Lecture 24: Model Selection with Lasso

In the last lecture we looked at the remarkable property of Lasso (more generally 1-norm penalty) to find a sparse solution to regression problems.

$$\min_{\beta} \|A\beta - Y\|^2 + \lambda\|\beta\|_1,$$

$$\hat{f}(x) = \sum_{j=0}^{J-1} \hat{\beta}_j \psi_j(x) \approx y(x)$$

Observe, if $\hat{\beta}$ is s-sparse, then only s entries in $\hat{\beta}$ are non-zero (active) & so

$$\hat{f}(x) = \sum_{j \in \{j \,|\, \hat{\beta}_j \neq 0\}} \hat{\beta}_j \psi_j(x),$$

In otherwords, $\hat{f}$ is particularly simple & depends only on s features.

This is an example of (automatic) model selection, ie, Lasso decides which features ψ_j are active our model for the data!

Model selection is a broad topic: The task of choosing the best model to describe a data set from a given collection of models.

- ex, what features ψ_j to use? wavelets, fourier, kernel, etc.
- How many features to use?

so we are using a narrow definition of model selection here.

ex diabetes data set.

We will now look at a benchmark data set to explain the idea in more detail.

Data Set Characteristics:

Number of Instances:	442
Number of Attributes:	First 10 columns are numeric predictive values
Target:	Column 11 is a quantitative measure of disease progression one year after baseline
Attribute Information:	• age age in years • sex • bmi body mass index • bp average blood pressure • s1 tc, total serum cholesterol • s2 ldl, low-density lipoproteins • s3 hdl, high-density lipoproteins • s4 tch, total cholesterol / HDL • s5 ltg, possibly log of serum triglycerides level • s6 glu, blood sugar level

We will fit two models to this data set, both models are linear

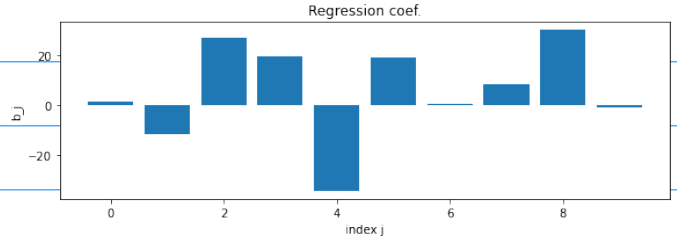
$$f(\underline{x}) = \beta_0 + \sum_{j=1}^{10} \beta_j \cdot x_j$$

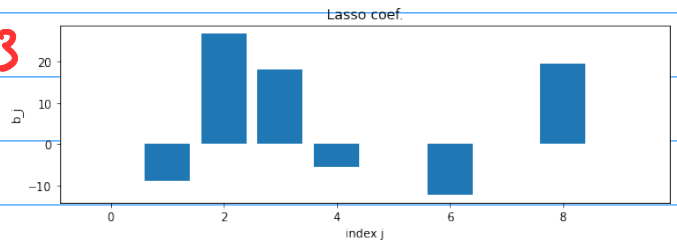
in the first case we simply solve linear regression (least squares)

$$\min_{\underline{\beta}} \|A\underline{\beta} - \underline{y}\|_2^2$$

in the second case we use Lasso:

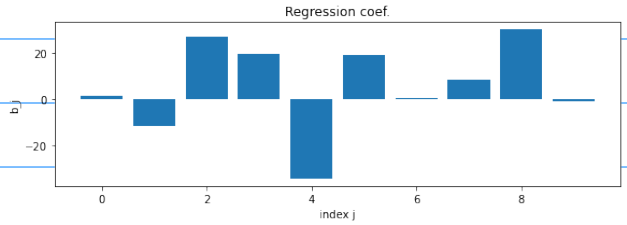
$$\min_{\underline{\beta}} \|A\underline{\beta} - \underline{y}\|_2^2 + \lambda\|\beta\|_1 .$$

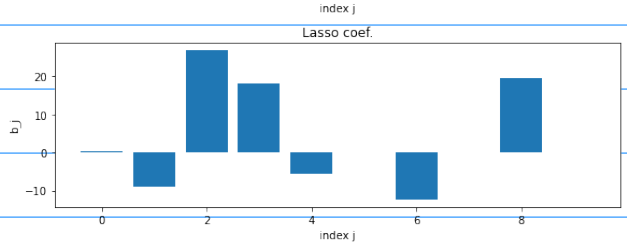




$\lambda = 1.3$

- Observe Lasso finds some β_j's that are identically $= 0$!





