1 Fill in the Blank 10 points

Consider fitting natural splines with 2, 8, and 20 degrees of freedom. Among these, df = 20 will have the lowest bias, df = 2 will have the lowest variance, df = 20 will have the lowest training error, and Not enough information will have the lowest expected test error.

df = 20 | df = 8 | Not enough information given | df = 2

df = 2 | df = 8 | df = 20 | Not enough information given

df = 8 | df = 20 | Not enough information given | df = 2

Not enough information given | df = 8 | df = 20 | df = 2

2 Multiple Answer 10 points

Consider fitting a spline with 10 degrees of freedom to a training dataset and evaluating its predictive performance on a test set. Which of the following are a cause or an effect of overfitting? Select all that apply.

- [x] Our model fits the training data very closely.
- [ ] Our model fits the test data very closely.
- [x] Our training dataset is small.
- [ ] Our test dataset is small.
- [ ] The true model complexity is large.

3 Multiple Answer 10 points

Which of the following models has the largest number of degrees of freedom? If there are ties, select all that apply. Note that "total knots" refers both to boundary and internal knots.

- [x] A degree 20 polynomial.
- [ ] A natural cubic spline with 20 total knots.
- [ ] A piece-wise constant fit with 21 total knots.
- [ ] A piece-wise linear spline with 11 total knots.

4 Fill in the Blank 10 points

Increasing the number of cross-validation folds [increases] the [sample size] of the learning sets, which tends to [increase] the complexity of the model chosen by cross-validation.

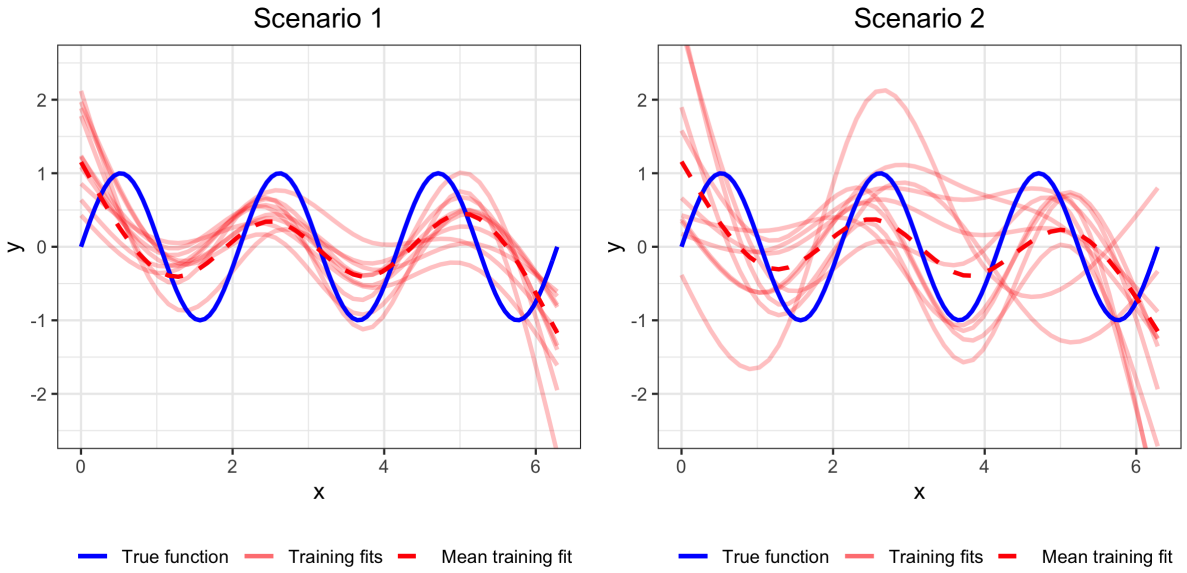
increases | decreases | sample size | noise level | true model complexity | fitted model complexity | increase | decrease | not impact

5 Multiple Answer 10 points

Consider two scenarios, where natural splines with five degrees of freedom are fit to training datasets.

- Scenario 1: The data distribution is $Y = \sin(3X) + \epsilon$, where $\epsilon \sim N(0, \sigma_1^2)$. The training set size is n_1.
- Scenario 2: The data distribution is $Y = \sin(3X) + \epsilon$, where $\epsilon \sim N(0, \sigma_2^2)$. The training set size is n_2.

Below are spline fits on ten training datasets drawn for each scenario, along with the mean training fit and true trend.



Which of the following relationships between the noise variances σ_1^2 and σ_2^2 and training sample sizes n_1 and n_2 are possible? Select all that apply.

- [x] $\sigma_1^2 > \sigma_2^2$ and $n_1 > n_2$
- [ ] $\sigma_1^2 > \sigma_2^2$ and $n_1 < n_2$
- [x] $\sigma_1^2 < \sigma_2^2$ and $n_1 > n_2$
- [x] $\sigma_1^2 < \sigma_2^2$ and $n_1 < n_2$

6 Numeric 10 points

Suppose $Y = \sin(3X) + \epsilon$, where $\epsilon \sim N(0, 1)$. What is the lowest possible expected test error, across all prediction rules?

1

7 Numeric 10 points

We use 10-fold cross-validation to train, tune, and evaluate a natural spline fit. We have 100 data points in total, which we split into 80 for training and 20 for testing. We try degrees of freedom values 1,2,...,15. How many total spline model fits does this entail? (A model fit refers to fitting coefficients to data, e.g. via a command like lm(y ~ ns(x, df = 5)).)

151

8 Multiple Answer 10 points

When viewed as a classifier, a COVID-19 test that always gives a negative result achieves the best possible value of which metric(s) below? Select all that apply.

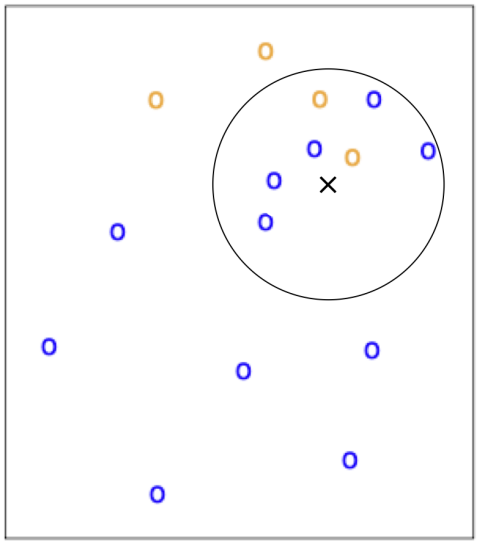
- [x] False positive rate
- [x] True negative rate
- [ ] False negative rate
- [ ] True positive rate
- [ ] F-score

9 Numeric 10 points

Consider predicting the class of the test point marked "×" via weighted K-nearest neighbors with K = 7, where the colored points are the training data. If $w_{\text{yellow}} / w_{\text{blue}} > C$, then weighted KNN will predict "yellow" for the test point. What is the minimum possible value of C?

Positive (rare)

Negative (common)



2.5

10 Numeric 10 points

Consider the confusion matrix from Lecture 4:

	Actually Positive	**Actually Negative**
Predicted Positive	10 True Positives (TP) (E.g. Sick people testing positive)	20 False Positives (FP) (E.g. Healthy people testing positive)
Predicted Negative	40 False negatives (FN) (E.g. Sick people testing negative)	30 True Negative (TN) (E.g. Healthy people testing negative)
Total	50 positives (P)	50 negatives (N)

What is the misclassification error?

0.6