Quiz 2

You have 30 minutes to complete this 10-question quiz. The questions, a mix of multiple choice, fill-in-the-blank, and numeric answers, are weighted equally. You can consult any course materials, the internet, or R. However, you must complete the quiz individually.

For fill-in-the-blank questions, please do not include any extra spaces in your answers. Please enter your numeric answers using numbers instead of letters (e.g. "10" instead of "ten"). Any references to degrees of freedom include all degrees of freedom, including intercept terms. For example, a straight line with a slope and an intercept has two degrees of freedom.

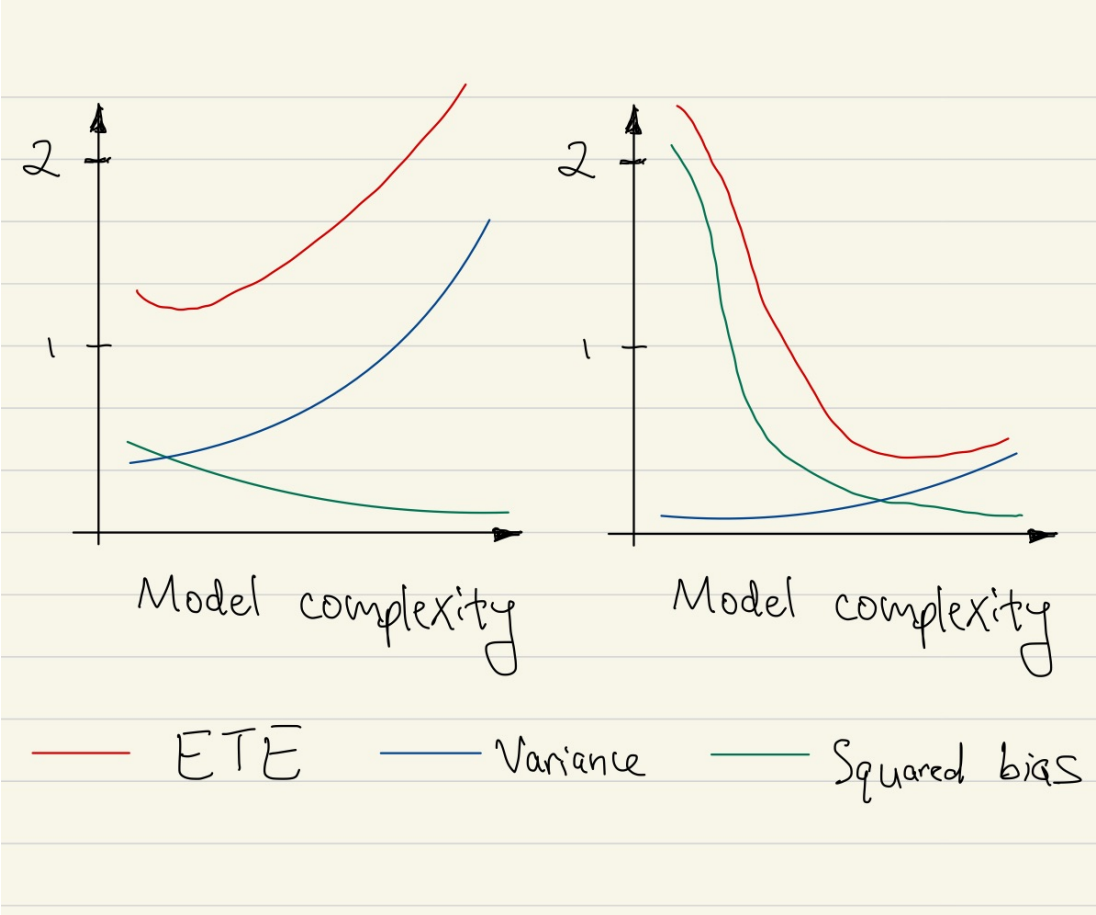
1 Numeric 0.5 points

Intending to fit a cubic spline with five internal knots to a dataset, we fit a separate cubic polynomial to each segment (so we did it incorrectly because these polynomials are not "stitched together" at their boundaries). How many degrees of freedom does this fit have?

24

2 Fill in the Blank 0.5 points

The following two plots show the expected test error, variance, and squared bias of a prediction method over a range of model complexity values in two problem settings (i.e. each plot has its own true model, error variance, and training sample size). The same prediction method is applied over the same range of model complexity values.



Compared to the plot on the left, the plot on the right has [lower] irreducible error, [higher] true model complexity, and [not enough information g] training sample size.

higher | not enough information given | lower

not enough information given | higher | lower

3 Multiple Answer 0.5 points

Suppose that $\mathrm{Wage} = f(\mathrm{Age}) + \epsilon$, where f is a natural cubic spline with 5 degrees of freedom. We would like to train a natural cubic spline $\hat{f}$ to a training set of 200 observations from this distribution. Which of the following values of degrees of freedom for $\hat{f}$ could lead to the lowest expected test error? Select all that apply. [Hint: A natural cubic spline fit with 5 or more degrees of freedom will be unbiased.]

- [x] df = 1
- [x] df = 3
- [x] df = 5
- [ ] df = 7

4 Numeric 0.5 points

In the setup from the previous problem, suppose additionally that $\sigma^2 = \mathrm{Var}[\epsilon] = 4$. We split up our training set into 10 groups of 20 observations each. We train a natural cubic spline with 5 degrees of freedom to each of the 10 groups of observations. For each value of Age, we take a sample variance of the 10 spline fits and then average across Age. What do we expect the result to be?

