Lecture 9: Feature Selection

Course: Biomedical Data Science

Parisa Rashidi
Fall 2018

Outline

- Wide data and its challenges
- Feature selection
  - Filter-based
  - Wrapper
  - Embedded

Wide Data

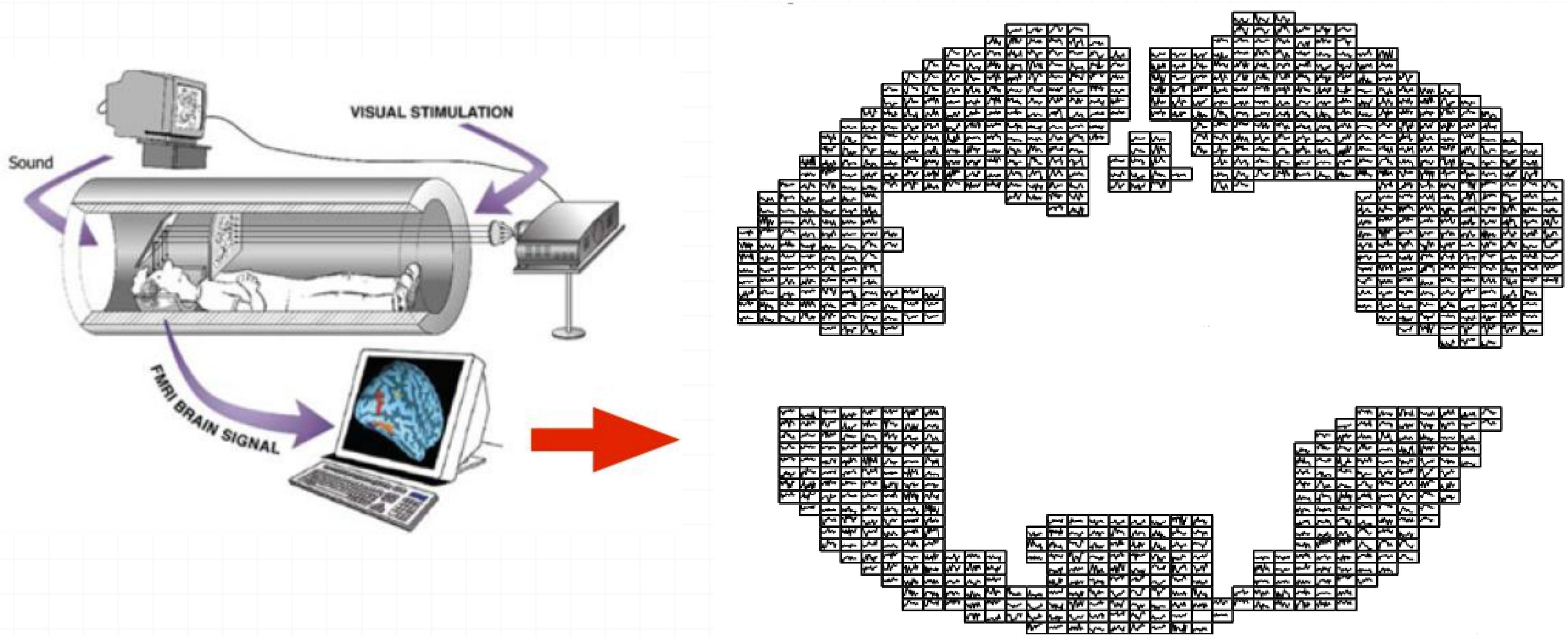
- Datasets with many more features than samples:

$$\text{Samples}\left\{\overbrace{\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ . & \cdots & . \\ x_{N1} & \cdots & x_{Np} \end{pmatrix}}^{\text{Features}}\right.$$

