmazra
- Mi a második feladat?
- Építeni kell egy alap-rendszert (baseline) és bemérni
- Hogyan mérünk?
- Ez osztályozási feladat, egyszerűen a sikeres jóslatok arányát nézzük
- Ha nagyon nem egyformák az osztályok méretei, akkor osztályonként mérünk és átlagolunk

Egy alaprendszer

- Minden fonémához tekintjük a tanítóadatok súlypontját
- Ez egy vektor, amit a fonéma modelljének hívunk
- Amikor tesztadat jön, akkor megnézzük melyik modellhez van a legközelebb
- Már ezzel is sokat lehet vacakolni: pl. bevegyük-e a *-os adatokat a tanító illetve a teszthalmazba?
- Kell-e minden adat? Lehet hogy F0 nélkül jobb lesz?
- Érdemes-e normalizálni valahogy az adatokat?
- Érdemes-e a súlypontba mindent beszámolni?
- Érdemes-e bonyolultabb modellekkel dolgozni?

GMM

- Gaussian Mixture Model
- Minden osztályt úgy tekintünk, mint egy n-dim valószinűségeloszlást (n a jegyek/dimenziók száma)
- Ezt lehet modellezni egy darab n-dim normális eloszlással
- Ennek paraméterei egy vektor (az átlag) és egy szórásmátrix $n(n-1)/2$ adat
- Fontos speciális eset: a mátrixot átlósnak tételezzük (a kovarianciákat elhanyagoljuk) akkor a szórás is csak n adat
- Lehet r darab normális eloszlás súlyozott keverékével dolgozni
- De csak ha van elég adat!