Advanced Machine Learning

András Kornai

AML 2024/10/02

Groups

- NLP: Acevedo Aktan Karoiu Tatrishvili
- Traffic signs: Bodai Oroszki Szőke Szűcs
- Fingerprint: Bárdos-Deák Boros Czakó Kránitz
- Flower: Békési Sooomro Szecskás Wiederschiz
- MRI: Hermán Kovács Nguyen Varga
- Simulation: Máth Nemes
- Sign lg: Barta Nagy Oroszki Szimonenk
- Speech: Gedeon Kövér
- WP: Juhász

Homework

Problem children

ATM8YB CEUQEJ DOC9GT IL75DE JLVG03 K1TO74 SKY4RU

Need to find project Send me email by the weekend! You need to find a project OR YOU WILL BE DROPPED FROM THE COURSE

Each project group must, by Tuesday the latest

- Name a contant person/group leader
- Set up a github project page, invite Levai and me
- Create a full project plan, with milestones, deliverables, responsibilities, ...
- Start writing a term paper that described the data sources, earlier work on this data, the models used, etc. If you use Overleaf rather than github, invite Levai and me

PCA and LDA

- Normalizing your data by PCA generally helps: nonlinearities are dampened
- On occasion, very significant dimension reduction (e.g. from 300 to 20) is achievable
- HW4.2 On PB data do any of the methods you used improve by running PCA first? No need to add methods you haven't tried yet, but if you are inspired by the leaderboard, you can
- Another classical method is Linear Discriminant Analysis (LDA)
- Not just for data reduction, can be used as a standalone classifier
- Invented by Fisher (1936), another classic dataset 'iris' was used (but we stay with PB)

LDA (binary case)

- We assume two classes $y = 0, 1$ and feature vectors $\vec{x}$. Also, we assume these are sampled from two normal distributions $p(\vec{x}|0)$ and $p(\vec{x}|1)$ which have means μ_0, μ_1 and covariances Σ_0, Σ_1 (for n dim we have a total of $n(n+1)$ parameters)
- Especially for biological distributions, normal assumption often makes very good sense
- $(\vec{x}-\vec{\mu}_0)^T\Sigma_0^{-1}(\vec{x}-\vec{\mu}_0)+\ln|\Sigma_0|-(\vec{x}-\vec{\mu}_1)^T\Sigma_1^{-1}(\vec{x}-\vec{\mu}_1)-\ln|\Sigma_1| > T$ (where T is the discriminant threshold)
- When we assume $\Sigma_0 = \Sigma_1$ (homoscedasticity) this simplifies to $\vec{x}\vec{w} > c$ where $\vec{w} = \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$ and $c = \vec{w} \cdot \frac{1}{2}(\vec{\mu}_1 + \vec{\mu}_0)$
- The hyperplane bisecting the vector connecting the means offers the best separation of the normal distributions (as long as they have the same variance)
- In high dimension, we may go for a *max margin* hyperplane instead, leading to SVMs
- Multiclass (m-way) is generally treated as a $\frac{1}{2}m(m-1)$ binaries