Introduction to Statistical Learning

ddebarr@uw.edu

2017-01-12

Administrative Stuff

- Pre-requisites: calculus, linear algebra
- Attendance: must attend 60% of classes
- On-site versus online: on-site students can do one online session [licensing]
- Homework: all assignments and due dates have been posted
  - <u>Only half credit awarded if turned in past due date</u>
  - For example: if you turn in a homework assignment late, and you would have scored 3 out of 3 points if you had turned it in on time, then you will be awarded 1.5 points
- Grading: must successfully complete 17 out of 28 possible homework points

Course Outline

1. Introduction to Statistical Learning
2. Linear Regression
3. Classification
4. Resampling Methods
5. Linear Model Selection and Regularization
6. Moving Beyond Linearity
7. Tree-Based Methods
8. Support Vector Machines
9. Unsupervised Learning
10. Neural Networks and Genetic Algorithms

Course Website

Assignments and Discussion

http://canvas.uw.edu/

Recordings

http://uweoconnect.extn.washington.edu/mlearn210/

Notes/Slides

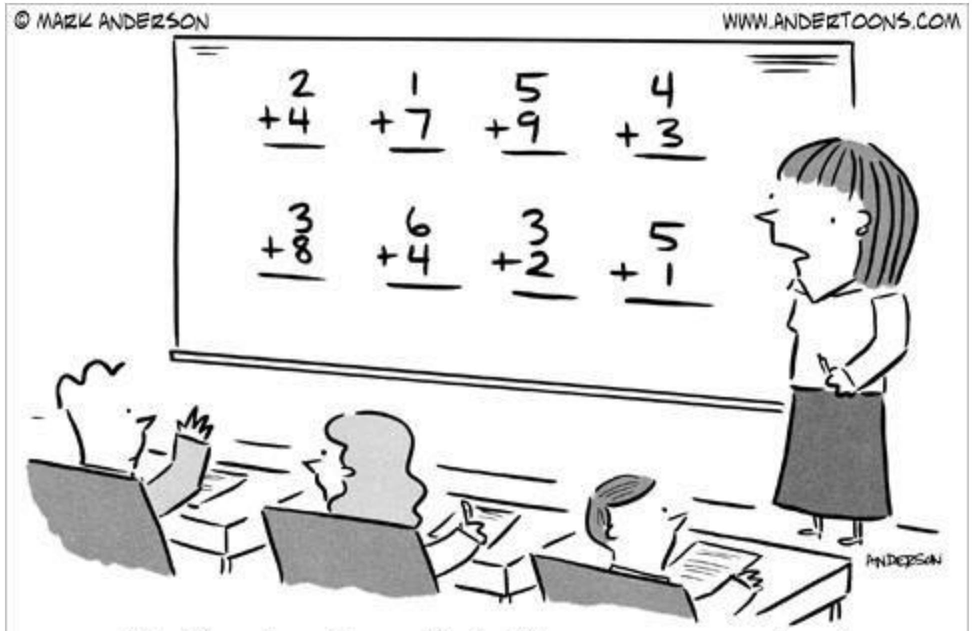
http://cross-entropy.net/ML210

Contact Info

- Dave DeBarr
  - ddebarr@uw.edu
  - Phone: (425) 679-2428

Considerations

- Remember to keep your sense of humor
- Keep up with the work every week
- Ask questions! If you have questions, others probably have the same questions!



"Let's solve these first. We can worry about data mining later."

Agenda

Machine Learning Definition

- Using data to create a model to map one-or-more input values to one-or-more output values
- Interest from many groups
  - Computer scientists: “machine learning”
  - Statisticians: “statistical learning”
  - Engineers: “pattern recognition”

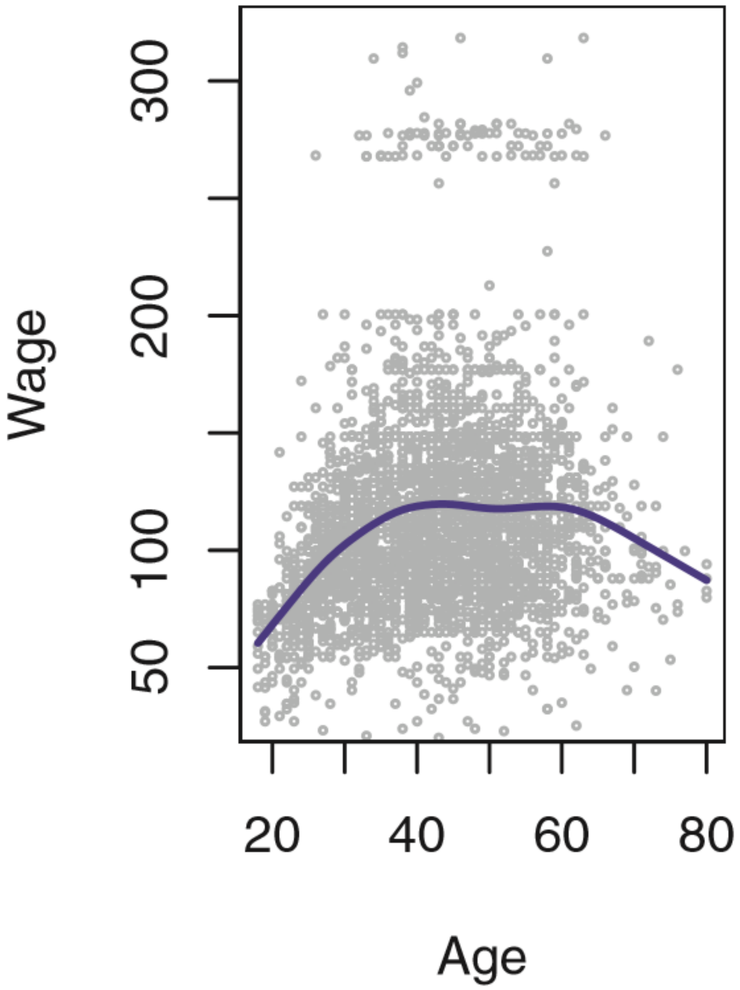
Applications

- E-Commerce: sentiment and trend analysis; dynamic pricing; predict which ad a user is most likely to click; customer segmentation
- Editing: spell correct
- Education: recommendations based on student’s aptitude
- Finance: predict whether an applicant will default on loan
- Genomics: predict gene function; personalized medicine
- Government: detect abusive tax avoidance transactions
- Healthcare: image analysis for diagnosis
- Manufacturing: predict when maintenance is needed
- Security: predict whether a transaction is fraudulent; biometrics recognition
- Translation: convert spoken language to another language

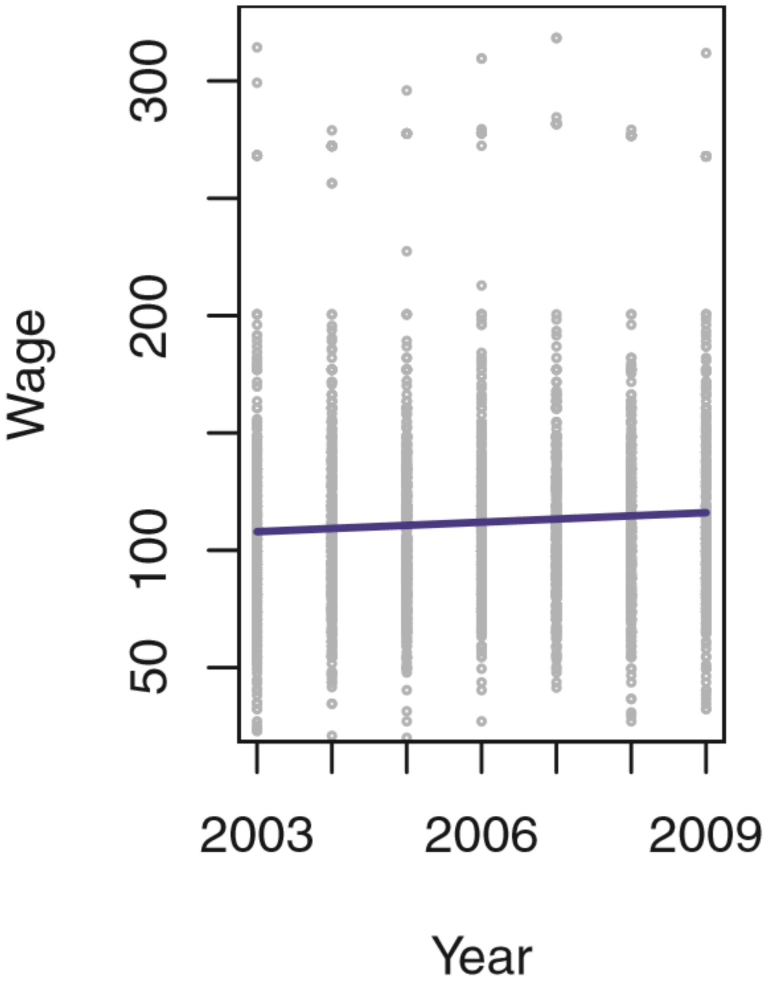
Examples of Learning Problems

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.
- Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data.
- Identify the numbers in a handwritten ZIP code, from a digitized image.
- Estimate the amount of glucose in the blood of a diabetic person, from the infrared absorption spectrum of that person's blood.
- Identify the risk factors for prostate cancer, based on clinical and demographic variables.

