Opening Theme: Flexibility vs. Stability

Patrick Breheny

August 25

Introduction

- We begin this course with a contrast of two simple, but very different, methods: the ordinary *least squares* regression model and the k-*nearest neighbor* prediction rule
- The linear model makes huge assumptions about the structure of the problem, but is quite stable
- Nearest neighbors is virtually assumption-free, but its results can be quite unstable
- Each method can be quite powerful in different settings and for different reasons

Simulation settings

- To examine which method is better in which setting, we will simulate data from a simple model in which y can take on one of two values: -1 or 1
- The corresponding x values are derived from one of two settings:
  - Setting 1: x values are drawn from a bivariate normal distribution with different means for $y = 1$ and $y = -1$
  - Setting 2: A mixture in which 10 sets of means for each class $(1, -1)$ are drawn; x values are then drawn by randomly selecting a mean from the appropriate class and then generating a random bivariate normal observation with that mean
- A fair competition between the two methods is then how well they do at predicting whether a future observation is 1 or -1 given its x values

The linear model: A review

- You are all familiar with the linear model, in which $\mathbf{y}$ is predicted via:

$$\hat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

- To review, the linear model is fit (*i.e.*, β is estimated) by minimizing the residual sum of squares criterion:

$$\text{RSS} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

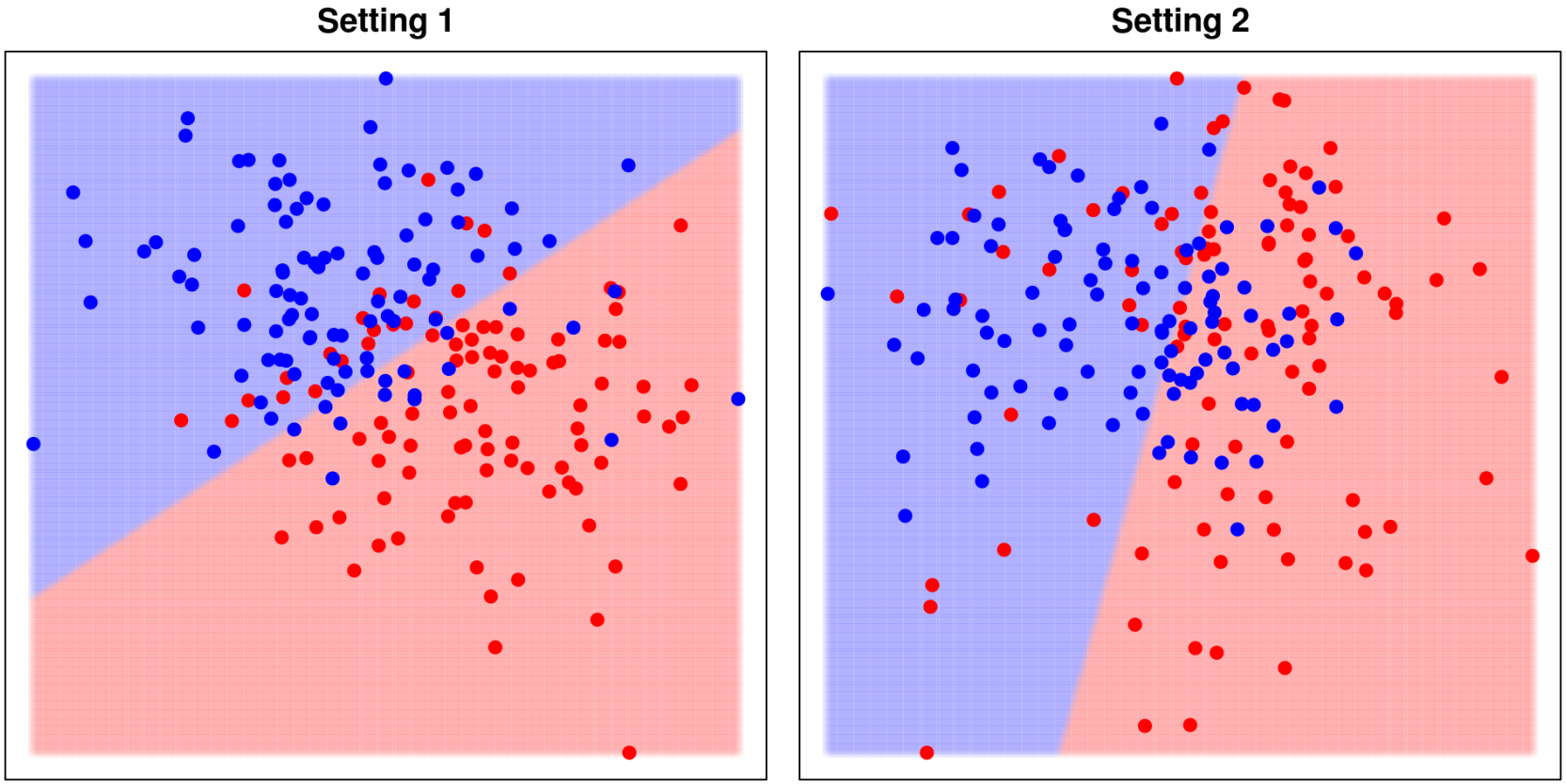
- Differentiating with respect to $\boldsymbol{\beta}$, we obtain the so-called *normal equations*:

