UNIVERSITY OF CALIFORNIA, BERKELEY

Lecture 13 - CS 189

Oscar Ortega

July 16, 2021

1 ISL, 6-6.1.2: LINEAR MODEL SELECTION AND REGULARIZATION

Recall how in a regression setting we try to describe the relationship between a response Y and a set of variables $X_0, ...X_P$. We learned least squares can help us try to approximate this quantity. But how can we improve on this algorithm?
Definitions:

- Prediction Accuracy: Things to note, if the relationship of of the response and the predictors is linear, least squares estimates, will have low bias. As the number of data points sampled increases, the variance will decrease as well. Having less data points than features renders us unable to generate a least-squares estimate. Shrinking the estimates often reduces the variance at the cost of a slight increase in bias.
- Model Interpretability: It is not always the case that all of the features of the data serve to predict the response. This leads us to want to remove these variables. Will go over techniques such as **subset-selection**, **shrinkage/regularization**, and **dimensionality reduction**

1.1 6.1.1: BEST SUBSET SELECTION

To perform this technique, we simply test the power set of features, and choose the member of that set that yields the best performance according to some metric. Note how this technique is infeasible as the number of features grows large $|\mathscr{P}\text{features}\}| = 2^{|\text{features}|}$

We can also perform **forward stepwise selection**, which is a dynamic programming solution of solving the problem. Where in each iteration k of subsets, we only consider

the predictors that augment the accuracy. Only viable subset method when p is very large.
Backward Step-wise Selection Will first perform least squares with all p predictors and iteratively removes the least useful predictors.

2 ISL, 6.2-6.2.1: Shrinkage Models

Because shrinking the coefficient estimates can reduce the variance of a given model, we perform shrinking of parameters by using techniques such as ridge and lasso regression.
We can define Ridge Regression as the minimizing the following objective:

$$\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j})^2 + \lambda \sum_{j=1}^{p} b_j^2$$

$$\|Xb - y\|^2 + \lambda\|b\|^2$$

- Note that unlike best subset selection, least squares models still include all of the features and in general leads to more complicated models.

2.1 LASSO Regression

We replace the squared sums of the lambda terms with absolute values.

$$\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j})^2 + \lambda \sum_{j=1}^{p} |b_j|$$

$$\|Xb - y\|^2 + \lambda\|b\|_1$$

We can also consider the two problems as finding the minimum value of the LS objective subject to the coefficients lying inside of a norm ball of size s. IE We can reformulate as follows:

$$\min_{\|w\|\leq s}\|Xw - y\| = \min\left(\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j})^2\right) \text{ s.t } \sum_{i=1}^{p} |b_j| \text{ or } b_j^2 \leq s$$

Note how best feature-subset selection would then correspond to the following equivalent minimization:

$$\min\left(\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j})^2\right) \text{ s.t } \sum_{i=1}^{p} \mathbf{1}(b_j \neq 0) \leq s$$

3 ESL, 3.4-3.4.3

3.1 Ridge Regression

Recall the formulation of the Ridge Regression formulation

$$\min_w \|Xw - y\|^2 + \lambda\|w\|^2$$

$$w^* = (X^T X + \lambda)^{-1} X^T y$$

3.2 Relationship with SVD

Consider the following expression, and recall that for any matrix A, there exists a Singular Value Decomposition $U\Sigma V^T$

$$X w^* = X(X^T X)^{-1} X^T y$$
$$= U\Sigma V^T (V\Sigma U^T U\Sigma V^T)^{-1} V\Sigma U^T y$$
$$= U\Sigma V^T (V\Sigma^2 V^T)^{-1} \Sigma U^T y$$
$$U\Sigma V^T (V\Sigma U^T U\Sigma V^T)^{-1} V\Sigma U^T y$$
$$= U\Sigma V^T V\Sigma^{-2} V^T V\Sigma U^T y$$
$$= UU^T y$$

And extrapolating this finding to the ridge solution, we know

$$X w^* \text{ridge} = X(X^T X + \lambda I)^{-1} X^T y$$
$$= U\Sigma(\Sigma^2 + \lambda I)^{-1} \Sigma U^T y$$
$$= \sum_{j=1}^{p} u_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} u_j^T y$$

Where we define each of the j terms as one of the **principle components of X**. Note how the lambda creates a soft shrinkage of the original diagonal terms of the centered matrix. We call this a 'soft thresholding' of the features, as opposed to a hard thresholding found by best subset selection.

