Subset Selection and Regularization

Leading discussion

Branislav Kveton
Intelligent Systems Program

Contexts

- Subset selection
- Continuous subset selection
- Ridge regression
- Ridge regression and some linear algebra
- Other methods than ridge regression
- The last word

Subset Selection

- One looks for a subset F ∈ {1, …, d} of features that is useful for prediction. The main reasons are
  - Prediction accuracy might be improved by sacrificing a bit of bias in exchange for reducing the variance
  - It easier to interpret a simple model than a complex one
- There are different variants of subset selection
  - Exhaustive version of the algorithm searches all possible subsets F. Even if correct, this approach is not feasible for the large number of features because the number of subsets grows as 2^d
  - Incremental approaches start with F = {} ({1, …, d}) and on the basis of some information criterion add (remove) a new feature to (from) F. These methods are guided by heuristics, which may be wrong

Continuous Subset Selection

- What is wrong with subset selection?
  - Features are either preserved in F or discarded. As this decision process is discrete, the prediction capability of the model may change significantly if a feature is preserved (discarded)
- Continuous version of subset selection is represented by
  - Shrinkage methods (regularization)
    - Ridge regression
    - Lasso regression
  - Methods using derived input directions
    - Principal components regression
    - Partial least squares

Ridge Regression

- Ridge regression is an extension of linear regression by adding a quadratic penalizing term

$$\beta^{ridge} = \arg\min_{\beta}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - x_i^T\beta\right)^2 + \lambda\beta^T\beta\right\}$$

- Intuitively, the larger the value of λ is, the larger is the shrinkage of the weights
- The optimalization problem can be equally formalized as

$$\begin{aligned}\beta^{ridge} &= \arg\min_{\beta} \quad \sum_{i=1}^{N}\left(y_i - \beta_0 - x_i^T\beta\right)^2 \\ & \quad \text{satisfying} \quad \beta^T\beta \le s\end{aligned}$$

- There exists one-to-one mapping between λ and s

Reparametrization Using Centered Inputs

- If we adopt the assumption that $x_{i,j}$ are centered and β_0 is approximated as the mean of y_is, we can rewrite the formula as

$$\begin{aligned}RSS(\lambda) &= \sum_{i=1}^{N}\left(y_i - (x_i - \bar{x})^T\beta\right)^2 + \lambda\beta^T\beta \\ &= \sum_{i=1}^{N}\left(y_i + \bar{x}\beta - x_i^T\beta\right)^2 + \lambda\beta^T\beta \\ &\approx \sum_{i=1}^{N}\left(y_i + \bar{y} - x_i^T\beta\right)^2 + \lambda\beta^T\beta\end{aligned}$$

- This indicates that ridge regression can be expressed in matrix notation similarly to the linear regression as

$$RSS(\lambda) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

Solving Ridge Regression

- Ridge regression can be solved exactly, which is similar to linear regression

$$
\begin{aligned}
RSS(\lambda) &= (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta \\
\nabla RSS(\lambda) &= -2X^T (y - X\beta) + 2\lambda I\beta &&= 0 \\
& -X^T (y - X\beta) + \lambda I\beta &&= 0 \\
& -X^T y + X^T X\beta + \lambda I\beta &&= 0 \\
& (X^T X + \lambda I)\beta &&= X^T y \\
& \beta &&= (X^T X + \lambda I)^{-1} X^T y
\end{aligned}
$$

- Nice property of the solution is that even if X^TX is singular, the addition of λI makes it nonsingular, which in turn means that an inverse matrix exists

Ridge Regression and Some Linear Algebra

- Every matrix X has singular value decomposition of the following form, where U spans the column space of X, and V spans the row space of X

$$
\underbrace{X}_{N\times d} = \underbrace{U}_{\substack{N\times d \\ \text{orthogonal}}} \cdot \underbrace{D}_{\substack{d\times d \\ \text{diagonal}}} \cdot \underbrace{V^T}_{\substack{d\times d \\ \text{orthogonal}}}
$$

- Least squares fit can be rewritten as

$$
\begin{aligned}
X\beta^{LS} &= X(X^T X)^{-1} X^T y = UDV^T \left(\underbrace{(UDV^T)^T}_{(AB)^T = B^T A^T} UDV^T \right)^{-1} \underbrace{(UDV^T)^T}_{(AB)^T = B^T A^T} y \\
&= UDV^T \left(\underbrace{VD^T U^T UDV^T}_{U^T U = I,\ VD^T D = D^T DV,\ VV^T = I} \right)^{-1} VD^T U^T y = UDV^T (D^T D)^{-1} VD^T U^T y \\
&= U \underbrace{(D^T D)(D^T D)^{-1}}_{AA^{-1} = I} \underbrace{V^T V}_{V^T V = I} U^T y \\
&= UU^T y
\end{aligned}
$$

