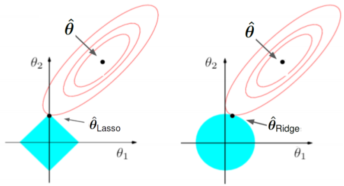
Introduction to Machine Learning

Regularization
Lasso vs. Ridge

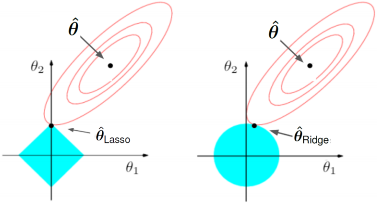


Learning goals

- Properties of ridge vs. lasso
- Coefficient paths
- What happens with corr. features
- Why we need feature scaling

LASSO VS. RIDGE GEOMETRY

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2 \qquad \text{s.t. } \|\boldsymbol{\theta}\|_p^p \leq t$$

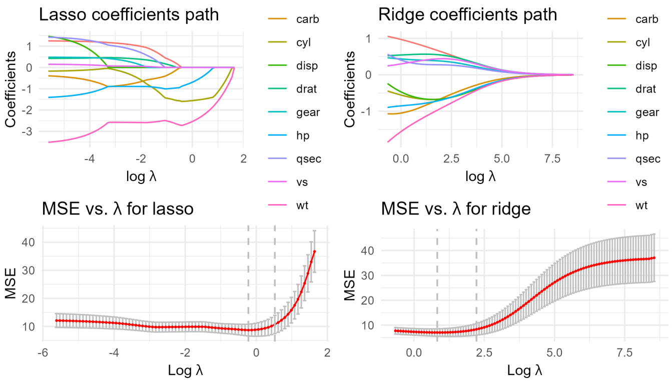


- In both cases (and for sufficiently large λ), the solution which minimizes $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$ is always a point on the boundary of the feasible region.
- As expected, $\hat{\boldsymbol{\theta}}_{\text{lasso}}$ and $\hat{\boldsymbol{\theta}}_{\text{ridge}}$ have smaller parameter norms than $\hat{\theta}$.
- For lasso, solution likely touches a vertex of constraint region. Induces sparsity and is a form of variable selection.
- For $p > n$: lasso selects at most n features ▸ Zou and Hastie 2005.

COEFFICIENT PATHS AND 0-SHRINKAGE

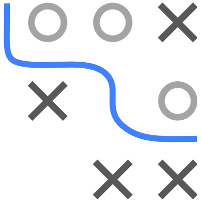
Example 1: Motor Trend Car Roads Test (mtcars)

We see how only lasso shrinks to exactly 0.



NB: No real overfitting here, as data is so low-dim.

REGULARIZATION AND FEATURE SCALING



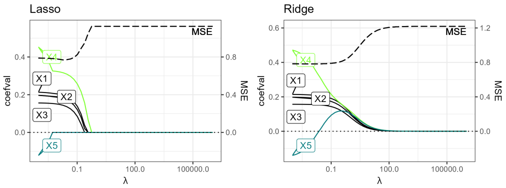
- Typically we omit θ_0 in penalty $J(\boldsymbol{\theta})$ so that the “infinitely” regularized model is the constant model (but can be implementation-dependent).
- Unregularized LM has **rescaling equivariance**, if you scale some features, can simply "anti-scale" coefs and risk does not change.
- Not true for Reg-LM: if you down-scale features, coeffs become larger to counteract. They are then penalized stronger in $J(\boldsymbol{\theta})$, making them less attractive without any relevant reason.
- **So: usually standardize features in regularized models, whether linear or non-linear!**

CORRELATED FEATURES: $L1$ VS $L2$

Simulation with $n = 100$:

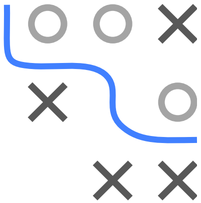
$$y = 0.2x_1 + 0.2x_2 + 0.2x_3 + 0.2x_4 + 0.2x_5 + \epsilon$$

x_1-x_4 are independent, but x_4 and x_5 are strongly correlated.



- L1 removes x_5 early, L2 has similar coeffs for x_4, x_5 for larger λ
- Also called “grouping property”: for ridge highly corr. features tend to have equal effects; lasso however “decides“ what to select
- L1 selection is somewhat “arbitrary”

SUMMARY ▸ Tibshirani 1996 ▸ Zou and Hastie 2005



- Neither ridge nor lasso can be classified as better overall
- Lasso can shrink some coeffs to zero, so selects features;
  ridge usually leads to dense solutions, with smaller coeffs
- Lasso likely better if true underlying structure is sparse
  ridge works well if there are many (weakly) influential features
- Lasso has difficulties handling correlated predictors;
  for high correlation, ridge dominates lasso in performance
- Lasso: for (highly) correlated predictors, usually an “arbitrary” one
  is selected, with large coeff, while the others are (nearly) zeroed
- Ridge: coeffs of correlated features are similar