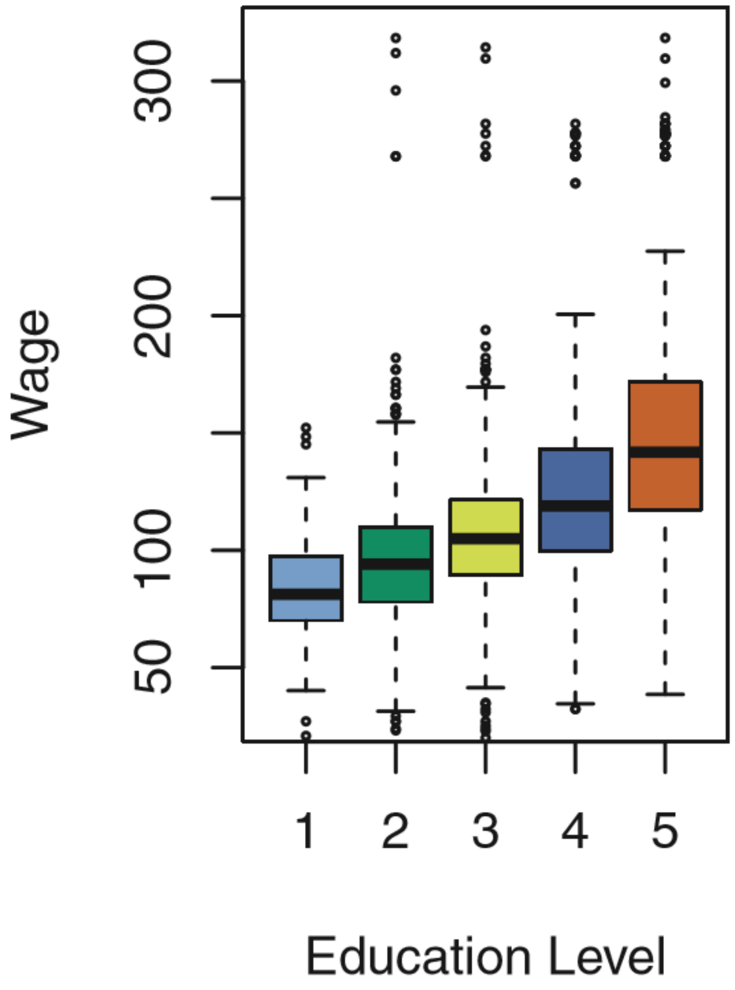
Wage Data



scatter plot

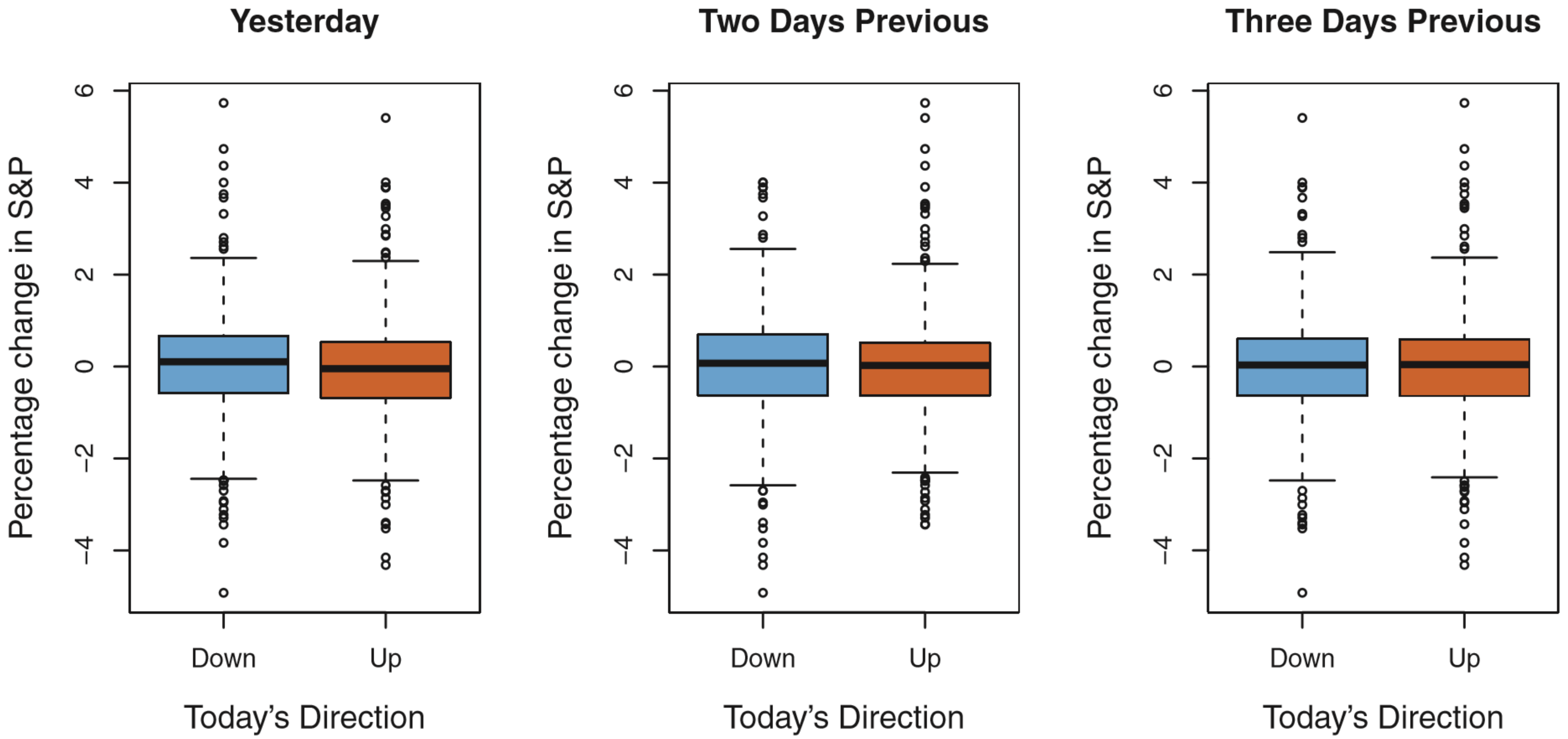


scatter plot

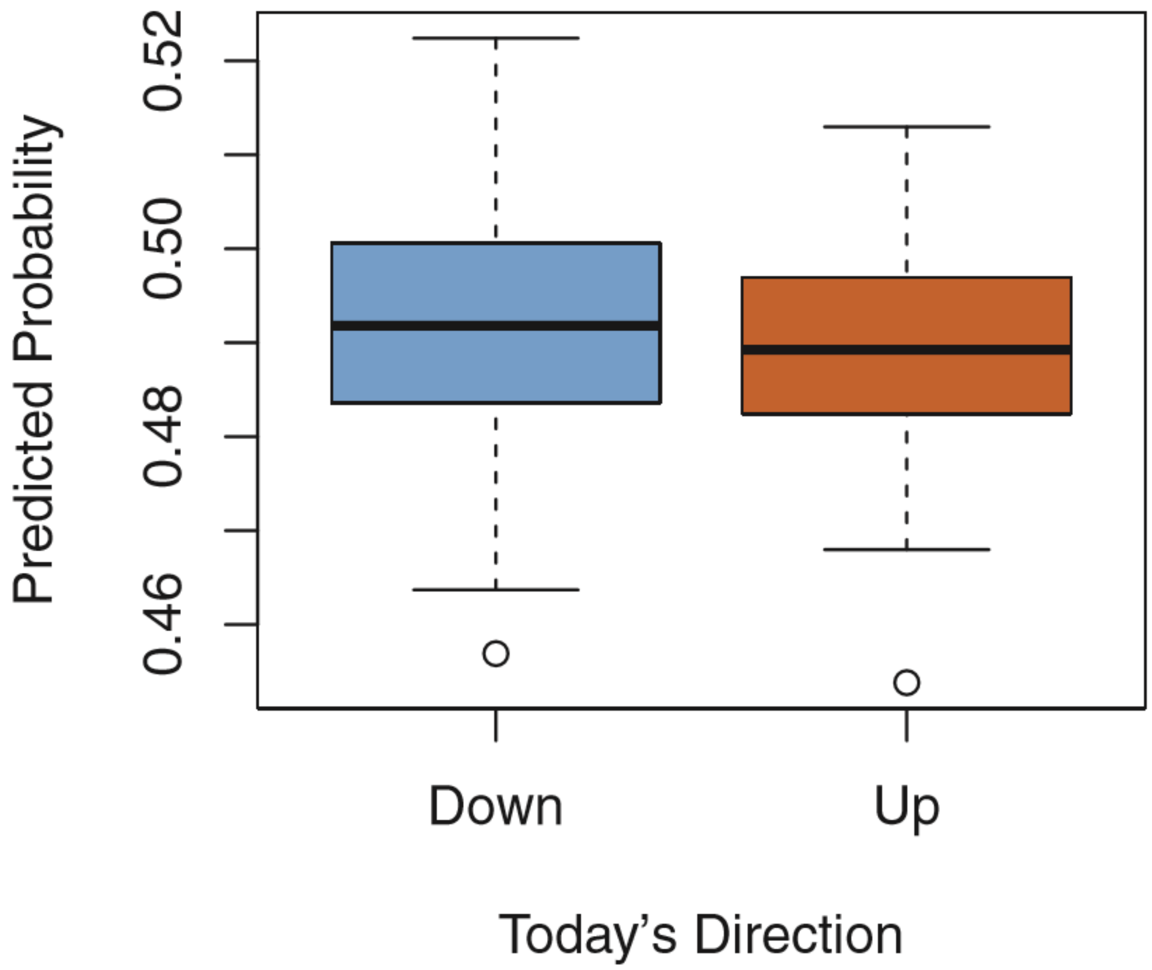


box plot

Change in Standard & Poor's Index



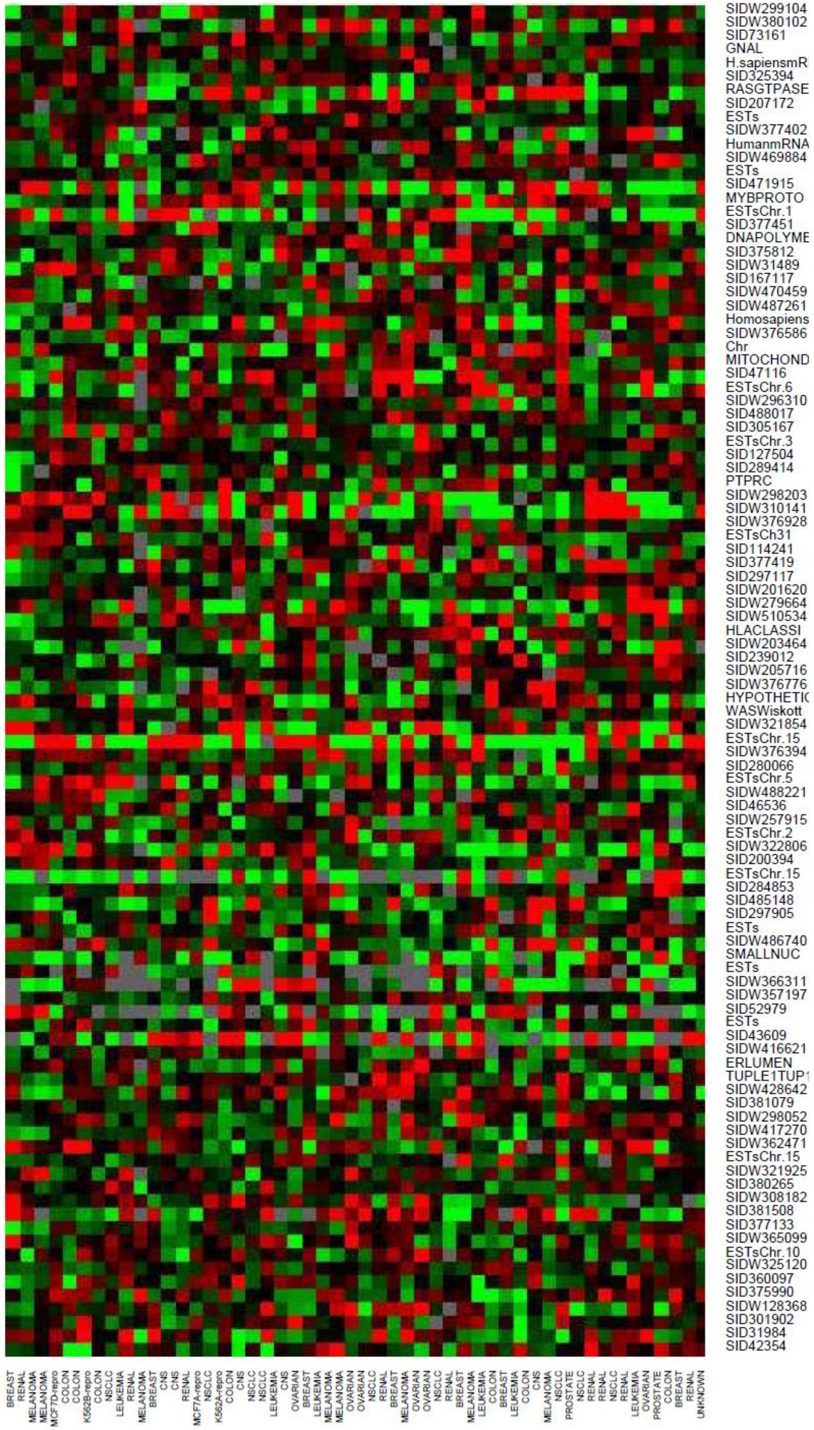
Predicted Probability of Decrease



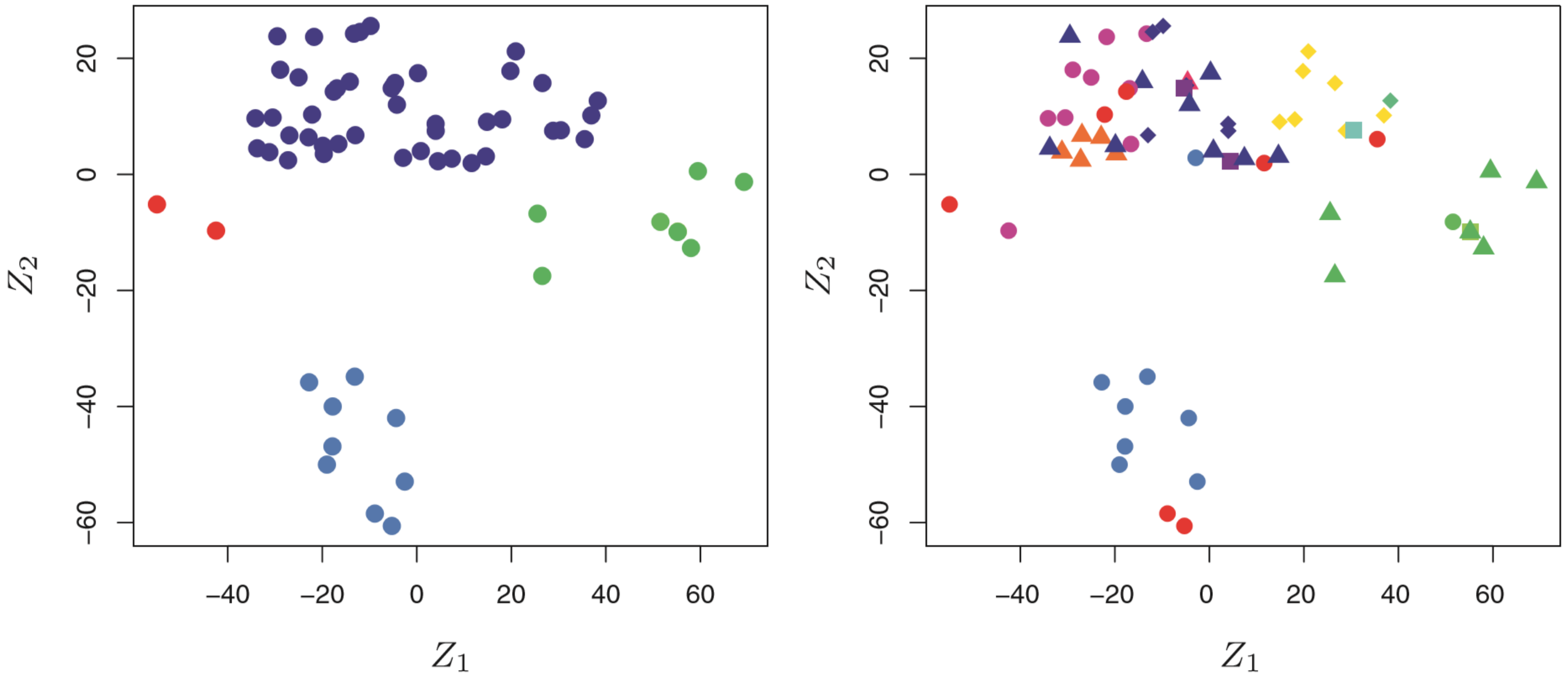
Slightly higher Predicted Probability of Decrease when there is an Actual Decrease

Gene Expression Data

- Genes are printed on a glass slide
- A target sample and a reference sample are labeled with red and green dyes
- The amount of messenger ribonucleic acid (mRNA) is measured for both the target and reference samples
- The log of the ratio of the two quantities typically ranges from -6 to 6



Gene Expression Data



Matrix Notation

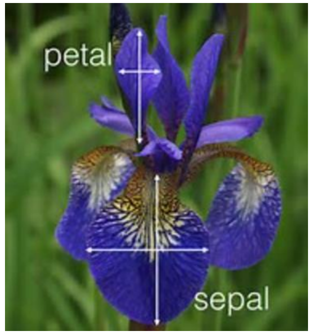
authors represent all vectors as columns

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \quad \mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

matrix: bold, upper-case **X**
each cell indexed by row and column

row: lower-case, script x:
values for an observation
i is an index for the row
p is the number of predictors

column: bold, lower-case **x**:
values for a variable
j is an index for the column
n is the number of observation



example: 150 x 4 matrix
sepal width, sepal length, petal width, petal length measurements for 150 flowers

x is used to identify input data

Output Vector

- An output vector is used for supervised learning
  - Numeric output values for regression
  - Nominal (categorical) output values for classification

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

y is used to identify output data

Alternative Names

- X
- Input Variable
- Predictor
- Covariate
- Independent
- Exogenous

- y
- Output Variable
- Response
- Target
- Dependent
- Endogenous

Counts

- ‘n’ is the number of observations in a data set (rows of the matrix)
- ‘p’ is the number of predictors in a data set (columns of the matrix)

Matrix Transposition

We just swap the row and column indices: $new_{j,i} = old_{i,j}$

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

$$\mathbf{X}^T = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & \dots & x_{np} \end{pmatrix}$$

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

$$x_i^T = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix}$$

Alternative Matrix Notation

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \end{pmatrix}$$

matrix expressed as a set of column vectors, where each column is a variable

$$\mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$

matrix expressed as a set of row vectors, where each row is an observation
[the authors are treating an observation
Vector as a column vector]