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$$

- Provided that $\mathbf{X}$ is of full rank, then the unique solution is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Linear model results



Linear model remarks

- The linear model seems to classify points reasonably in setting 1
- In setting 2, on the other hand, there are some regions which seem questionable
- For example, in the lower left hand corner of the plot, does it really make sense to predict "blue" given that all of the nearby points are "red"?

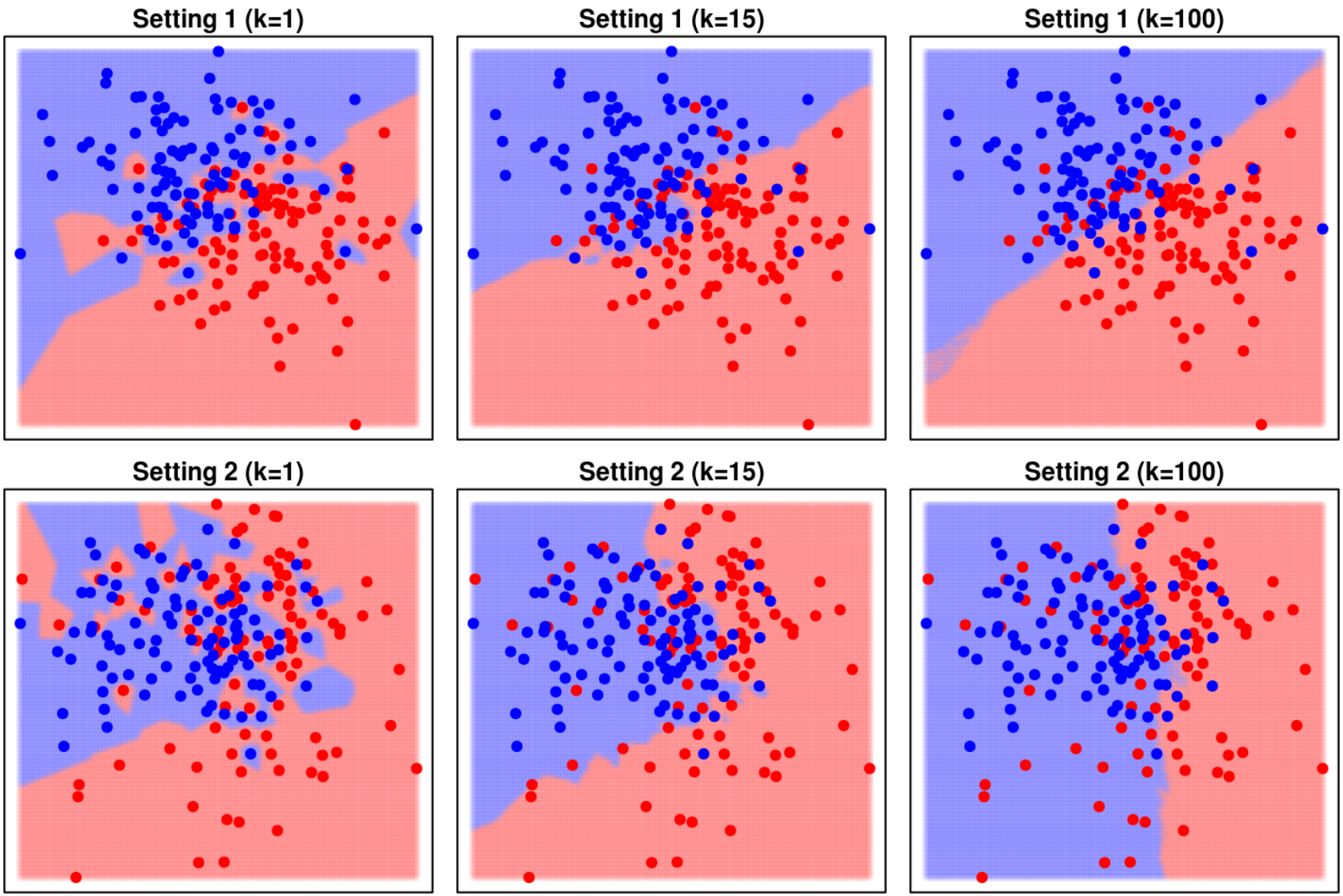
Nearest neighbors

- Consider then a completely different approach in which we don't assume a model, a distribution, a likelihood, or anything about the problem: we just look at nearby points and base our prediction on the average of those points
