Regularization

Jong-Han Kim

EE787 Fundamentals of machine learning
Kyung Hee University

Sensitivity

- we have a linear predictor $\hat{y} = g(x) = \theta^{\mathsf{T}} x$
- if $|\theta_i|$ is large, then the prediction is very sensitive to x_i
  (*i.e.*, small changes in x_i lead to large changes in the prediction)
- large sensitivity can lead to overfit, poor generalization
  (which would turn up in validation)
- for $x_1 = 1$ (the constant feature), there is no sensitivity, since the feature does not change
- suggests that we would like θ (or $\theta_{2:d}$ if $x_1 = 1$) not too large

Regularizer

- we will measure the size of θ using a *regularizer* function $r : \mathbf{R}^d \to \mathbf{R}$
- $r(\theta)$ is a measure of the size of θ (or $\theta_{2:d}$)

- *quadratic regularizer* (a.k.a. ℓ_2 or sum-of-squares):

$$r(\theta) = ||\theta||^2 = \theta_1^2 + \cdots + \theta_d^2$$

- *absolute value regularizer* (a.k.a. ℓ_1):

$$r(\theta) = ||\theta||_1 = |\theta_1| + \cdots + |\theta_d|$$

Regularized empirical risk minimization

- predictor should fit the given data well, *i.e.*, we want empirical risk

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta^{\mathsf{T}} x^i, y^i)$$

  to be small

- predictor should not be too sensitive, *i.e.*, we want $r(\theta)$ small

- to trade off these two objectives, form *regularized empirical risk*

$$\mathcal{L}(\theta) + \lambda r(\theta)$$

  where $\lambda \geq 0$ is the *regularization parameter* (or *hyper-parameter*)

- *regularized empirical risk minimization* (RERM): choose θ to minimize regularized empirical risk

- an optimization problem

Regularized empirical risk minimization

- for $\lambda = 0$, RERM reduces to ERM
- RERM produces a *family* of predictors, one for each value of λ
- in practice, we choose a few tens of values of λ, usually logarithmically spaced over a wide range
- use validation to choose among the candidate predictors
- we choose the largest value of λ that gives near minimum test error (*i.e.*, least sensitive predictor that generalizes well)

Ridge regression

- *ridge regression*: square loss and regularizer $r(\theta) = ||\theta||^2$ (or $||\theta_{2:d}||^2$ if $x_1 = 1$)
- also called *Tykhonov regularized least squares*
- regularized empirical risk is

$$\begin{aligned} \mathcal{L}(\theta) + \lambda r(\theta) &= ||X\theta - y||^2 + \lambda||\theta||^2 \\ &= \left\| \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} \theta - \begin{bmatrix} y \\ 0 \end{bmatrix} \right\|^2 \end{aligned}$$

- so optimal θ is

$$\theta^\star = \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix}^\dagger \begin{bmatrix} y \\ 0 \end{bmatrix} = (X^T X + \lambda I)^{-1} X^T y$$

- (how do you modify this to handle $r(\theta) = ||\theta_{2:d}||^2$?)

