L_1-Regularization for Supervised Learning

Diego Vidaurre

Computational Intelligence Group
Universidad Politécnica de Madrid, Spain

Madrid
November, 2010

Preliminaries

Definition: Regularization is to introduce additional information to solve an ill-posed problem or to avoid overfitting.

- The goal is to get from data a model characterized by a certain set of parameters.
- For simplicity, we deal here with:
  - A continuous response (regression).
  - Linear models, where the response is linear in the inputs.

Notice that regularization techniques can be also applied to supervised classification, unsupervised classification, covariance matrix estimation, etc., and they are not limited to linear models.

- **Data**: $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ and $\boldsymbol{y} = (y_1, ..., y_n)^T$, where $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$ and $y_i = f(\boldsymbol{x}_i) + \epsilon$, $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$.
- **Model parameters**: Regression coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$.
- **Aim**: Find $\boldsymbol{\beta}$ that minimize a loss function $\sum_{i=1}^{n} l(y_i, \boldsymbol{x}_i)$.

In regression, the typical loss function is the residual sum of squares

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$$

The least squares estimator is the minimizer of the residual sum of squares.

The Bias-Variance Tradeoff

- Many models have a complexity parameter that has to be determined.
- We cannot use the training data to determine it because of overfitting.
- Instead, we want to minimize the expected prediction error, test error or generalization error.

$$\begin{aligned} E(Error(\beta)) &= \sigma^2 + E(y - \boldsymbol{x}^T\beta)^2 \\ &= \sigma^2 + [y - E(\boldsymbol{x}^T\beta)]^2 + Var(\boldsymbol{x}^T\beta) \end{aligned}$$

The second and third terms make up the mean squared error.

- Hence, The expected prediction error can be decomposed into observational error, model bias and estimation bias.
- The Gauss-Markov theorem states that least squares estimator has the lowest expected prediction error of all unbiased linear estimators.
- Traditional approaches in statistics seek unbiased predictions with the smallest variance possible.
- However, they may well exist a biased estimator with smaller expected prediction error.
- Regularization, by imposing certain restrictions, trades a **little bias in exchange of a larger reduction in variance** and avoids overfitting.
- Furthermore, it often allows to solve problems that are not directly solvable (ill-posed problems).

L_1-Regularization

Regularization is a constrained optimization:

$$min_{\beta}\Big(\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\beta)^2 + \lambda J(\beta)\Big)$$

$$= min_{\beta}\Big(\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\beta)^2\Big) \quad s.t. \quad J(\beta) \leq s.$$

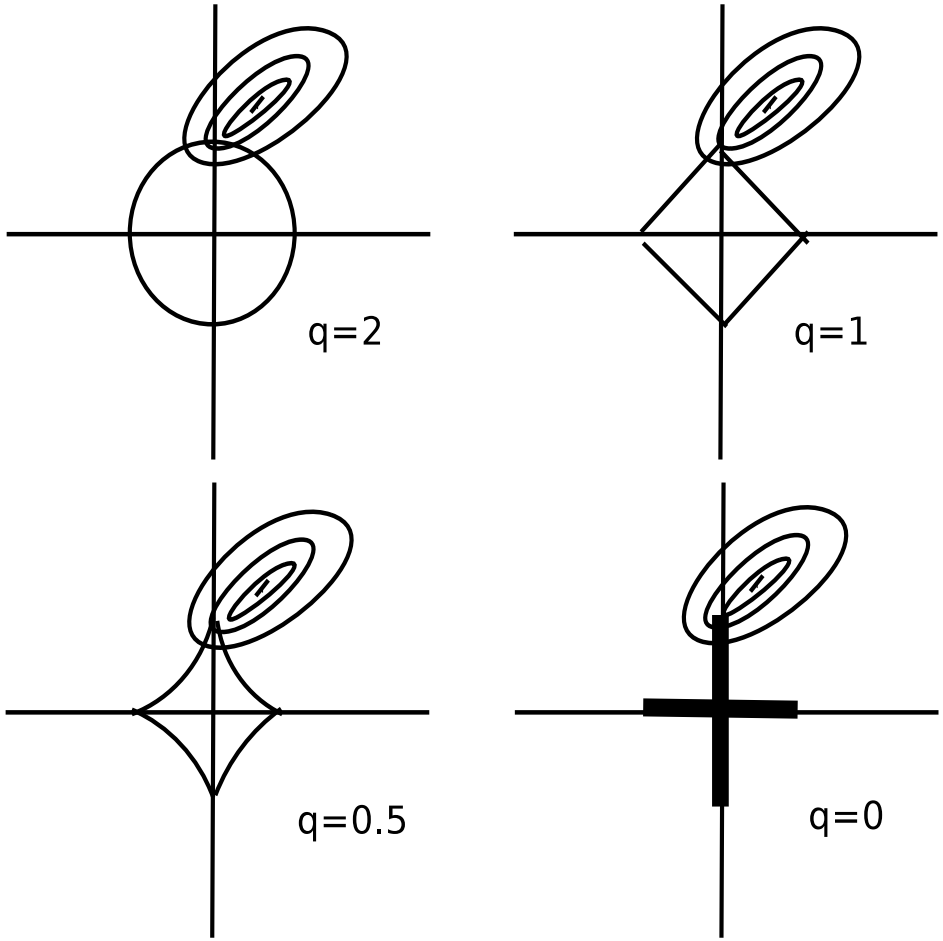
- In least squares estimation, when n is small with regard to p or there are correlated variables, the regression coefficients exhibit high variance and are unstable.
- Regularization helps to keep regression coefficients variance under control.
- $\lambda \geq 0$ controls the amount of regularization. There is a one-to-one correspondence between λ and s.

