Machine Learning Network-Constrained Regression of Epigenetic Data

Sivo Vladimirov Daskalov

Corpus Christi College

28 June 2017

UNIVERSITY OF
CAMBRIDGE

Outline

Epigenetic background

Project goals

Penalized regression methods

Composite voting regression

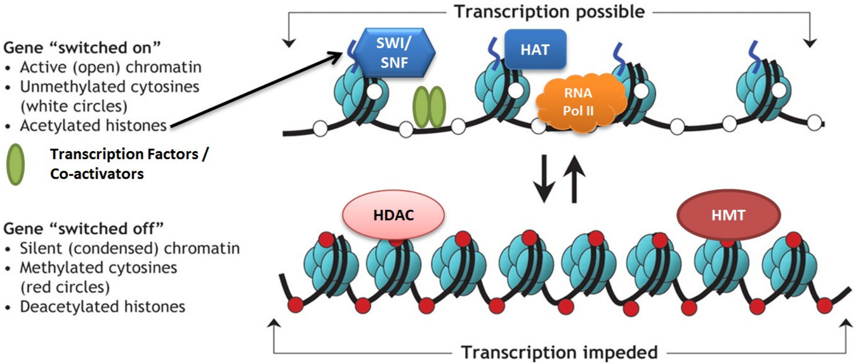
Orchestrated hyperparameter tuning

Experiments on synthetic datasets

Breast cancer dataset analysis

Summary

Epigenetic background



Gene “switched on”

- Active (open) chromatin
- Unmethylated cytosines (white circles)
- Acetylated histones

Transcription Factors / Co-activators

Gene “switched off”

- Silent (condensed) chromatin
- Methylated cytosines (red circles)
- Deacetylated histones

*Figure is adapted from Luong, P. Basic Principles of Genetics

Project goals

Question:
How is the expression of each gene affected by the methylation of related genes?

Approach:

$$\text{Linear regression} \begin{cases} \text{Predictors: methylation levels for all genes} \\ \text{Target variable: expression level for gene of interest} \end{cases}$$

Penalized regression methods

Lasso	$\lambda \sum_{i=1}^{p} \lvert\beta_i\rvert$
Elastic Net	$\lambda_1 \sum_{i=1}^{p} \lvert\beta_i\rvert + \lambda_2 \sqrt{\sum_{i=1}^{p} \beta_i^2}$
Grace	$\lambda_1 \sum_{i=1}^{p} \lvert\beta_i\rvert + \lambda_2 \sum_{u\sim v} \left(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}}\right)^2 w(u,v)$
aGrace	$\lambda_1 \sum_{i=1}^{p} \lvert\beta_i\rvert + \lambda_2 \sum_{u\sim v} \left(\frac{sign(\tilde{\beta}_u)\beta_u}{\sqrt{d_u}} - \frac{sign(\tilde{\beta}_v)\beta_v}{\sqrt{d_v}}\right)^2 w(u,v)$
GBLasso	$\lambda \sum_{u\sim v} \left[\left(\frac{\lvert\beta_u\rvert}{\sqrt{d_u}}\right)^{\gamma} + \left(\frac{\lvert\beta_v\rvert}{\sqrt{d_v}}\right)^{\gamma}\right]^{1/\gamma}$
Linf	$\lambda \sum_{u\sim v} \max\left(\frac{\lvert\beta_u\rvert}{\sqrt{d_u}}, \frac{\lvert\beta_v\rvert}{\sqrt{d_v}}\right)$
aLinf	$\lambda \sum_{u\sim v} \left\lvert \frac{sign(\tilde{\beta}_u)\beta_u}{\sqrt{d_u}} - \frac{sign(\tilde{\beta}_v)\beta_v}{\sqrt{d_v}} \right\rvert$
TTLP	$\lambda_1 \sum_{i=1}^{p} J_\tau \lvert\beta_i\rvert + \lambda_2 \sum_{u\sim v} \left\lvert J_\tau\left(\frac{\lvert\beta_u\rvert}{w_u}\right) - J_\tau\left(\frac{\lvert\beta_v\rvert}{w_v}\right) \right\rvert$
LTLP	$\lambda_1 \sum_{i=1}^{p} \lvert\beta_i\rvert + \lambda_2 \sum_{u\sim v} \left\lvert J_\tau\left(\frac{\lvert\beta_u\rvert}{w_u}\right) - J_\tau\left(\frac{\lvert\beta_v\rvert}{w_v}\right) \right\rvert$