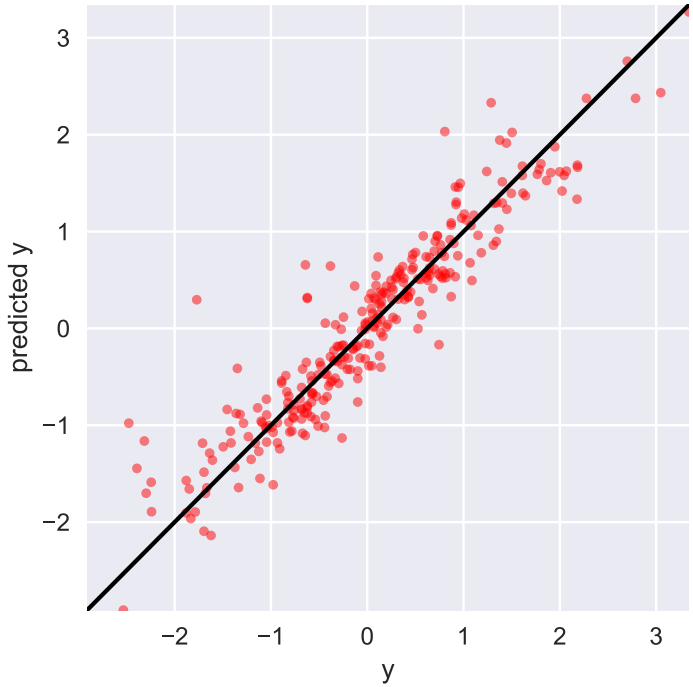
Example: House prices

- sale prices of 2930 homes in Ames, Iowa from 2006 to 2010
- 80 features
- we use 16 features

Example: Regression

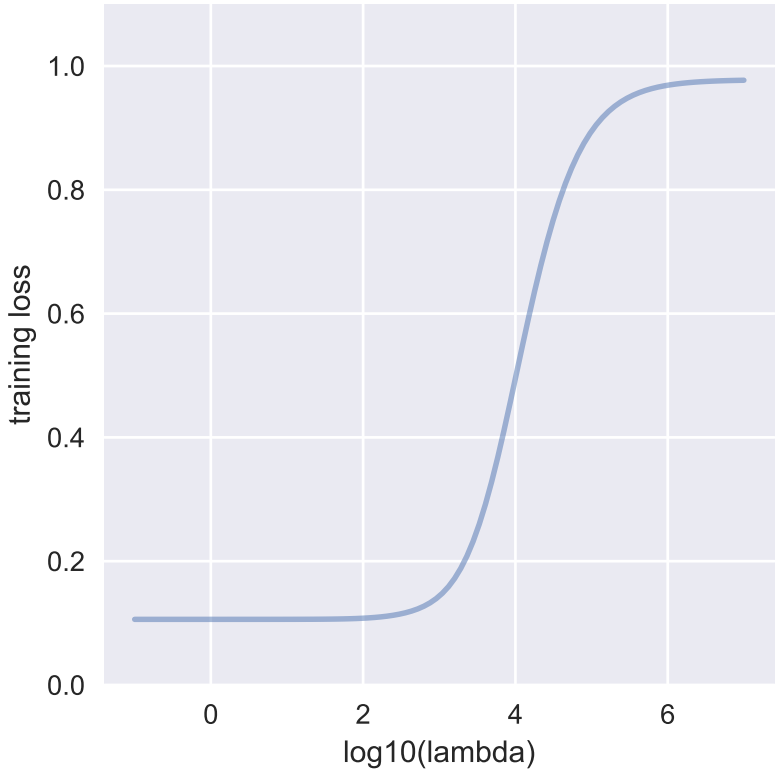
- we manually remove 4 outliers with area > 4000 (we'll see later how to detect outliers)

