COMP 790-124: Goals for today

- Projects

- Notation
- Elastic Net penalty
- Choosing λ, α
- Gene expression modeling: penalized regression application

Projects: LaTeX BibTeX PubMed

LaTeX: `http://tobi.oetiker.ch/lshort/lshort.pdf`
PubMed: `http://www.ncbi.nlm.nih.gov/pubmed`
PubMed/BibTex: `http://www.bioinformatics.org/texmed/`

ML and applications

Computer Science favors conferences

- Computer Vision conferences: ECCV, CVPR, ICCV ...
- Natural Language processing: ACL, EMNLP, CoNLL ...
- CompBio conferences: ISMB, RECOMB, ACM BCB, PSB
- ML conferences/workshops: ICML, AIStats, UAI, NIPS
- Data mining conferences: KDD, ICDM

Browse through conference proceedings for your area and ML in the past couple of years.

This will give you an idea of popular problems, sources of data, etc. Reading conclusions of papers can help you formulate an idea.

Project proposals

Pick an area, read a review article and related papers in the area, obtain data.

Write a 1pg proposal in format available from the webpage.

If not sure about an area, come talk to me.

A one page proposal by 10/1, LaTeX format is posted.

Questions?

Notation

Recall our trajectory last time:

- write probabilistic model of our data
- write objective down in terms of parameters θ (log likelihood or log posterior)
  - write optimization problem
- optimize objective with respect to the parameters θ

Notation

There is a standard way to write optimization problems

$$\begin{array}{ll} \underset{\boldsymbol{\theta} \in \mathbf{R}^n}{\text{minimize}} & f(\boldsymbol{\theta}) \\ \text{subject to} & g(\boldsymbol{\theta}) = 0 \\ & h(\boldsymbol{\theta}) \leq 0 \end{array}$$

and also a standard way to denote minimizers (optimal values for $\boldsymbol{\theta}$)

$$\underset{\boldsymbol{\theta} \in \mathbf{R}^n, g(\boldsymbol{\theta})=0, h(\boldsymbol{\theta}) \leq 0}{\text{argmin}} \quad f(\boldsymbol{\theta})$$

Notation – Lasso optimization problem

Recall the objective, a likelihood function restricted to terms involving $\boldsymbol{\beta}$

$$-1/2 \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i' \boldsymbol{\beta})^2 - \lambda \sum_{j=1}^{p} |\beta_j|$$

In the standard form for optimization problems we write

$$\underset{\beta_0 \in \mathbf{R}, \boldsymbol{\beta} \in \mathbf{R}^p}{\text{minimize}} \quad \tfrac{1}{2} \textstyle\sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i' \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Note the sign flip: we maximize likelihoods/probabilities but we minimize costs/losses.

Questions?

Elastic Net [2]

Ridge regression as a means to deal with correlated predictors.

- Problem: too many predictors are non-zero

Lasso as means to deal with large number of predictors.

- Problem: arbitrary decisions between highly correlated predictors

How to combine sparsity of lasso and "equal" weighting of ridge?

Elastic Net

Interpolate between the two costs/priors

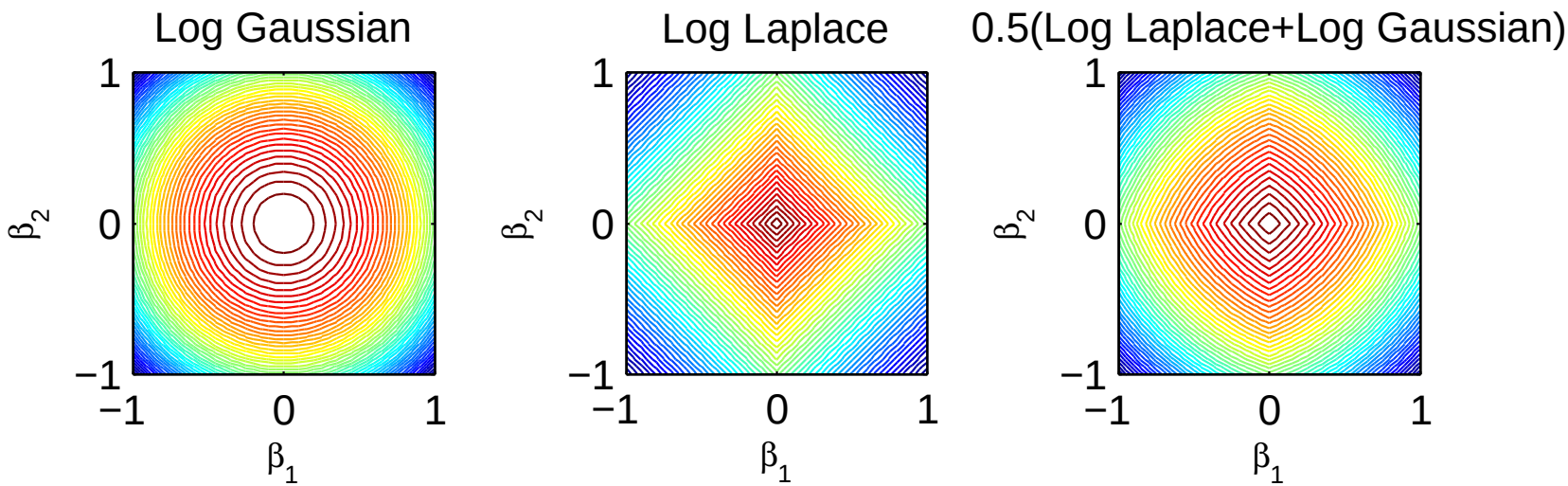
$$-1/2 \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i' \boldsymbol{\beta})^2 - \alpha \frac{\lambda}{2} \sum_{j=1}^{p} \beta_j^2 - (1-\alpha)\lambda \sum_{j=1}^{p} |\beta_j|$$

