Example: House Prices

Jong-Han Kim

EE787 Fundamentals of machine learning
Kyung Hee University

Price data

- sale prices of 2930 homes in Ames, Iowa from 2006 to 2010
- data contains 80 features

Features

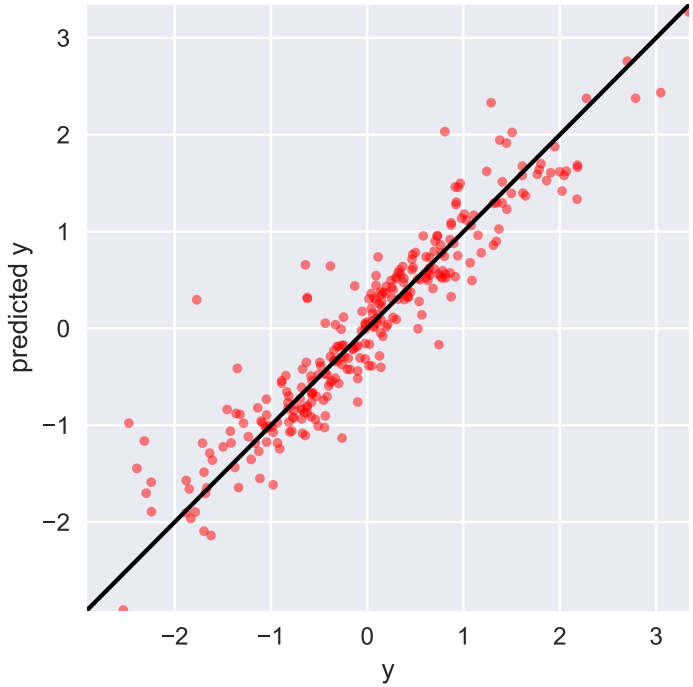
fit with 16 features

- area of lot
- year built
- year of last remodel
- area of basement
- area of living space (above ground)
- area of first floor
- area of second floor
- number of bedrooms (above ground)
- number of kitchens (above ground)
- number of fireplaces
- area of garage
- area of wooden deck
- number of half bathrooms
- number of rooms (above ground)
- overall condition (scored 1-10)
- overall quality of materials and finish (scored 1-10)

Outliers

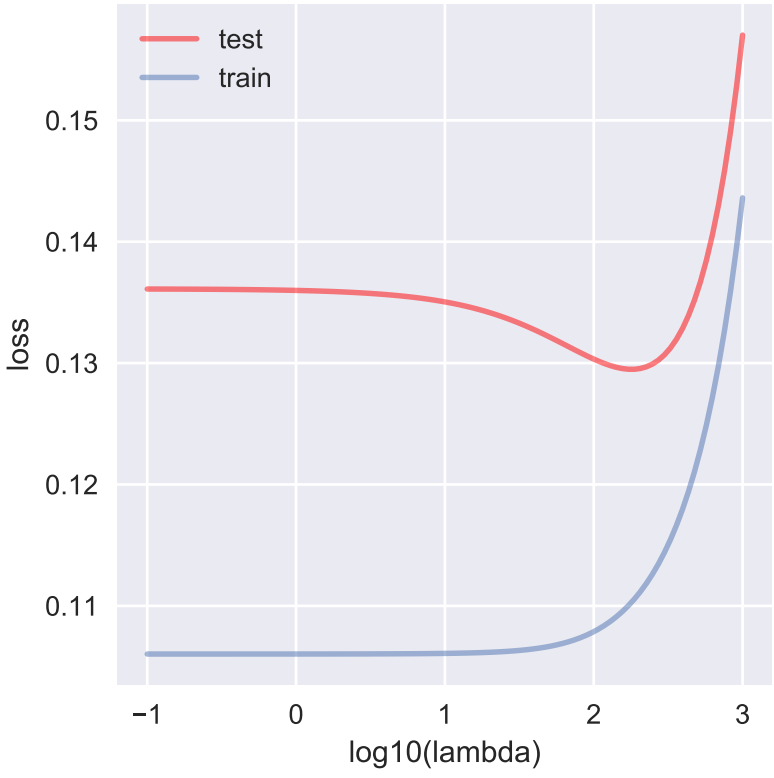
- we manually remove 4 outliers with area > 4000 (we'll see later how to detect outliers)