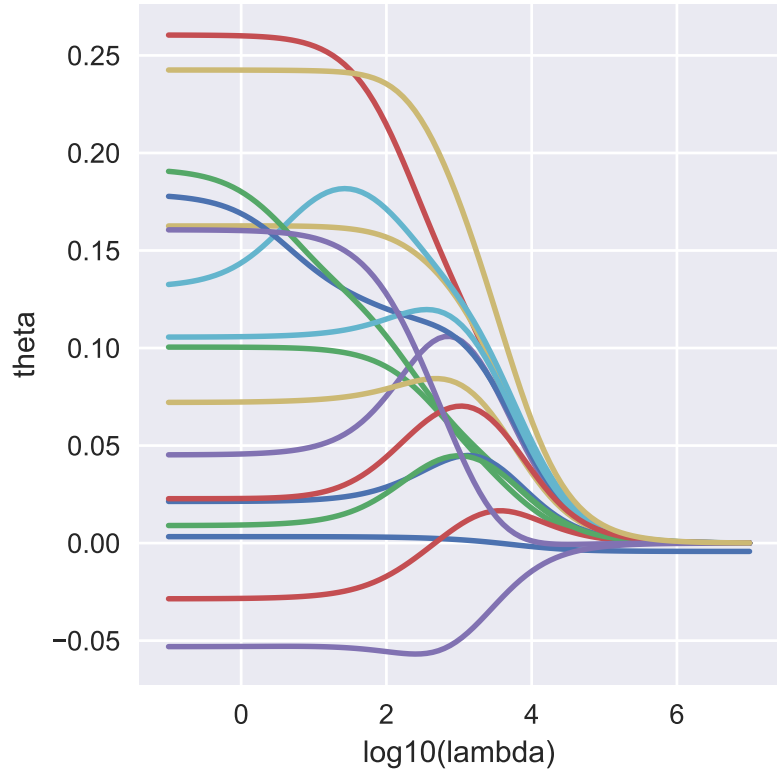
Example: Regression



- split data randomly into 1164 training, 291 test
- target is log(price)
- standardize all features (and log(price))
- training error 0.1060, test error 0.1361
- plot shows all test points

