Graph-Sparse Logistic Regression

Alexander LeNail[1], Ludwig Schmidt[2], Jonathan Li[1], Tobias Ehrenberger[1], Karen Sachs[1], Stefanie Jegelka[2], Ernest Fraenkel[1]

[1]MIT BE, [2]MIT CSAIL

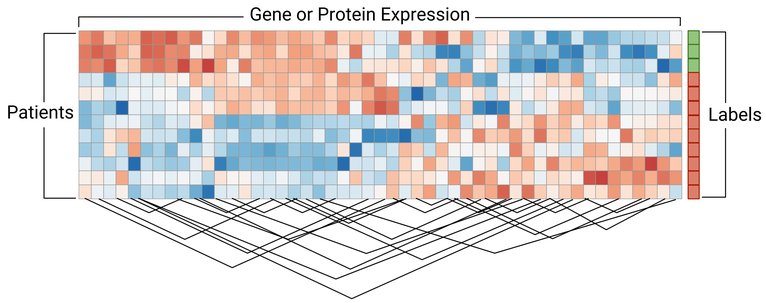
Introduction

Scientists often use machine learning to **learn more about their data**, not to predict it. They want to **interpret** their models.
The simplest way to build an interpretable model is to keep it **small**.
The classic way to accomplish this is with **LASSO** (L1-regularization).

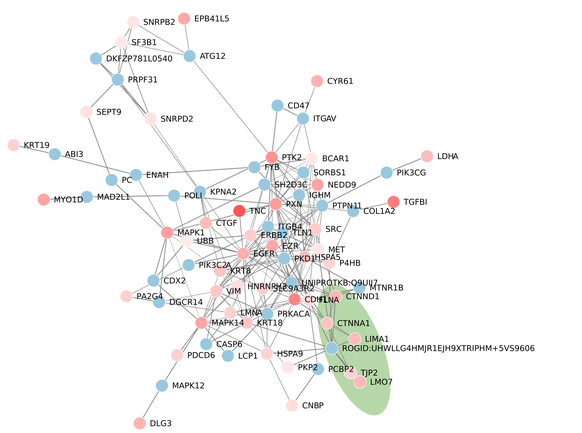
Problem setup

The LASSO does not always find the "right" model. In some cases, we can leverage side-information to help find the right model.

In our case, side-information is graphical structure among the features.



We can re-draw each instance as a labeling of the graph's nodes:



The "right" model is a "graph-sparse" model, i.e. a model with a sparse support which is connected on a graph. The goal is therefore to find the **most predictive connected subgraph**.
(possibly add the logistic loss function here?)

Our approach

function $\mathrm{GSLR}(X,\ y,\ G,\ s,\ \eta,\ k)$
  Let $f(X, y, \theta)$ be the logistic loss.
  $\hat{\theta}^0 \leftarrow 0$
  for $i \leftarrow 0, \ldots, k-1$ **do**
    $\tilde{\theta}^{i+1} \leftarrow \hat{\theta}^i - \eta \cdot \nabla f(X, y, \hat{\theta}^i)$
    $\hat{\theta}^{i+1} \leftarrow P_{G,s}(\tilde{\theta}^{i+1})$ ▷ Graph-sparse projection
  return $\hat{\theta}^k$

Main idea:

Gradient Update
via softmax loss function

Sparse Projection
via Prize-Collecting Steiner Forest

Sparse Projection Operator:
Given an arbitrary vector $p \in \mathbb{R}^d$, the projection operator $P_{G,s}$ returns a vector $q \in \mathbb{R}^d$ satisfying the following two properties:

- **Approximate projection:** The vector q is an approximate projection, i.e., instead of achieving the smallest distance to the input point p among points in the constraint set, the distance achieved by q is within a small constant factor:

$$\|p - q\|_2^2 \ \leq\ 2 \cdot \min_{q' \text{ is } (G,s)\text{-sparse}} \|p - q'\|_2^2 \,. \qquad (1)$$

- **Approximate graph sparsity:** The support of the vector q forms a connected component of size at most $6s + 1$ in the graph G.