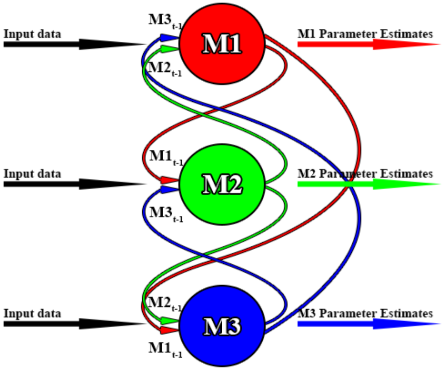
Composite voting regression

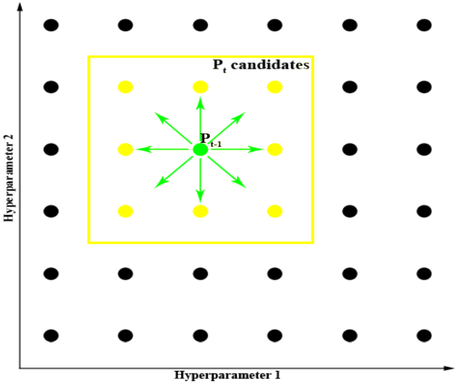
	X_1	X_2	...	X_p
Method 1	$M_1(\beta_1)$	$M_1(\beta_2)$	...	$M_1(\beta_p)$
Method 2	$M_2(\beta_1)$	$M_2(\beta_2)$	...	$M_2(\beta_p)$
...	...	...	...	...
Method *k*	$M_k(\beta_1)$	$M_k(\beta_2)$	...	$M_k(\beta_p)$

$$X_j = \begin{cases} \textit{important}, & \text{if } \frac{\sum_{i=1}^{k}[M_i(\beta_j) \neq 0]}{k} \geq \text{fraction of votes threshold} \\ \textit{unrelated}, & \text{otherwise} \end{cases}$$

Final model obtained from OLSE on the set of important predictors

Orchestrated hyperparameter tuning





Synthetic dataset generation and setup

Synthetic dataset generation

- Designed to be similar to real epigenetic datasets
- 20 datasets with 550 predictors and differently generated responses
- Training and test sets of size 300 and 100 respectively

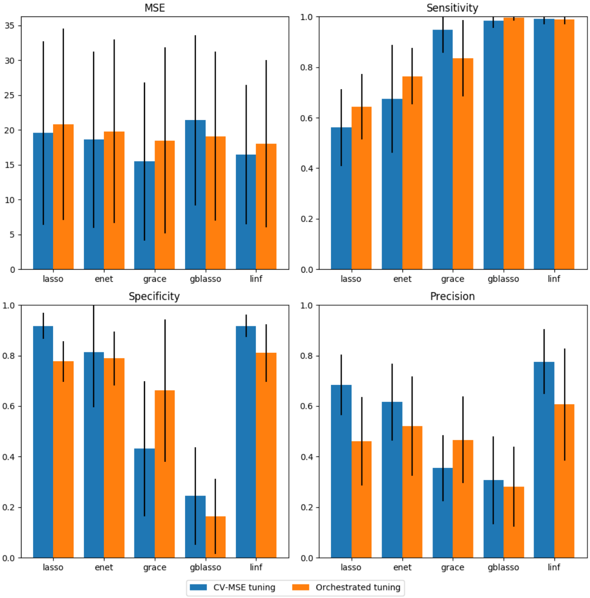
Hyperparameter tuning setup

Search space: Predefined parameter grids for all regression methods

CV-MSE tuning: Traditional 5-fold cross-validated mean squared error

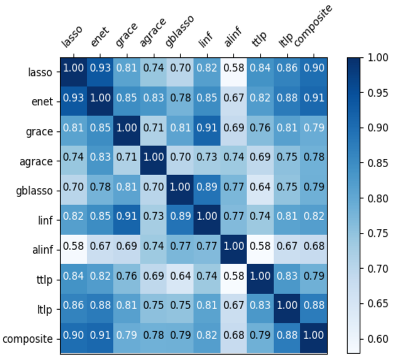
Orchestrated tuning: Starting points obtained from the CV-MSE tuning

