CSE 446
Linear Regression

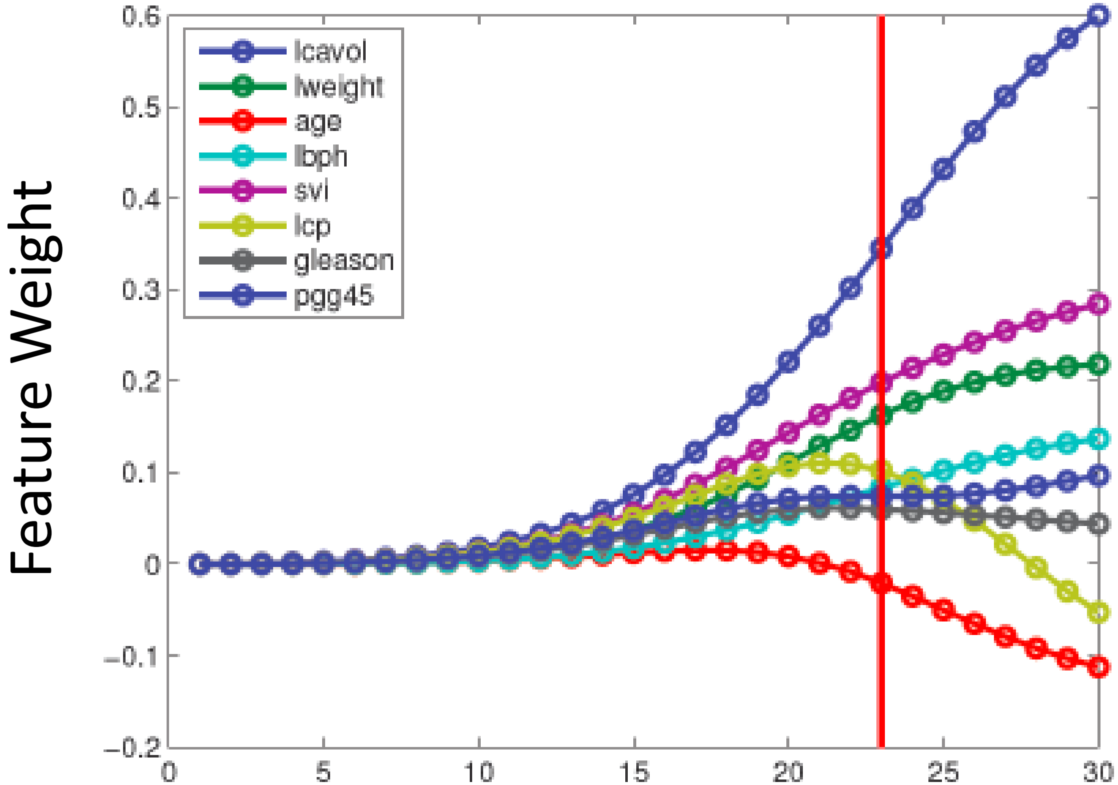
Administrative

- Linear algebra review session tomorrow

Lecture Notes

- Regularization (put a prior on the weights)
- See lecture notes

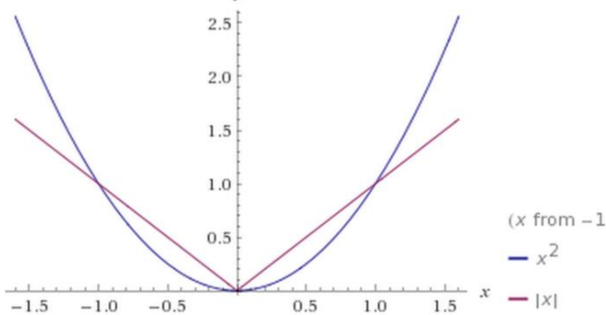
Ridge Coefficent Path



Larger λ ← $||\hat{w}||_2$ → Smaller λ

From
Kevin Murphy
textbook

Why Gaussian prior?



- Ridge:

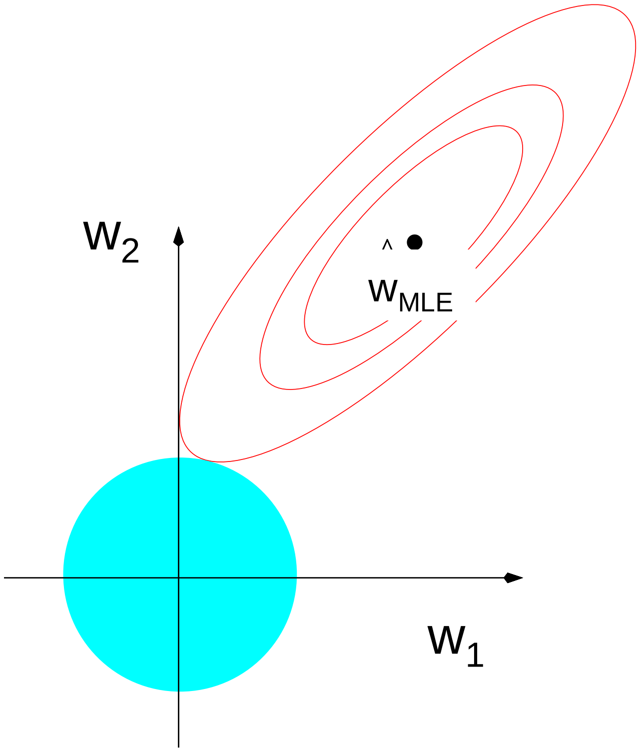
$$\hat{w}_{\text{ridge}} = \arg\min_{w} \sum_{i=1}^{N} (x_i \cdot w - y_i) + \lambda \sum_{j=1}^{d} w_j^2$$