- This approach is called the *nearest-neighbor* method, and is defined formally as

$$\hat{y}(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i,$$

where $N_k(\mathbf{x})$ is the neighborhood of $\mathbf{x}$ defined by its k closest points in the sample

Nearest neighbor results



Nearest neighbor remarks

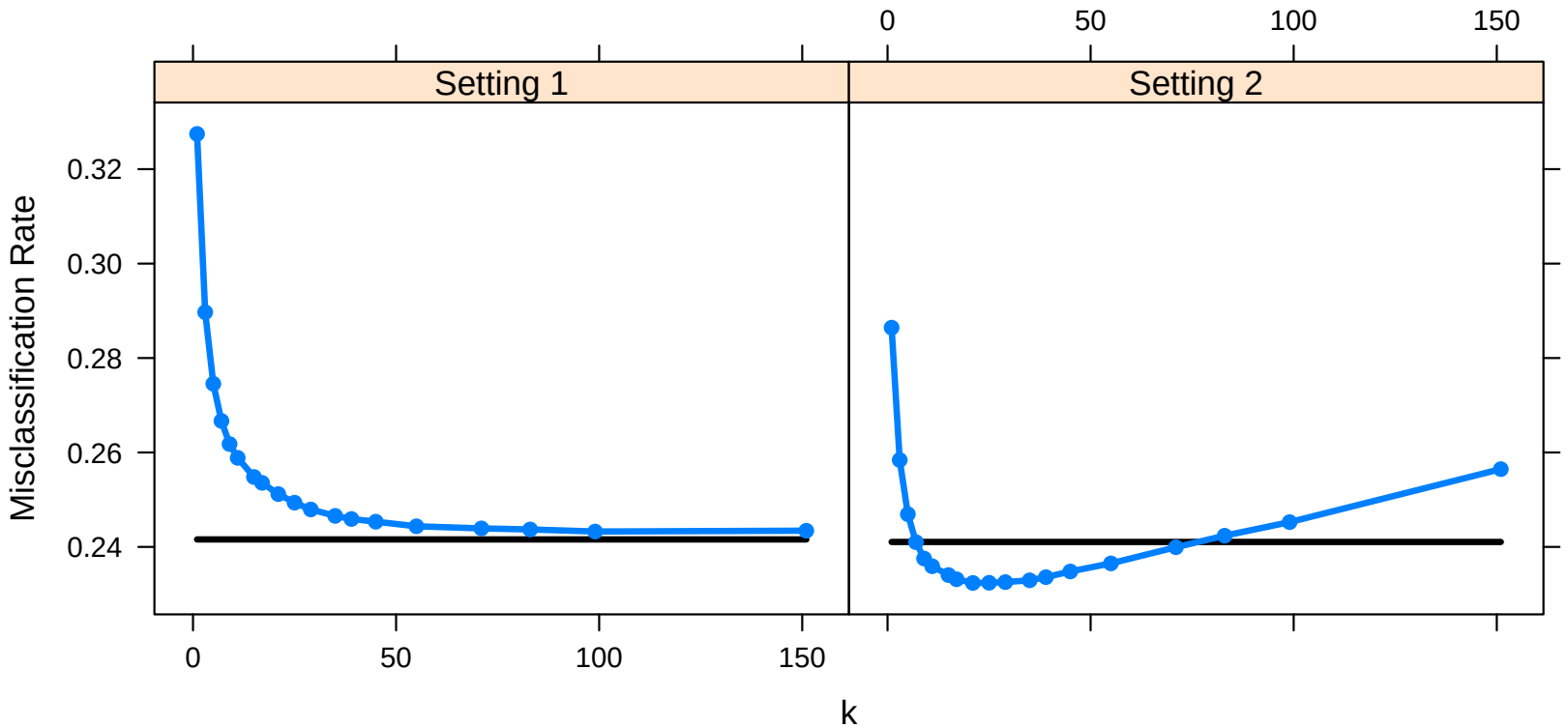
- Nearest neighbor seems not to perform terribly well in setting 1, as its classification boundaries are unnecessarily complex and unstable
- On the other hand, the method seemed perhaps better than the linear model in setting 2, where a complex and curved boundary seems to fit the data better
- Furthermore, the choice of k plays a big role in the fit, and the optimal k might not be the same in settings 1 and 2

