Feature selection and regularization

Fraida Fund

Contents

Feature selection

Problem: given high dimensional data $\mathbf{X} \in R^{N \times p}$ and target variable y,

Select a subset of $k << p$ features, $\mathbf{X}_S \in R^{N \times k}$ that is most relevant to target y.

Motivation for feature selection problem

- Limited data
- Very large number of features
- Examples: spam detection using "bag of words", EEG, DNA MicroArray data

Feature selection

Many possible models

- Given n features, there are 2^n possible feature subsets
- Feature selection is model selection over 2^n models - too expensive for large n

Feature selection methods

- **Wrapper methods**: use learning model on training data, and select relevant features based on the performance of the learning algorithm.
- **Filter methods**: consider only the statistics of the training data, don't actually fit any learning model. Much cheaper!
- **Embedded methods**: use something built-in to learning method (e.g. coefficient magnitude in linear regression)

Univariate feature selection

- Score each feature x_i according to its importance in predicting target y
- Pick k features that are most important (use CV to choose k?)
- Problem: features may not be independent (remember attractiveness rankings in linear regression lab?)

Greedy feature selection

- Let S^{t-1} be the set of selected features at time $t-1$
- Compute the score for all combinations of current set + one more feature
- For the next time step S^t, add the feature that gave you the best score.

(Alternatively: start with all features, and "prune" one at a time.)

Scoring by mutual information (1)

How to score features? One way is to use **mutual information**:

For continuous variables:

$$I(X;Y) = \int_X \int_Y p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy$$

For discrete variables:

$$I(X;Y) = \sum_X \sum_Y p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy$$

Scoring by mutual information (2)

Determines how similar the joint distribution $p(x, y)$ is to the products of the marginal distributions $p(x)p(y)$.

If X and Y are independent, $p(x, y) = p(x)p(y)$ and then the integral will be zero.

Scoring by mutual information (3)

For feature selection: choose $\mathbf{X}_S$ to maximize mutual information between $\mathbf{X}_S$ and y.

$$\tilde{S} = \underset{S}{\text{argmax}}\, I(\mathbf{X}_S; y), \quad s.t. |S| = k$$

where k is the number of features we want to select.

Scoring by mutual information (4)

Greedy method: Let S^{t-1} be the set of selected features at time $t - 1$. Select feature f_t so that

$$f_t = \arg \max_{i \notin S^{t-1}} I(\mathbf{X}_{S^{t-1} \cup i}; y)$$

Scoring by mutual information (5)

Basic intuition: MI is a measure of **relevancy** of new feature minus **redundancy** of new feature vs. features already in the set.

Other scoring metrics

- Correlation coefficient between feature and target
- F-test: measures whether a feature is significant. F-test for one features is difference in MSE for single feature vs. prediction by mean.

$$F = (N - 2)\frac{R2}{1 - R2}$$

Illustration: scoring features

Regularization

Penalty for model complexity

With no bounds on complexity of model, we can always get a model with zero training error on finite training set - overfitting.

Basic idea: apply penalty in loss function to discourage more complex models

Regularization vs. standard LS

Least squares estimation:

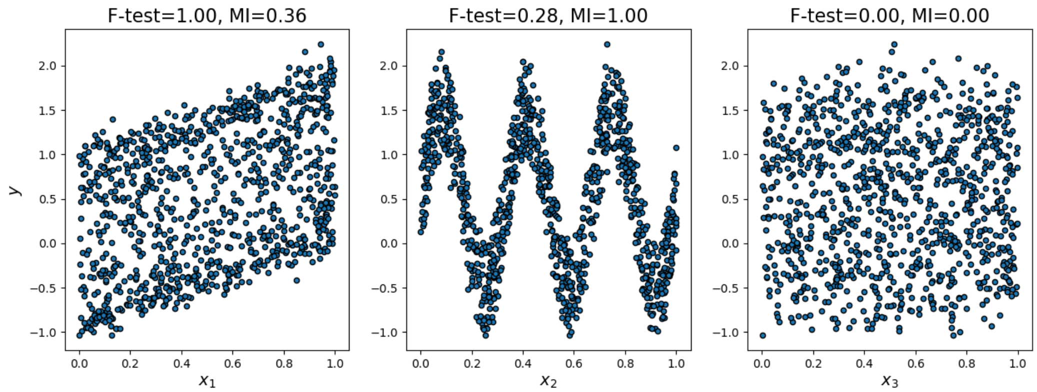

Figure 1: F-test selects x_1 as the most informative feature, MI selects x_2.

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\, RSS(\beta), \quad RSS(\beta) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Regularized estimation w/ **regularizing function** $\phi(\beta)$:

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\, J(\beta), \quad J(\beta) = RSS(\beta) + \phi(\beta)$$

Common regularizers: Ridge and LASSO

Ridge regression (L2):

$$\phi(\beta) = \alpha \sum_{j=1}^{d} |\beta_j|^2$$

LASSO regression (L1):

$$\phi(\beta) = \alpha \sum_{j=1}^{d} |\beta_j|$$

Graphical representation

Common features: Ridge and LASSO

- Both penalize large β_j
- Both have parameter α that controls level of regularization
- Intercept β_0 not included in regularization sum (starts at 1!), this depends on mean of y and should not be constrained.