Matrix Multiplication

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$$

$$(\mathbf{AB})_{ij} = \sum_{k=1}^{d} a_{ik} b_{kj}$$

$$\mathbf{AB} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 1 \times 5 + 2 \times 7 & 1 \times 6 + 2 \times 8 \\ 3 \times 5 + 4 \times 7 & 3 \times 6 + 4 \times 8 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$$

$$\boldsymbol{A} \in \mathbb{R}^{n\ x\ p} \qquad \boldsymbol{B} \in \mathbb{R}^{p\ x\ k} \qquad \boldsymbol{AB} \in \mathbb{R}^{n\ x\ k}$$

$\mathbb{R}$: a value from the real number line

Vector Multiplication

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \qquad x = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix}$$

$$\beta^T x = \beta_0 * x_0 + \beta_1 * x_1 + \beta_2 * x_2$$

[sometimes called a dot product]

Terminology Note

- Scalar: a single numeric value
- Vector: a 1-dimensional array of values
- Matrix: a 2-dimensional array of values
- Tensor: an array of values with 3 or more dimensions [e.g. an array of images]

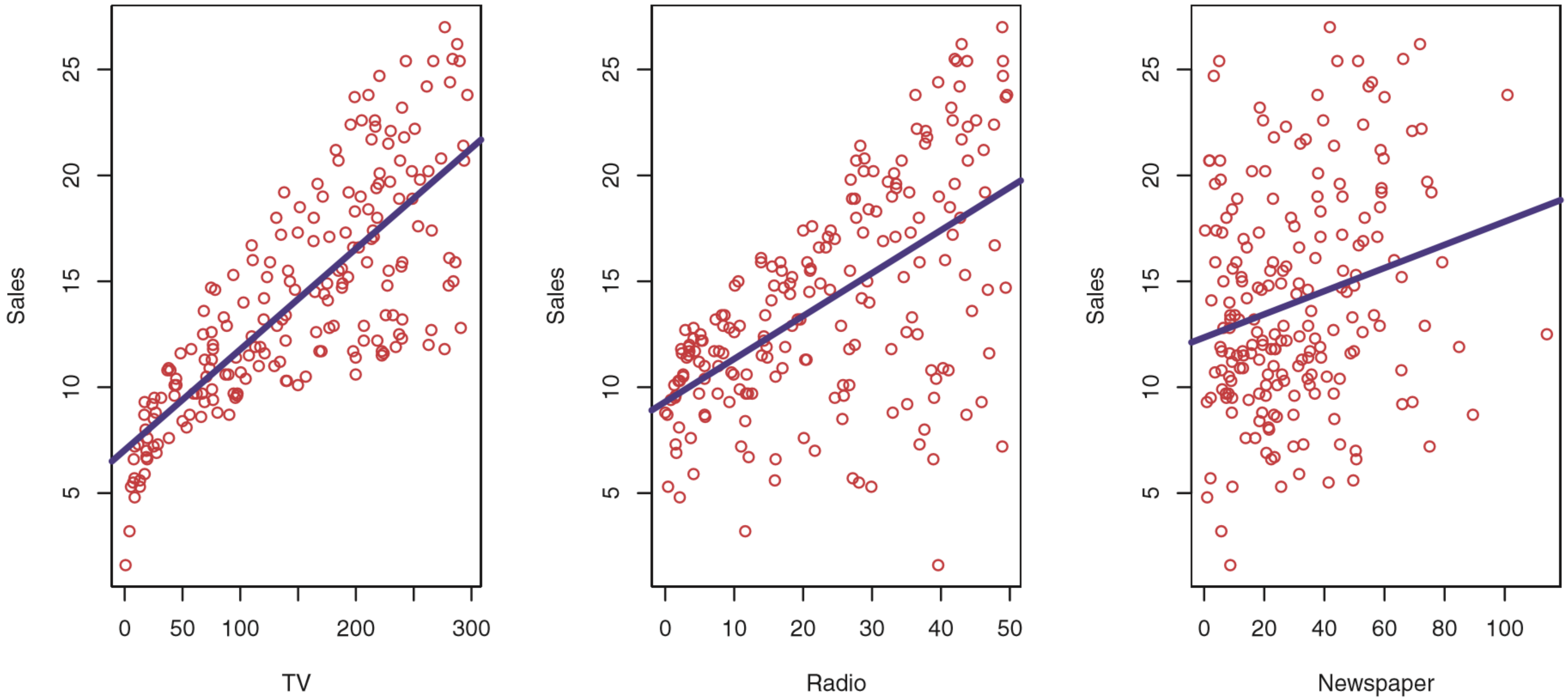
Organization of the Book

- Statistical Learning Terminology and Concepts, plus ‘k’ nearest neighbor
- Regression: Linear Regression
- Classification: Logistic Regression and Linear Discriminant Analysis
- Resampling: Cross Validation and the Bootstrap
- Regression Revisited: Stepwise Selection, Ridge Regression, Principal Components Regression, Partial Least Squares, and the LASSO
- Non-Linear Regression
- Tree-Based Classification: Bagging, Boosting, and Random Forests
- Support Vector Machines
- Unsupervised Learning: Principal Component Analysis, k-Means Clustering, and Hierarchical Clustering

Data Sets Referenced by the Textbook

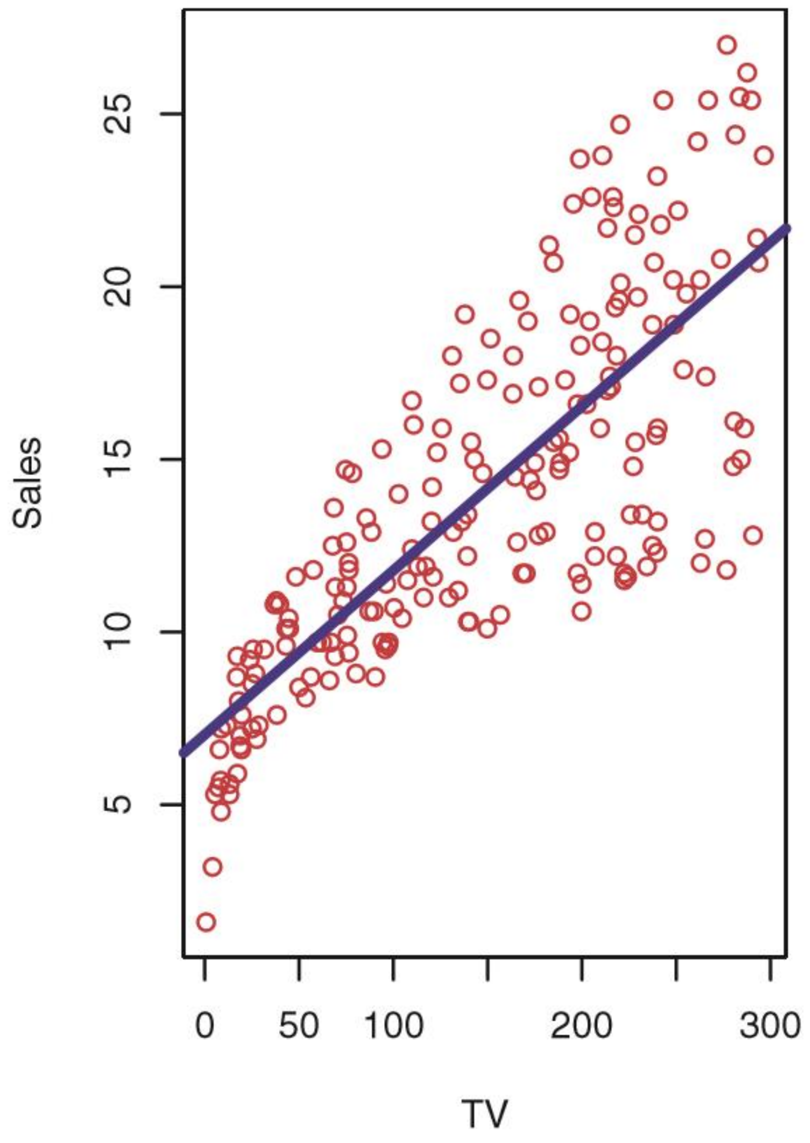
Name	Description
Auto	Gas mileage, horsepower, and other information for cars.
Boston	Housing values and other information about Boston suburbs.
Caravan	Information about individuals offered caravan insurance.
Carseats	Information about car seat sales in 400 stores.
College	Demographic characteristics, tuition, and more for USA colleges.
Default	Customer default records for a credit card company.
Hitters	Records and salaries for baseball players.
Khan	Gene expression measurements for four cancer types.
NCI60	Gene expression measurements for 64 cancer cell lines.
OJ	Sales information for Citrus Hill and Minute Maid orange juice.
Portfolio	Past values of financial assets, for use in portfolio allocation.
Smarket	Daily percentage returns for S&P 500 over a 5-year period.
USArrests	Crime statistics per 100,000 residents in 50 states of USA.
Wage	Income survey data for males in central Atlantic region of USA.
Weekly	1,089 weekly stock market returns for 21 years.