Inference

- Of course, it is potentially misleading to judge whether a method is better simply because it fits the sample better
- What matters, of course, is how well its predictions generalize to new samples
- Thus, consider generating new data sets of size 10,000 and measuring how well each method does at predicting these 10,000 new results
- By repeatedly simulating sample and prediction data sets, we can estimate the long-run prediction accuracy of each method in each of the two settings

Simulation results

Black line = least squares; blue line = nearest neighbors



Remarks

- In setting 1, linear regression was always better than nearest neighbors
- In setting 2, nearest neighbors was usually better than linear regression
- However, it wasn't *always* better than linear regression – when k was too big or too small, the nearest neighbors method performed poorly
- In setting 1, the bigger k was, the better; in setting 2, there was a "Goldilocks" value of k (about 25) that proved optimal

The regression function

- Let us now develop a little theory in order to provide insight as to the reason for our simulation results
- The nearest neighbors and least squares methods have the common goal of estimating the *regression function*:

$$f(\mathbf{x}) = \mathrm{E}(Y|\mathbf{x}),$$

although they go about it very differently:

  - Nearest neighbors conditions on those observations closest to $\mathbf{x}$ and then estimates the expected value by taking the average
  - Least squares starts by making a strong assumption about f and then uses *all* the data to estimate f

Expected prediction error

- If our model is asked to predict y given $\mathbf{x}$, we can calculate the *prediction error*
- As we will see later on in the course, prediction error can be measured in many ways; for now, it is most convenient to consider *squared error loss* $(y - \hat{f}(\mathbf{x}))^2$
- A very reasonable criterion by which to judge a model is therefore the expected value of the prediction error; with squared error loss,