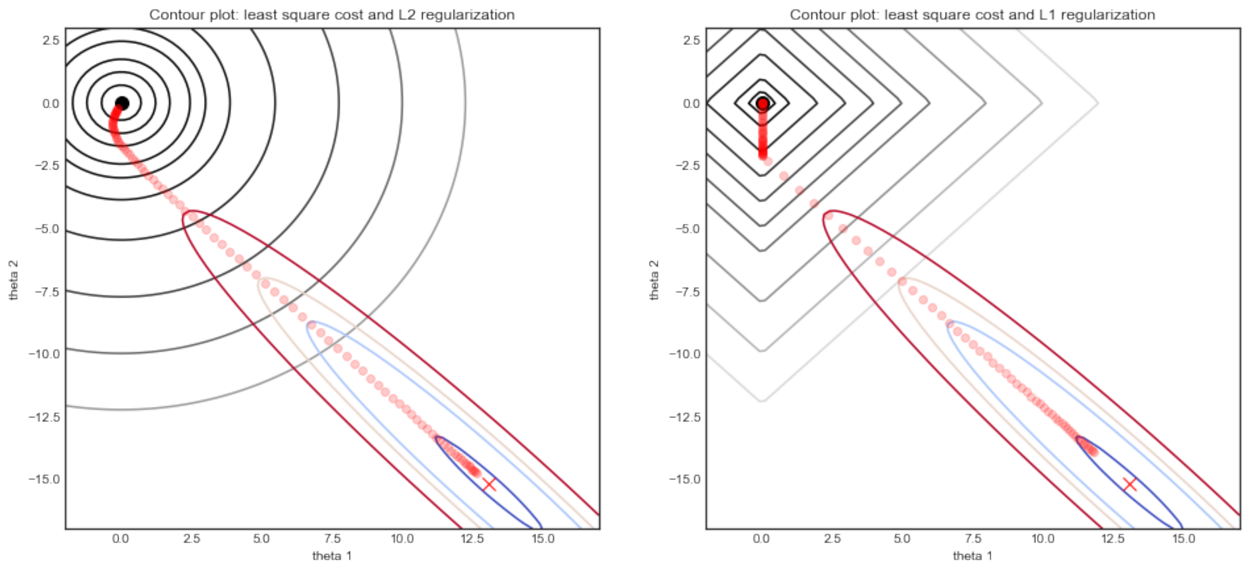

Figure 2: LS solution (+), RSS contours. As we increase α, LASSO solution moves from the LS solution to 0.

Differences: Ridge and LASSO (1)

Ridge (L2):

- minimizes $|\beta_j|^2$,
- does not penalize small non-zero coefficients
- heavily penalizes large coefficients
- tends to make many "small" coefficients
- Not for feature selection

Differences: LASSO (2)

LASSO (L1)

- minimizes $|\beta_j|$
- tends to make coefficients either 0 or large (sparse!)
- does feature selection (setting β_j to zero is equivalent to un-selecting feature)

Standardization (1)

Before learning a model with regularization, we typically *standardize* each feature and target to have zero mean, unit variance:

- $x_{i,j} \rightarrow \frac{x_{i,j}-\bar{x}_j}{s_{x_j}}$
- $y_i \rightarrow \frac{y_i-\bar{y}}{s_y}$

Standardization (2)

Why?

- Without scaling, regularization depends on data range

- With mean removal, no longer need β_0, so regularization term is just L1 or L2 norm of coefficient vector

L1 and L2 norm with standardization (1)

Assuming data standardized to zero mean, unit variance:

- Ridge cost function:

$$J(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{d} |\beta_j|^2 = ||\mathbf{A}\boldsymbol{\beta} - \mathbf{y}||^2 + \alpha||\boldsymbol{\beta}||^2$$

L1 and L2 norm with standardization (2)

- LASSO cost function ($||\boldsymbol{\beta}||_1$ is L1 norm):

$$J(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{d} |\beta_j| = ||\mathbf{A}\boldsymbol{\beta} - \mathbf{y}||^2 + \alpha||\boldsymbol{\beta}||_1$$

Ridge regularization

Why minimize $||\boldsymbol{\beta}||^2$?

Without regularization:

- large coefficients lead to high variance
- large positive and negative coefficients cancel each other for correlated features (remember attractiveness ratings in linear regression lab…)

Ridge term and derivative

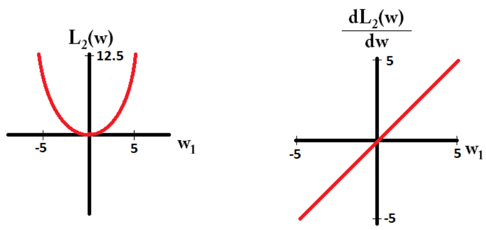

Figure 3: L2 term and its derivative for one parameter.

