DS 100/200: Principles and Techniques of Data Science **Date: April 10, 2020**

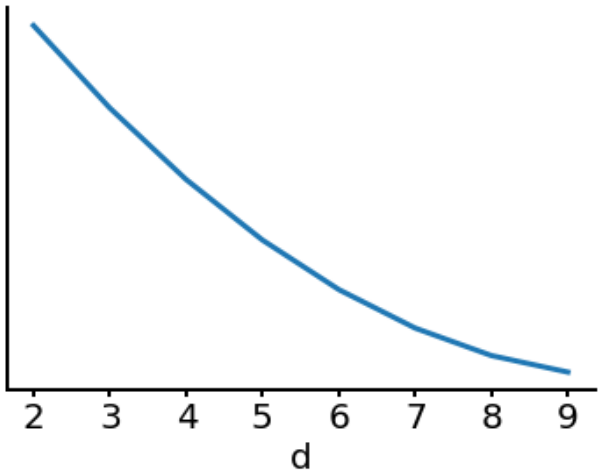
Discussion #10

Name:

Bias-Variance Trade-Off

1. 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 $m \cdot d$. You're not sure how many videos to consider.

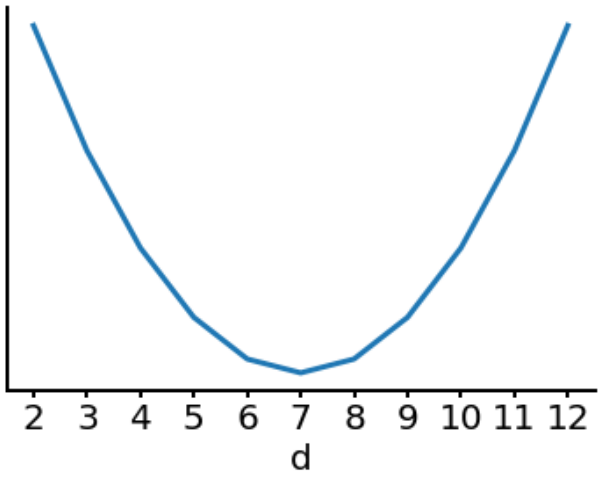
   (a) Your colleague generates the following plot, where the value d is on the x axis. However, they forgot to label the y-axis.



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

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

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



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

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

2. 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

$$\underbrace{\mathbb{E}\left[(Y - f_{\hat{\theta}}(x))^2\right]}_{\text{obs. variance}} = \underbrace{\sigma^2}_{} + \underbrace{(g(x) - \mathbb{E}[f_{\hat{\theta}}(x)])^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[(f_{\hat{\theta}}(x) - \mathbb{E}[f_{\hat{\theta}}(x)])^2\right]}_{\text{model variance}}$$

(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?

Y: depends on ϵ
- noise in Y (from ϵ)

$f_{\hat{\theta}}(x)$: random
- generated from a sample of Ys

$E[\epsilon] = 0$
$\text{Var}[\epsilon] = \sigma^2$

(c) True or false and explain. $\mathbb{E}[\epsilon f_{\hat{\theta}}(x)] = 0$

True: ϵ and $f_{\hat{\theta}}(x)$ are independent.
For independent RVs, $E[XY] = E[X]E[Y]$.
Then, $E[\epsilon f_{\hat{\theta}}(x)] = E[\epsilon]\,E[f_{\hat{\theta}}(x)] = 0 \cdot E[f_{\hat{\theta}}(x)] = 0$.

(d) Suppose you lived in a world where you could collect as many data sets you would like. Given a fixed algorithm to fit a model f_θ to your data e.g. linear regression, describe a procedure to get good estimates of $\mathbb{E}\left[f_{\hat{\theta}}(x)\right]$

repeat as many times as possible: → then, take average of all fitted models
- sample data
- fit model

(e) If you could collect as many data sets as you would like, how does that affect the quality of your model $f_\theta(x)$? → good estimate of $E[f_{\hat{\theta}}(x)]$
→ but if you chose a poor model to begin with, doesn't improve it

Ridge and LASSO Regression

3. Earlier, we posed the linear regression problem as follows: Find the $\vec{\theta}$ value that minimizes the average squared loss. In other words, our goal is to find $\hat{\vec{\theta}}$ that satisfies the equation below:

$$\hat{\vec{\theta}} = \underset{\vec{\theta}}{\operatorname{argmin}}\, L(\vec{\theta}) = \underset{\vec{\theta}}{\operatorname{argmin}}\, \frac{1}{n}||\vec{y} - \mathbb{X}\vec{\theta}||_2^2$$

Here, $\mathbb{X}$ is a $n \times d$ matrix, $\vec{\theta}$ is a $d \times 1$ vector and $\vec{y}$ is a $n \times 1$ vector. As we saw in lecture, the optimal $\hat{\vec{\theta}}$ is given by the closed form expression $\hat{\vec{\theta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^t\vec{y}$.