Graph-Sparse Projection through PCSF:
The Graph-Sparse projection operator $P_{G,s}$ is a carefully-tuned set of Prize-Collecting Steiner Forest instances.
The PCSF objective is to minimize:

$$f(F) = \beta \sum_{v \notin V_F} p(v) + \sum_{e \in E_F} c(e) + \omega \cdot \kappa, \qquad (2)$$

Intuitively, the goal is to pay as little edge cost to connect the highest prize nodes.
We project the parameter vector *theta* onto the graph by setting it as the prize for PCSF.

Synthetic Data

Since we don't have the ground truth subgraphs for the real data, we create synthetic data by sampling from the multivariate gaussian defined by the real data, then translating "positive" examples by a "perturbation" vector. $\textit{negative} = \mathcal{N}(\vec{\mu}, \Sigma)$ $\quad \textit{positive} = \mathcal{N}(\vec{\mu}, \Sigma) + \vec{x}$

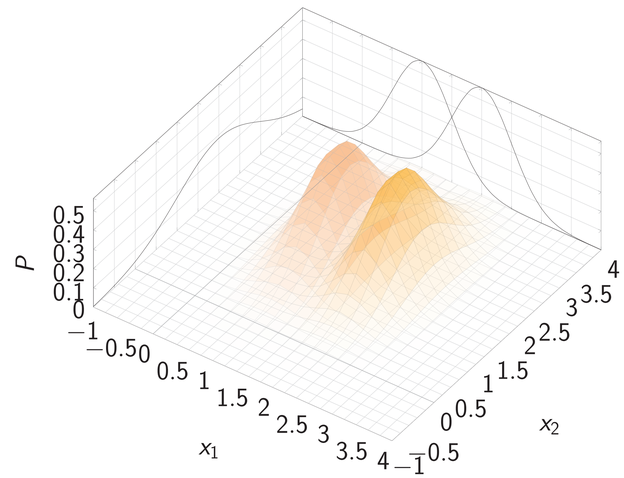
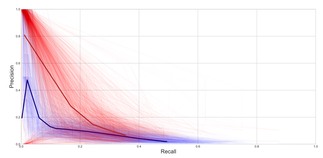


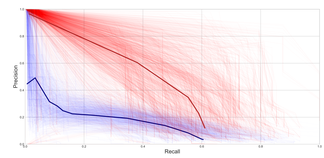
Figure: A low dimensional cartoon of our synthetic data generation strategy. Here, the pink gaussian represents our original data, from which we sample our negative examples. We sample our positive examples from the orange gaussian by first sampling them from the pink gaussian and translating them by the perturbation vector (in this case $< 0, \mu_{x2} - \mu_{x2}* >$, with x_2 in the pathway and x_1 not.

Experiments

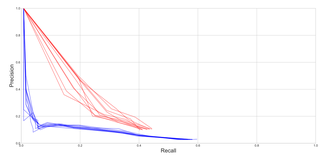
We benchmark our technique against the LASSO by how many of the truly "perturbed" features each uses in its support.



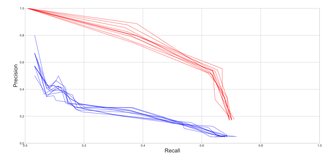
(a) scheme 1 – all datasets, all folds



(b) scheme 2 – all datasets, all folds



(c) scheme 1 – random dataset, all folds



(d) scheme 2 – random dataset, all folds

Figure: L1-Regularized Logistic Regression (blue) versus Graph-Sparse logistic Regression (red). Each trace represents one fold of one dataset, varying sparsity. The bolded trace is the average.

We then use our technique on real Ovarian Cancer data, and find that the support chosen by GSLR is qualitatively superior.