In the extremes of α:

- $\alpha = 0$ recovers lasso
- $\alpha = 1$ recovers ridge

Since you get to choose α, you should not do worse than lasso or ridge.

Elastic Net Level Curves[1]



$$-1/2 \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 - \alpha \frac{\lambda}{2} \sum_{j=1}^{p} \beta_j^2 - (1-\alpha)\lambda \sum_{j=1}^{p} |\beta_j|$$

[1] make your own: use `meshgrid` and `contour` in Matlab

Elastic Net Optimization: Coordinate Ascent [1]

We are still in the world of coordinate ascent and assuming normalized predictors.
Taking derivatives of the objective, setting them to zero, and reasoning by cases, we get:

$$\begin{aligned}
\beta_0 &= \frac{1}{n}\sum_i y_i \\
\beta_j &= \frac{1}{1+(\alpha\lambda)} S\left(\sum_i y_i^{(-j)} x_{i,j}\ ,\ (1-\alpha)\lambda\right),
\end{aligned}$$

where

$$\begin{aligned}
S(x,\lambda) &\equiv \operatorname{sign}(x)\max(|x|-\lambda, 0) \\
y_i^{(-j)} &\equiv y_i - \beta_0 - \sum_{k\neq j} x_{i,k}\beta_k
\end{aligned}$$

Demo on some real data

Questions?

Choosing λ, α

Given data $\{(\mathbf{x}_i, y_i) : i = 1..n\}$, λ, α we know how to produce $\beta_0, \boldsymbol{\beta}$.

It is fair to say that $\beta_0, \boldsymbol{\beta}$ are functions of data, λ, and α

Data is given but λ, α are chosen.

Choosing λ, α

Assume the training dataset is drawn from $p(x, y)$

Ideally we want to choose λ and α such that the resulting $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ give small error on future data.

$$E_p[(y - \hat{\beta}_0 - \mathbf{x}'\hat{\boldsymbol{\beta}})^2] = \int_{\mathbf{x}} \int_{y} p(\mathbf{x}, y)(y - \hat{\beta}_0 - \mathbf{x}'\hat{\boldsymbol{\beta}})^2 d\mathbf{x} dy$$

We don't have the capability to compute this, since the true distribution $p(\mathbf{x}, y)$ is hidden from us.

But, we can emulate future draws.

Choosing λ, α

Randomly split data into train and test (50/50,80/20).

Let Test denote indices of data that are in the test set and Train indices of data that are in the train set. $\mathrm{Test} \bigcap \mathrm{Train} = \emptyset$

Learn $\hat{\beta}_0, \hat{\boldsymbol{\beta}}$ on $\{(\mathbf{x}_i, y_i) : i \in \mathrm{Train}\}$ using some λ, α

The estimated test error

$$\mathrm{Err}_{\mathrm{Test}}(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \frac{1}{|Test|} \sum_{i \in Test} [(y_i - \hat{\beta}_0 - \mathbf{x}_i' \hat{\boldsymbol{\beta}})^2]$$

in the limit of the test set size, tends to

$$E_p[(y - \hat{\beta}_0 - \mathbf{x}' \hat{\boldsymbol{\beta}})^2]$$

Choosing λ, α

All this to say that we can use a test set error to evaluate our performance on unseen data.

We can choose α,λ so as to minimize this proxy (test error).