- The penalization function $J(\beta)$ can take several forms.
- Typically, we set

$$J(\beta) = ||\beta||_q^q = \sum_{j=1}^{p} |\beta_j|^q$$

- $q = 2$ (ridge or l_2-regularization) shrinks the parameters (Hoerl and Kennard, 1970).
- $q = 1$ (lasso or l_1-regularization) shrinks the parameters and selects variables (Tibshirani, 1996).
- $q = 0$ is a classical feature selection approach, where $J(\beta)$ is the number of free parameters.
- In general, variable selection is produced when $q \leq 1$. The problem is convex only when $q \geq 1$.
- Lasso is the intersection of both conditions.

The reason of the lasso variable selection ability from a geometric interpretation.



A Bayesian perspective

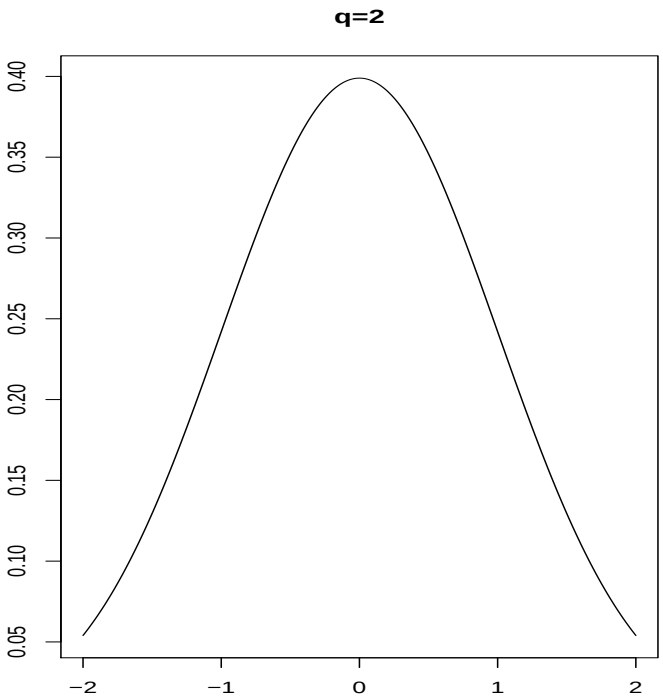
- Inferences are then based on the posterior distribution:

