Data100 Sp22 Disc 9

Cross Validation/Bias Variance

Attendance:

https://tinyurl.com/disc9michelle

Announcements

Due Dates

- Proj 1B (Housing) due **TONIGHT**

Other

-Have a great spring break! ଘ(੭´ᵕ`)੭

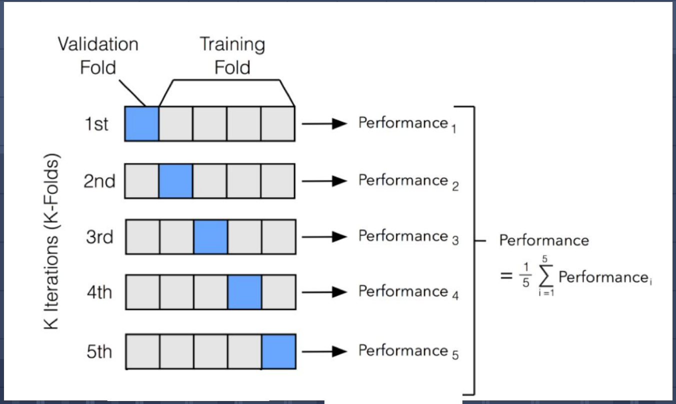
Cross Validation

1

Motivation

- Test sets are often unseen
- Help us use our data better

5-fold is generally used



Q1

1. After running 5-fold cross validation, we get the following mean squared errors for each fold and value of λ when using Ridge regularization:

↑ minimize

Fold Num	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	Row Avg
1	80.2	70.2	91.2	91.8	83.4
2	76.8	66.8	88.8	98.8	82.8
3	81.5	71.5	86.5	88.5	82.0
4	79.4	68.4	92.3	92.4	83.1
5	77.3	67.3	93.4	94.3	83.0
Col Avg	79.0	68.8	90.4	93.2	

How do we use the information above to choose our model? Do we pick a specific fold? a specific λ? or a specific fold-λ pair? Explain.

$\lambda = 0.2$

Q2

2. In the typical setup of k-fold cross validation, we use a different parameter value on each fold, compute the mean squared error of each fold and choose the parameter whose fold has the lowest loss.

   ○ A. True

   ✓ B. False

Probability

2

Binomial Distribution

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Q3

3. A multiple choice test has 100 questions, each with five possible answers of which one is right. The grading scheme is as follows:

   - 4 points are awarded for each right answer.
   - For each other answer (wrong, missing, etc), one point is taken off; that is, -1 points are awarded.

   A student hasn't studied at all and therefore guesses each answer uniformly at random, independently of all the other answers.

   iid ⇒ binomial

   Define the following random variables:

   - → R: the number of answers the student gets right
   - → W: the number of answers the student does not get right
   - S: the student's score on the test

   We analyze the random variable R, which denotes the number of answers the student got right.

   (a) What is the distribution of R? Provide the name and parameters of the appropriate distribution. Explain your answer.

   $R \sim \text{Binom}(n=100,\ p=0.2)$

   Binom expectation $= n \cdot p$

   (b) Find $\mathbb{E}(R)$.

   $100 \cdot \frac{1}{5} = 20$

Q4

4. True or False: $\mathbb{SD}(R) = \mathbb{SD}(W)$.

Q5

5. Find $\mathbb{E}(S)$, the student's expected score on the test.

$$E[S] = E[4 \cdot R + (-1) \cdot W] = 4E[R] - 1E[W] = 4 \cdot 20 - 1 \cdot 80 = 0$$

Q6

6. Find $\mathbb{SD}(S)$.

Bias Variance Tradeoff

* very important *

3

What is bias/variance?