Advertising Data



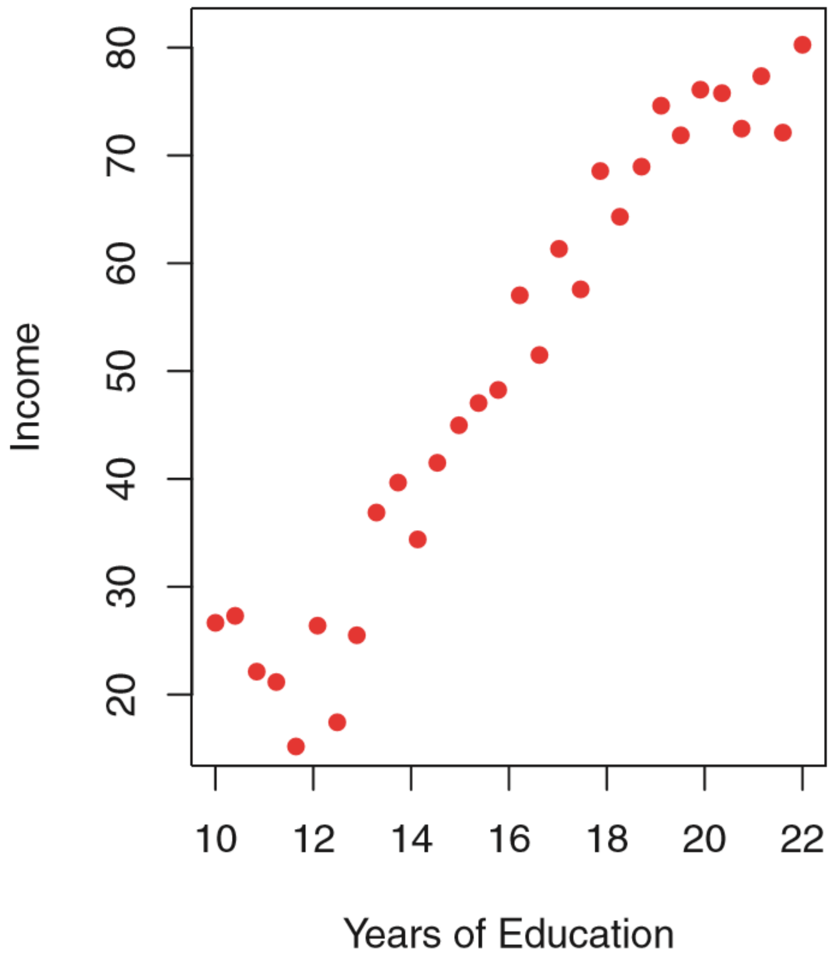
Our First Equation

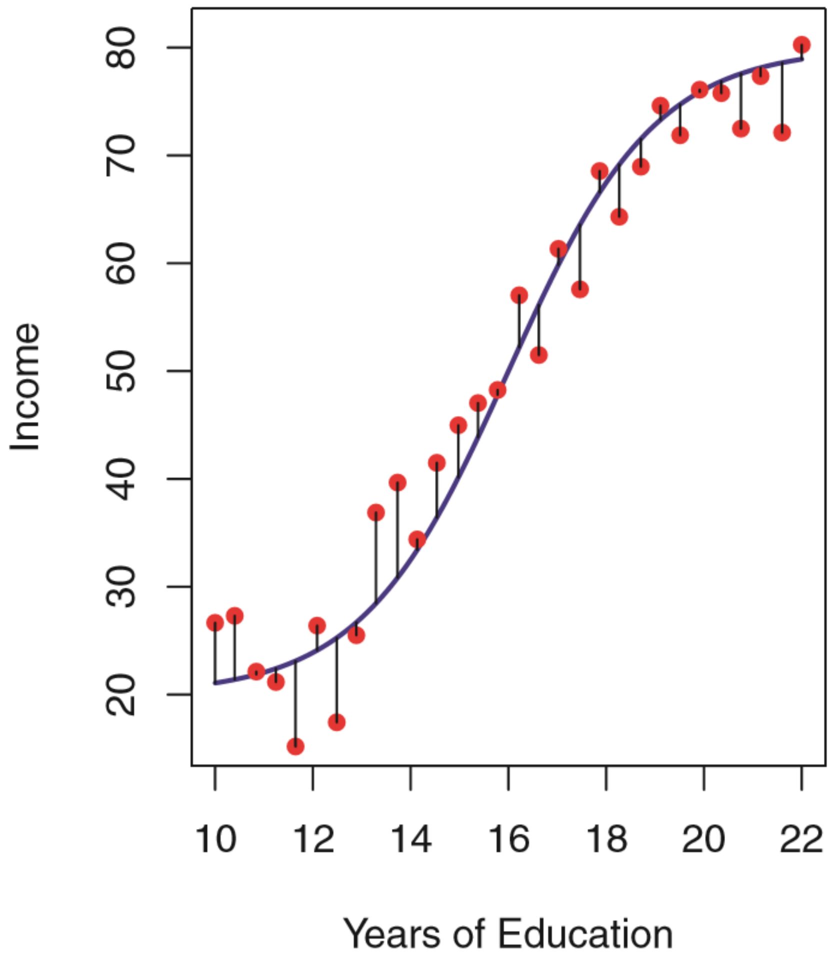
- $Y = f(X) + \epsilon$
- Y is an output Sales value
- $f(X)$ is a function of TV budget
  - $f(X) = 0.05 * X + 7$
    - `Slope: (22 – 7) / (300 – 0) = 0.05`
    - `Intercept: 22 - 0.05 * 300 = 7`
  - `f(  0) = 0.05 *   0 + 7 =  7`
  - `f(100) = 0.05 * 100 + 7 = 12`
  - `f(200) = 0.05 * 200 + 7 = 17`
  - `f(300) = 0.05 * 300 + 7 = 22`
- ϵ is a residual “error” term (Greek letter “epsilon”)



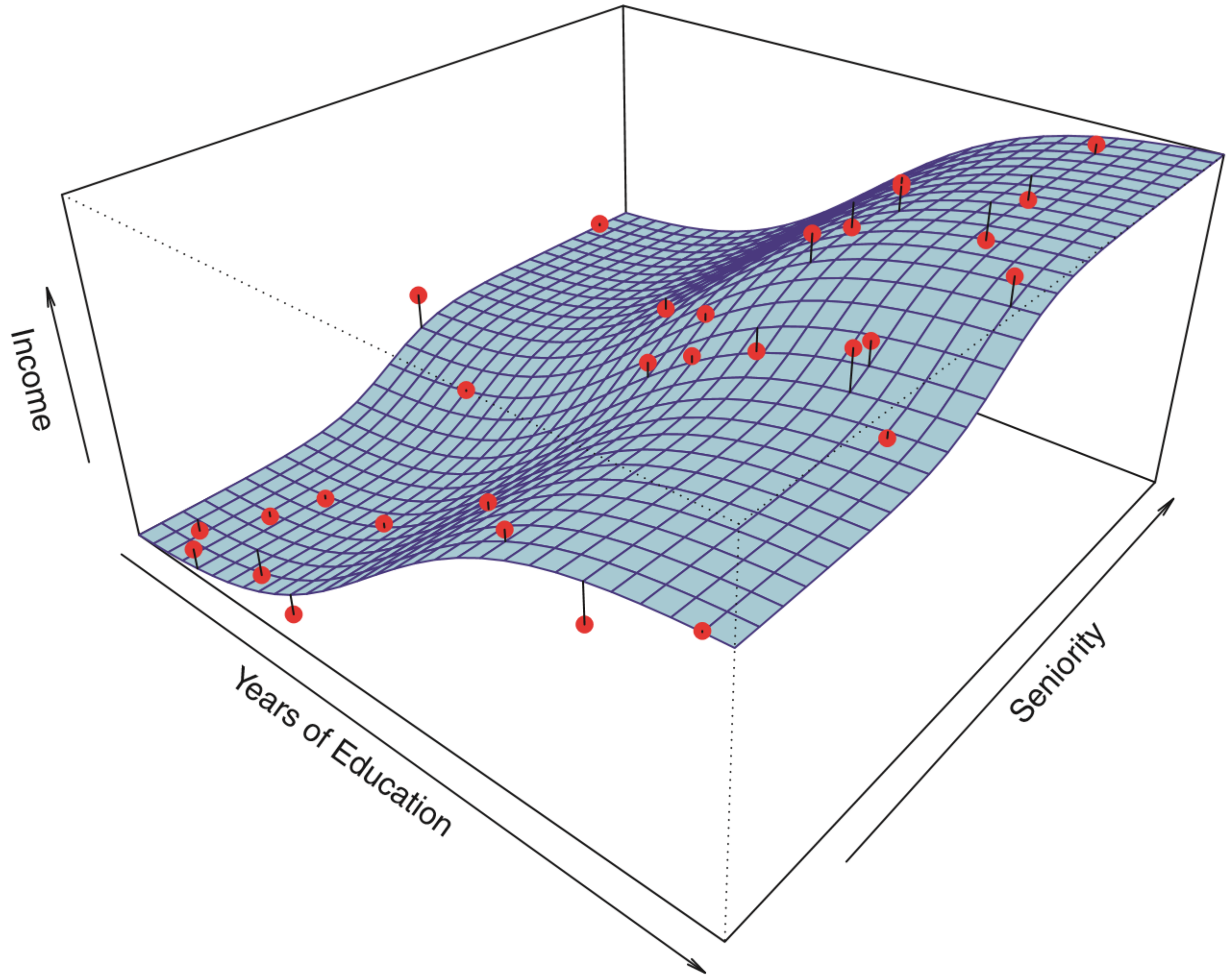
Data = read.csv("http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv")

Income as a Function of Education





Income as a Function of Education and Seniority



Why Estimate $f(X)$?

$\hat{Y} = \hat{f}(X)$

- The hats (circumflex characters: '^') indicate we're talking about estimates rather than some notion of absolute truth
- $\hat{f}(X)$ is the function we learned from data: our function is a model that maps an input to an output
- $\hat{Y}$ is our prediction

- Reasons:
  - To predict an outcome
  - To understand the influence of the predictors on the outcome

Prediction [Our First Loss Function: Squared Error]

- A loss function measures how well a model is able to map inputs to outputs
- $E\left(Y - \hat{Y}\right)^2 = E\left[f(X) + \epsilon - \hat{f}(X)\right]^2 = E\left[f(X) - \hat{f}(X)\right]^2 + Var(\epsilon)$
- $E\left[f(X) - \hat{f}(X)\right]^2$ is referred to as reducible error: we could reduce the error if we had better features
- $Var(\epsilon)$ is referred to as irreducible error, because we believe the process is stochastic rather than deterministic
- $E(\quad)$ indicates we're talking about an expected value (average value)
- $Var(\quad)$ indicates we're talking about variance, the expected squared deviation from the mean
  - Since we believe our residual error has a mean of zero $\mathrm{E}\left(\epsilon^2\right) = Var(\epsilon)$

Inference [Understanding]

