Advanced Machine Learning

András Kornai

AML 2024/09/25

Test results

- Relax, it doesn't count towards your grade!
- Problem 1: only a handful of good solutions (Benedek Máth)
- Problem 2: much easier, except some people defined independence
- Problem 3: a bit harder (Máté Szőke)
- Problem 4: again easy
- Problem 5: few people had time (Tamás Szűcs)
- If you have $\leq 50\%$ (19 points or less, 6 people) you need to study the parts where you didn't get points

Valuable prizes!

- Botond Bárdos Deák, Judit Hermán (PB)
- Máté Szőke, Benedek Máth, Blanka Kövér (test)
- Problem children JLVG03 DOC9GT BDQD5E K1TO74 ATM8YB S3VYH3 https://bit.ly/47CCfp0
- People should refine the PB problem (this will be HW)
- Leaderboard?

Linear stuff (Highleyman 1962)

A quote

One may rightly ask just why he should consider a linear decision function. Is there any guarantee that it will work? In general, this question can only be answered by designing the categorizer, and then deciding whether the resulting system is good enough.

- Even when you just take an algorithm off the shelf, working with linear methods is a good first step
- They are easy to visualize/understand
- Already give a good idea of where the problems are
- Help in selecting the "true" model

Data normalization

- Data matrix D has rows (observations, 1520 in PB data) and columns (measurements, 4 in PB)
- We begin with *means centering*, subtracting the mean of each column from each entry in the column
- Optionally, we may divide by the variance as well. (Note that constant column causing DIVZERO could be omitted to begin with!)
- Covariance matrix C is formed by arranging scalar products of col i and col j in (symmetric) square marix $C = D^T D$
- C is positive semidefinite (definite if data rows were linearly independent), variance in an arbitrary direction $\vec{x}$ is given by $\vec{x}^T C \vec{x}$
- To maximize this we need to solve

$$\frac{d}{d\vec{x}} \vec{x}^T C \vec{x} - \lambda \vec{x}^T \vec{x}$$

SVD to the rescue

- The Lagrange multiplier $-\lambda \vec{x}^T \vec{x}$ makes clear the critical points are solutions to $C\vec{x} = \lambda \vec{x}$
- So the solutions λ_i are, by definition, the eigenvalues and the x_i are the corresponding eigenvectors
- Let the SVD of D be UGV^T. The columns of V are the eigenvectors of C, and the positive singular values found in the diagonal matrix G (conventionally arranged to run from larger to smaller) are the square roots of the eigenvalues λ_i of C
- We use these to measure the "goodness" of the principal components. If $\Lambda = \sum_{i=1}^{c} \lambda_i$, we say that each λ_i *accounts for* a fraction λ_i/Λ of the total variance
- **Theorem** (Eckart–Young 1936) the first r columns of U, G, V can be used to form $U_r G_r V_r^T$ which is the best rank r approximation of D (in Frobenius norm)
- HW: do this for PB data, plot goodness as function of r
- Next time: LDA

1st homework

- People interested in speech should read Klautau (2002). Weka people can dig into the details
- Start stealing good ideas from the best notebooks! Everybody needs to call their own HW1_XY.ipynb, where XY are the first two letters of the authur's Neptun code. When you just took someone else's code, give credit in comments. (If you substantially reworked it, but the idea wasn't original with you, still give credit.)
- Your 1st homework is to build the best *prototype* classifier for the Peterson-Barney data, and comparing it to the (typically more advanced) method you used before under all four conditions (all train/all test; nonstarred train/all test; all train/nonstarred test; nonstarred train/nonstarred test)
- You can perform any kind of data manipulation you like, as long as your model is truly a prototye-based model
- This homework **will** be part of your grade. Due Sunday night

Project discussion

- Read Highleyman (1962) (everybody). 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 next week, if you can't come up with a plan a project will be assigned to you