Bias -> How "accurate"/"well" our model is doing

Variance -> How our model performs on other data

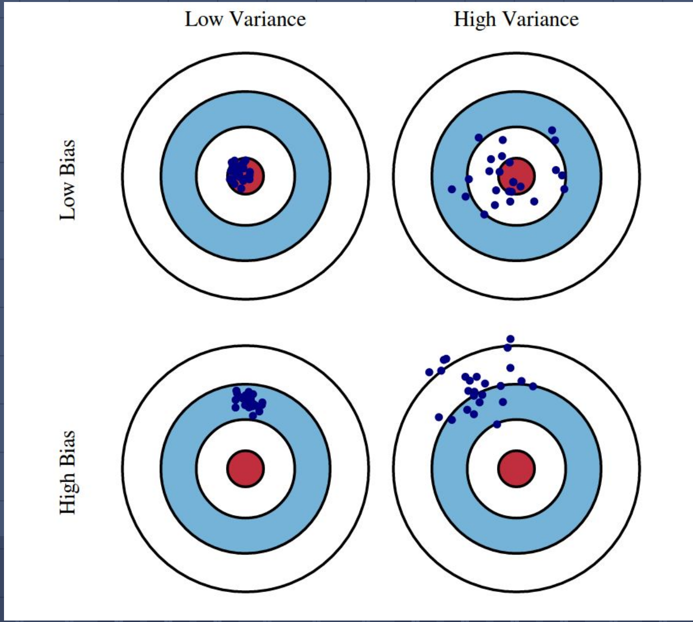
Example: Predicting gas prices

constant model: predicts gas price is always $1/gallon

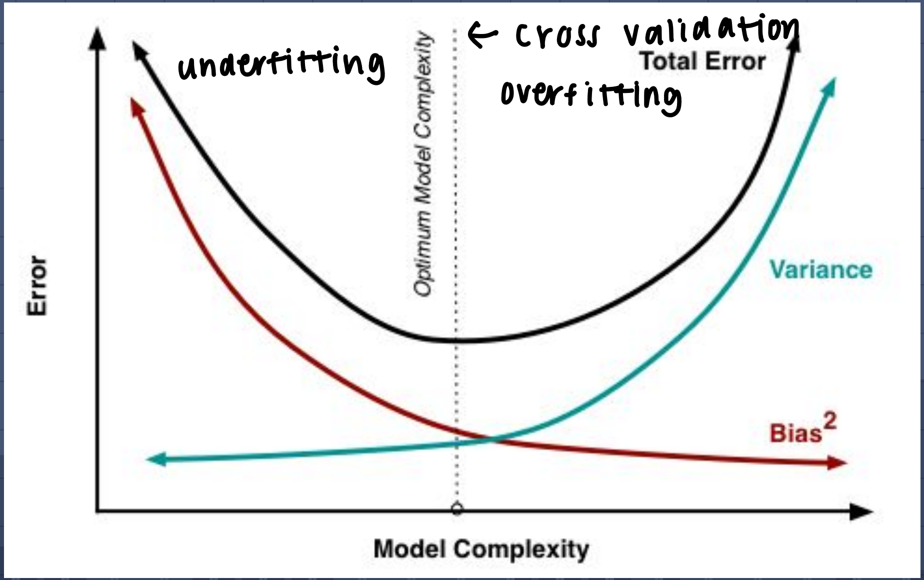
Bias ⇒ High

variance ⇒ Low

Feature: If war/not


Low Variance
High Variance
Low Bias
High Bias

Bias variance Tradeoff

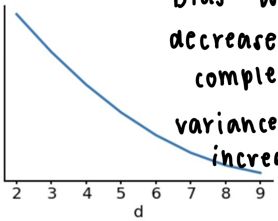


Q7

7. Your team would like to train a machine learning model in order to predict the next YouTube video that a user will click on based on the videos the user has watched in the past. We extract m attributes (such as length of video, view count etc) from each video and our model will be based on the previous d videos watched by that user.

Hence the number of features for each data point for the model is $m \cdot d$. Currently, you're not sure how many videos to consider.

(a) Your colleague generates the following plot, where the value d is on the x-axis. However, they forgot to label the y-axis.



Bias will always decrease with complexity

variance will always increase with complexity

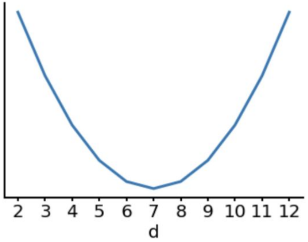
# of features ⟶

Which of the following could the y-axis represent? Select all that apply.

- ☑ A. Training Error
- ☑ B. Validation Error
- ☑ C. Bias
- ☐ D. Variance

Bias ⟺ Training Error

(b) Your colleague generates the following plot, where the value d is on the x-axis. However, they forgot to label the y-axis again.



Which of the following could the y axis represent? Select all that apply.

- ☐ A. Training Error
- ☑ B. Validation Error
- ☐ C. Bias
- ☐ D. Variance

Q8

8. We randomly sample some data $(x_i, y_i)_{i=1}^n$ and use it to fit a model $f_{\hat{\theta}}(x)$ according to some procedure (e.g. OLS, Ridge, LASSO). We then sample a new point that is independent from our existing points, but sampled from the same underlying truth as our data. Furthermore, assume that we have a function $g(x)$ and some noise generation process that produces ϵ such that $\mathbb{E}[\epsilon] = 0$ and $\text{var}(\epsilon) = \sigma^2$. Every time we query mother nature for Y at a given a x, she gives us $Y = g(x) + \epsilon$. (The true function for our data is $Y = g(x) + \epsilon$.) A new ϵ is generated each time, independent of the last. In class, we showed that

$$\underbrace{\mathbb{E}\left[(Y - f_{\hat{\theta}}(x))^2\right]} = \underbrace{\sigma^2} + \underbrace{(g(x) - \mathbb{E}[f_{\hat{\theta}}(x)])^2} + \underbrace{\mathbb{E}\left[(f_{\hat{\theta}}(x) - \mathbb{E}[f_{\hat{\theta}}(x)])^2\right]}$$

(a) Label each of the terms above.

Word Bank: observation variance, model variance, observation bias^2, model bias^2, model risk, empirical mean square error.

(b) What is random in the equation above? Where does the randomness come from?

(c) Calculate the value of $\mathbb{E}[\epsilon f_{\hat{\theta}}(x)]$.