Wide data, i.e.
$\#\text{features} \gg \#\text{Samples}$

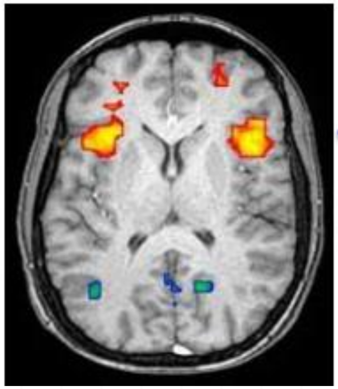
[Bio] fMRI Data

- A series of images, each containing many voxels



Wide Data Examples

- Genomics, microarray studies: p = 40K genes are measured for each of N = 100 subjects.
- Genome-wide association studies: p = 1~2M SNPs measured for N = 2000 case-control subjects.
- Predicting word stimulus from fMRI data



Thinking about a house? or a Dog?

Feature Selection

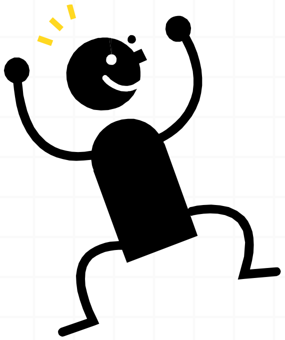
- Reduce dimensionality (curse of dimensionality) by removing:
  - Redundant or highly correlated features
    - Purchase price and sales tax paid
  - Irrelevant features
    - Student ID number for predicting student's GPA
  - Noisy features
    - signal-to-noise ratio too low to be useful for discriminating

Feature Selection Approaches

- Filter approaches:
  - Features are selected before applying the machine learning techniques, typically using a score
- Wrapper approaches:
  - Use machine learning algorithm as a black box to find the best subset of features
- Embedded:
  - Feature selection occurs naturally as part of the learning process
    - L1/L2 regularized linear regression

Benefits of Feature Selection

- Curse of dimensionality will be alleviated
- Model becomes easier to interpret
- Generalization power will be improved
- Learning will be sped up



Individual vs. Subset Feature Selection

- Individual feature selection
  - It selects one feature at a time
- Subset selection
  - It tries to select a subset of features, as opposed to single features

Filter Approach

Filter Approach

- Heuristic measures are used to calculate the relevance of a feature.
  - E.g. correlation of a feature with label
- The feature are ranked and sorted.
- We choose the top k features.

Filter Approach

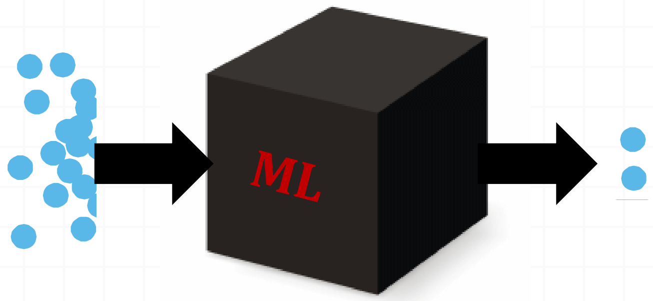
- Fast, very inexpensive
- With the cost of computation going down, it is best to include the learner in the loop.
  - There is no guarantee that the heuristic will match the bias of the learner.

Wrapper Approach

*http://tinyurl.com/q3jlklu

Wrapper Approaches

- Use machine learning algorithm as a black box to find the best subset of features
- There are 2^d subsets of d features (so a brute force approach will not work!)
- Subset selection methods
  - Forward selection
  - Backward selection
  - Floating search (Add k, remove l)



Forward Selection

- Start with no features.
- Add another feature at each step
  - Which feature: the one that decreases the error the most
  - Error: MSE or misclassification rate
- Continue adding features until:
  - Adding the next feature does not decrease the error.

