CIS 490 Machine Learning

Lecture 7

Instructor: (Julia) Hua Fang

1

2

Last time

Linear Regression (2):

- Running R for linear regression

You are expected to

- Interpret outputs
- Understand Training and Testing sets
- Understand Training and Testing errors

Adapted from Jeff Howbert, Greg Shakhnarovich, Patrick Breheny , M. Magdon-Ismail, Patrick Breheny, Jeff Schneider

2

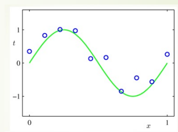
3

Answers in LS6 Suppl.: Input X for linear regression

- Types of the inputs **X** for linear regression:
  - **Original quantitative inputs**: "raw" data without transformation
  - **Transformation of quantitative inputs:** using log (), exp (), square root (), square (), etc.

    e.g., Income ($), use Log (income)

    - Polynomial linear regression



$$y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \ldots \beta_h \cdot x^h$$

x = time (e.g., hours, weeks, years); β_1 Slope;

x^2 = time^2; β_2: acceleration/ deceleration rate, etc.

When h =2, what x^2 is called?? Answer: Quadratic

When h =3, what x^3 is called ?? Answer: Cubic

https://online.stat.psu.edu/stat462/node/158/

(Why linear?)

Answer: since it is linear in the regression coefficients, $\beta_1, \beta_2, \ldots, \beta_h$

3

4

Quiz:

1. Can you use the same set as training set and testing set? Y/N
2. Testing error is not the same as training error and sometimes larger than training error. T/F

4

Outline

5

- Supported Learning
  - ➢Regression
    - ❖ Linear regression
    - ❖ Regularized linear regression

You are expected to understand:

- ✓ Overfitting: Random error vs. deterministic error; Bias-Variance tradeoff
- ✓ Regularization for linear regression:
  - Ridge regression
  - Lasso regression
    - -- Cross Validation

Adapted from Jeff Howbert, Greg Shakhnarovich, Patrick Breheny , M. Magdon-Ismail, Patrick Breheny, Jeff Schneider

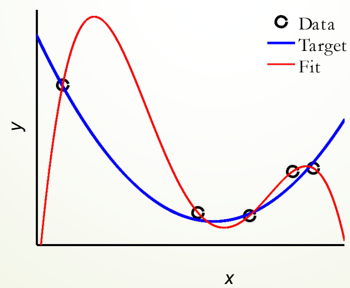
5

Overfitting

6

Overfitting: When

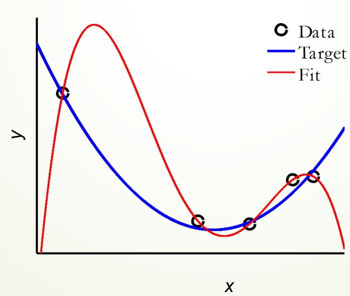
- Overfitting occurs when
  1. A statistical model describes random error or noise, instead of the underlying relationship



7

Overfitting: When

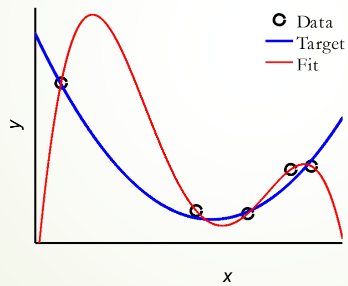
- Overfitting occurs when
  2. A model is excessively complex, such as having too many parameters relative to the number of observations.



8

Overfitting: Why

- Because it overreacts to minor fluctuations in the training data.



9

Overfitting:

Random Error vs. Deterministic Error

10

Random Error

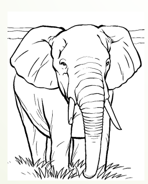
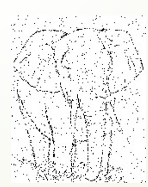
Random error is also called stochastic noise

11

Random Error/ stochastic noise

- What is random error?

