Haladó Gépi Tanulás 2

Kornai András

HGT 9/15

Milyen eredmények voltak 2020-ban?

- Gradient Boosted Tree (0.89, 0.89)
- Decision Trees (0.84)
- Logistic Regression (0.84, 0.83)
- Linear Regression (0.83, 0.82)
- Random Forest (0.914, 0.926)
- k-Nearest Neighbor Manhattan: (0.83, 0.90)
- feedforward NN (left as exercise to the reader!!)

Ki nyeri az értékes díjakat, és mire?

- "M" csoport (Medgyes Csaba, Matuz Máté, Mészáros Bálint) 0.958
- Baseline (súlyponttal és többféle távolsággal számolva) Euclidean (0.52, 0.58); scaled (0.68, 0.75); Canberra (0.74,0,81); log (0.79,0.81)
- A "shared task" paradigma
- Megteremtője Allen Sears vadászpilóta

Az AI (CL) hosszútávú céljai

- Nem egyszerűen beszélő gépeket akarunk (Turing), hanem olyan gépeket amik okosabban beszélnek az embereknél!
- Számos részterületen már most jobban gondolkodnak (nemcsak sakk, go, számolás)
- Bizonyos dolgokat már egész jól csinálunk, pl. viszonylag hatékonyan tudunk *ki, hol, mikor, kit, mivel* jellegű kérdésekre válaszolni (de nem *hogyan* és *miért* jellegű kérdésekre)
- Valamennyire el tudjuk dönteni, hogy egy kijelentés egy dologról azt pozitív vagy negatív szinben tünteti-e fel (sentiment analysis)
- Nagyon jól tudjuk klaszterezni a hírforrásokat bias szerint
- Troll-detekció, szerző-beazonosítás

CL ahogy Asimov gondolta 1952-ben

"And just how do you arrive at that remarkable conclusion, Mr. Mayor?"
"In a rather simple way. It merely required the use of that much-neglected commodity – common sense. You see, there is a branch of human knowledge known as **symbolic logic**, which can be used to prune away all sorts of clogging deadwood that clutters up human language."
"What about it?" said Fulham.
"I applied it. Among other things, I applied it to this document here. I think I can explain it more easily to five physical scientists by symbols rather than by words." Hardin removed a few sheets of paper from the pad under his arm and spread them out. "I didn't do this myself, by the way," he said. "Muller Holk of the Division of Logic has his name signed to the analyses, as you can see."
"As you see, gentlemen, something like ninety percent of the treaty boiled right out of the analysis as being meaningless, and what we end up with can be described in the following interesting manner:
"Obligations of Anacreon to the Empire: None!
"Powers of the Empire over Anacreon: None!"

A klasszikus CL program

- A feladatokat ugyanúgy bontjuk komponensekre mint a nyelvészek: fonológia (hangtan); morfológia (szótan); szintaxis (mondattan); szemantika (jelentéstan)
- Mindegyikhez írunk egy programot ami bemenettől kimenetig visz
- Ezeket sorbakötjük (ha kell mint pl. MT-nél akkor a túloldalon is)
- Profit

A lineáris eset

1. Az elméleti minimum: átlag, szórás, várható érték, feltételes valószínűség
2. Elsőnek `highleyman_1962.pdf`
3. Terminológiai változások az elmult 60 évben: *receptor* $\rightarrow$ *feature extractor*; *measurement space* $\rightarrow$ *feature space*; etc.

Régi idők

Even if the optimal decision function were known, its implementation would require, in general, the use of a digital computer or other complex equipment. The cost of such equipiment may, in many cases, outweigh the advantages of mechanized categorization.

What are the issues?

- Categories are numbered $1, 2, \ldots, p$, 0 reserved for *reject*, occurrence probabilities ω_i.
- $\beta(m|i)$ the probability of measurment m from true class i, generally not known, often assumed (n-dim) Gaussian
- $\delta(d_j|m)$ is the classifier making decision d_j given input m
- $C_{i,j}$ is the cost of making decision d_j given input from class i, $C_{ij} > C_{i0} > c_{ii} (= 0)$
- Expected loss $C(\delta) = \sum_{ij} c_{ij}\omega_i\beta(m|i)\delta(d_j|m)dm$
- Chow (1957) proved that indicator functions (1 at a region, 0 elsewhere) are the best δ, optimum decision function depends only on $\omega_i\beta(m, i)$
- The notion of a *decision boundary* is introduced, we are looking for approximating these by hyperplanes given by $\langle \alpha, x \rangle + \alpha_0 \geq 0$