CV

20-fold

- Similar results if we use CV to find a good λ.

We can now consider the lasso solution $\underline{\beta}^{lasso}$ & take $|\beta_j^{lasso}|$ as a measure of importance of the corresponding feature x_j!

$|\beta_j| =$

age	sex	bmi	bp	s1	s2
[0.23034532	9.12140468	26.91925512	18.1047688	5.56576932	0.
12.32976343	0.	19.5820362	0.	]	
s3	s4	s5	s6		

we observe that β_6, β_8 & β_{10} are identically 0 & can be deemed "not important"!

We can then infer that the features x_5, x_2, x_4 are "not important" in training of our model & can be removed to simplify the model.

WARNING: Note that the above observations should be taken with a huge grain of salt! The Lasso will help you identify unimportant features for a given model & for a given data set. These might change drastically if you change your model or even the training data set. So be careful in drawing conclusions!

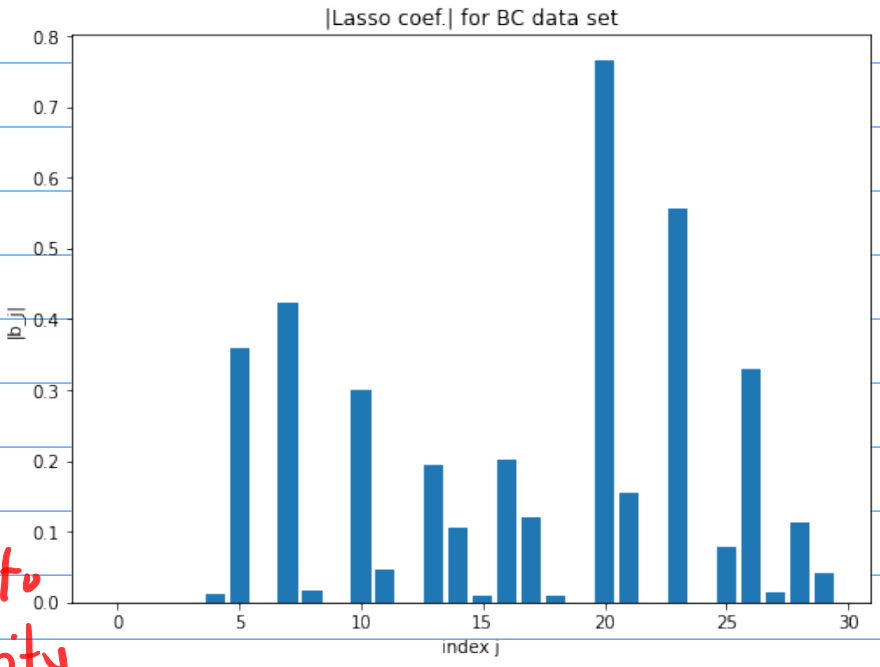
This approach is particularly useful if the data set has a lot of features.

Another ex (Breast cancer dataset)

Number of Instances:	569
Number of Attributes:	30 numeric, predictive attributes and the class
Attribute Information:	• radius (mean of distances from center to points on the perimeter) • texture (standard deviation of gray-scale values) • perimeter • area • smoothness (local variation in radius lengths) • compactness (perimeter^2 / area - 1.0) • concavity (severity of concave portions of the contour) • concave points (number of concave portions of the contour) • symmetry • fractal dimension ("coastline approximation" - 1) The mean, standard error, and "worst" or largest (mean of the three worst/largest values) of these features were computed for each image, resulting in 30 features. For instance, field 0 is Mean Radius, field 10 is Radius SE, field 20 is Worst Radius. • **class:** ◦ WDBC-Malignant

This is a classification data set with 30 features! As a first step in building a classifier we may wish to discard unimportant features.

* in this instance there are 10 input features that appear to be unimportant.



** modify the code & see sensitivity to λ & level of sparsity of β!