To prevent overfitting, we saw that we can instead minimize the sum of the average squared loss plus a regularization function $\alpha\mathcal{S}(\vec{\theta})$. If use the function $\mathcal{S}(\vec{\theta}) = ||\vec{\theta}||_2^2$, we have "ridge regression". If we use the function $\mathcal{S}(\vec{\theta}) = ||\vec{\theta}||_1$, we have "LASSO regression". For example, if we choose S($\vec{\boldsymbol{\theta}}$) = $||\vec{\theta}||_2^2$, our goal is to find $\hat{\vec{\theta}}$ that satisfies the equation below:

$\alpha = \lambda$

$$\hat{\theta} = \underset{\vec{\theta}}{\operatorname{argmin}}\, L(\vec{\theta}) = \underset{\vec{\theta}}{\operatorname{argmin}}\, \frac{1}{n}||\vec{y} - \mathbb{X}\vec{\theta}||_2^2 + \alpha||\vec{\theta}||_2^2 = \underset{\vec{\theta}}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{X}_{i,\cdot}^T\vec{\theta})^2 + \alpha\sum_{j=1}^{d}\theta_j^2$$

Recall that α is a hyperparameter that determines the impact of the regularization term. Though we did not discuss this in lecture, we can also find a closed form solution to ridge regression: $\hat{\vec{\theta}} = (\mathbb{X}^T\mathbb{X} + n\alpha\mathbf{I})^{-1}\mathbb{X}^T\vec{y}$. It turns out that $\mathbb{X}^T\mathbb{X} + n\alpha\mathbf{I}$ is guaranteed to be invertible (unlike $\mathbb{X}^T\mathbb{X}$ which might not be invertible).

(a) As model complexity increases, what happens to the bias and variance of the model?

complexity ↑

bias ↓

var ↑

(b) In terms of bias and variance, how does a regularized model compare to ordinary least squares regression?

decreased complexity → increased bias → lower variance

(c) In ridge regression, what happens if we set $\alpha = 0$? What happens as α approaches ∞?

$\alpha : 0$: same as OLS

$\alpha \to \infty$: $\theta \to 0$

(d) How does model complexity compare between ridge regression and ordinary least squares regression? How does this change for large and small values of α?

answered above

(e) If we have a large number of features (10,000+) and we suspect that only a handful of features are useful, which type of regression (Lasso vs Ridge) would be more helpful in interpreting useful features?

LASSO : encourages sparsity

(f) What are the benefits of using ridge regression?

guaranteed solution, more general model

Random Variables

4. The average response time for a question on Piazza this semester was 11 minutes. As always, the number of questions answered by each TA is highly variable, with a few TAs going above and beyond the call of duty. Below are the number of contributions for the top four TAs (out of $20,000$ total Piazza contributions):

TA	# contributions
Daniel	2000
Suraj	1800
Manana	700
Allen	500

Suppose we take a sample with replacement of size $n = 500$ contributions from the original $20,000$ contributions. We will also define some random variables:

- $D_i = 1$ when the i^{th} contribution in our sample is made by Daniel; else $D_i = 0$.

- $S_i = 1$ when the i^{th} contribution in our sample is made by Suraj; else $S_i = 0$.
- $M_i = 1$ when the i^{th} contribution in our sample is made by Manana; else $M_i = 0$.
- $A_i = 1$ when the i^{th} contribution in our sample is made by Allen; else $A_i = 0$.
- $O_i = 1$ when the i^{th} contribution is made by anyone other than Daniel, Suraj, Manana, or Allen; else, $O_i = 0$

what type of RV

Bernoulli

(a) i. What is $P(A_1 = 1)$?

$P(A_1 = 1) =$ $\frac{500}{20000}$

ii. What is $\mathbb{E}[S_1]$?

$\mathbb{E}[S_1] =$ $\frac{1800}{20000}$

iii. What is $\mathbb{E}[M_{100}]$?

$\mathbb{E}[M_{100}] =$ $\frac{700}{20000}$

iv. What is $\text{Var}[D_{50}]$?

$p = \frac{2000}{20000} = \frac{1}{10}$

$\text{Var}[D_{50}] =$ $\frac{1}{10} \cdot \left(1 - \frac{1}{10}\right)$

v. What is $D_{400} + S_{400} + A_{400} + M_{400} + O_{400}$?

$D_{400} + S_{400} + A_{400} + M_{400} + O_{400} =$ 1

(b) For parts b.i and b.ii, let:

- $N_D = \sum_{i=1}^{500} D_i$
- $N_S = \sum_{i=1}^{500} S_i$
- $N_M = \sum_{i=1}^{500} M_i$
- $N_A = \sum_{i=1}^{500} A_i$
- $N_O = \sum_{i=1}^{500} O_i$

$$E[N_A] = E\left[\sum_{i=1}^{500} A_i\right] = \sum_{i=1}^{500} E[A_i] = \sum_{i=1}^{500} \frac{500}{20000}$$

$$= 500 \cdot \frac{500}{20000} = \frac{25}{2}$$

i. What is $\mathbb{E}[N_A]$?

$\mathbb{E}[N_A] =$ $\frac{25}{2}$

ii. What is $\text{Var}(N_D + N_S + N_A + N_M + N_O)$?

$= 500$

$\text{Var}(500)$

$\text{Var}(N_D + N_S + N_A + N_M + N_O) =$ 0

(c) Now, suppose we take a sample with replacement of 20 contributions, what is the probability that 7 were by Daniel?

distribution?

Probability = $\binom{20}{7}\left(\frac{1}{10}\right)^7\left(\frac{9}{10}\right)^{13}$ (n = 20)

$p = \frac{1}{10}$

(d) Finally, suppose we take a sample with replacement of 10 contributions. What is the probability that 3 were by Daniel, 3 were by Suraj, and 4 were by Manana? (Note: Refer to Lecture 2 to refresh your knowledge on how to calculate this type of probability)

Probability = $\frac{10!}{3!3!4!}\left(\frac{1}{10}\right)^3\left(\frac{1800}{20000}\right)^3\left(\frac{700}{20000}\right)^4$

⇓ # of rearrangements

$\frac{10!}{3!3!4!} = \binom{10}{3}\cdot\binom{7}{3}$

$p_D = \frac{1}{10}$

$p_S = \frac{1800}{20000}$

$p_M = \frac{700}{20000}$