Comparison of model metrics



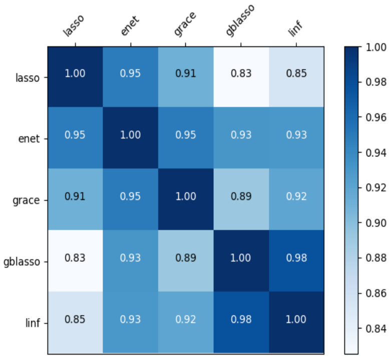
Regression method similarity evaluation

Cosine similarity between estimated coefficient vectors



	lasso	enet	grace	agrace	gblasso	linf	alinf	ttlp	ltlp	composite
lasso	1.00	0.93	0.81	0.74	0.70	0.82	0.58	0.84	0.86	0.90
enet	0.93	1.00	0.85	0.83	0.78	0.85	0.67	0.82	0.88	0.91
grace	0.81	0.85	1.00	0.71	0.81	0.91	0.69	0.76	0.81	0.79
agrace	0.74	0.83	0.71	1.00	0.70	0.73	0.74	0.69	0.75	0.78
gblasso	0.70	0.78	0.81	0.70	1.00	0.89	0.77	0.64	0.75	0.79
linf	0.82	0.85	0.91	0.73	0.89	1.00	0.77	0.74	0.81	0.82
alinf	0.58	0.67	0.69	0.74	0.77	0.77	1.00	0.58	0.67	0.68
ttlp	0.84	0.82	0.76	0.69	0.64	0.74	0.58	1.00	0.83	0.79
ltlp	0.86	0.88	0.81	0.75	0.75	0.81	0.67	0.83	1.00	0.88
composite	0.90	0.91	0.79	0.78	0.79	0.82	0.68	0.79	0.88	1.00

CV-MSE tuning



	lasso	enet	grace	gblasso	linf
lasso	1.00	0.95	0.91	0.83	0.85
enet	0.95	1.00	0.95	0.93	0.93
grace	0.91	0.95	1.00	0.89	0.92
gblasso	0.83	0.93	0.89	1.00	0.98
linf	0.85	0.93	0.92	0.98	1.00

Orchestrated tuning

Breast cancer dataset and setup

Dataset properties:

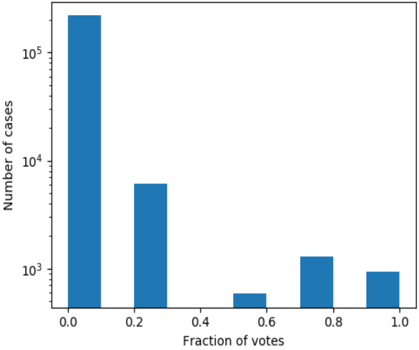
- Obtained from The Cancer Genome Atlas (TCGA)
- Methylation and expression data for 215 breast cancer patients
- Selected subset of genes associated with breast cancer
- Samples divided in 3/4 training and 1/4 test sets

Methylation data from the promoter and gene body regions considered separately

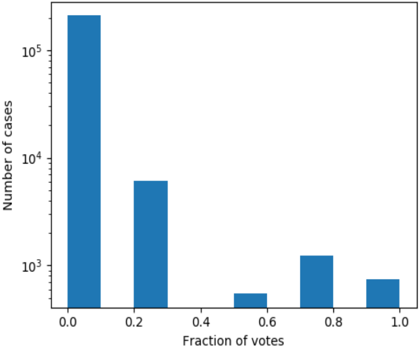
Regression methods used: Lasso, Elastic Net, Grace, Linf and the proposed Composite Voting Regression

Vote fraction distribution

Threshold of 0.75 chosen (3 out of 4 methods must agree)



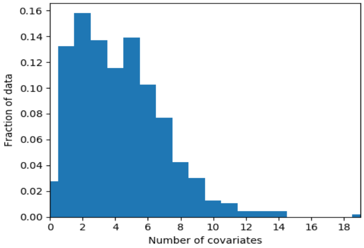
Gene body region

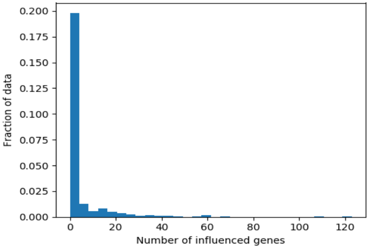


Gene promoter region

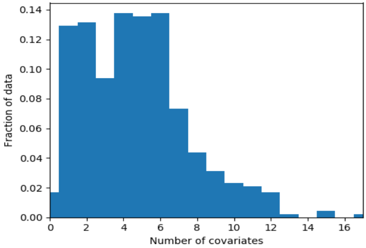
Distribution of dependencies

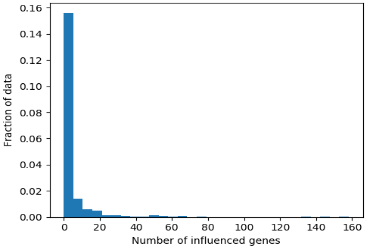
Composite voting regression on the gene promoter region





Composite voting regression on the gene body region





Summary

- Implementation of 9 regression methods found in literature
- Composite voting regression
- Orchestrated hyperparameter tuning
- Comparison and evaluation on synthetic datasets
- Exploration of a real breast cancer dataset

Contact details:
sivodaskalov@gmail.com