- LASSO (“least absolute shrinkage and selection operator”):

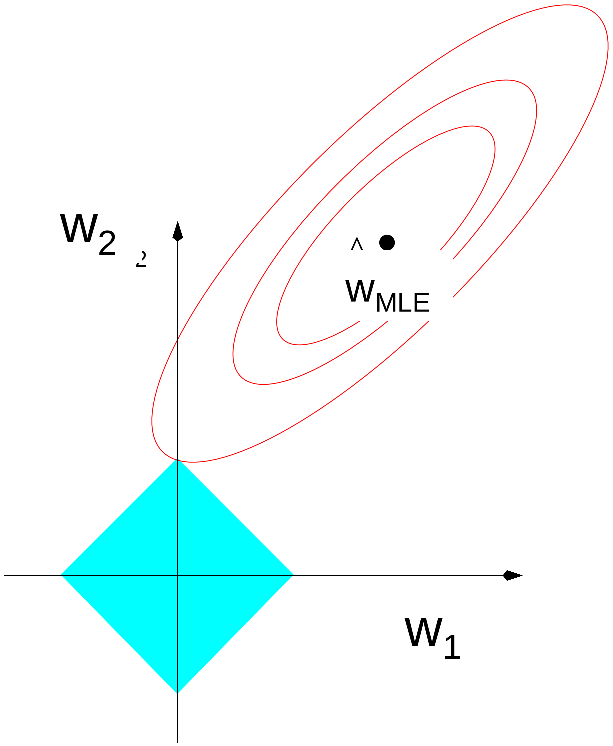
$$\hat{w}_{\text{LASSO}} = \arg\min_{w} \sum_{i=1}^{N} (x_i \cdot w - y_i) + \lambda \sum_{j=1}^{d} |w_j|$$

  – Linear penalty pushes more weights to zero
  – Allows for a type of *feature selection*
  – But, not differentiable and no closed form solution….