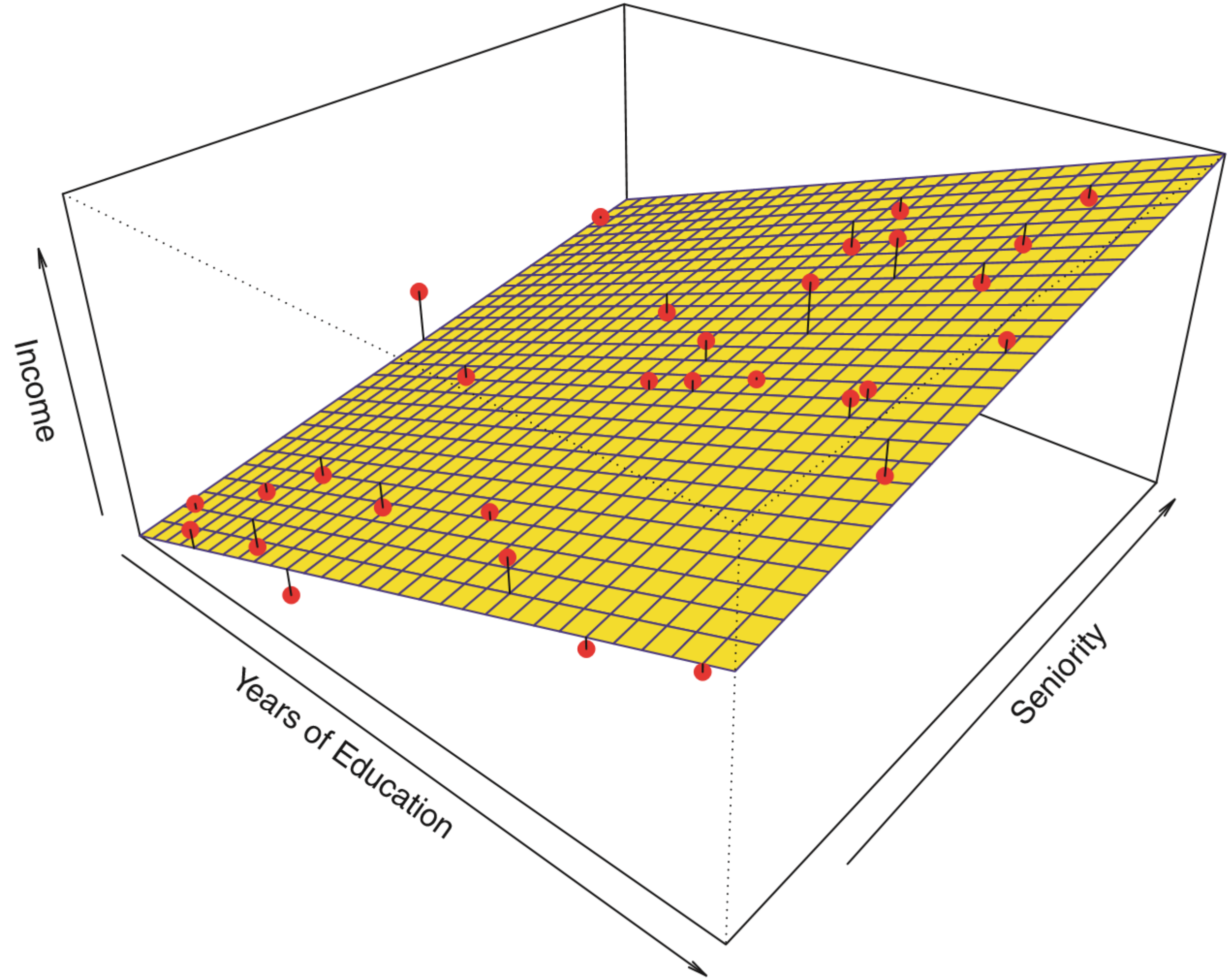
- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between the inputs and outputs be summarized adequately using a linear model, or is the relationship more complex?

- Examples:
  - Which media contribute to sales?
  - Which media generate the biggest boost in sales?
  - How much increase in sales is associated with a given increase in TV advertising?

How Do We Estimate f?

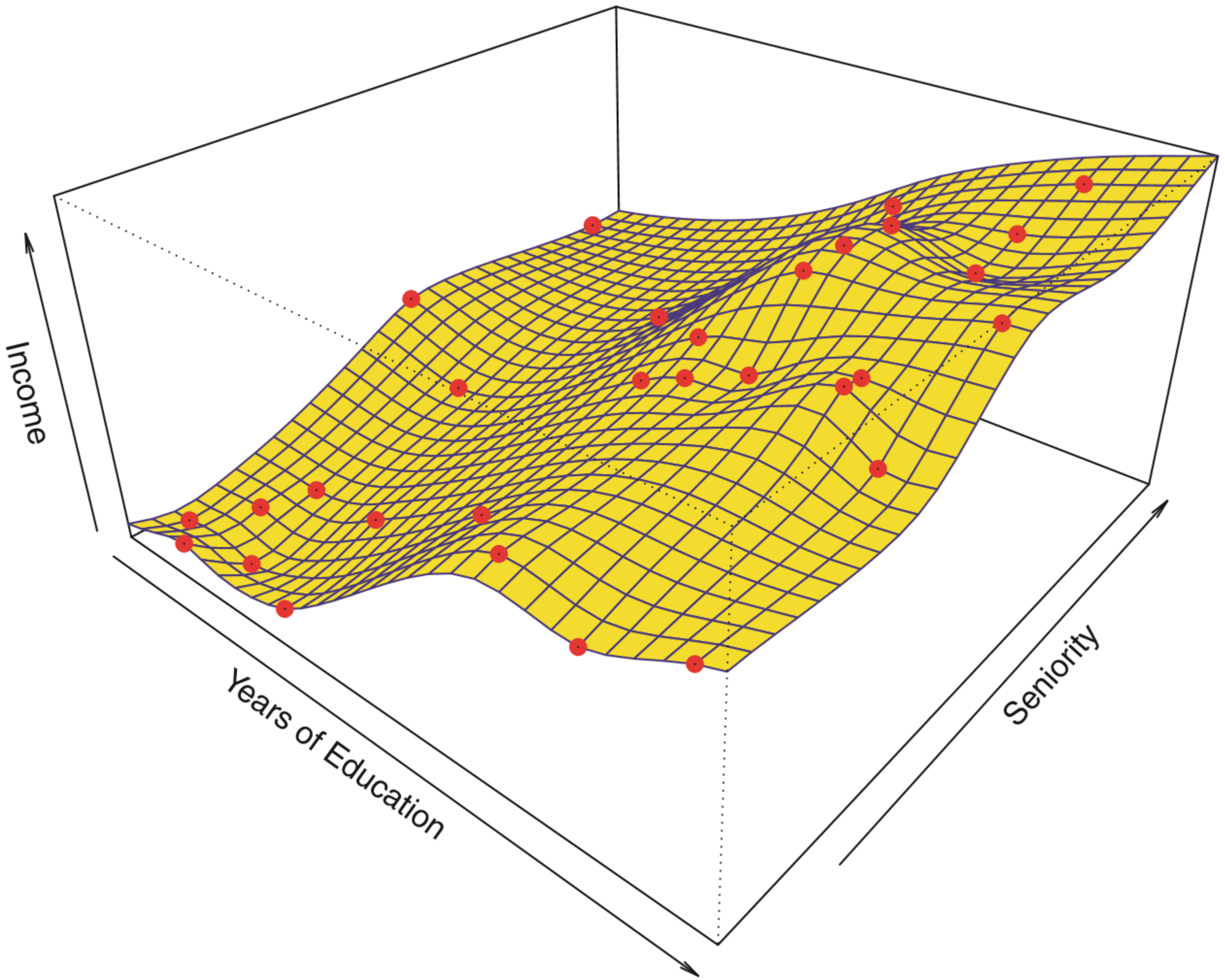
- Parametric methods: the size of the model is fixed; e.g. linear regression, polynomial regression, logistic regression, neural network
- Non-Parametric methods: the size of the model can grow with the amount of training data; e.g. nearest neighbor, random forests, gradient boosting, support vector machines

