MC886/MO444 Machine Learning and Pattern Recognition

Assignment #1 — 2017s2 — Prof. Sandra Avila

Objective

Explore linear regression alternatives and come up with the best possible model to the problems, avoiding overfitting. In particular, predict the release year of a song from audio features. Songs are mostly western, commercial tracks ranging from 1922 to 2011, with a peak in the year 2000s.

Activities

1. Perform Linear Regression (LR) as the baseline (first solution) and devise LR-based alternative (more powerful) solutions.
2. Use the specified training/test data for providing your results and avoid overfitting.
3. Devise and test more complex models.
4. Plot the cost function vs. number of iterations in the training set and analyze the model complexity. What are the conclusions? What are the actions after such analyses?
5. Use different Gradient Descent (GD) learning rates when optimizing. Compare the GD-based solutions with Normal Equations if possible (perhaps you should try with smaller sample sizes for this task). What are the conclusions?
6. Prepare a 4-page (max.) report with all your findings. It is UP TO YOU to convince the reader that you are proficient on linear regression and the choices it entails.

Dataset

This data is a subset of the Million Song Dataset: `http://labrosa.ee.columbia.edu/millionsong/` a collaboration between LabROSA (Columbia University) and The Echo Nest. Prepared by T. Bertin-Mahieux <tb2332@columbia.edu>.

Dataset Information:

- You should respect the following traininig/test split: $463,715$ training examples (year-prediction-msd-train.txt, and $36,285$ test examples (year-prediction-msd-test.txt).
- There are 90 attributes as follows: $12 =$ timbre average, $78 =$ timbre covariance. The first value is the year (target), ranging from 1922 to 2011. Features extracted from the 'timbre' features from The Echo Nest API. We take the average and covariance over all 'segments', each segment being described by a 12-dimensional timbre vector.
- The data is available at:
  `https://www.dropbox.com/s/21v6bnz0lkcv2us/year-prediction-msd-train.txt.zip`
  `https://www.dropbox.com/s/h1tavv5fwvi56nw/year-prediction-msd-test.txt.zip`

Deadline

Friday, **September 1st** in the beginning of the class, 7 pm. There will be no deadline extension.

Submission

On the deadline day, bring your 4-page printed report. The template for report is available at `https://www.dropbox.com/s/nc6d89otr8ekvjd/report-model.zip`. Please, print on both sides of the page.
The report should be written in Portuguese or English.
Submit a zip file, with the code and the report (PDF file), via Moodle.
This activity is NOT individual, it must be done in pairs (two-person group).