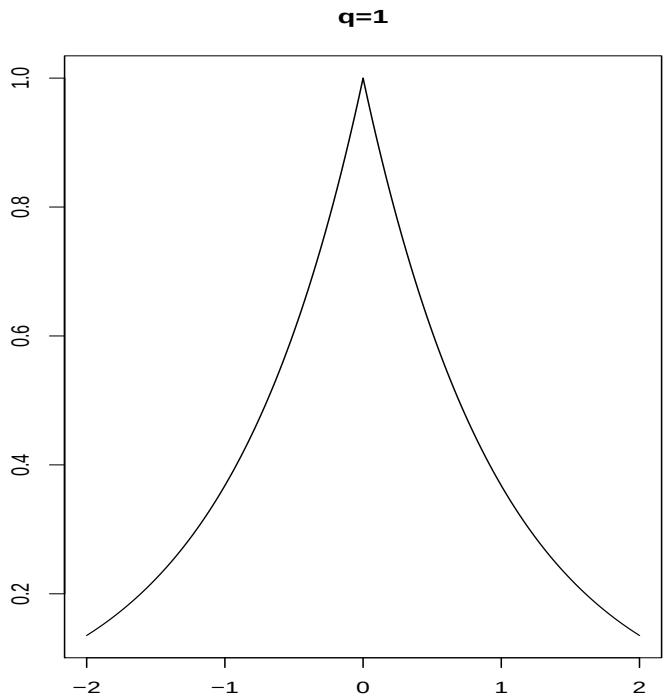
$$\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \frac{\pi(\boldsymbol{\beta})L(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{X})}{\int \pi(\boldsymbol{\beta})L(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{X})}$$

- If we treat the parameters as random variables we can establish a prior distribution over $\boldsymbol{\beta}$, $\pi(\boldsymbol{\beta})$.
- Under a Bayesian point of view, least squares is the maximum a posteriori estimate with a non-informed prior.
- A *shrinkage prior* centered at zero for the parameters leads to more stable estimates $\rightarrow$ regularization.
- Markov chain Monte Carlo provides an approach for generating samples from the posterior distribution.

Ridge and lasso corresponds to specific priors.

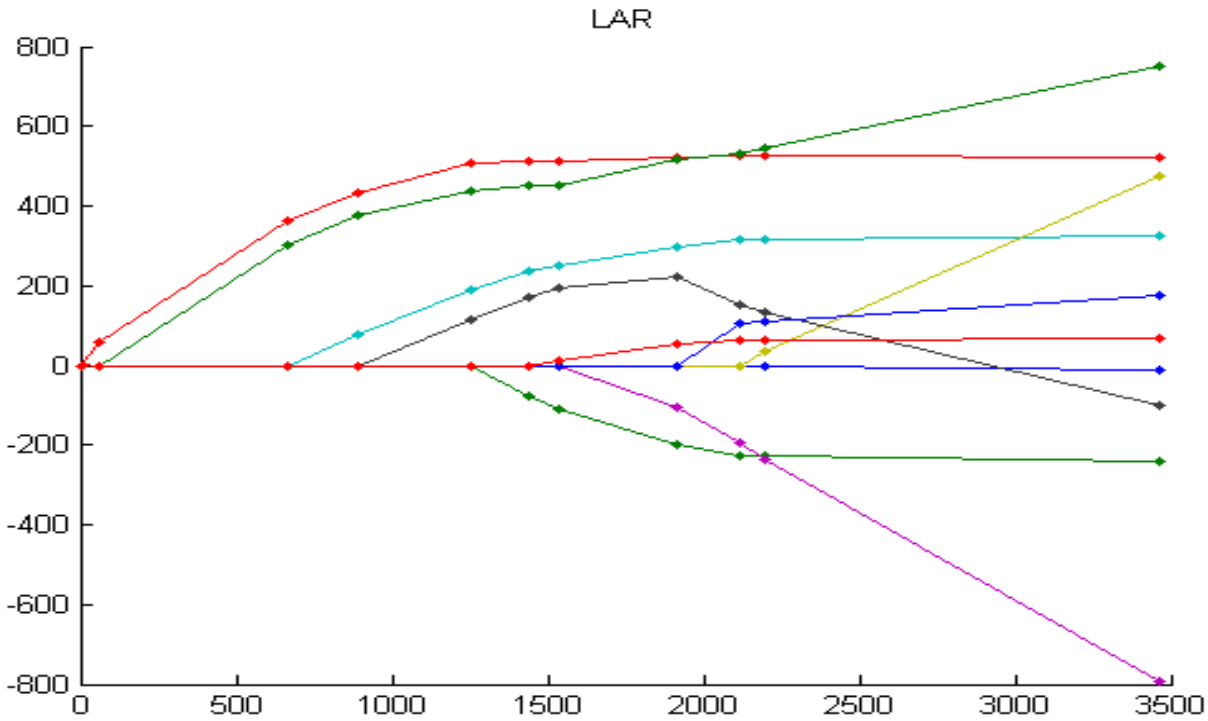
- $q = 2 \longrightarrow \beta_j \sim \mathcal{N}(0, \lambda)$.
- $q = 1 \longrightarrow \beta_j \sim \mathcal{N}(0, t), t \sim \mathcal{E}(\lambda^2/2)$
  (Park and Casella, 2008).