Parametric Linear Model for Income

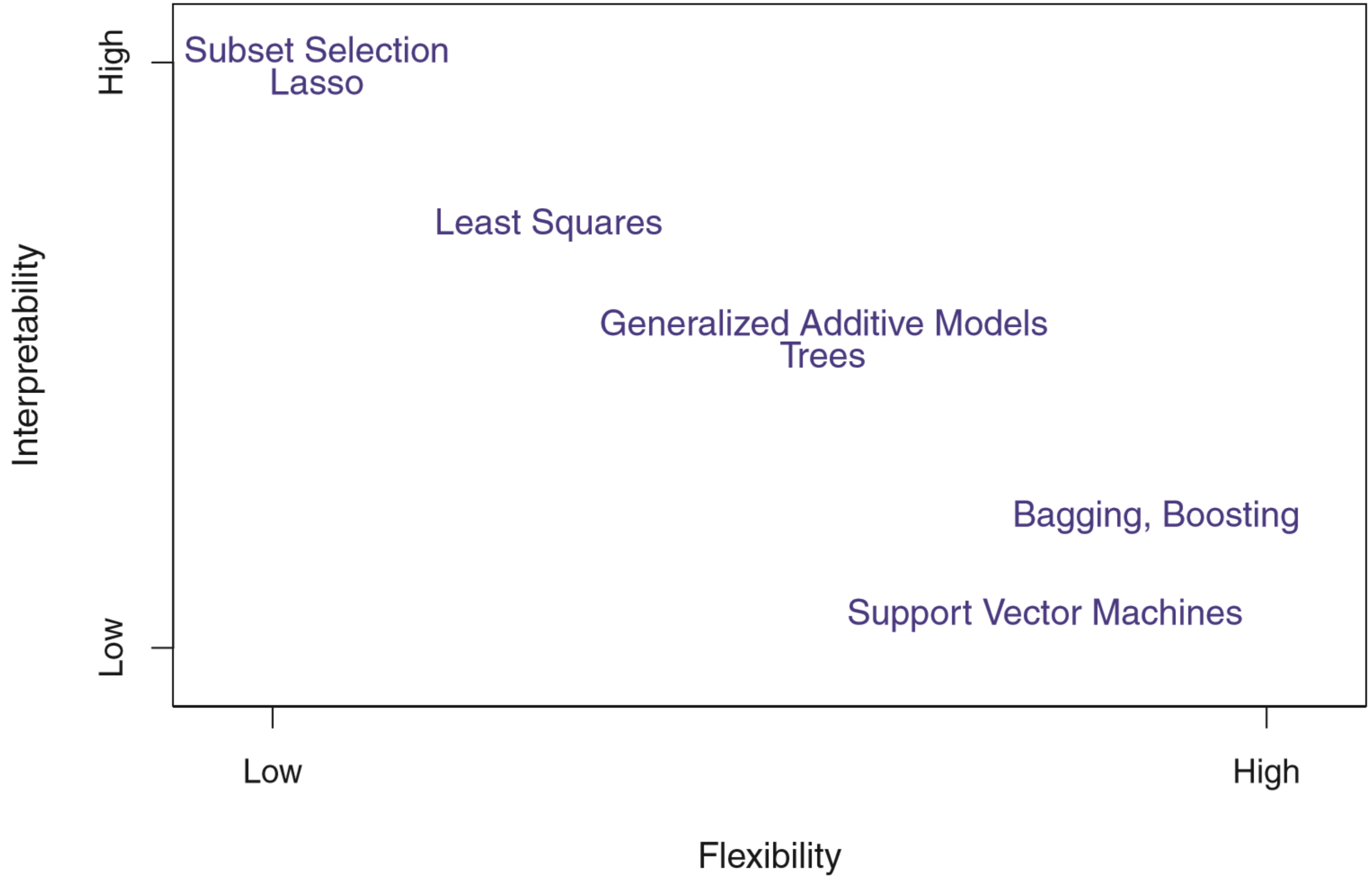


$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}$$

Non-Parametric Non-Linear Model for Income



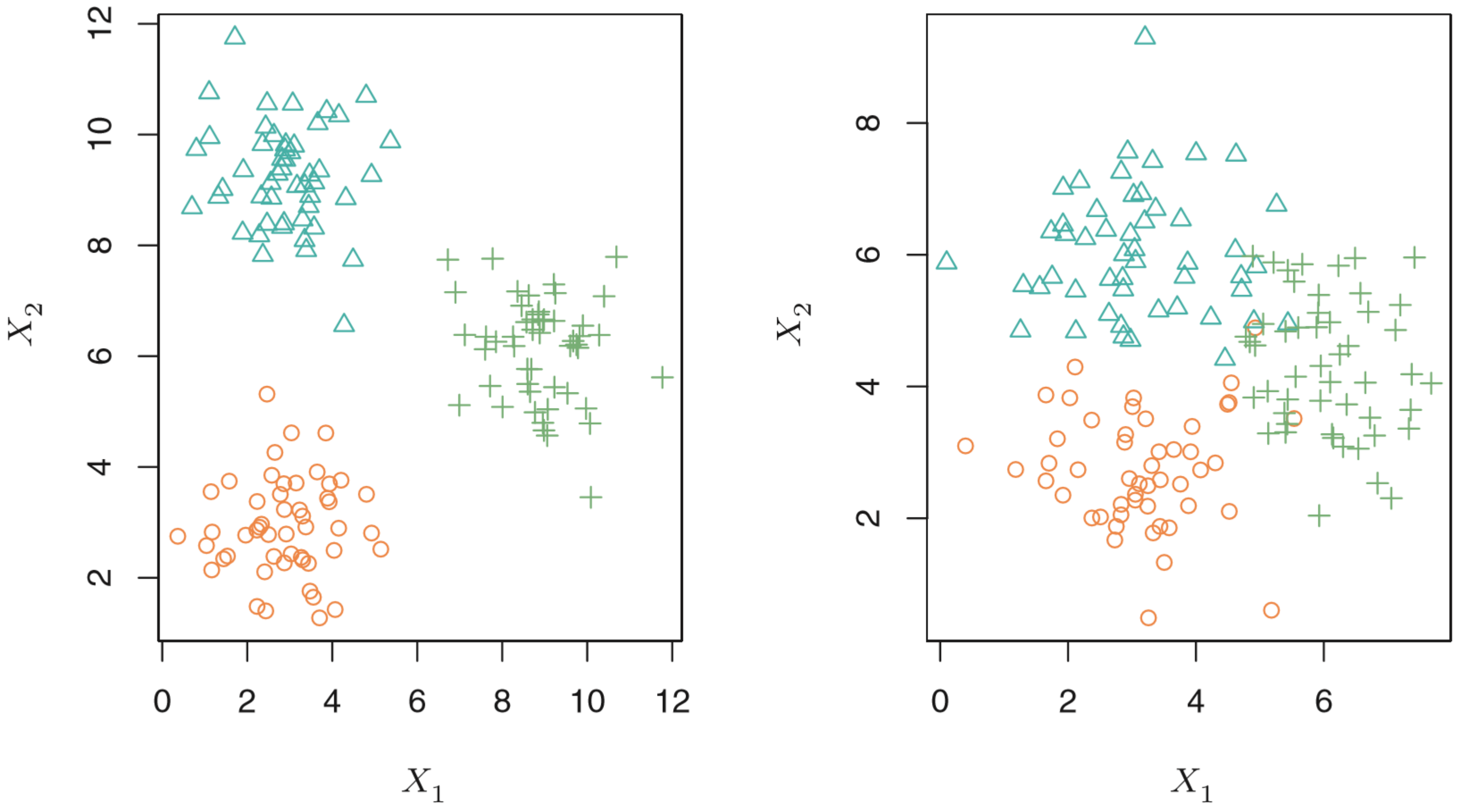
Trade-Off Between Prediction Accuracy and Model Interpretability



Supervised versus Unsupervised Learning

- Supervised Learning
  - The learning algorithm is given a target output variable
    - Classification: the output variable is nominal (categorical, qualitative)
    - Regression: the output variable is numeric (quantitative)
- Unsupervised Learning
  - The learning algorithm is *not* given a target output variable
    - Clustering
    - Principal Component Analysis

Unsupervised Learning and Class Overlap



Measuring the Quality of the Model

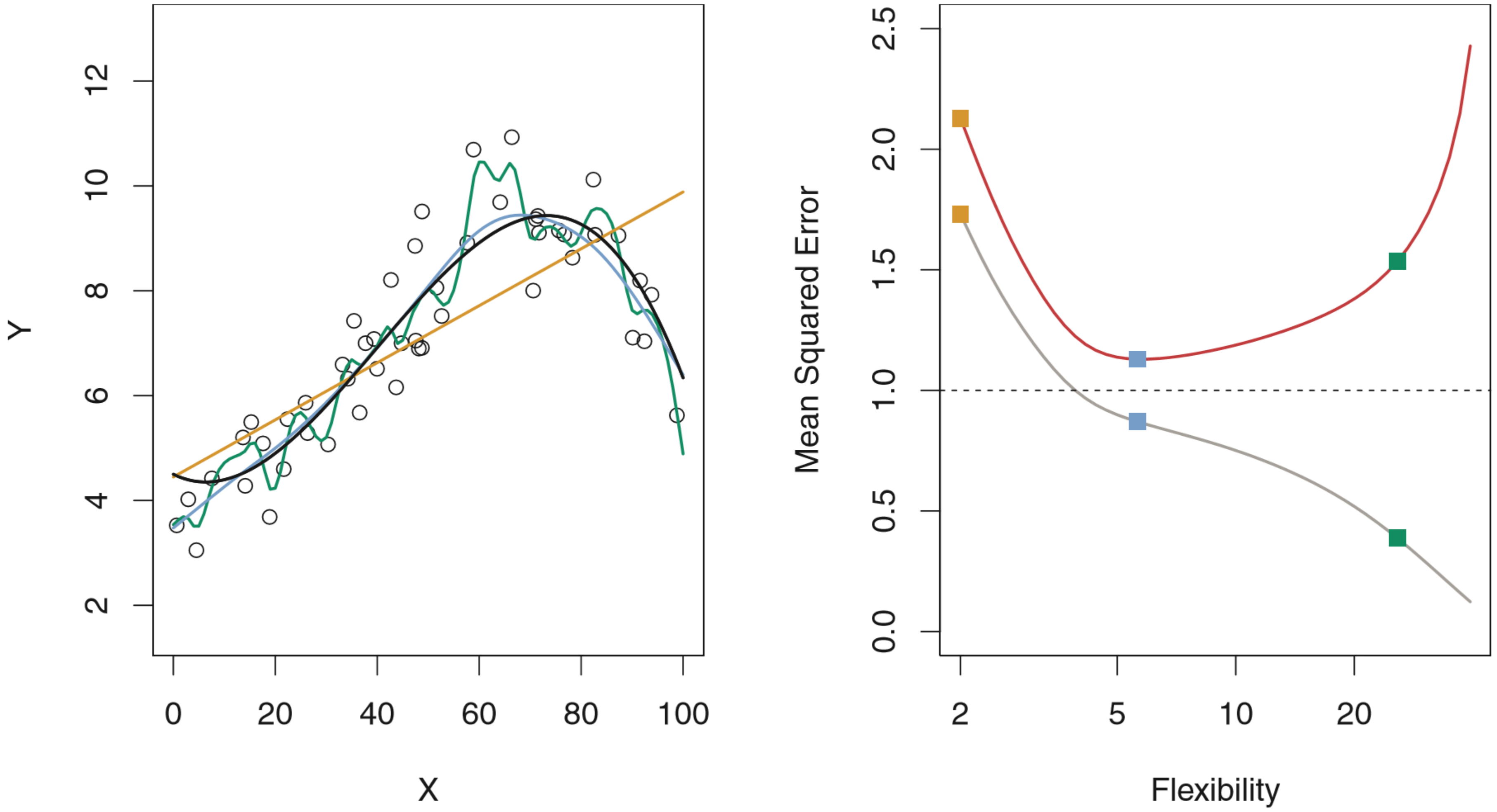
Common Loss functions

- Regression
  - Gaussian loss (mean squared error)
  - Laplacian loss (mean absolute error)
- Classification
  - Log loss
  - Hinge loss

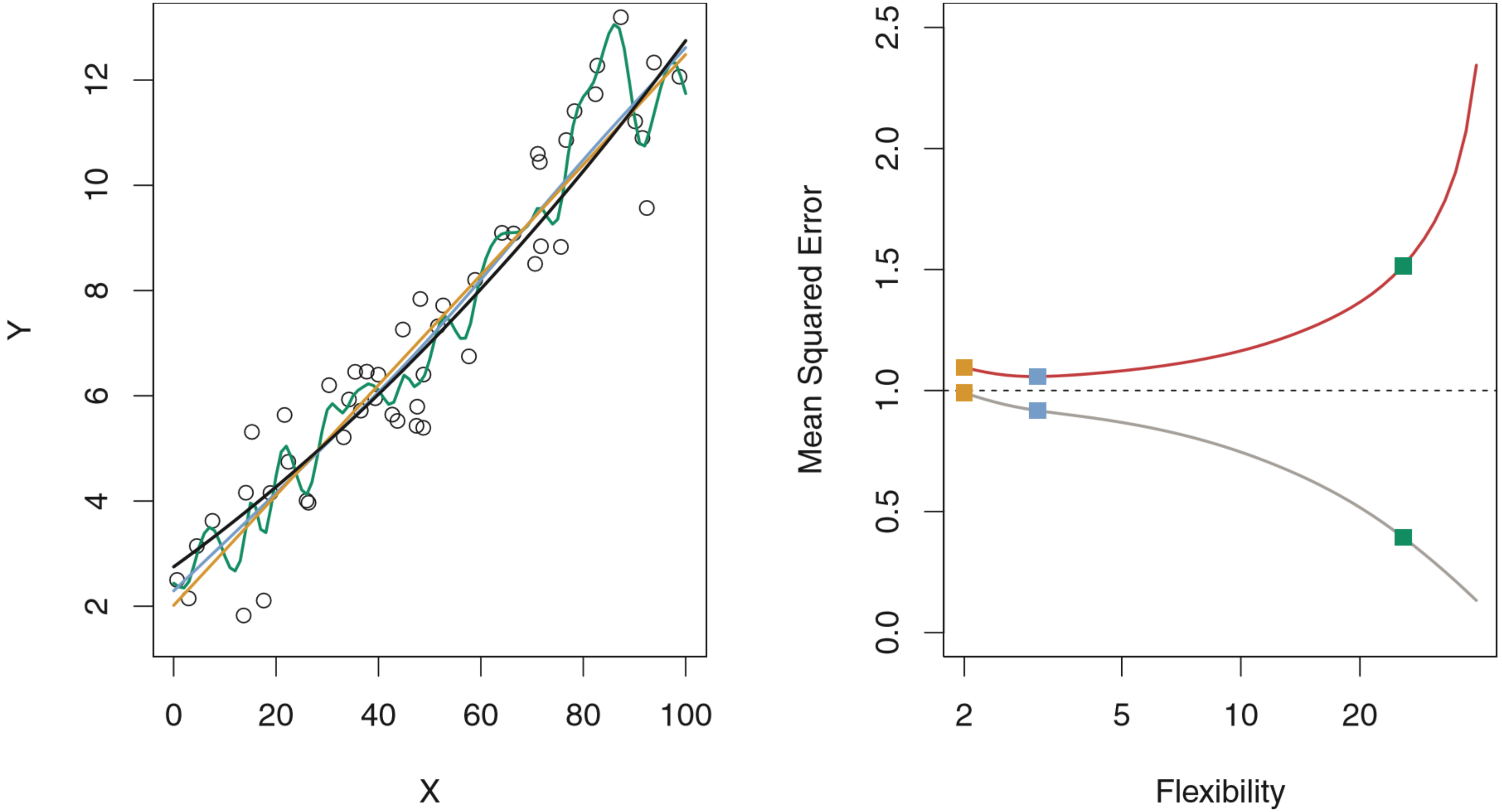
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \qquad \text{Ave}(y_0 - \hat{f}(x_0))^2$$

Example: High Bias (underfitting) versus High Variance (overfitting)

