L1L2 Regression in Oncogenomics

Michele Filosi

FBK-MPBA: Predictive Models for Biomedicine and Environment

http://mpba.fbk.eu

Join with: C. Furlanello, G. Guzzetta, D. Albanese, G. Jurman, R. Visintainer, S. Riccadonna
Regression in MLPY
A Simulation study
  L1L2 vs LARS comparison
Application on Real Data
  Quantitative Phenotype
Regression in MLPY

- Ordinary Least Square (1)
- Ridge Regression (1)
- Kernel Ridge Regression
  - Linear
  - Gaussian
  - Polynomial
- Least Angle Regression (2)
- LASSO (3)
- Elastic Net (4,5)
- Gradient Descend

```python
>>> import numpy as np
>>> import mlpy
>>> ols = mlpy.RidgeRegression(alpha=0.0)
>>> rr = mlpy.RidgeRegression(alpha=alpha_opt)
>>> krr = mlpy.KernelRidgeRegression(kernel,alpha)
>>> krr.learn(x,y)
>>> kernel = mlpy.KernelLinear()
>>> kernel = mlpy.KernelGaussian(sigma)
>>> kernel = mlpy.KernelPolinomial(d)
>>> lars = mlpy.Lar(m)
>>> lasso = mlpy.Lasso(m)
>>> en = mlpy.ElasticNet(tau, mu)
>>> gd = mlpy.GradientDescendent(kernel,t, stepsize)
```

Regression model with embedded feature selection (De Mol et al 09)

1. Feature selection step

\[ \beta_{L_1L_2} = \arg \min \beta \frac{1}{n} \| Y - X \beta \|^2 + \mu \| \beta \|_2^2 + \tau \| \beta \|_1 \]

- A variant of the elastic net model (Zou&Hastie 05)
- Optimization problem:
  
  $\beta$: regression weights; $Y$: observed output; $X$: input data matrix; $\mu$, $\tau$: regularization parameters

- $\mu$ and $\tau$ modulate the selection of the features:
  - $\mu$ preserves correlation among selected features (Ridge Regression)
  - $\tau$ enforces the sparsity of the solution (LASSO)

2. Regression step

- Correction of the weight bias by an OLS regression on selected features (No parameters)
3. A New Implementation

- L1L2 with double optimization implemented in C, Python/NumPy, now a component of the mlpy package (https://mlpy.fbk.eu), using its functions for data import, handling and cross-validation. Parallelized for HPC.

4. Previous Results

- Classification from real predictions (DeMol et al., 2009)
- Gene expression data of:
  - leukaemia (n=72, p=7,000)
  - lung cancer (n=181, p=12,000)
  - prostate cancer (n=102, p=12,000)
- Classification from real predictions (Fardin et al., 2009)
- Gene expression data of:
  - neuroblastoma cell lines (n=18, p=50,000)
  - A signature for hypoxia
A Simulation Study

Compare the feature selection performances and prediction ability

DATASETS:
- SAMPLES n=100
- FEATURES 70000... 70

Parameter tuning
Montecarlo 20x
Split ¼ – ¾

L1L2
Feature selection

Lars/Lasso
Feature selection

OLS Weights

Stability
Steps
Prediction
MSE

L1L2
VS
Lars

N=70
N=30
N feat.
A Simulation Study

Compare the feature selection performances and prediction ability

L1L2 vs LARS comparison:
- MSE curve almost the same
- Selected feature: rel_feat / sel_feat
- Union Stability: t_feat_sel / n_resampl.

How the features were built:
- 2.5% of relevant features
- Relevant features grouped by 5 correlated
- Correlation within group >= 95%
- Noise added to label with \( \sigma = 1e-2 \)
Protocol inspired by the **MAQC-II project guidelines** (Shi et al., 2010)

- Split dataset in **development/validation**
- For each experiment run a **10x Montecarlo 1/4 – 3/4** on the development set
- Select **optimal parameters**, \((\mu^*, \tau^*)\) according to average prediction performances across the 10 test sets
  - **Accuracy**: MSE or \(R^2\)
- Run **L1L2** on the **whole development set** with the best parameters \((\mu^*, \tau^*)\)
- Select **features** with L1L2
- Compute the **weights** with **OLS**
- Predict on validation dataset
Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome


A. Diagnostic classifier: MDS vs AML by Gene Expression Profiling (GEP)

B. Prognostic Quantitative phenotype: time of AML transformation from diagnosis (originally a classification: within 18 mo. or later)
From classifiers to eQTL

1. AML vs MDS Diagnostic Classifier
2. eQTL on the best feature
   - Prediction of correlated features
   - Gene Pathways

DAP: FDA-MAQC-II
Classifier: SRDA

MCC
BORDA List
Best Feature

eQTL
L1L2 Prediction of Time to AML

Prediction of risk to transformation

N = 22 Transformed to AML

P = 54614 features

DATASET

L1L2

Feature Selection

Prediction

Cutoff value

Gene Signature

Cutoff

Time to AML

Gene Expression
Regression Approaches to Quantitative Phenotype

Open Problems:

- Using OLS to correct regression weight?
- Does a RLS perform better?
- How to find function correlated features to the most discriminant?
- Application on SNP data:
  - Which encoding do we use?