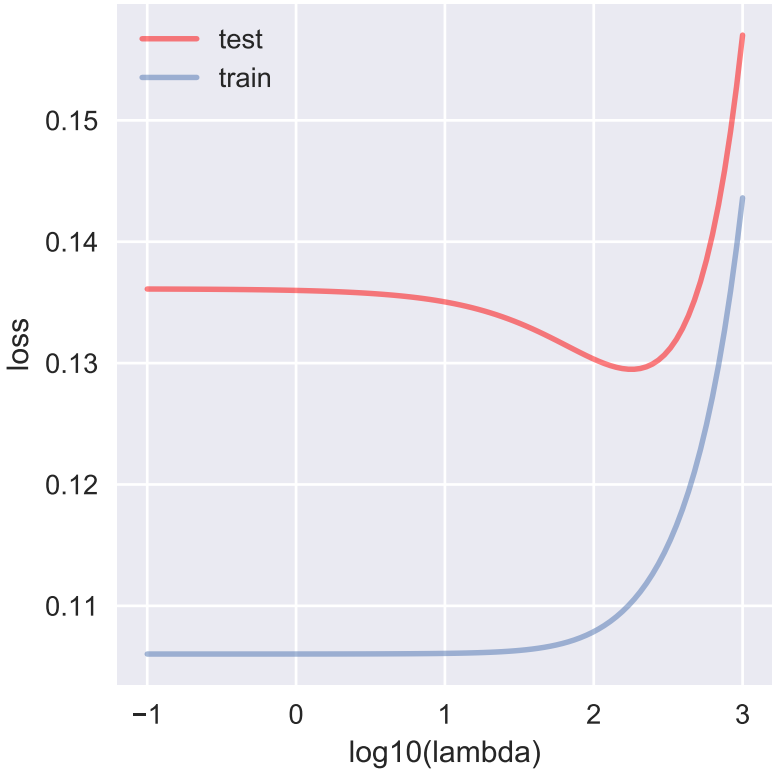
Example: Ridge regression

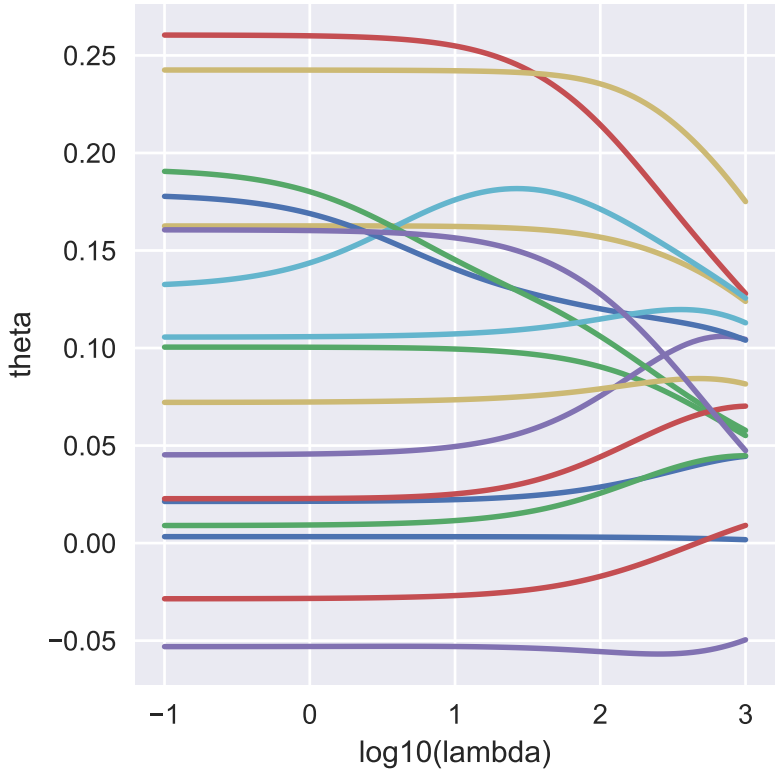




- leftmost error is training error with no regularization: 0.1060
- rightmost error is variance of training data: 0.9787
- plot of θ_i versus λ (on right) is called *regularization path*
- rightmost θ has $\theta_0 = -0.0043$, the mean of training y values