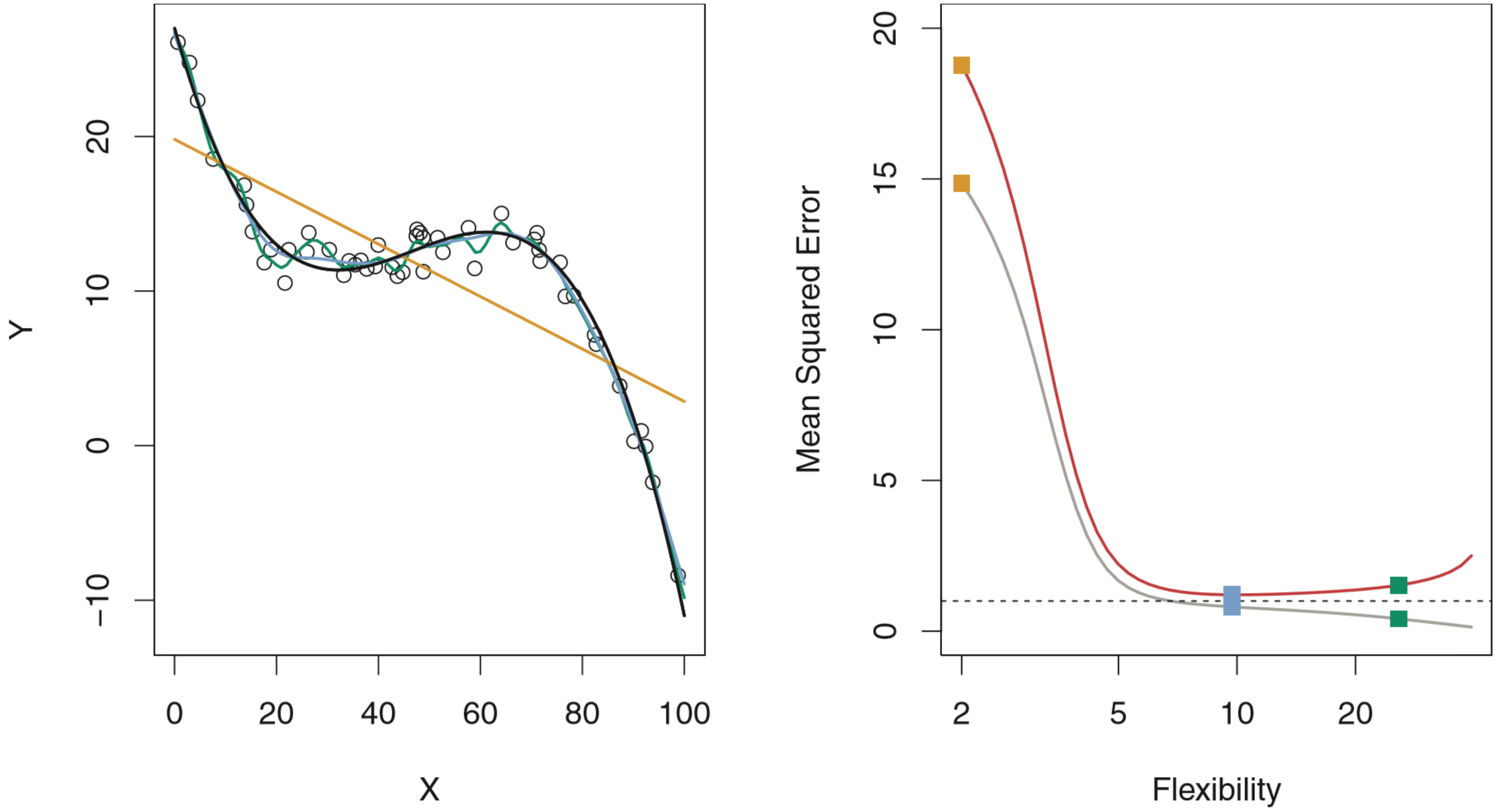
Overfitting: the region of flexibility where the loss increases for the testing data but decreases for the training data



Example: Overfitting



Bias versus Variance Trade-Off



Bias Variance Decomposition

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = E\left(f(x_0) + \epsilon - \hat{f}(x_0)\right)^2 = E\left(f(x_0) - \hat{f}(x_0)\right)^2 + Var(\epsilon)$$

$$= E\left(f(x_0) - E\left(\hat{f}(x_0)\right) + E\left(\hat{f}(x_0)\right) - \hat{f}(x_0)\right)^2 + Var(\epsilon)$$

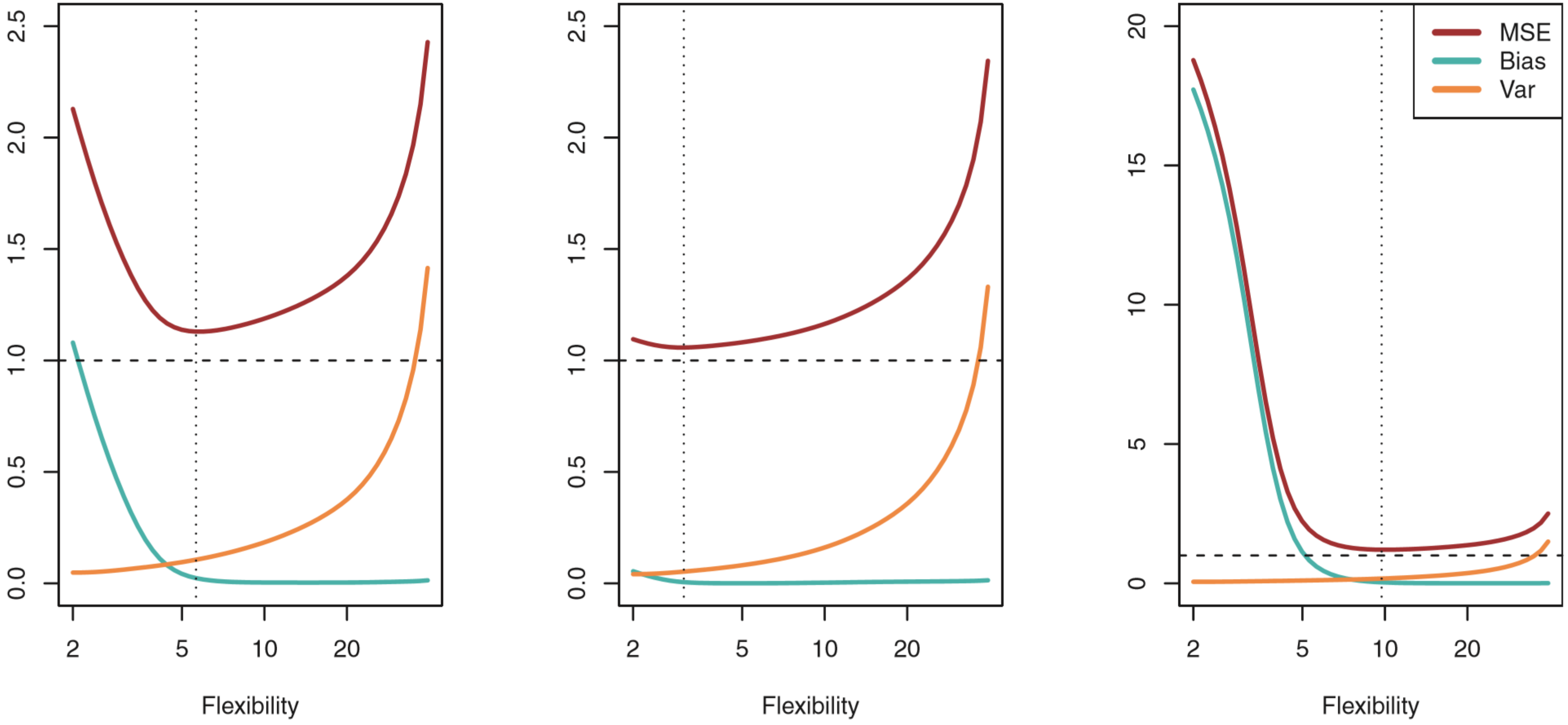
$$= E\left(\left(f(x_0) - E\left(\hat{f}(x_0)\right)\right)^2 + 2 * \left(f(x_0) - E\left(\hat{f}(x_0)\right)\right) * \left(E\left(\hat{f}(x_0)\right) - \hat{f}(x_0)\right) + \left(E\left(\hat{f}(x_0)\right) - \hat{f}(x_0)\right)^2\right) + Var(\epsilon)$$

$$= E\left(f(x_0) - E\left(\hat{f}(x_0)\right)\right)^2 + 0 + \mathrm{E}\left(E\left(\hat{f}(x_0)\right) - \hat{f}(x_0)\right)^2 + Var(\epsilon)$$

$$= \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var\left(\hat{f}(x_0)\right) + Var(\epsilon)$$

- We’re adding and subtracting the same value (zero) on line 2
- We’re grouping pairs of terms and multiplying on line 3
- We’re using $\mathrm{E}\left(E\left(\hat{f}(x_0)\right) - \hat{f}(x_0)\right) = 0$ on line 4

Optimal Flexibility Varies by Problem



Variance Increases and Bias Decreases as Model Flexibility Increases

Classification Error

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

Accuracy = 1 - Error

I() is an indicator function which returns 1 iff (if and only if) the condition is true; e.g. the actual class label is not equal to the predicted class label

$$\text{Ave}\left(I(y_0 \neq \hat{y}_0)\right)$$

Bayes Classifier

The Bayes classifier picks the class 'j' that maximizes the probability

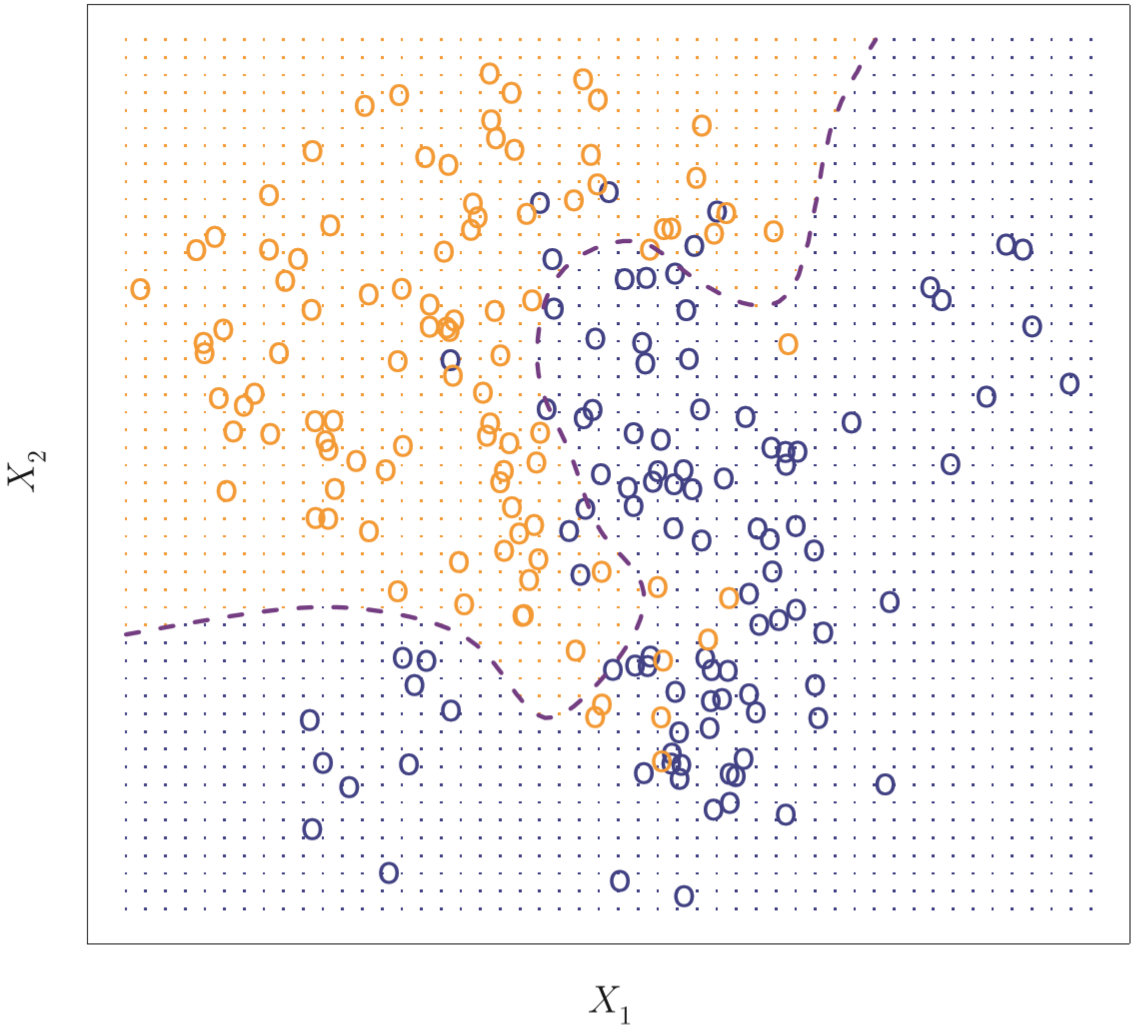
$$\Pr(Y = j | X = x_0)$$

Read "probability that Y is equal to j given that X is equal to x_0"