Backward Selection

- Start with all features.
- Remove another feature at each step
  - Which one: the one that decreases the error the most
  - Error: MSE or misclassification rate
- Continue removing features until
  - Removing the next features does not decrease the error

Notes on Subset Selection

- Computational complexity higher than filter approaches
- A greedy approach
  - Local search, adding features one by one, so there is no guarantee for finding the optimal solution
  - Variations
    - Add multiple features at a time
    - Add backtracking
    - Floating search (Add k, remove l)

Embedded Approaches

The following slides are partially based on: glmnet Webinar, Trevor Hastie, Stanford Statistics, Link

Linear Regression

- We try to minimize the difference between predicted values and the actual values.

$$\boxed{\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2}$$

Actual Value

Predicted Value

Minimize their difference (Loss Function)

Linear Models for Wide Data

- The linear model has regained favor as the tool of choice for wide data.
- However, Since $p \gg N$, we cannot fit these models using standard approaches.
  - We need to consider some constraints.
    - Regularization

Regularized Models (Embedded Approaches)

- Ridge Regression
- Lasso
- Elastic Net

Ridge Regression

- Ridge regression is similar to linear regression, but adds the constraint $\sum_{j=1}^{p} \beta_j^2 \leq t$
- Shrinks coefficients toward zero, and hence controls variance.

$$\boxed{\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2}$$

Loss Function (function of β)

Constraints on β

$$\boxed{\sum_{j=1}^{p} \beta_j^2 \leq t}$$

Lasso

- Lasso regression is similar to linear regression, but adds the constraint $\sum_{j=1}^{p} |\beta_j| \leq t$
- Lasso does variable selection and shrinkage, while ridge only shrinks.

$$\boxed{\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2}$$

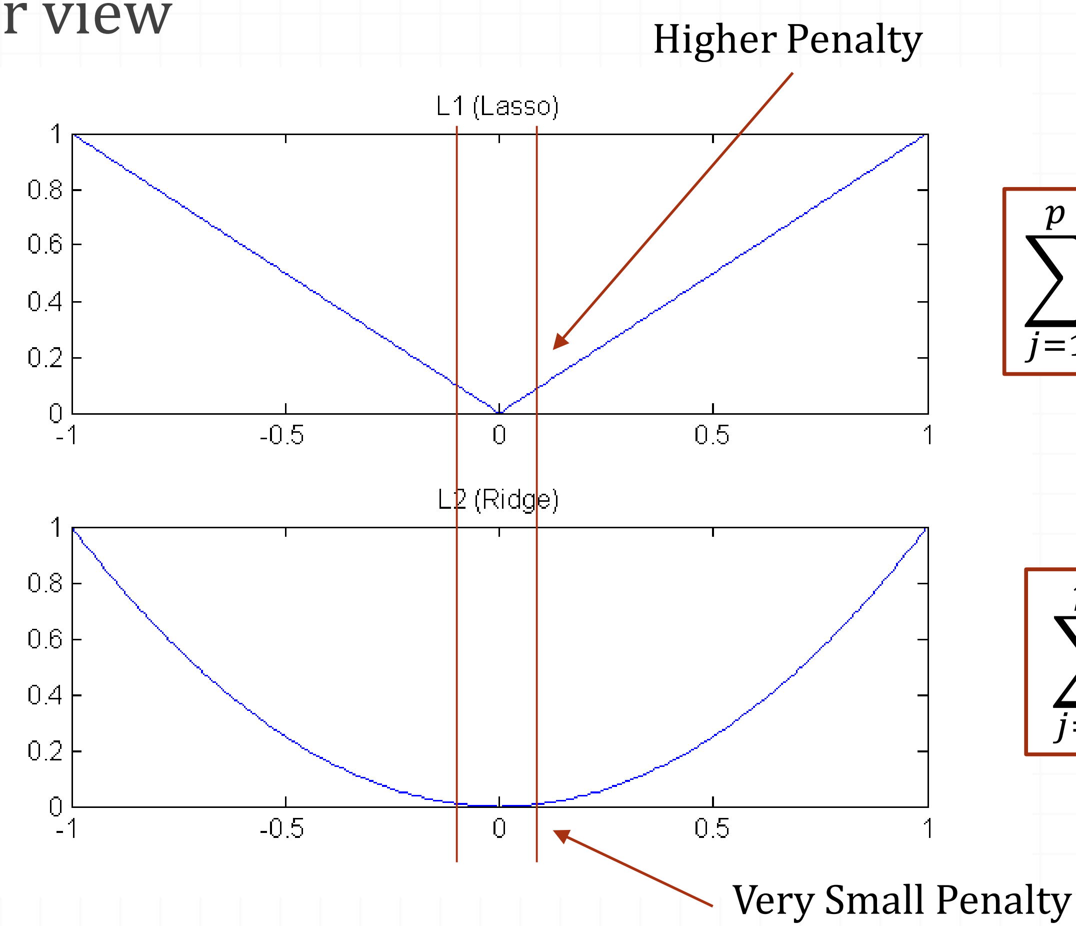
Loss Function (function of β)