Coordinates of y with respect to the orthonormal basis U

Ridge Regression and Some Linear Algebra

- Ridge regression fit can be rewritten as well as

$$
\begin{aligned}
X\beta^{ridge} &= X\left(X^TX+\lambda I\right)^{-1}X^Ty = UDV^T\left(\underbrace{\left(UDV^T\right)^T}_{(AB)^T=B^TA^T}UDV^T+\lambda I\right)^{-1}\underbrace{\left(UDV^T\right)^T}_{(AB)^T=B^TA^T}y \\
&= UDV^T\left(\underbrace{VD^TU^TUDV^T}_{U^TU=I,\,VD^TD=D^TDV,\,VV^T=I}+\lambda I\right)^{-1}VD^TU^Ty = UDV^T\left(D^TD+\lambda I\right)^{-1}VD^TU^Ty \\
&= U\left(D^TD\right)\left(D^TD+\lambda I\right)^{-1}\underbrace{V^TV}_{V^TV=I}U^Ty = UD^2\left(D^2+\lambda I\right)^{-1}U^Ty \\
&= \sum_{j=1}^{d} u_j \frac{d_j^2}{d_j^2+\lambda}u_j^Ty
\end{aligned}
$$

Coordinates of y with respect to the orthonormal basis U

- D is a diagonal matrix with entries $d_1 \geq d_2 \geq \ldots \geq d_d \geq 0$

Ridge Regression and Some Linear Algebra

- If $d_i < d_j$, then for any $\lambda \geq 0$

$$\frac{d_i^2}{d_i^2+\lambda} < \frac{d_j^2}{d_j^2+\lambda} \leq 1$$

- D^2 is a matrix of eigenvalues and V is a matrix of eigenvectors for the covariance matrix X^TX (Eigen Decomposition Theorem)

$$
\begin{aligned}
X &= UDV^T \\
X^TX &= \underbrace{\left(UDV^T\right)^T}_{(AB)^T=B^TA^T}UDV^T \\
X^TX &= VD^T\underbrace{U^TU}_{U^TU=I}DV^T \\
X^TX &= VD^2V^T
\end{aligned}
$$

Ridge Regression and Some Linear Algebra

- The first principal component z_1, which preserves the most of the variance, can be expressed as

$$z_1 = Xv_1 = u_1 d_1$$

and the latter equality holds because of

$$\begin{aligned} X &= UDV^T = UDV^{-1} \\ XV &= UD \end{aligned}$$

- As principal components are perpendicular to each other, and u_j can be viewed as a normalized version of z_j, we can conclude that the shrinkage of d_j affects how much are the coordinates regarding a principal component shrunken

Ridge Regression as the Mean of a Posterior Distribution

$$\begin{aligned} y_i &\approx N\left(\beta_0 + x_i^T\beta, \sigma^2\right) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - \beta_0 - x_i^T\beta\right)^2\right] \end{aligned}$$

$$\begin{aligned} \beta &\approx N\left(0, \tau^2\right) \\ &= \frac{1}{(2\pi)^{d/2}\left|\tau^2 I\right|^{1/2}} \exp\left[-\frac{1}{2}\beta^T\left(\tau^2 I\right)^{-1}\beta\right] \end{aligned}$$

$$\begin{aligned} \ell(y; X, \beta) &= \ln\left(\begin{array}{l} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \beta_0 - x_i^T\beta\right)^2\right] \\ \frac{1}{(2\pi)^{d/2}\left|\tau^2 I\right|^{1/2}} \exp\left[-\frac{1}{2}\beta^T\left(\tau^2 I\right)^{-1}\beta\right] \end{array}\right) \\ &\approx \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(y_i - \beta_0 - x_i^T\beta\right)^2 + \frac{1}{\tau^2}\beta^T\beta \\ &= \sum_{i=1}^{N}\left(y_i - \beta_0 - x_i^T\beta\right)^2 + \frac{\sigma^2}{\tau^2}\beta^T\beta \end{aligned}$$

Lasso Regression

- Lasso regression has penalty defined as the sum of the absolute values of the weights β as

$$\beta^{lasso} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T \beta\right)^2$$
$$\text{satisfying} \quad \sum_{j=1}^{d} \left|\beta_j\right| \le t$$

- Absolute value in lasso penalty makes the problem of weights' estimation non-linear
- The penalty tends to drive less important weights to zero faster than the one in ridge regression

Methods Using Derived Input Directions

- Principal components regression uses $M \le d$ vectors selected by PCA to do regression on them
- As these vectors are orthogonal, regression problem is divided into M independent regression problems
- As opposing to PCR, partial least squares technique takes into account y when features are selected

The Last Word

- Regularization encompassed more general problems of the form

$$\min_{f \in H} \left\{ \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \right\}$$

  where L(y, f(x)) is a loss function, J(f) is penalty for the parameterization, and H is a space where J(f) is defined
- In addition to linear regression, another useful application of regularization is in neural networks