A1

a

Bias refers to the error introduced by approximating a complex real-world problem with a simpler model. High bias means the model is underfitting the data. Variance is the error introduced by the model's sensitivity to small changes in the training data. High variance indicates overfitting. The bias-variance tradeoff involves balancing these two types of errors. As model complexity increases, bias decreases but variance increases, and vice versa. The goal is to find a balance that minimizes total error.

b

When model complexity increases, bias decreases because the model can better fit the training data, but variance increases because the model may overfit. When model complexity decreases, variance decreases because the model becomes less sensitive to the training data, but bias increases because the model might underfit.

c

False. While reducing features can help prevent overfitting and improve generalization, this is not always true. If we remove important features, the model may lose predictive power and underfit instead.

d

False. Hyperparameters should not be tuned on the test set because it would lead to overfitting the test data, making it no longer a reliable measure of generalization performance.

Procedure for Hyperparameter Tuning:

Split the data into training, validation, and test sets. Train the model on the training set and use the validation set to tune hyperparameters. Once the best parameters are selected, evaluate the model on the test set.

e

False. The training error typically underestimates the true error because the model has already seen the training data and is optimized for it. The true error is better estimated using validation or test data.

A2(a)

First, we find the likelihood function of the parameter λ for the given data. Let $x_1, x_2, \ldots, x_n$ represent the observed goal counts for n games, where the number of goals in each game follows a Poisson distribution with parameter λ. The probability mass function for a single observation is:

$$P(x_i \mid \lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}.$$

Since the observations are independent, the likelihood function for the entire dataset is the product of the probabilities:

$$L(\lambda) = \prod_{i=1}^{n} P(x_i \mid \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}.$$

Next, we take the natural logarithm of the likelihood function to simplify the optimization process. The log-likelihood function is:

$$\log L(\lambda) = \log \left(\prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right).$$

Expanding the logarithm of the product:

$$\log L(\lambda) = \sum_{i=1}^{n} \log \left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right).$$

Now, apply the logarithm of a quotient:

$$\log \left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right) = \log \left(\lambda^{x_i} e^{-\lambda} \right) - \log(x_i!).$$

For the term $\log \left(\lambda^{x_i} e^{-\lambda} \right)$, apply the logarithm of a product:

$$\log \left(\lambda^{x_i} e^{-\lambda} \right) = \log \left(\lambda^{x_i} \right) + \log \left(e^{-\lambda} \right).$$

Simplify each term: 1. $\log \left(\lambda^{x_i} \right) = x_i \log \lambda$, 2. $\log \left(e^{-\lambda} \right) = -\lambda$.
Substitute these back:

$$\log \left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right) = x_i \log \lambda - \lambda - \log(x_i!).$$

Combining everything:

$$\log L(\lambda) = \sum_{i=1}^{n} \left(x_i \log \lambda - \lambda - \log x_i! \right).$$

To find the maximum likelihood estimate of λ, we differentiate $\log L(\lambda)$ with respect to λ:

$$\frac{\partial}{\partial\lambda}\log L(\lambda) = \frac{\partial}{\partial\lambda}\left(\sum_{i=1}^{n} x_i \log\lambda - \sum_{i=1}^{n}\lambda - \sum_{i=1}^{n}\log x_i!\right).$$

Taking derivatives term by term:

$$\frac{\partial}{\partial\lambda}\log L(\lambda) = \sum_{i=1}^{n}\frac{x_i}{\lambda} - \sum_{i=1}^{n} 1 + \frac{\partial}{\partial\lambda}\left(-\sum_{i=1}^{n}\log x_i!\right).$$

Since $\sum_{i=1}^{n}\log x_i!$ does not depend on λ, its derivative is zero:

$$\frac{\partial}{\partial\lambda}\log L(\lambda) = \sum_{i=1}^{n}\frac{x_i}{\lambda} - n.$$

Set the derivative equal to zero:

$$\sum_{i=1}^{n}\frac{x_i}{\lambda} - n = 0.$$

Rearranging:

$$\frac{\sum_{i=1}^{n} x_i}{\lambda} = n.$$

Multiply through by λ:

$$\sum_{i=1}^{n} x_i = n\lambda.$$

Solve for λ:

$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Thus, the maximum likelihood estimate for λ is:

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

■

A2(b)

To find the numerical estimate of λ, we use the maximum likelihood estimate:

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The number of goals in the first five games is given as $[2, 4, 6, 0, 1]$. Substituting these values:

$$\hat{\lambda} = \frac{2+4+6+0+1}{5} = \frac{13}{5} = 2.6.$$

Thus, the estimated λ is:

$$\hat{\lambda} = 2.6.$$

Next, we calculate the probability that the team scores 6 goals in their next game. The number of goals per game follows a Poisson distribution, with probability mass function:

$$P(X = x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Substituting $x = 6$ and $\lambda = 2.6$:

$$P(X = 6 \mid \lambda = 2.6) = \frac{2.6^6 e^{-2.6}}{6!}.$$

We compute this step by step: 1. Compute 2.6^6:

$$2.6^6 = 308.915776.$$

2. Compute $e^{-2.6}$:

$$e^{-2.6} \approx 0.07427.$$

3. Compute 6!:

$$6! = 720.$$

Substitute these values back:

$$P(X = 6 \mid \lambda = 2.6) = \frac{308.915776 \cdot 0.07427}{720}.$$

Simplify:

$$P(X = 6 \mid \lambda = 2.6) \approx \frac{22.93885}{720} \approx 0.03187.$$

Thus, the probability that the team scores 6 goals in their next game is approximately:

$$P(X = 6 \mid \lambda = 2.6) \approx 0.03187 \quad (3.187\%).$$

■

A2(c)