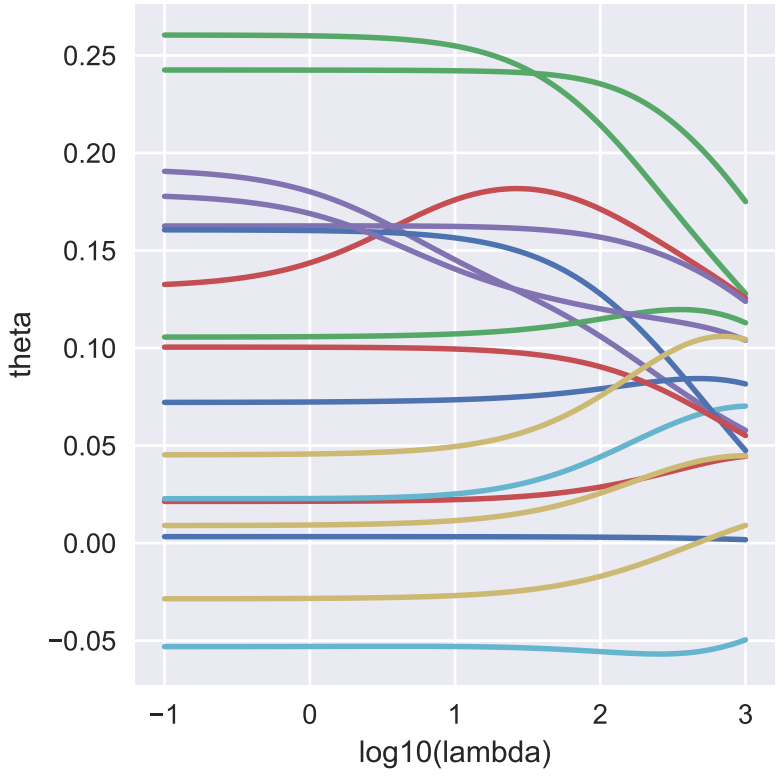
Regression



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

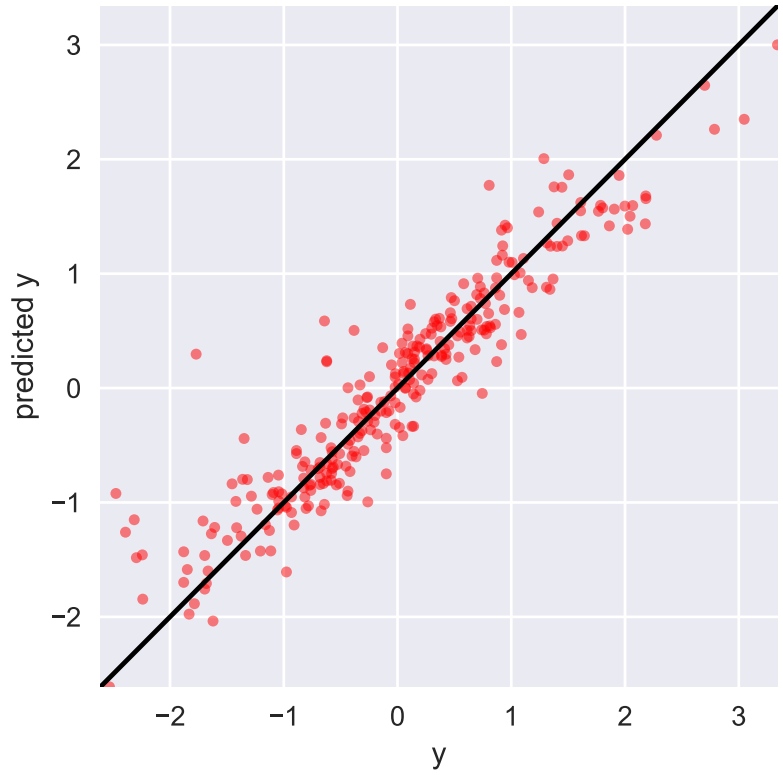
Ridge regression





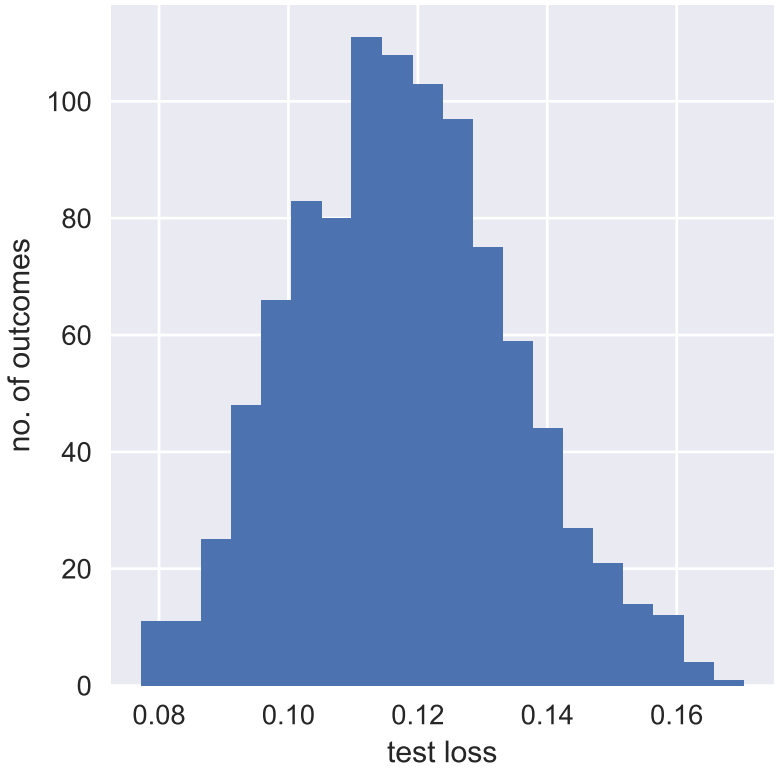
- regularization $\lambda = 187$ is optimal; improves test performance a bit
- θ is shrunk by regularization, so model is less sensitive

Ridge regression



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

Repeated train/test



- mean test loss 0.118

Some extra features

- 25 different neighborhoods, one-hot embedded
- 5 different building types, one-hot embedded

