Advanced Machine Learning

András Kornai

AML 2024/11/06

Groups

- Speech emotion recognition: *Gedeon* Kövér
- Wikipedia and other data harvest *Juhász*

Language endangerment and extinction

- Digital language death (Kornai 2013, 2025?)
- There are over 7k languages spoken today, 2.5k (35%) were considered endangered (in the traditional sense – 100 year time horizon). We investigated digital survival.
- Four classes: Thriving, Vital, Heritage, Still. Few dozen manually selected examples of each. E.g. French, Spanish, Chinese are T; Czech, Finnish are V; Latin, Classical Chinese are H; Rerau, Terik S.
- Measurements collected along 30+ dimensions: number of speakers, existence of spellchcker, size of wikipedia, ...
- 6 features kept. Result: over 95% of languages not just endangered, but already digitally dead. It's not "there will be an extinction", the extinction is done.

What does MaxEnt buy us?

- No need to select arbitrary thresholds
- No need to decide which features matter
- Can flexibly set classes, e.g. use only 2 (Live/Dead) or 3, or as many as you wish (assuming you can set up manual seeds reliably)
- Strong internal consistency checking
- External consistency checking is hard (because of politics) but it doesn't matter!

Genetic algorithms

- Super-attractive idea: take algorithms that are described by some genotype, mutate, cross, keep the most fit, rinse, repeat
- If it's good enough for Mother Nature it should be good enough for us
- Has serious problems in practice
- Results not stable
- Results rely on accidental properties of testbed
- Not interpretable
- Check out `thompson_1996.pdf` for an early description
- Pitfalls: too little dev data set aside, too much exploitation of train data

Simulated annealing

- More modest in its goals, inasmuch you don't craft the fitness function, it is given to you ahead of time
- Controlled random mutation, no crossover – can be thought of as directed semi-random walk over parameter space
- We start with random model at high temperature T
- Every model i (setting of model parameters) has an *energy* (unfitness) E_i, and there is a global function that describes the probability P of transition from model s_i to s_j as a function $P(T, E_i, E_j) > 0$ that depends only on temperature and the energy of these models. Usually if $E_j > E_i$ we assume $P = 1$.
- If $E_j < E_i$ we still assume $P > 0$ unless $T = 0$ (in which case the algorithm reduces to greedy search)
- A lot depends on the size and structure of the neighborhood we inspect
- Wikipedia has great animation!

Boltzmann machines

- We use simulated annealing in the special case when we wish to learn a connection matrix W whose element w_{ij} measures the strength of the connection from j to i. States are on (1) or off (0) and energy is given as $E = -\left(\sum_{i<j} w_{ij}\, s_i\, s_j + \sum_i \theta_i\, s_i\right)$
- Energy of a state is proportional to neg log prob of state: $\Delta E_i = -k_B\, T \ln(p_{\text{i=off}}) - (-k_B\, T \ln(p_{\text{i=on}}))$
- $-\frac{\Delta E_i}{T} = \ln\left(\frac{1}{p_{\text{i=on}}} - 1\right)$
- Training involves clamping the "visible layer" to observed (gold) data, followed by running the entire system to thermal equilibrium
- Connection update, based on minimizing KL distance between true (observed) and predicted (thermal equilibrium) distributions: $\frac{\partial G}{\partial w_{ij}} = -\frac{1}{R}[p_{ij}^{+} - p_{ij}^{-}]$
- Similar update for biases $\frac{\partial G}{\partial \theta_i} = -\frac{1}{R}[p_i^{+} - p_i^{-}]$

More on being "brain-inspired"

- Realistic assumptions: there are lots of neurons, update must be local
- Assume states are ± 1, look at entire state space sized $2^{10^{11}}$
- sigmoid quashing $\sigma_\beta(r) = \frac{1}{1+e^{-\beta r}}$, uniform θ, so probability of neuron i firing at time t is $\sigma_\beta(\Sigma_j W_{ij} \frac{s_j+1}{2} - \theta)$
- "Thought vector" $\Psi(t) = |s_1, \ldots, s_n\rangle$ follows path on hypercube determined by 2^n by 2^n transition matrix P which gives the the scalar product $\langle \Psi(t+1)|P|\Psi(t)\rangle$ With probability 1, P is diagonalizable
- From here we follow **Little:1974**. Temporal updates are very fast (millisecond scale) and we are interested in permenanet engrams (year scale)

Permanent memory

- Express Ψ in the normalized basis of eigenvectors ϕ_r (eigenvalues λ_r initially all assumed different) as $\psi(\Psi) = \sum_r \phi_r(\Psi)$
- Since these are orthonormal, the scalar product simplifies to $\langle \Psi(t+1)|P|\Psi(t)\rangle = \Sigma_r \lambda_r \phi_r(\alpha(t+1))\phi_r(\alpha(t))$
- States cycle trough $M = 2^n$ steps (every state is reachable). The time average of the probability of the system being in state α is $\Gamma(\alpha) = \frac{\sum_r \lambda_r^M \phi_r^2(\alpha)}{\sum_r \lambda_r^M}$
- As long as there is a unique largest eigenvalue λ_1, for large M the contributions of all the other eigenvectors and eigenvalues will be negligible both in the numerator and the denominator and we are left with $\Gamma(\alpha) = \phi_1^2(\alpha)$
- $\Gamma(\alpha, \beta) = \phi_1^2(\alpha)\phi_1^2(\beta) = \Gamma(\alpha)\Gamma(\beta)$ i.e. the probabilities are independent, *there are no persistent states*
- Suppose there are two largest eigenvalues, or λ_2 is very close to λ_1. This gives $\Gamma(\alpha, \beta) = \frac{\lambda_1^M \phi_1^2(\alpha) + \lambda_2^M \phi_2^2(\alpha)}{\lambda_1^M + \lambda_2^M}$

Conclusion from the Little model

we thus have the possibility of states occurring (. . .) which are correlated over arbitrarily long periods of time. It is worth noting too that the characteristics of the states which so persist are describable in terms of the eigenvectors associated only with the degenerate maximum eigenvalues. In this sense these persistent states are very much simpler to describe than an arbitrary state (. . .) for they involve only that small set of eigenvectors associated with the degenerate maximum eigenvalues, whereas other states (require) the full set of 2^n *eigenvectors.*