1

5 Fill in the Blank 0.5 points

We use the validation set approach to train, tune, and evaluate a natural cubic spline fit. We have 100 data points in total, which we split into 50 for training, 25 for validation, and 25 for testing. We try out degrees of freedom values 1,2,...,15. How many spline model fits did the training data contribute to? How about the validation data? How about the test data? (A model fit refers to fitting coefficients to data, e.g. via a command like lm(y ~ splines::ns(x, df = 5)).)

The training data contributed to a total of 16 spline model fit(s); the validation data contributed to a total of 1 spline model fit(s); the test data contributed to a total of 0 model fits(s).

6 Multiple Answer 0.5 points

When tuning degrees of freedom for a natural cubic spline based on cross-validation, we apply the one-standard-error rule to select df = 5. Which of the following degrees of freedom could have minimized the cross-validation error estimate?

- [ ] df = 1
- [ ] df = 3
- [x] df = 5
- [x] df = 7

7 Fill in the Blank 0.5 points

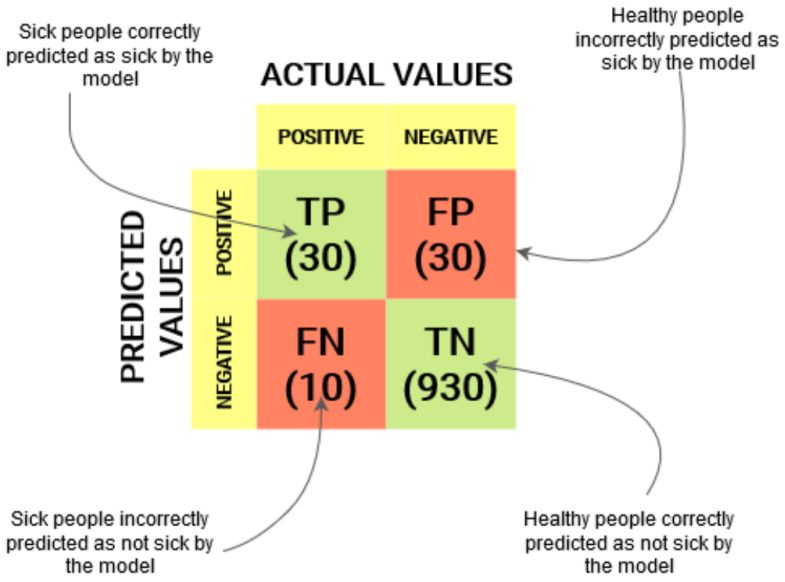
Consider a useless binary classifier that does not look at the feature vector; instead it guesses "negative" with probability 1/3 and "positive" with probability 2/3. This classifier has false positive rate of

2/3 and a false negative rate of 1/3.

Express your answers as simplified fractions with no extra spaces.

8 Fill in the Blank 0.5 points

Below is the confusion matrix for a classifier on a test set.



The false negative rate, expressed as a simplified fraction, is

1/4 . To bring this rate down, we would need to

decrease

the probability threshold for positive classification. Doing so would

increase the false positive rate.

⣿ increase ⣿ decrease ⣿ not enough information

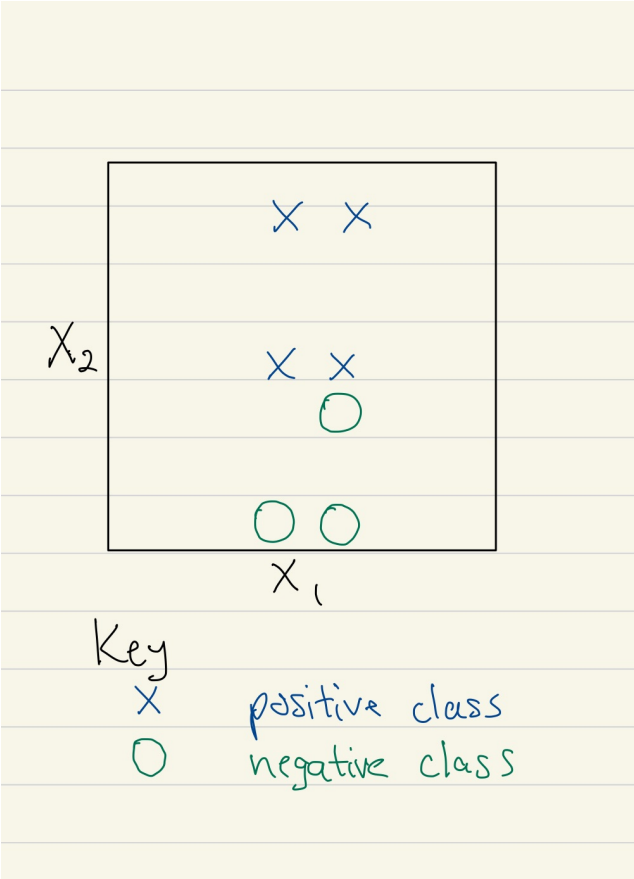
9 Multiple Choice 0.5 points

Suppose we apply KNN classification to a dataset. Which of the following value of K leads to the highest model flexibility/complexity?

- (●) K = 1
- ( ) K = 10
- ( ) K = 100
- ( ) Not enough information given

10 Fill in the Blank 0.5 points

We apply K-nearest neighbors classification with K = 3 and majority voting to the training dataset below, which contains 7 observations.



The training misclassification error is 1/7

and the training false positive rate is 1/3 .

Express your answers as simplified fractions, with no extra spaces.