Cross-validation

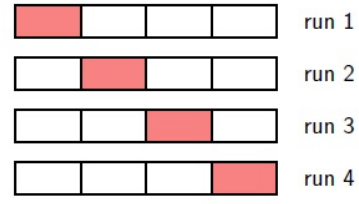
K-fold cross-validation: divide dataset into k subsets, call them $\mathrm{Set}_1, \ldots \mathrm{Set}_k$

Foreach i=1:k

1. Set $\mathrm{Test} = \mathrm{Set}_i$ and $\mathrm{Train} = \bigcup_{k \neq i} \mathrm{Set}_k$
2. Learn β_0 and $\boldsymbol{\beta}$ on $\{(\mathbf{x}_i, y_i) : i \in \mathrm{Train}\}$ using α, λ
3. $\mathrm{CVErr}_i(\alpha, \lambda) = \frac{1}{|\mathrm{Test}|} \sum_{i \in \mathrm{Test}} [(y_i - \hat{\beta}_0 - \mathbf{x}_i' \hat{\boldsymbol{\beta}})^2]$

$\mathrm{CVErr}(\alpha, \lambda) = \frac{1}{k} \sum_k \mathrm{CVErr}_k(\alpha, \lambda)$

4-fold illustration



Extreme $k = n$, size of dataset, then we obtain Leave-One-Out (LOO) scheme. And resulting estimate of error is LOO CV estimate.

CVErr above is a mean of fold specific errors; report standard deviations as well.

Choosing λ, α

So we can score a pair λ, α in a number of ways

How do we do search?

A universal approach: grid of values

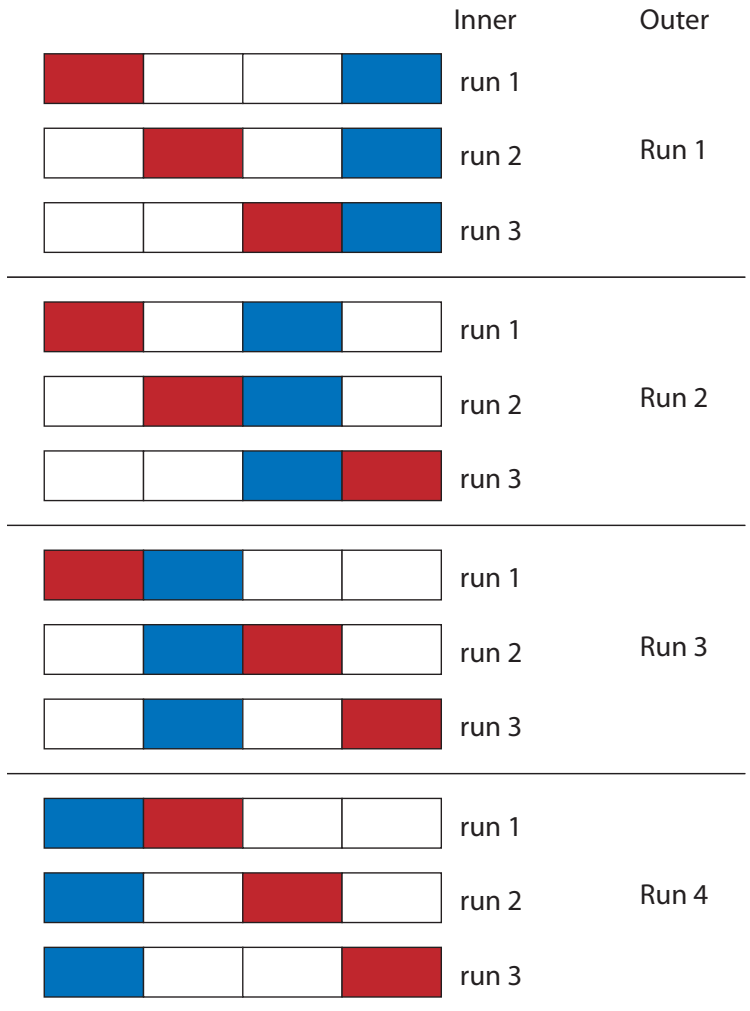
Question for you

Observation: The elastic net fitting method consists of fitting $\beta_0, \boldsymbol{\beta}$ *and* choosing α, λ via cross validation.

How do we estimate the performance of a method that involves cross validation as a part of training?

Question for you

Wrap it in another cross validation: $\overbrace{\underbrace{\text{Train}, \text{Validate}}_{\text{inner}}, \text{Test}}^{\text{outter}}$



Questions?

We did ...

- Elastic Net
- Prediction error and cross validation
- Saw a vignette for Elastic Net


Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani.
Regularization paths for generalized linear models via coordinate descent.
Journal of Statistical Software, 33(1):1–22, 2 2010.


Hui Zou and Trevor Hastie.
Regularization and variable selection via the elastic net.
Journal of the Royal Statistical Society Series B, 67(2):301–320, 2005.