Ridge closed-form solution

$$J(\boldsymbol{\beta}) = ||\mathbf{A}\boldsymbol{\beta} - \mathbf{y}||^2 + \alpha||\boldsymbol{\beta}||^2$$

Taking derivative:

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2\mathbf{A}^T(\mathbf{y} - \mathbf{A}\boldsymbol{\beta}) + 2\alpha\boldsymbol{\beta}$$

Setting it to zero, we find

$$\beta_{ridge} = (\mathbf{A}^T\mathbf{A} + \alpha\mathbf{I})^{-1}\mathbf{A}^T\mathbf{y}$$

LASSO term and derivative

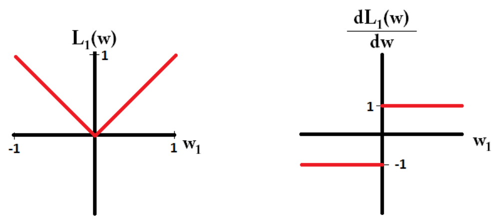

Figure 4: L1 term and its derivative for one parameter.

- No closed-form solution: derivative of $|\beta_j|$ is not continuous
- But there is a unique minimum, because cost function is convex, can solve iteratively

Effect of regularization level

Greater α, more complex model.

- Ridge: Greater α makes coefficients smaller.
- LASSO: Greater α makes more weights zero.

Effect of regularization - LASSO

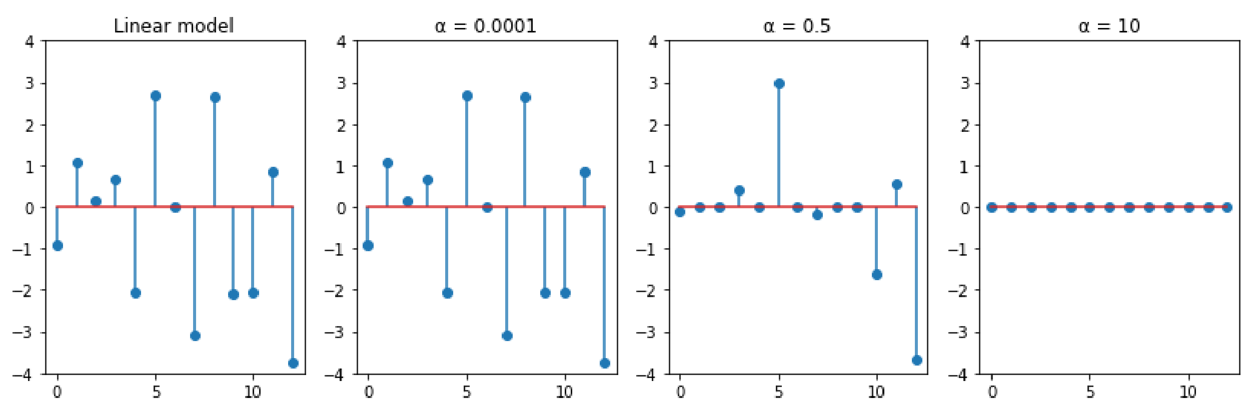

Figure 5: Increasing α

Effect of regularization - Ridge

Selecting regularization level

How to select α? by CV!

- Outer loop: loop over CV folds
- Inner loop: loop over α

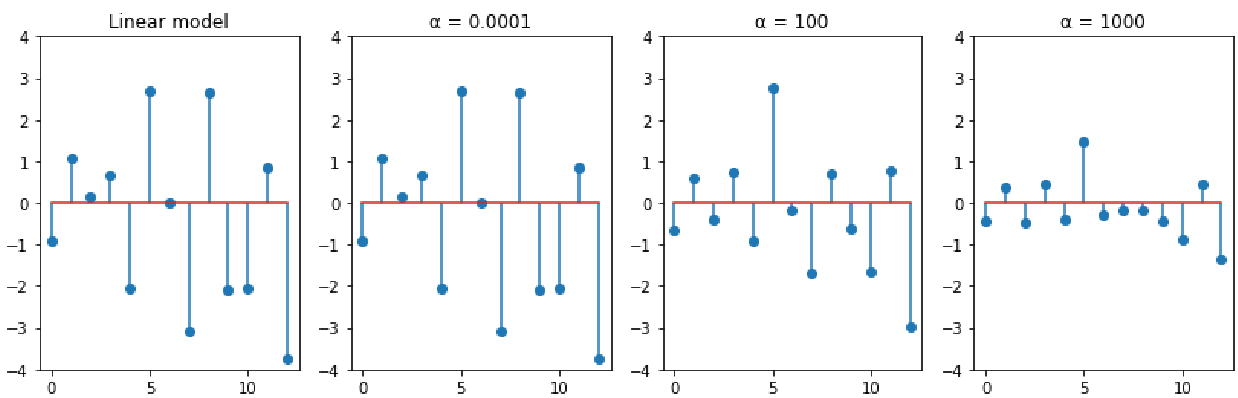

Figure 6: Increasing α