Data 100, Spring 2024

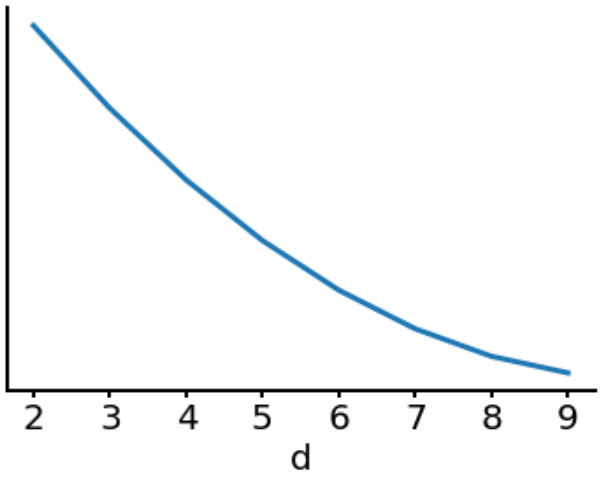
Discussion #10

Note: Your TA will probably not cover all the problems. This is totally fine, the discussion worksheets are not designed to be finished within an hour. They are deliberately made slightly longer so they can serve as resources you can use to practice, reinforce, and build upon concepts discussed in lectures, labs, and homework.

Bias-Variance Trade-off

1. Your team would like to train a machine learning model to predict the next YouTube video a user will click on based on the videos the user has watched. We extract up to d attributes (such as length of video, view count, etc.) from each video, and our model will be based on the previous m videos watched by that user. Hence, the number of features for each data point for the model is $m \times d$. Currently, you're not sure how many videos to consider.

   (a) Your colleague, Lillian, generates the following plot, where the value d on the x-axis denotes the number of features used for a particular model. However, she forgot to label the y-axis. Assume that the features are added to the model in decreasing levels of importance: More important features are first, and less important ones are after.



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

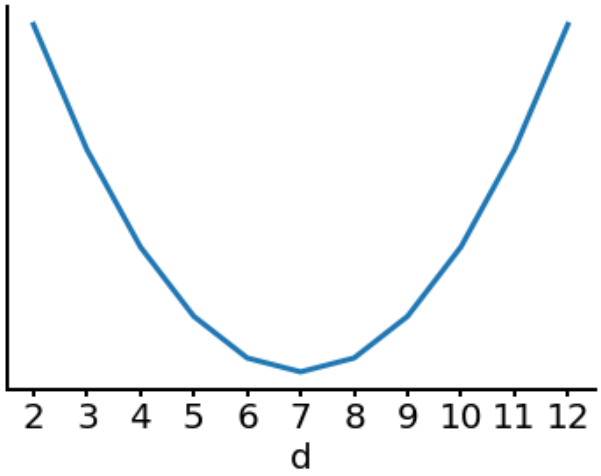
   ☐ A. Training Error

   ☐ B. Validation Error

   ☐ C. Bias

   ☐ D. Variance

(b) Lillian generates the following plot, where the value d is on the x-axis. However, she 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

(c) Explain what happens to the error on the holdout set as we increase d. Why?

2. We randomly sample n data points, (x_i, y_i), and use them to fit a model $f_{\hat{\theta}}(x)$ according to some procedure (e.g. OLS, Ridge, LASSO). Then, we sample a new data point (independent of our existing points) from the same underlying data distribution. Furthermore, assume that we have a function $g(x)$ and some noise generation process that produces ϵ such that $\mathbb{E}[\epsilon] = 0$ and $\text{Var}(\epsilon) = \sigma^2$. Whenever we query Y at a given x, we are given $Y = g(x) + \epsilon$, a corrupted version of the real ground truth output. A new ϵ is generated each time, independent of the last. We showed in the lecture 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 using the following word bank. Not all words will be used.
   - observation variance
   - model variance
   - $(\text{observation bias})^2$
   - $(\text{model bias})^2$
   - model risk
   - empirical mean square error

   (b) What quantities are random variables in the above equation? In our assumed data-generation process, where is the randomness in each variable coming from (i.e., which part of the assumed underlying model makes each random variable 'random')?

   (c) Calculate the value of $\mathbb{E}[\epsilon f_{\hat{\theta}}(x')]$, where $f_{\hat{\theta}}(x')$ is some predicted value of the response variable at some new fixed x' using a model trained on a random sample, and ϵ is the observation error for a new data point at this fixed value of x'.

Regularization and Bias-Variance Trade-off

3. We will use a simple constant model $f_\theta(x) = \theta$ to show the effects of regularization on bias and variance. For the sake of simplicity, we will assume that there is no noise or observational variance, so the ground truth output is equal to the observed outputs: $Y = g(x)$.

   (a) Recall that the optimal solution for the constant model with an MSE loss and a dataset $\mathcal{D}$ with $y_1, y_2, ..., y_n$ is the mean $\bar{y}$.

   We use L2 regularization with a regularization penalty of $\lambda > 0$ to train another constant model. Derive the optimal solution to this new constant model **with L2 regularization** to minimize the objective function below.

$$R(\theta) = \arg\min_{\theta} \left[\left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \theta)^2 \right) + \lambda \theta^2 \right]$$

   Note: As mentioned in the lecture, we do not impose a regularization penalty on the bias term and this problem only serves as a practice.

(b) Use the bias-variance decomposition to show that for a constant model **with L2 regularization**, its optimal expected loss on a sample test point (x, Y) in terms of the training data y is equal to the following.

$$\mathbb{E}_{\mathcal{D}}[(Y - f_{\hat{\theta}}(x))^2] = (Y - \frac{1}{1+\lambda}\mathbb{E}_{\mathcal{D}}[\bar{y}])^2 + \frac{1}{(1+\lambda)^2}\text{Var}_{\mathcal{D}}(\bar{y})$$

What expected loss do we obtain when $\lambda = 0$, and what does that mean in terms of our model?

Note: The subscript next to the expectation and variance lets you know what is random inside the expectation (i.e., what is the expectation taken over?). In this case, we calculate the expectation and variance of $\bar{y}$ across the dataset $\mathcal{D}$.

(c) Remark on how regularization has affected the model bias and variance as λ increases. Consider what would happen to these quantities as $\lambda \to \infty$.