Regularization paths

- The complete solution of the lasso for all values of λ forms the lasso regularization path.
- The lasso regularization path is piecewise linear: we only need to compute the solution at a finite set of knots. Furthermore, this knots are placed exactly when a variable appears into or disappears from the model.



The LAR algorithm(Efron *et. al*, 2004) cheaply obtains the whole regularization path:

- 1. Let $\boldsymbol{r} = \boldsymbol{y}$ and $\beta = \mathbf{0}$.
- 2. Find variable $\boldsymbol{X}_{\cdot j_1}$ more correlated with $\boldsymbol{r}$.
- 3. Move β_{j_1} in the direction of $sign(corr(\boldsymbol{r}, \boldsymbol{X}_{\cdot j_1}))$ until any other variable $\boldsymbol{X}_{\cdot j_2}$ reachs the same correlation with $\boldsymbol{r}$ than $\boldsymbol{X}_{\cdot j_1}$.
- 4. Move $(\beta_{j_1}, \beta_{j_2})$ in the joint least squares direction until any other variable $\boldsymbol{X}_{\cdot j_3}$ has the same correlation with the current residual $\boldsymbol{r}$.
- 5. Repeat step 4 until all variables have been included into the model.

Advantages of lasso

- Lasso is less biased than ridge and in addition it performs variable selection.
- The LAR algorithm, with small modifications, can solve the Lasso at the cost of a least squares fit. The best solution within the regularization path can be found by minimizing a C_p-like statistic:

$$C_p = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} RSS(\beta)^2 - n + 2k,$$

  where k is the number of nonzero β_j.
- Under appropriate assumptions, consistency is possible as long as $log(p) = O(n)$ (Zhao and Yu, 2006).

Lasso extensions

- Lasso variable selection has been shown to be consistent under certain conditions. However, there exist certain scenarios where the lasso is inconsistent.
- The adaptive lasso (Zou, 2006) is consistent under softer conditions.
- Basically, it reformulates the lasso penalty as

$$\lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

where $\boldsymbol{w} = (w_1, ..., w_p)$ are weights based on the "importance" of the variable.

- Besides (unlike lasso), the adaptive lasso enjoys the *oracle properties*.

- In cases of known groups of variables, it may be needed to drop or select entire groups.
- Now, β_j is a vector of coefficients. The group lasso (Yuan and Lin, 2006) has a penalty of the form:

$$\lambda \sum_{j=1}^{p} ||\beta_j||_{\boldsymbol{K}_j},$$

where

$$||\beta_j||_{K_j} = \sqrt{\beta^T \boldsymbol{K}_j \beta},$$

- $\boldsymbol{K}_j$ is typically chosen to be the identity matrix.
- This is useful for example for groups of dummy variables from a categorical predictor.

- Sometimes, there are unknown groups of variables.
- Lasso tends to discard all but one redundant variables. What if we want to keep the entire group of redundant (but relevant) variables?
- The elastic net (Zou and Hastie, 2005) combines a lasso penalty with a ridge penalty:

$$\lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2.$$

- Moreover, the elastic net gives better predictions when variables are considerable correlated and allows to select more than n variables.
- It is useful for example in brain imaging data, where we would want to know all the brain areas involved in some task (maybe more than n), even if they are correlated.

- In other ocasions, there are some significant ordering in the variables and we are interested to have similar coefficients for (spatially) close variables.
- The fused lasso (Tibshirani and Saunders, 2005) penalizes both the coefficients and the difference between adjacent coefficients:

$$\lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|.$$

- This has been extended to the *hinge* loss function for support vector machines.
- The fused lasso is motivated by problems of analyzing protein mass spectroscopy data, where closer sites are known to be jointly relevant or irrelevant.