  -- Fluctuations/measurement errors that we cannot model.

y   $f(x)$

12

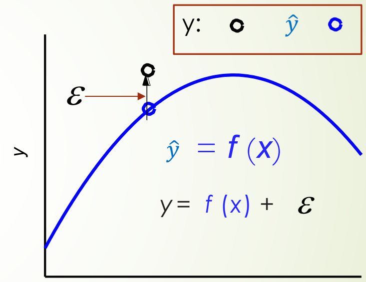
Random Error/ stochastic noise

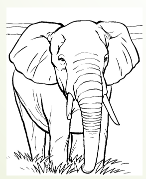
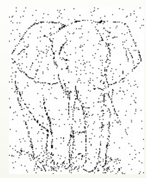
- What is random error? Mathematically,

Population: $y = f(x) + \varepsilon$ $\quad\quad \varepsilon \sim N(0,\sigma^2)$

Given a learner/estimated model

$\hat{y} = f(x)$



y   $f(x)$

13

Deterministic Error

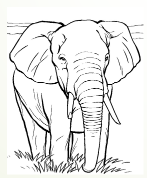
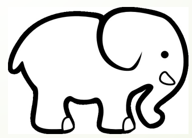
Deterministic Error is also called Deterministic noise

14

Deterministic Error/ Noise

What is Deterministic error?

--Model Error/Bias

y   $f(x)$

15

Deterministic Error/ Noise

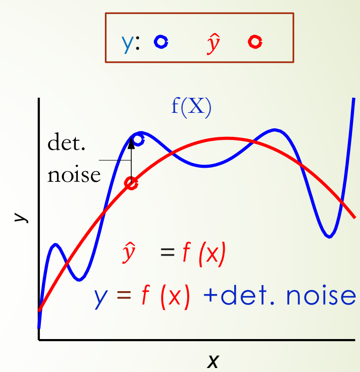
What is Deterministic error?

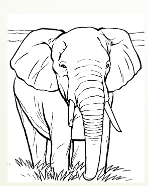
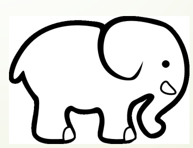
Population

$y = f(x)$ + 'deterministic noise'

Given a learner/estimated model :

$\hat{y} = f(x)$



y   $f(x)$

16

Overfitting:

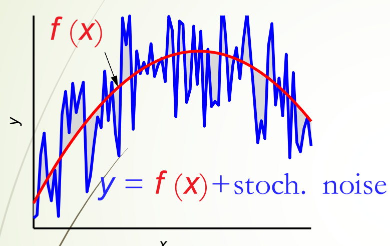
Random Error vs. Deterministic Error

(in-class exercises)

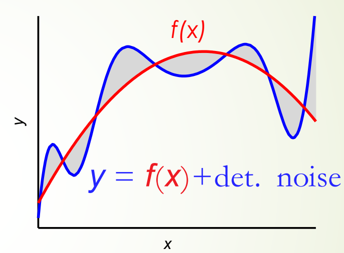
17

In-class exercise (1)

Random Error/Stochastic Noise



Deterministic Error/Noise

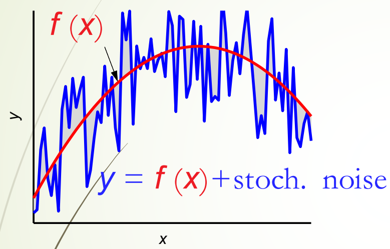


Where does Stochastic or Deterministic Noise come from?

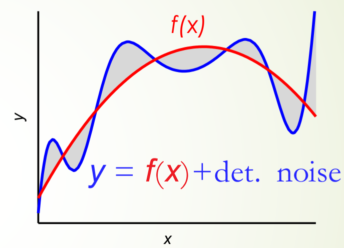
18

In-class exercise (2)

Random Error/Stochastic Noise



Deterministic Error/Noise



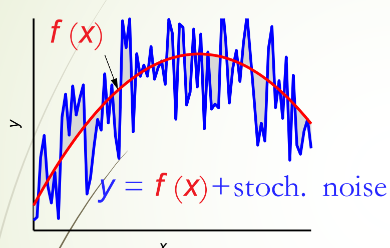
If we remeasure y, would Stochastic Error change?

If we remeasure y, would Deterministic Noise change?

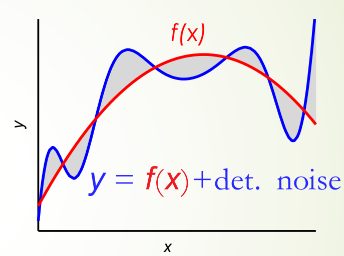
19

In-class exercise (3)

Random Error/Stochastic Noise



Deterministic Error/Noise



If we change f(x), would Stochastic Error change?