The number of goals scored in six games is now given as $[2, 4, 6, 0, 1, 8]$. To find the updated numerical estimate of λ, we use the maximum likelihood estimate:

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Substituting the values:

$$\hat{\lambda} = \frac{2 + 4 + 6 + 0 + 1 + 8}{6} = \frac{21}{6} = 3.5.$$

Thus, the updated λ is:

$$\hat{\lambda} = 3.5.$$

Next, we calculate the probability that the team scores 6 goals in their 7th game. The number of goals per game follows a Poisson distribution, with probability mass function:

$$P(X = x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Substituting $x = 6$ and $\lambda = 3.5$:

$$P(X = 6 \mid \lambda = 3.5) = \frac{3.5^6 e^{-3.5}}{6!}.$$

We compute this step by step: 1. Compute 3.5^6:

$$3.5^6 = 1838.265625.$$

2. Compute $e^{-3.5}$:

$$e^{-3.5} \approx 0.030197.$$

3. Compute $6!$:

$$6! = 720.$$

Substitute these values:

$$P(X = 6 \mid \lambda = 3.5) = \frac{1838.265625 \cdot 0.030197}{720}.$$

Simplify:

$$P(X = 6 \mid \lambda = 3.5) \approx \frac{55.459}{720} \approx 0.0771.$$

Thus, the probability that the team scores 6 goals in their 7th game is approximately:

$$P(X = 6 \mid \lambda = 3.5) \approx 0.0771 \quad (7.71\%).$$

■

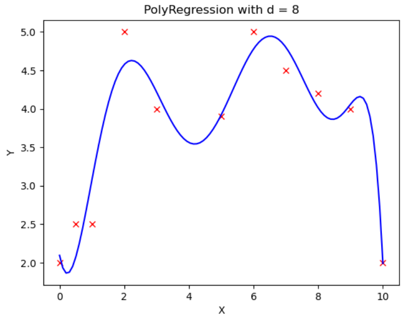

Figure 1: Plot before increase in regularization ($\lambda = 0$)

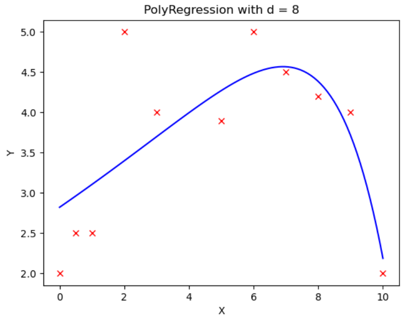

Figure 2: Plot after increase in regularization ($\lambda = 1$)

A3(b)

To analyze the effect of regularization on polynomial regression, we first fit a polynomial of degree $d = 8$ with no regularization ($\lambda = 0$). The resulting plot is shown in Figure 1. Next, we increase the regularization parameter to $\lambda = 1$, which penalizes large coefficients and smooths the curve. The resulting plot is shown in Figure 2. Increasing regularization reduces overfitting, resulting in a smoother function that generalizes better to new data points.

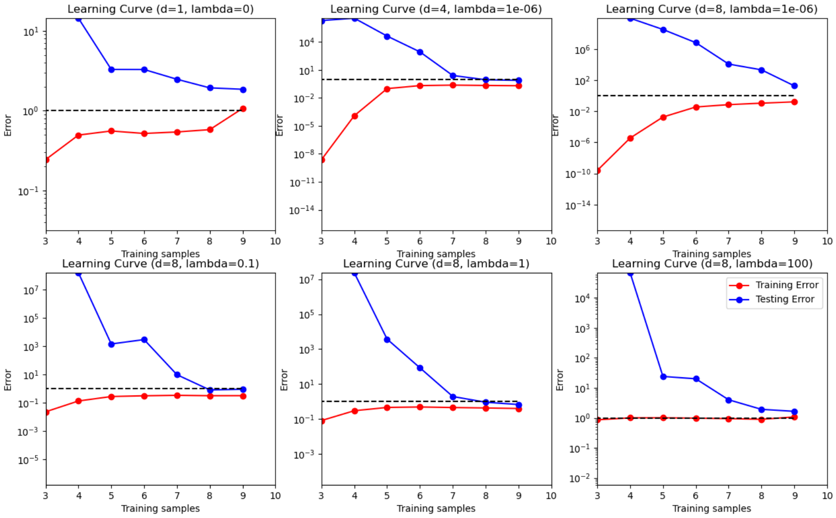

Figure 1: Plots of learning curves for various values of λ and d.

A4

A5(b)

The training and testing errors of the Ridge Regression classifier trained on the MNIST dataset with $\lambda = 10^{-4}$ are as follows:

$$\text{Train Error: } 14.805\%$$

$$\text{Test Error: } 14.66\%$$

These errors demonstrate the performance of the classifier on both the training and testing datasets.

A6

The total time spent on this homework was approximately 10–12 hours.