- In general, many lasso-based penalties and many loss functions are applicable.
- However, high dimensionality in real problems makes computational efficiency a crucial matter.
- The question is: Under which conditions a LAR efficient procedure can be used? In other words, when is the regularization path piecewise linear?
- There are two sufficient conditions (Rosset and Zhu, 2007):
  - The loss function is piecewise quadratic as a function of β along the regularization path.
  - The penalty function is piecewise linear as a function of β along the regularization path.

A bit on my work

- If $f(\cdot)$ is not linear, we could include high-order terms, but which ones?
- Instead, we can make a weighted locally regression for a specific point of interest, where closer points are given more importance than further ones (Cleveland, 1979).
- Can we directly plug lasso into weighted locally regression?
- We base on distances to know which points are close or far, should all variables be included in the distance calculation?
- Our proposal is an iterative algorithm that alternates between distance calculation, weighting and LARS (Vidaurre *et. al*, 2010).

- Besides theoretical aspects of lasso and further extensions, applications of the l_1 penalty are a hot research topic.
- Assuming the data to be jointly Gaussian, the lasso can be used for undirected graph induction. In (Meinshausen and Bühlmann, 2006), the authors use Lasso regression individually with every variable against the rest, in such a way that an edge is created when the regression coefficient is not zero.
- However, a direct use of Lasso fails to recover the true sparsity pattern when variables are highly correlated.
- Focusing on Gaussian Bayesian networks, we proposed a greedy search heuristic on the equivalence class space using the lasso as help to guide the search (Vidaurre *et. al*, 2009).

- Bayesian network classifiers perform classification by selecting the class that maximizes the joint likelihood.
- We can use l_1-regularization to train a Bayesian network classifier.
- We focus on naïve Bayes (or at least we stem from the naïve Bayes assumptions).
- The likelihood of the Bayesian network classifier over the training data set is not a adequate loss function.
- Basing on the results of Rosset and Zhu, we formulate a loss function and devise a LAR procedure to get a classifier (based on naïve Bayes) than can discard redundant and irrelevant variables (Vidaurre *et. al*, work on progress).

Bibliography

- A. Hoerl and R. Kennard: “Ridge problems: Biased estimates for nonorthogonal problems”, Technometrics, 1970
- R. Tibshirani. “Regression shrinkage and selection via the lasso”, Journal of the Royal Statistics Society: Series B, 1996.
- T. Park and G. Casella. “The Bayesian lasso”, Journal of the Royal Statistics Society, 2008.
- B. Efron, J. Johnstone, T. Hastie and R. Tibshirani. “Least angle regression”, Annals of Statistics, 2004.
- P. Zhao and B. Yu. “On model selection consistency of lasso”, Journal of Machine Learning Research, 2006
- H. Zou. “The adaptive lasso and its oracle properties”, Journal of the Royal Statistics Society: Series B, 2006.

- M. Yuan and Y. Lin. “Model selection and estimation in regression with grouped variables”, Journal of the Royal Statistical Society: Series B, 2006.
- H. Zou and T. Hastie. “Regularization and variable selection via the elastic net”, Journal of the Royal Statistical Society: Series B, 2005.
- R. Tibshirani and M. Saunders. “Sparsity and smoothness via the fused lasso”, Journal of the Royal Statistical Society: Series B, 2005.
- S. Rosset and J. Zhu. “Piecewise linear regularized solution paths”, Annals of Statistics, 2007.
- W. Cleveland. “Robust locally weighted regression and smoothing scatter-plots”, Journal of the American Statistical Association, 1979.

- N. Meinshausen and P. Bühlmann, “High-dimensional graphs and variable selection with the Lasso”, Annals of Statistics, 2006.
- D. Vidaurre, C. Bielza and P. Larrañaga. “Learning an L1-regularized Gaussian Bayesian Network in the Equivalence Class Space”, IEEE Transactions on Systems, Man and Cybernetics, Part B, 2009.
- D. Vidaurre, C. Bielza and P. Larrañaga. “Lazy lasso for local regression”, TEST, Sent.
- D. Vidaurre, C. Bielza and P. Larrañaga. “L_1-regularization on a Bayes-based classifier for continuous and discrete predictors”, work on progress.