If we change f(x), would Deterministic Noise change?

20

Overfitting: Bias – Variance tradeoff

21

Bias – Variance tradeoff

$$SSE/N = MSE = \underbrace{\text{Variance}}_{\substack{\text{Random Error/} \\ \text{Stochastic Noise}}} + \underbrace{\text{Bias}^2}_{\substack{\text{Deterministic} \\ \text{Error/Noise}}}$$

"It might be wise to use a biased estimator so long as it reduces our variance, assuming our goal is to minimize squared error." (Murphy, Section 6.4.4, a must read)

"Murphy": Means the reference book--Machine Learning: A Probabilistic Perspective. (2012) Kevin P. Murphy. The MIT press. Cambridge, Massachusetts. ISBN 978-0-262-01802-9.
Will use "Murphy" for simplicity for later lectures.

22

Overfitting: Bias – Variance tradeoff
(in-class exercises)

23

24

In class exercise

1. Given the amount of MSE is fixed, if variance increases, will bias increase or decrease?
2. Given the amount of MSE is fixed, if bias increases, will variance increase or decrease?
3. Given the amount of bias is fixed, if variance increases, will MSE increase or decrease?

Recall SSE/N =MSE = Variance + Bias^2

24

What should we do with overfitting?

25

Regularization!

26

What is regularization?

- refers to a process of introducing additional information in order to prevent overfitting

How does it work?

27

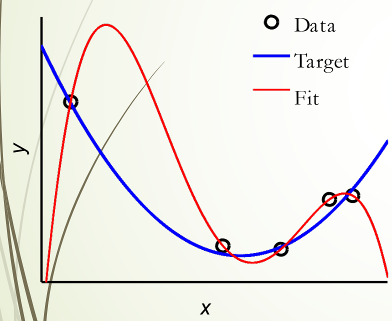
28

Constraining/penalizing the model!

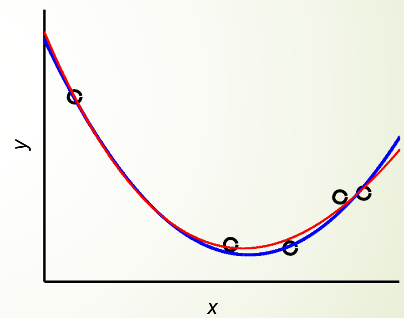
28

Effect of Regularization: Illustrative

no regularization



regularization!

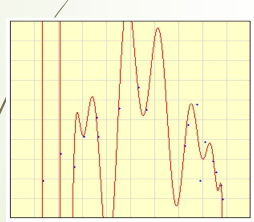


29

Effect of Regularization: Illustrative

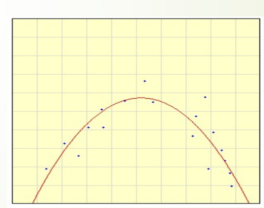
- Overfitting usually leads to very large parameter estimate choices, e.g.:

-1.1 + 4,700,910.7 X – 8,585,638.4 X^2 + …



no regularization

-2.2 + 3.1 X – 0.30 X^2

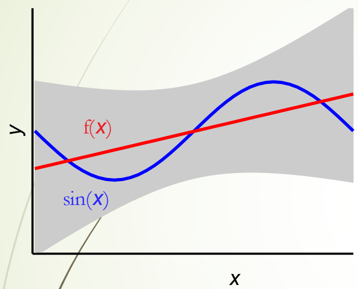


regularization!

30

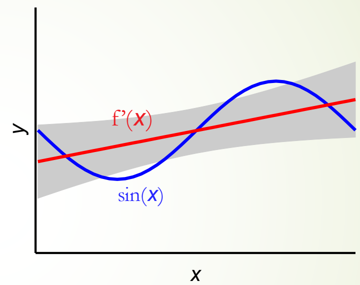
Effect of Regularization: Bias-Variance Trade-off

No regularization



Example: bias = 0.21
var = 1.69

regularization



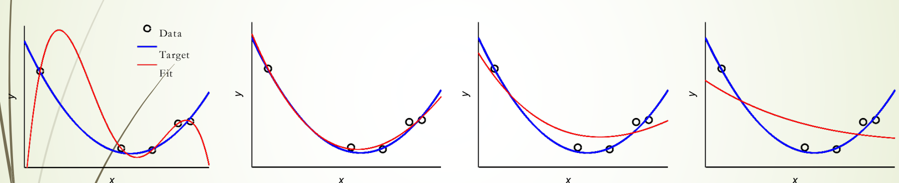
Example: bias = 0.23
var = 0.33

Recall SSE/N =MSE = Variance + Bias^2

31

Effect of Regularization; Don't Overdose (illustrative graph)

λ = 0 | λ = 0.0001 | λ = 0.01 | λ = 1



Overfitting → → Underfitting

What is λ?