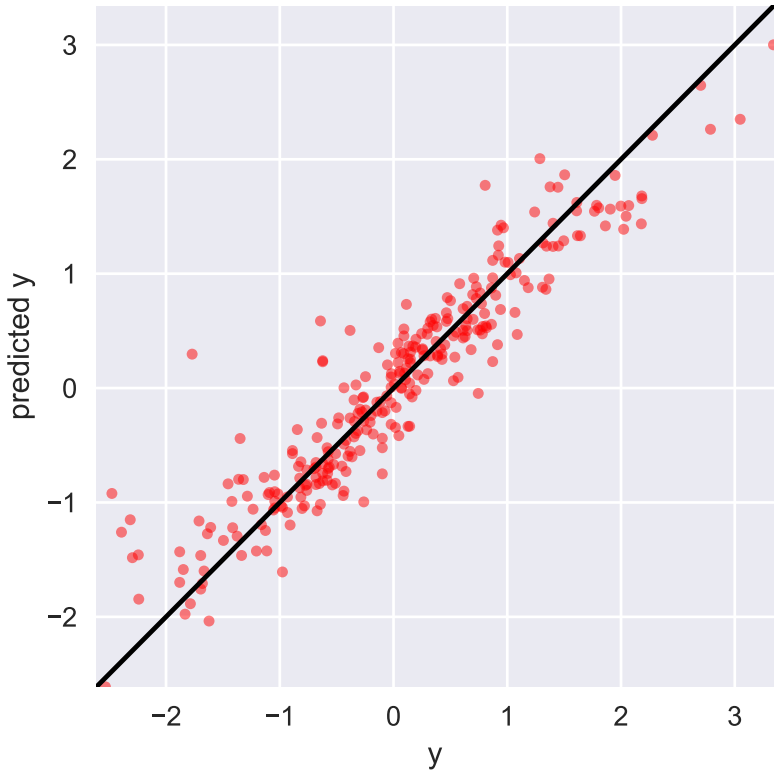
Example: Ridge regression





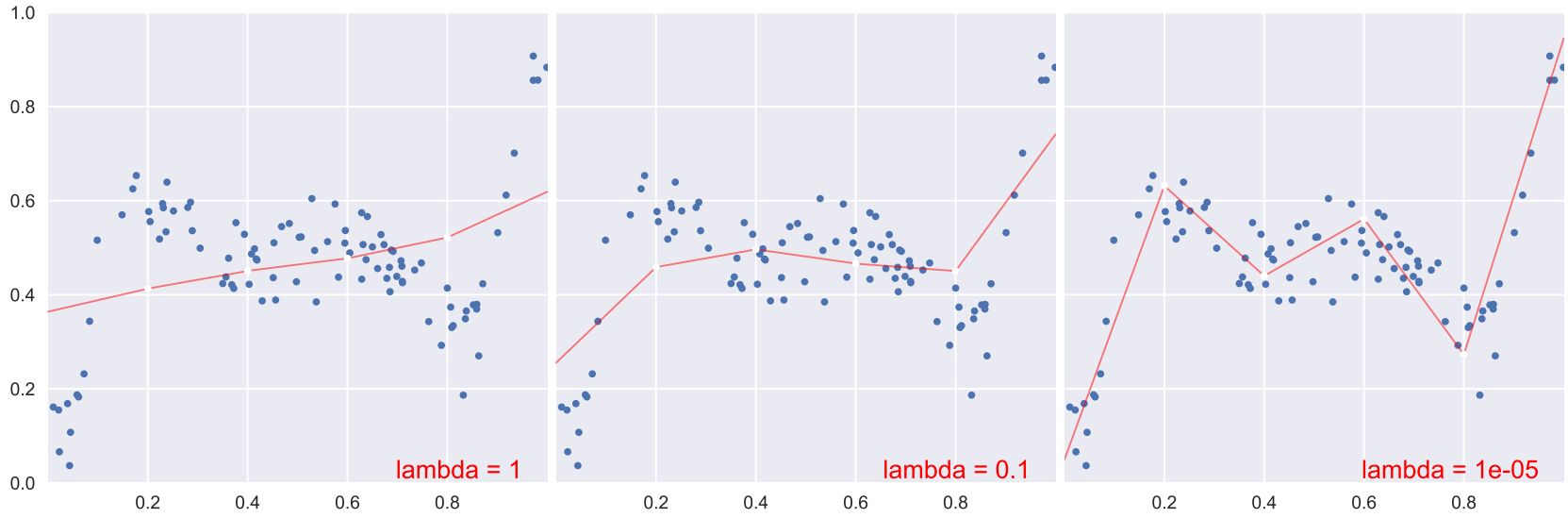
- regularization $\lambda = 187$ is optimal; improves test performance a bit
- θ is shrunk by regularization, so predictor is less sensitive

Example: Ridge regression



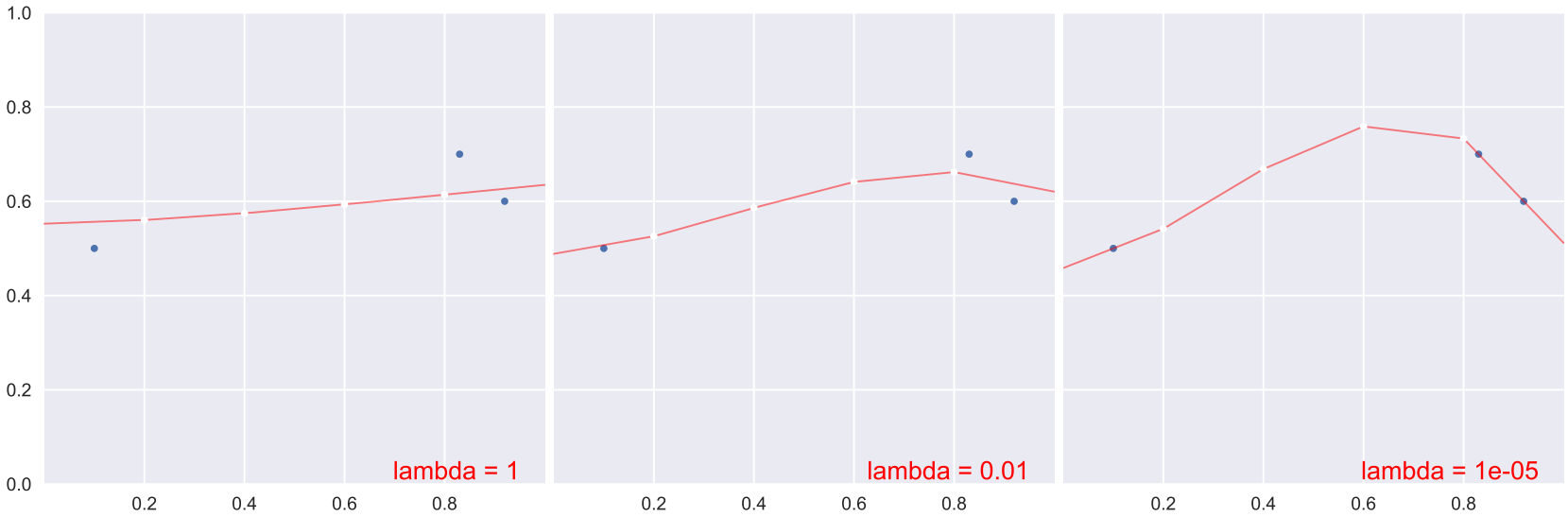
- least squares test error is 0.1361, with $||\theta|| \approx 0.55$
- ridge regression test error (with $\lambda = 178$) is 0.1295 with $||\theta|| \approx 0.46$
- ridge regression predictor is less sensitive

