MSSE Data C200, Spring 2021

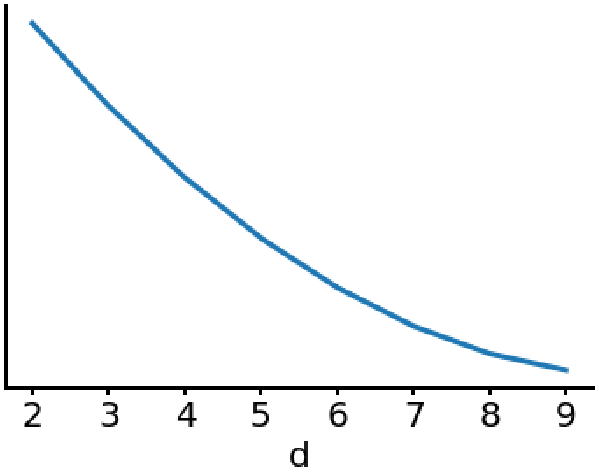
Discussion #10

Name: *Monday March 8th*

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 $m \cdot d$. Currently, you're not sure how many videos to consider.

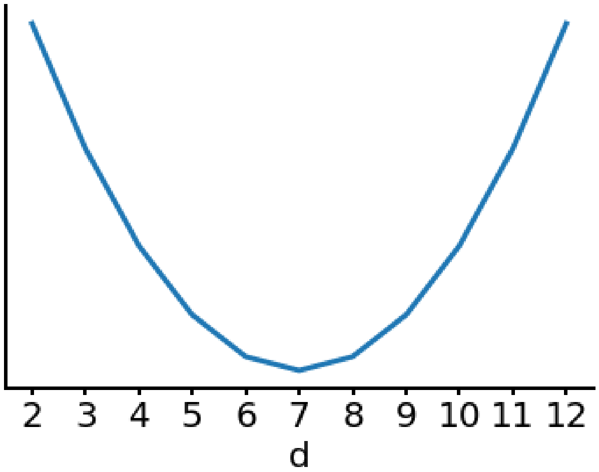
   (a) Your colleague generates the following plot, where the value d is on the x-axis. However, they forgot to label the y-axis.



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 again.



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 $\mathrm{var}(\epsilon) = \sigma^2$. 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) True or false and explain. $\mathbb{E}\left[\epsilon f_{\hat{\theta}}(x)\right] = 0$

(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 $\mathbb{E}\left[f_{\hat{\theta}}(x)\right]$

(e) If you could collect as many data sets as you would like, how does that affect the quality of your model $f_\theta(x)$?

Ridge and LASSO Regression

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

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

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

To prevent overfitting, we saw that we can instead minimize the sum of the average squared loss plus a regularization function $\lambda\mathcal{S}(\theta)$.

- If use the function $\mathcal{S}(\theta) = ||\theta||_2^2$, we have “ridge regression”.
- If we use the function $\mathcal{S}(\theta) = ||\theta||_1$, we have “LASSO regression”.

For example, if we choose $\mathcal{S}(\theta) = ||\theta||_2^2$, our goal is to find $\hat{\theta}$ that satisfies the equation below:

$$\begin{aligned}\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, L(\theta) &= \underset{\theta}{\operatorname{argmin}}\, \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2 + \lambda||\theta||_2^2 \\ &= \underset{\theta}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{X}_{i,\cdot}^T\theta)^2 + \lambda\sum_{j=0}^{p}\theta_j^2\end{aligned}$$

Recall that λ 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: $\hat{\theta} = (\mathbb{X}^T\mathbb{X} + n\lambda\mathbf{I})^{-1}\mathbb{X}^T\mathbb{Y}$. It turns out that $\mathbb{X}^T\mathbb{X} + n\lambda\mathbf{I}$ 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?

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

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

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

(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?

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

Cross Validation

4. After running 5-fold cross validation, we get the following mean squared errors for each fold and value of λ:

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 lambda? or a specific fold-lambda pair? Explain.

5. You build a model with two regularization hyperparameters λ and γ. You have 4 good candidate values for λ and 3 possible values for γ, and you are wondering which λ, γ pair will be the best choice. If you were to perform five-fold cross-validation, how many validation errors would you need to calculate?

6. 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