$$\boxed{\sum_{j=1}^{p} |\beta_j| \leq t}$$

Constraints on β

Lasso vs. Ridge

- Another view

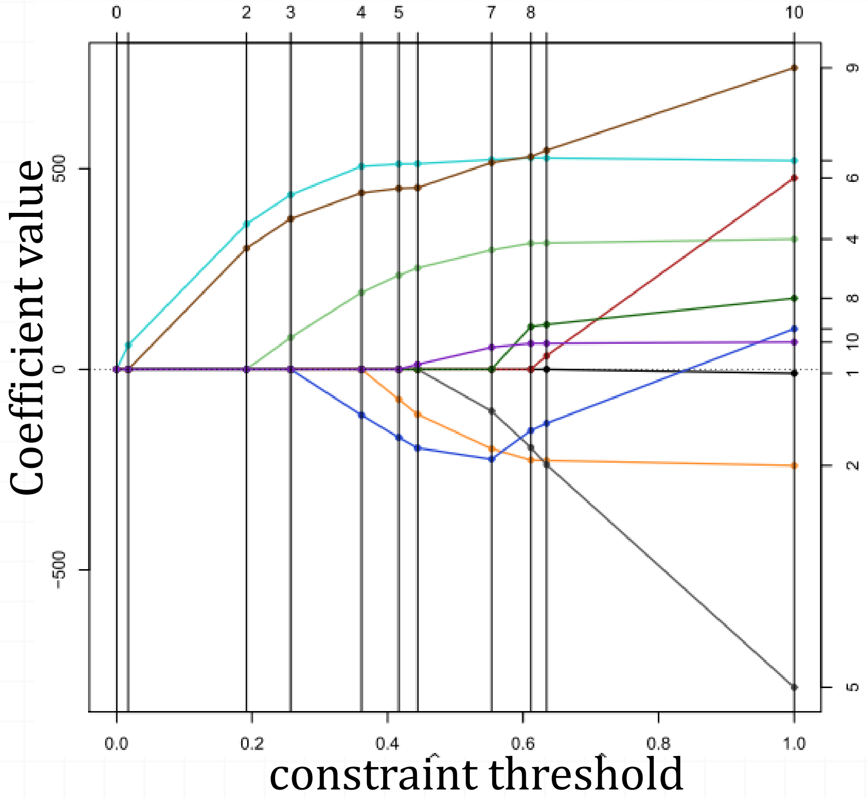


$$\sum_{j=1}^{p} |\beta_j| \leq t$$

$$\sum_{j=1}^{p} \beta_j^2 \leq t$$

Coefficient Path

- Each path shows the value of a coefficient vs. constraint threshold



Elastic Net

- A combination of Lasso and Ridge

$$\min_\beta \frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} P_\alpha(\beta_j)$$

with $P_\alpha(\beta_j) = \frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|.$

Norm 2
(Ridge)

Norm 1
(Lasso)

α creates a compromise between the lasso and ridge.

All Three

- Elastic provides a compromise between Lasso and Ridge