Example: Piecewise linear fit



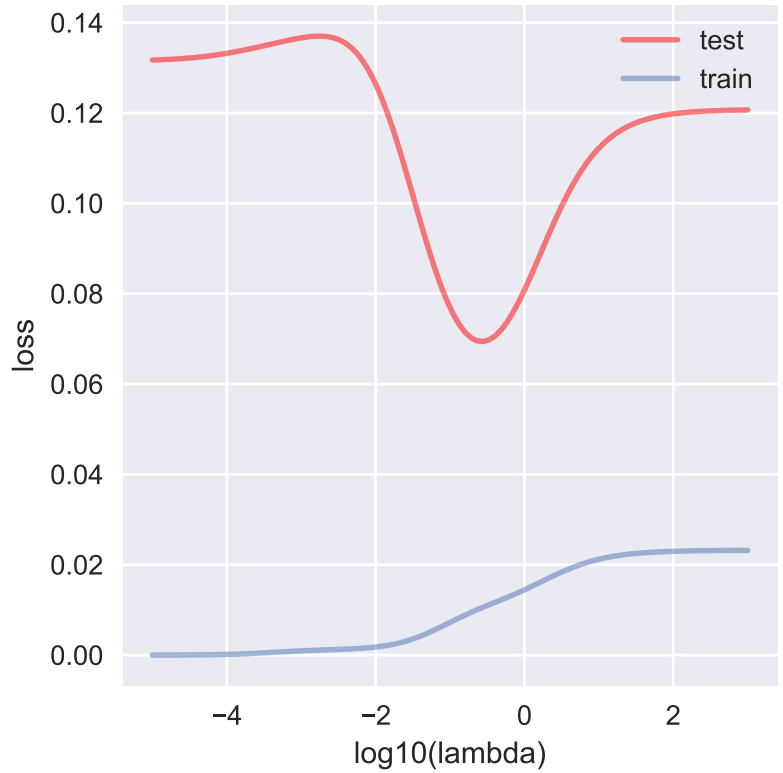
- features $x = (1, u, (u - 0.2)_+, (u - 0.4)_+, (u - 0.6)_+, (u - 0.8)_+)$
- $\lambda = 1$ gives $\theta = (0.36, 0.25, -0.057, -0.056, 0.089, 0.26)$
- $\lambda = 10^{-5}$ gives $\theta = (0.05, 2.9, -3.9, 1.6, -2, 4.8)$

Fitting predictors with more parameters than data points



- this makes no sense in general
- but with regularization, you can do this
- $\lambda = 1$ gives $\theta = (0.55, 0.039, 0.033, 0.022, 0.011, -0.0007)$
- $\lambda = 10^{-5}$ gives $\theta = (0.46, 0.42, 0.22, -0.18, -0.58, -0.98)$

Fitting predictors with more parameters than data points



- minimum point balances fitting training data versus sensitivity