32

Regularized regression:
a.k.a penalized regression

33

Regularized regression

- Regularized regression aims to impose a “complexity” penalty by penalizing (ie. regularizing) coefficient estimates (also called “weights”).

Equivalently, by “shrinking the coefficient estimates towards Zero”.

34

Regularized regression

- Regularized regression: we focus on
  - Regularized linear regression:
    - Ridge regression
    - Lasso regression

Advantages of Ridge and Lasso:

- ✓ Can help reduce variance: rooted in the bias-variance trade-off.
- ✓ Computational advantage: include all predictors in the model and run once,

35

36

Ridge regression:

L_2 regularization

36

Ridge Regression: L_2 regularization

Recall: Least squares estimation for linear regression: find $\hat{\beta}$ that minimize

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

- Ridge regression: a.k.a. L_2 regularization.

Goal: Find $\hat{\beta}^R$ which minimizes:

$$\underbrace{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{RSS}} + \lambda \underbrace{\sum_{j=1}^{p}\beta_j^2}_{\text{Shrinkage Penalty Term}} = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2 \qquad \text{Equation (1)}$$

where $\lambda \geq 0$ is the regularization/tuning parameter; λ controls the amount of regularization

37

Ridge Regression: L_2 regularization

- Least squares estimation for linear regression: generates only one set of coefficient estimates, $\hat{\beta}$
- Ridge regression: produce a different set of coefficient estimates, $\hat{\beta}^R_\lambda$, for each value of λ,

so another formulation for Ridge

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \quad \text{subject to} \quad \sum_{j=1}^{p}\beta_j^2 \leq s. \qquad \text{Equation (2)}$$

for every value of λ there is a corresponding *s, that* gives $\hat{\beta}^R_\lambda$
Note: Equation (2) is an alternative formation of Equation (1)
(see ITSL, "*An Introduction To Statistical Learning,*" 6.2)

The notation $\|\beta\|_2$ denotes the ℓ_2 norm (pronounced "ell 2") of a vector, and is defined as $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p}\beta_j{}^2}$.

$\|\beta\|_2$ Measures the distance of $\hat{\beta}$ to 0.

38

Ridge Regression: L_2 regularization

- Recall Least squares solution (with calculus) for $\hat{\beta}$:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- The solution to the ridge regression problem is given by

$$\hat{\beta}^R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Note the similarity to the (ordinary) least squares solution, but with the addition of a "ridge" down the diagonal.

39

Ridge Regression: L_2 regularization

- Recall $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$

$$\hat{\beta}^R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Can you tell?

If $\lambda = 0$, $\hat{\beta}^R$ =?

If $\lambda \to \infty$, $\hat{\beta}^R$ = ?

40

Ridge Regression: L_2 regularization

- **Corollary:** As $\lambda \to 0$, $\hat{\beta}^R \to \hat{\beta}^{\text{OLS}}$
- **Corollary:** As $\lambda \to \infty$, $\hat{\beta}^R \to 0$

Shrinkage: the ridge regression penalty has the effect of shrinking the estimates toward zero

Interpretation of Shrinkage in the context of bias-variance trade-off:

- $\lambda = 0$, $\hat{\beta}^R = \hat{\beta}^{\text{OLS}}$, the variance is high but no bias.
- λ increases, the shrinkage of $\hat{\beta}^R$ leads to a substantial reduction in the variance of the predictions, at the expense of a slight increase in bias

41

42

Next time

- Lasso
- Running ridge and lasso: R
  - Introduce Cross-validation

42

43

Reading assignments

- (ITSL)6.2.1
- (ITSL)6.2.2: read through "Comparing the Lasso and Ridge Regression"

ITSL: "An Introduction to statistical learning with applications in R"

43