Geometric Intuition



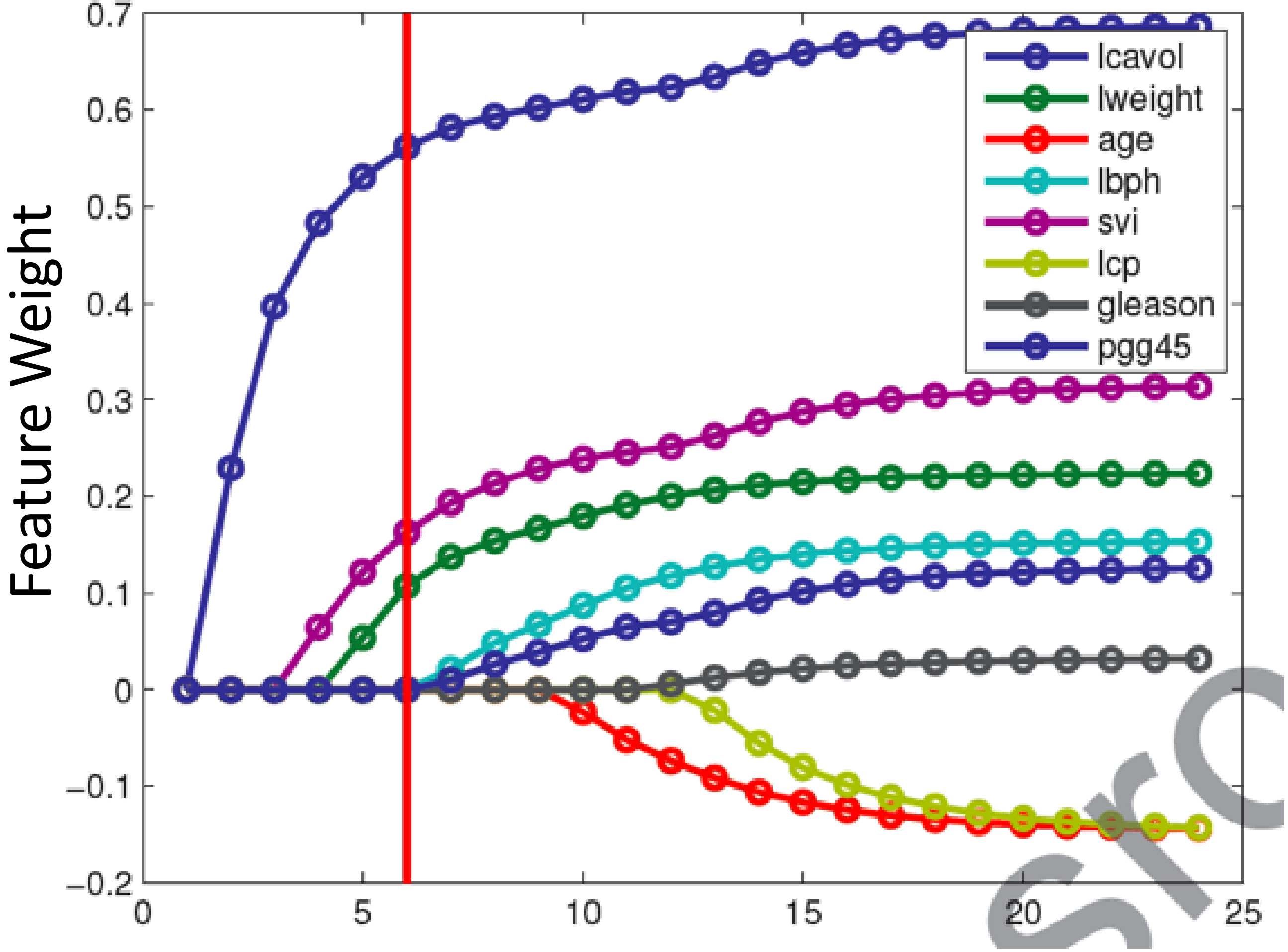
Ridge Regression



Lasso

From Rob Tibshirani slides

LASSO Coefficent Path



From
Kevin Murp
textbook

Larger λ ←

$$||\hat{w}||_1$$

→ Smaller λ

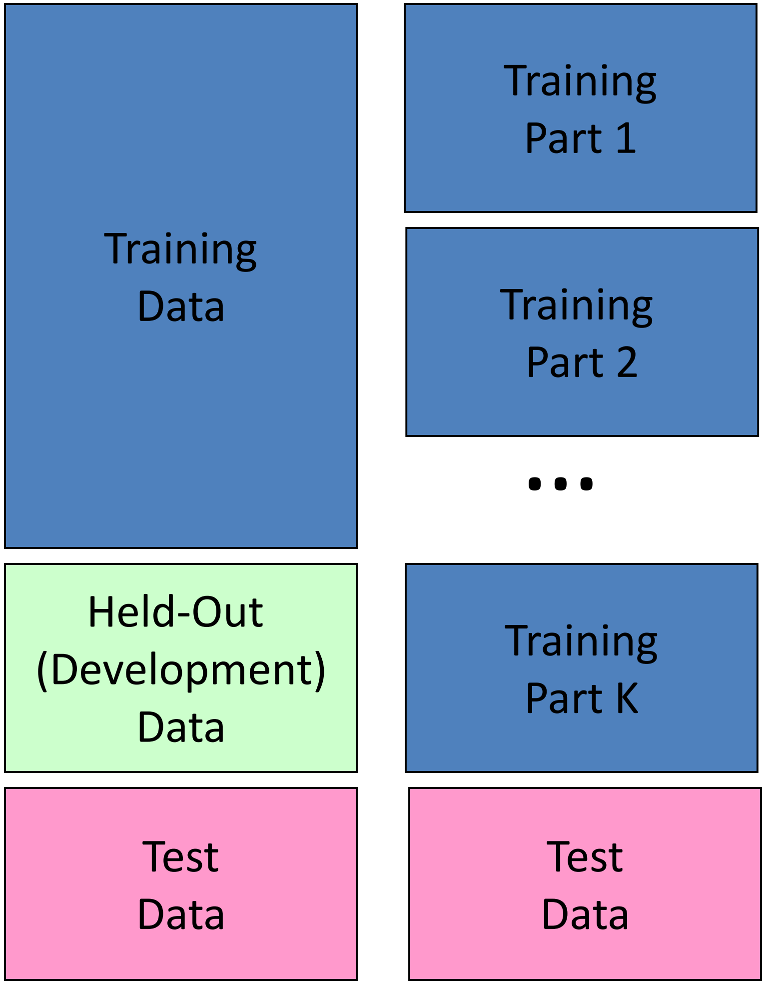
How does varying lambda change w?

$$\hat{w}_{\text{ridge}} = \arg\min_{w} \sum_{i=1}^{N} (x_i \cdot w - y_i) + \lambda \sum_{j=1}^{d} w_j^2$$

- Larger λ? Smaller λ?
- As λ →0?
  - Becomes same a MLE, unregularized
- As λ →∞?
  - All weights will be 0!

How to pick lambda?

- Experimentation cycle
  - Select a hypothesis *f* to best match training set
  - Tune hyperparameters on held-out set
    - Try many different values of lambda, pick best one
- Or, can do k-fold cross validation
  - No held-out set
  - Divide training set into k subsets
  - Repeatedly train on k-1 and test on remaining one
  - Average the results



What you need to know

- Regression
  - Basis function/features
  - Optimizing sum squared error
  - Relationship between regression and Gaussians
- Regularization
  - Ridge regression math & derivation as MAP
  - LASSO formulation
  - How to set lambda (hold-out, K-fold)
- Bias-Variance trade-off (covered on Friday)