4 Lecture

4.1 Ridge Regression aka. Tikhonov Regularization

Here is the formulation of the optimization problem:

$$\min_w \|Xw - y\|_2^2 + \lambda \|w\|_2^2$$

Where w' is we still use the 'fictitious dimension' trick, but we replace the component α in the fictitious dimension with a 0. This implies that X has fictitious dimension but we don't penalize α. Adds a regularization term, also known as a penalty term. this is useful for the shrinkage of weights in terms of their absolute values.
Why do we do this?

- guarantees that we have a positive definite matrix for the normal equations. Recall that this implies unique solutions. Consider how this will definitely happen when $d > n$. We consider a PSD matrix as returning an 'ill posed' problem.

- Reduces over-fitting by reducing variance.Why? Without constraints, a small fit in the data might create a an optimal predictor with wildly different values. So, we penalize large weights.

4.2 Geometric interpretation of regularization

.

What is the solution for Ridge Regression?

$$(X^T X + \lambda I)w = X^T y$$

Where I' is identity matrix with bottom right corner set to zero is because we do not want to penalize the bias term, we can modify this if the matrix is not PSD. We can solve this by solving the normal equations for w. From there, we return the function $h(z) = w^T z$ Increasing λ corresponds to more regularization, and smaller weights.

Recall our data model $y = Xv + e$, where e is noise. We can define the variance of ridge regression as equal to the following:

$$var(z^T (X^T X + \lambda I')X^T e)$$

As $\lambda \to \infty$, variance approaches 0, but bias increases. We find the soft-spot where variance decreases by a large portion, but bias increases by only a small amount.

An alternative would be to use an asymmetric penalty by replacing I' with another diagonal matrix.

4.3 Bayesian Justification for Ridge Regression

Assign a prior probability on w' : Gaussian $P(w') = \mathcal{N}(0, \sigma^2)$ Apply MLE to the posterior:

Use bayes theorom: posterior $P(w|X, y) = \frac{\mathcal{L}(w)P(w')}{P(X,y)}$.

We then maximize the log posterior of this likelihood:

$$\ln(\mathcal{L}(w)) + \ln(P(w')) - c$$

$$= -c\|Xw - y\|^2 - b\|w'\|^2 - d$$

$$\to \|Xw - y\|^2 + \lambda\|w'\|^2$$

This method(using likelihood, but maximizing posterior) is called maximum a posteriors (MAP).

4.4 Subset Selection

Intuitively all features will increase variance, but not all features reduce bias. Therefore we want to remove the poorly predictive features, so you just ignore them. This will give less over-fitting, smaller test errors and easier interpretation of models since they will be simpler. This principle is useful in all classification and regression methods.

This however can be hard as different features can partially encode same information, combinatorially hard to choose best feature subsets.
Alg: Best subset selection. Try all $2^d - 1$ of non-empty subsets of features. Slow.
Heuristic 1: Forward step-wise selection. Start with null model (0 features) repeatedly add best feature until validation errors start increasing. Repeatedly add best feature until validation errors start increasing (due to over-fitting) instead of decreasing. Requires training $O(d^2)$ models.
Heuristic 2: Backward step-wise selection, Start with all d features; repeatedly remove whose removal gives best reduction in validation error. Also trains $O(d^2)$. As an additional heuristic, only try to remove features with small weights.
Z-score of weight w_i is $z_i = \frac{w_i}{\sigma\sqrt{v_i}}$ where v_i is the ith diagonal entry of $(X^T X)^{-1}$.

4.5 LASSO

Find w that minimizes:

$$\|Xw - y\|^2 + \lambda \|w'\|_1$$

The unit cross-polytope is the convex hull of all the positive negative unit coordinate vectors.