SINGLE-FAMILY TOWNHOUSE TWO-FAMILY-CONVERSION

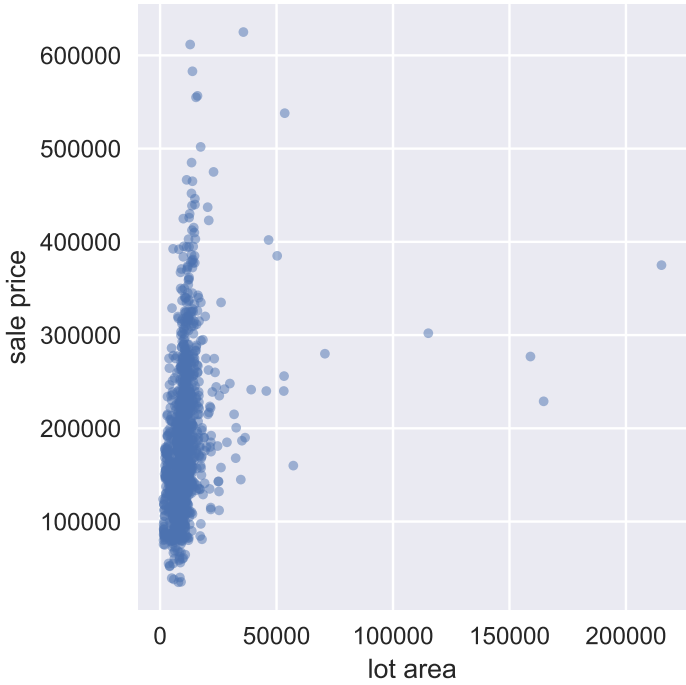
TOWNHOUSE DUPLEX

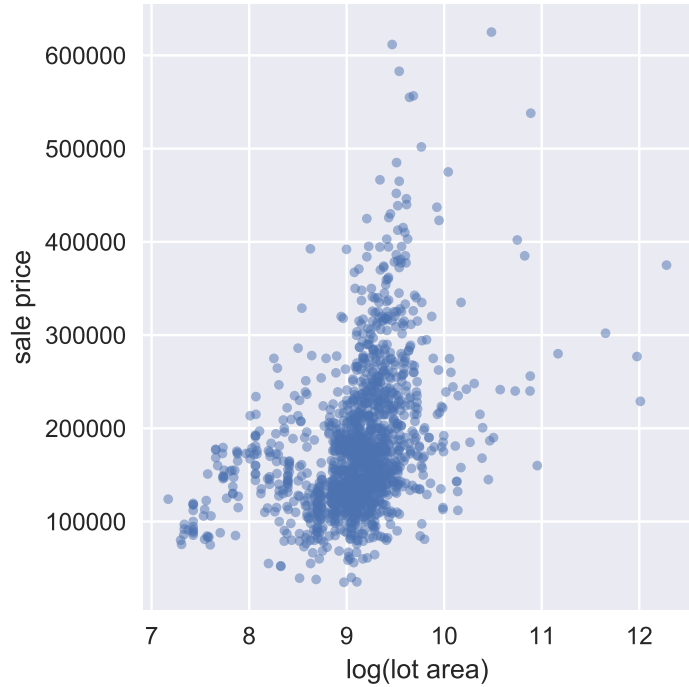
- kitchen quality, one-hot embedded (validated better than real embedding)

EXCELLENT, GOOD, TYPICAL, FAIR

- garage capacity, number of cars $\{0, 1, 2, 3, 4\}$, embedded as real (validated better than one-hot embedding)
- repeated test/train gives average test loss 0.101

Feature engineering





- additional feature $x_{\text{new}} = \log(\text{lot area})$
- additional feature $x_{\text{new}} = \log(\text{living area})$
- repeated test/train gives average test loss 0.0982

Feature engineering

- Boolean feature

$$\phi_{a+}(u) = \begin{cases} 1 & \text{if } u > a \\ 0 & \text{otherwise} \end{cases} \qquad \phi_{a-}(u) = \begin{cases} 1 & \text{if } u < a \\ 0 & \text{otherwise} \end{cases}$$

- add new features
  - ϕ_{1000+}(living area) and ϕ_{600-}(living area)
  - ϕ_{6000+}(lot area) and ϕ_{4000-}(lot area)
- repeated test/train gives average test loss 0.0973
- corresponds to mean percentage house price error $\approx 8\%$

Important features

OverallQual	0.172	KitchenQual-Ex	0.0543
YearBuilt	0.146	Neighborhood-Crawfor	0.054
TotalBsmtSF	0.124	Neighborhood-NridgHt	0.0522
OverallCond	0.1	Neighborhood-StoneBr	0.0484
GrLivArea	0.0994	loglot	0.0431
logliv	0.0887	KitchenAbvGr	-0.0408
1stFlrSF	0.0763	Neighborhood-Somerst	0.0371
YearRemodAdd	0.073	Neighborhood-NoRidge	0.0362
GarageArea	0.0656	Neighborhood-Edwards	-0.035
Fireplaces	0.0601	liv-	-0.0345
GarageCars	0.0595	WoodDeckSF	0.0305
2ndFlrSF	0.0584	HalfBath	0.0293
Neighborhood-IDOTRR	-0.0569	LotArea	0.0292
Neighborhood-OldTown	-0.0565	lot-	-0.0284