$$\text{EPE} = \text{E}\{(Y - \hat{f}(\mathbf{x}))^2 | \mathbf{x}\}$$

The bias-variance decomposition

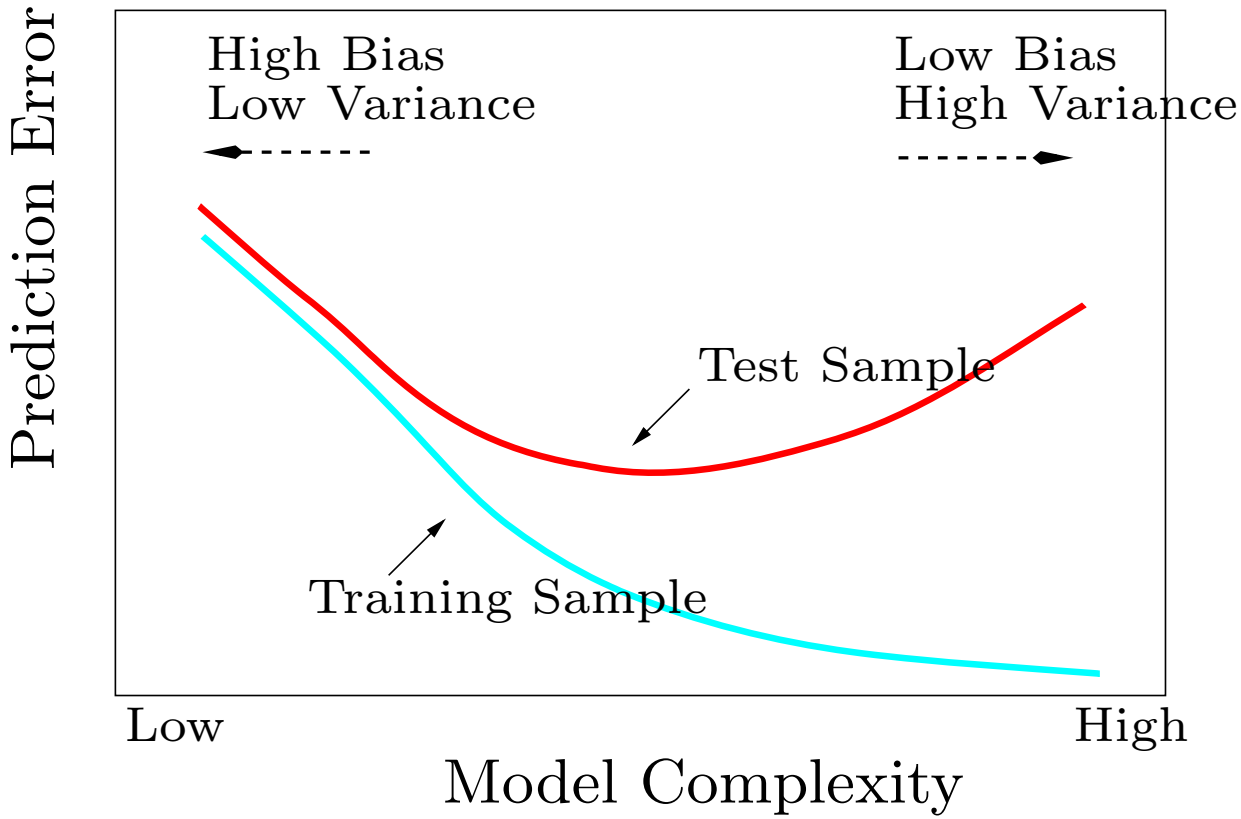
- **Theorem:** At a given point $\mathbf{x}_0$,

$$\text{EPE} = \sigma^2 + \text{Bias}^2(\hat{f}) + \text{Var}(\hat{f}),$$

  where σ^2 is the variance of $Y|\mathbf{x}_0$
- Thus, expected prediction error consists of three parts:
  - *Irreducible error*; this is beyond our control and would remain even if we were able to estimate f perfectly
  - *Bias* (squared); the difference between $\text{E}\{\hat{f}(\mathbf{x}_0)\}$ and the true value $f(\mathbf{x}_0)$
  - *Variance*; the variance of the estimate $\hat{f}(\mathbf{x}_0)$
- The second and third terms make up the *mean squared error*

The bias-variance decomposition: An illustration

An illustration from our textbook:



Limitations of nearest neighbor methods

- Our earlier simulation results seem to suggest that if we choose k appropriately, nearest neighbor methods can always do at least roughly as well as linear regression, and sometimes much better
- To some extent, this holds true, provided that the dimension of $\mathbf{x}$ is small
- However, the statistical properties of nearest neighbor methods worsen rapidly as p grows

The curse of dimensionality

- For example, suppose that x follows a uniform distribution on the unit p-cube; how many points will be within a ball centered at $\mathbf{x}_0$ of radius 0.2?
- When $p = 2$ and $n = 120$, we can expect 15 points in the neighborhood of x_0
- When $p = 3$, we need 448 observations to get 15 observations in a neighborhood of the same size
- When $p = 10$, we need over 57 million observations
- This phenomenon is commonly referred to as the *curse of dimensionality*, and we will return to it again in the course

Conclusions

- So where do we stand?
- Fitting an ordinary linear model is rarely the best we can do
- On the other hand, nearest-neighbors is rarely stable enough to be used, even in modest dimensions, unless our sample size is very large

Conclusions (cont'd)

- These two methods stand on opposite sides of the methodology spectrum with regard to assumptions and structure
- Many of the methods we will discuss in this course involve bridging the gap between these two methods – making linear regression more flexible, adding structure and stability to nearest neighbor ideas, or combining concepts from both
- One of the main themes of this course will be the need to find, develop, and apply methods that bring the right mix of flexibility and stability that is appropriate for the data