OK boomer (actually, silent gen)

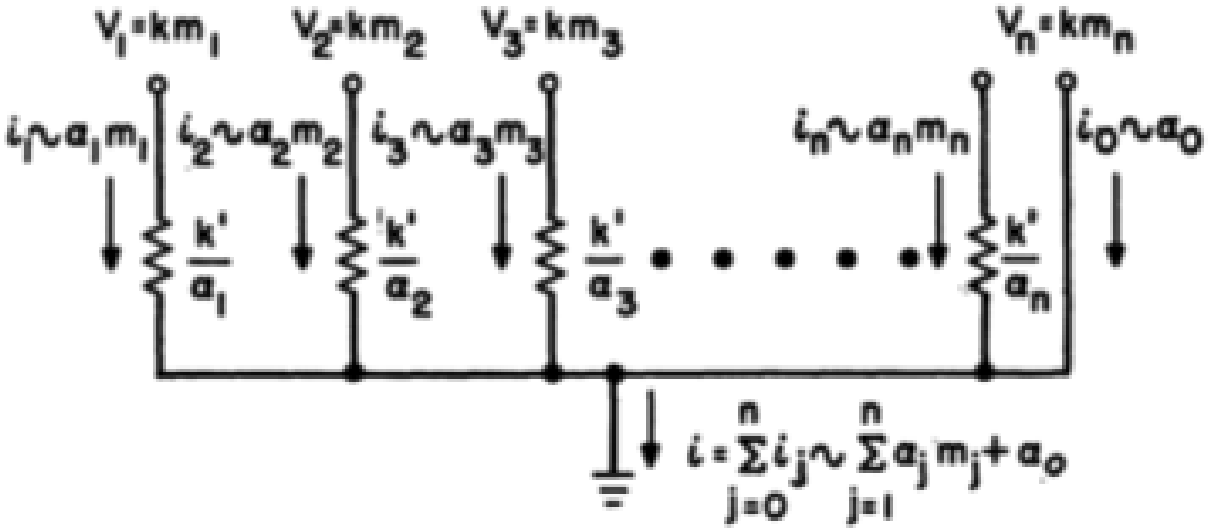

Fig. 2—Implementation of a hyperplane.

The money quote

In order to classify a point m (that is, to recognize an input pattern), it is only necessary to evaluate quantities like $\langle \alpha, x \rangle + \alpha_0$. But such a calculation can be done with several varieties of very inexpensive networks, such as the resistive adder shown in Fig. 2.

Prototype-based classification

- Assume we have k_i samples $m_{i,\ldots,m_{ik_i}}$ from class i
- We are looking for a *prototype* M_i from each class, which will be our model
- Classification is performed by comparing a new sample m to each of the models M_i, and choosing the one with the best fit
- The classifier is characterized by the M_i and the fit (distance) function d
- Assume everything is linear, and d is euclidean distance, how do we select the best M_i
- First, let's implement a prototype model where the M_i are just the average of the $m_{i,}$ samples.

Project discussion

- Read Highleyman (1962) further. Maybe you want to present the stuff in class (this is one kind of project)?
- Think of a problem domain: ASR; OCR; biometric identification; pattern classification; ranking/recommendation; info extraction; info retrieval; natural language processing (NLP); financial; medical... You can Bring Your Own Data
- Beating SOTA on any standard task guarantees an A in this course (doesn't guarantee publication these days, but it's a good step)
- Initial project plans due by 3rd week, if you can't come up with a plan a project will be assigned to you

Tudunk-e tanulni?

- Modell-osztályból választunk modellt: nincs olyan, hogy "nyílt modellosztály" vagy "tetszőleges modell"
- Csak a felügyelt tanulás ér el használható eredményeket
- A transzformerek sikere az olcsó felügyelésnek köszönhető
- Nem nagyon értjük a mélytanulásos modelleket, de gyanús, hogy ezek memorizáció+legközelebbi szomszéd alapúak
- Túl sok a paraméter, vagy mégse?

Szemantika

- 1. Logic-based: the Frege–Russell–Tarski–Montague mainstream, henceforth MG (including lineal descendants like Discourse Representation Theory, Dynamic Predicate Logic, Inquisitive Semantics, etc)
- 2. Based on (hyper)graphs: Traditional AI/KR, AMR, 4lang
- 3. Based on linear algebra: distributional semantics (CVS)
- 4. Based on automata theory: Finite State models (operational semantics, cf. Tim Fernando's minicourse)
- 5. Based on rejection of formal apparatus: cognitive semantics \ Jackendoff