The Bayes error rate is

$$1 - E\left(\max_j \Pr(Y = j | X)\right)$$

Bayes Classifier for Simulated Problem



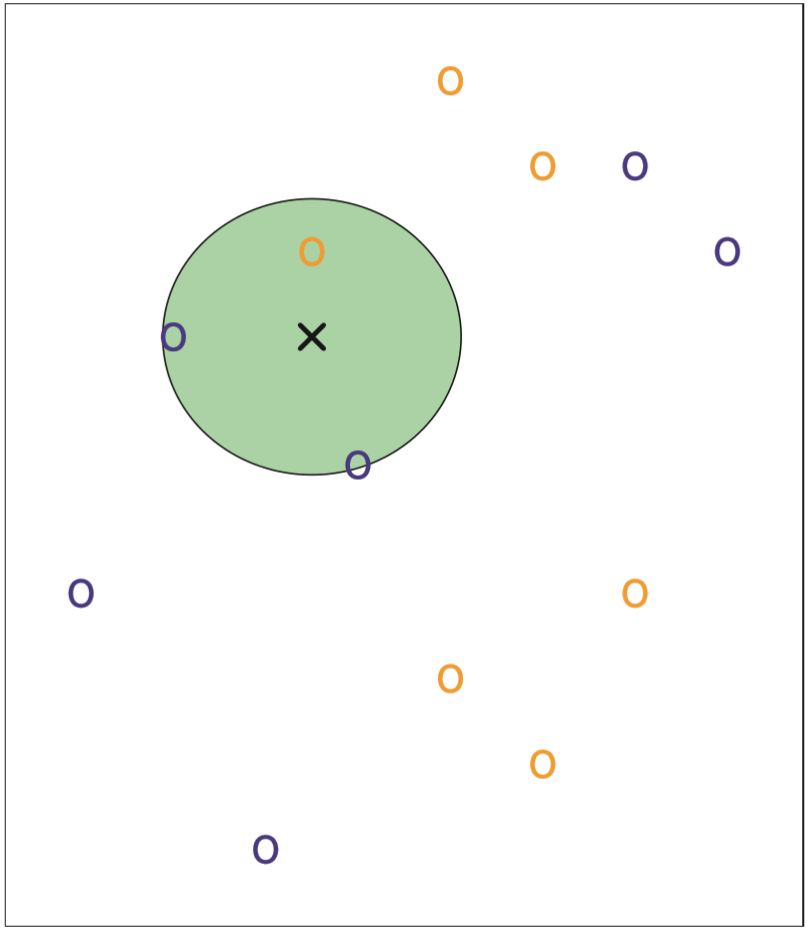
K Nearest Neighbors

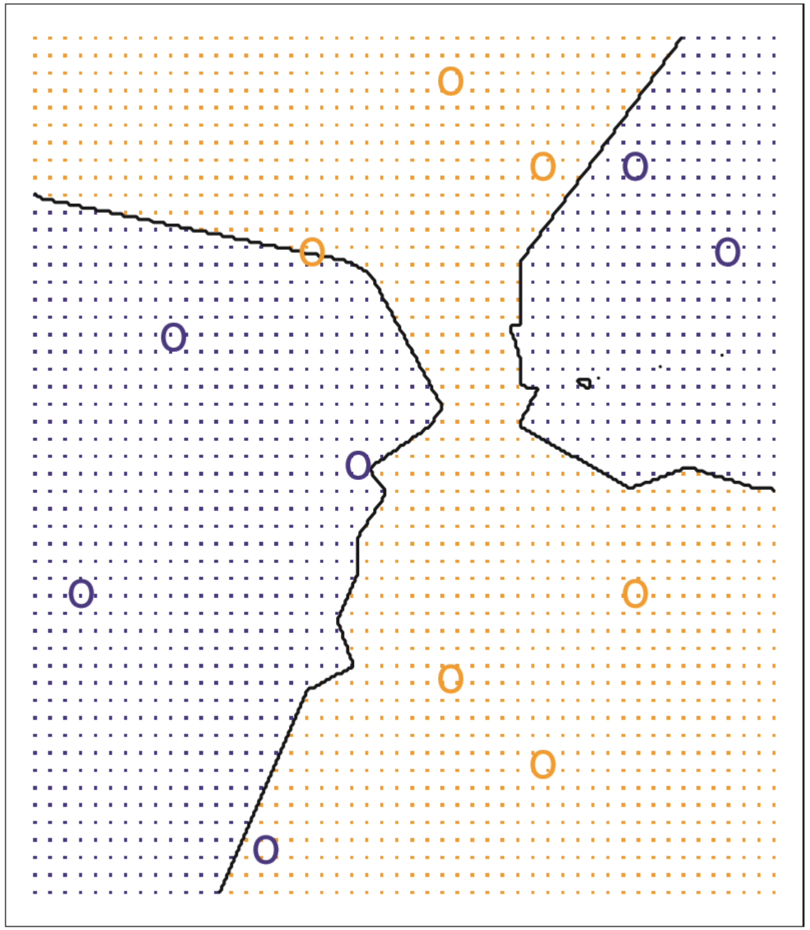
$$\Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

$where\ \mathcal{N}_0\ is\ the\ set\ of\ indices\ for\ the\ 'K'\ nearest\ neighbors\ of\ x_0$

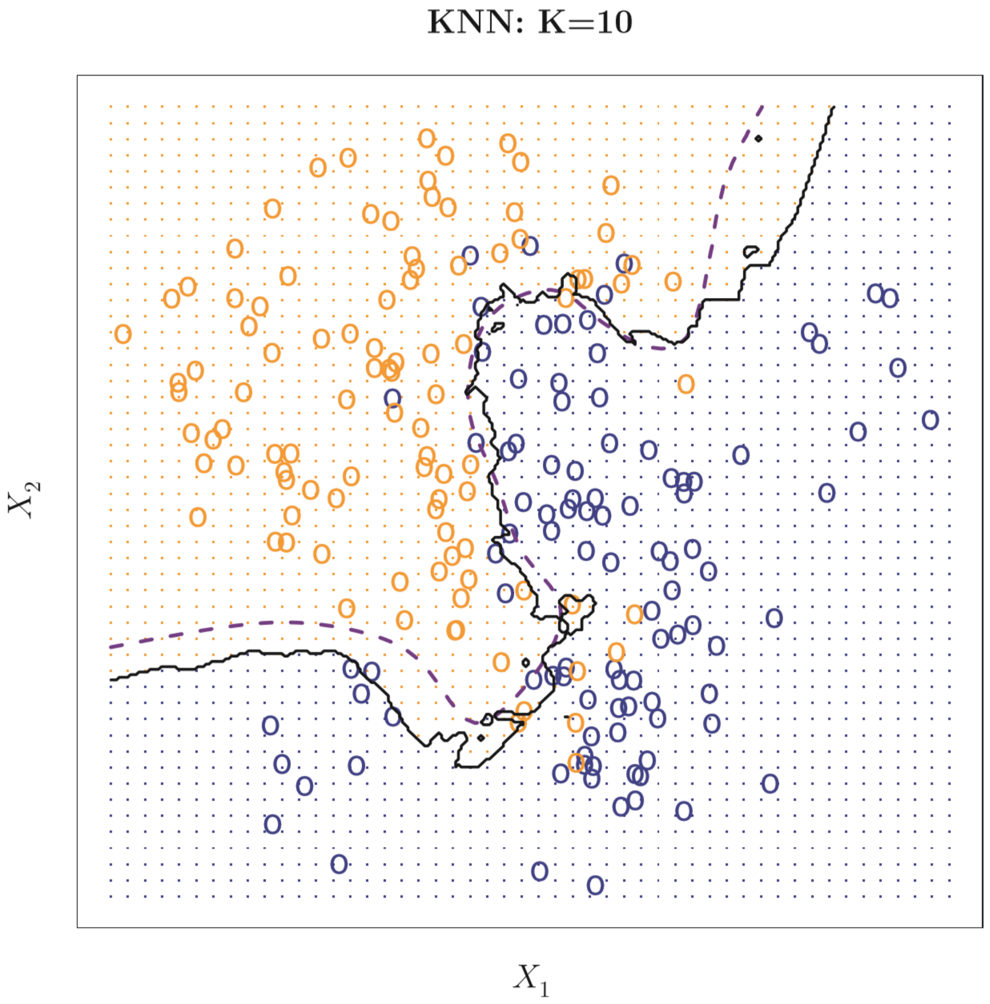
For classification using K nearest neighbors, we're estimating the proportion of nearest neighbors that belong to class 'j'

K Nearest Neighbor Classifier Example (k=3)

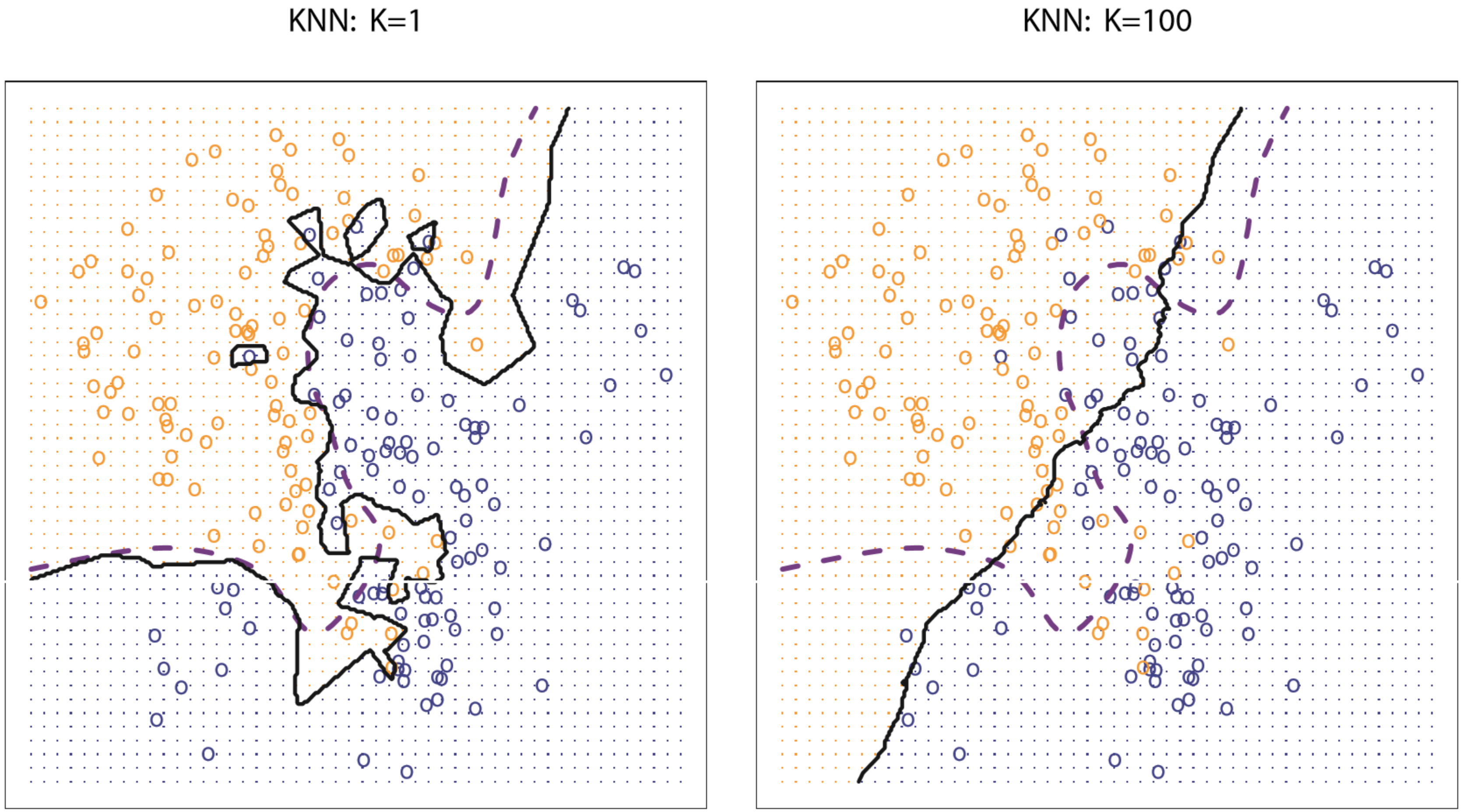




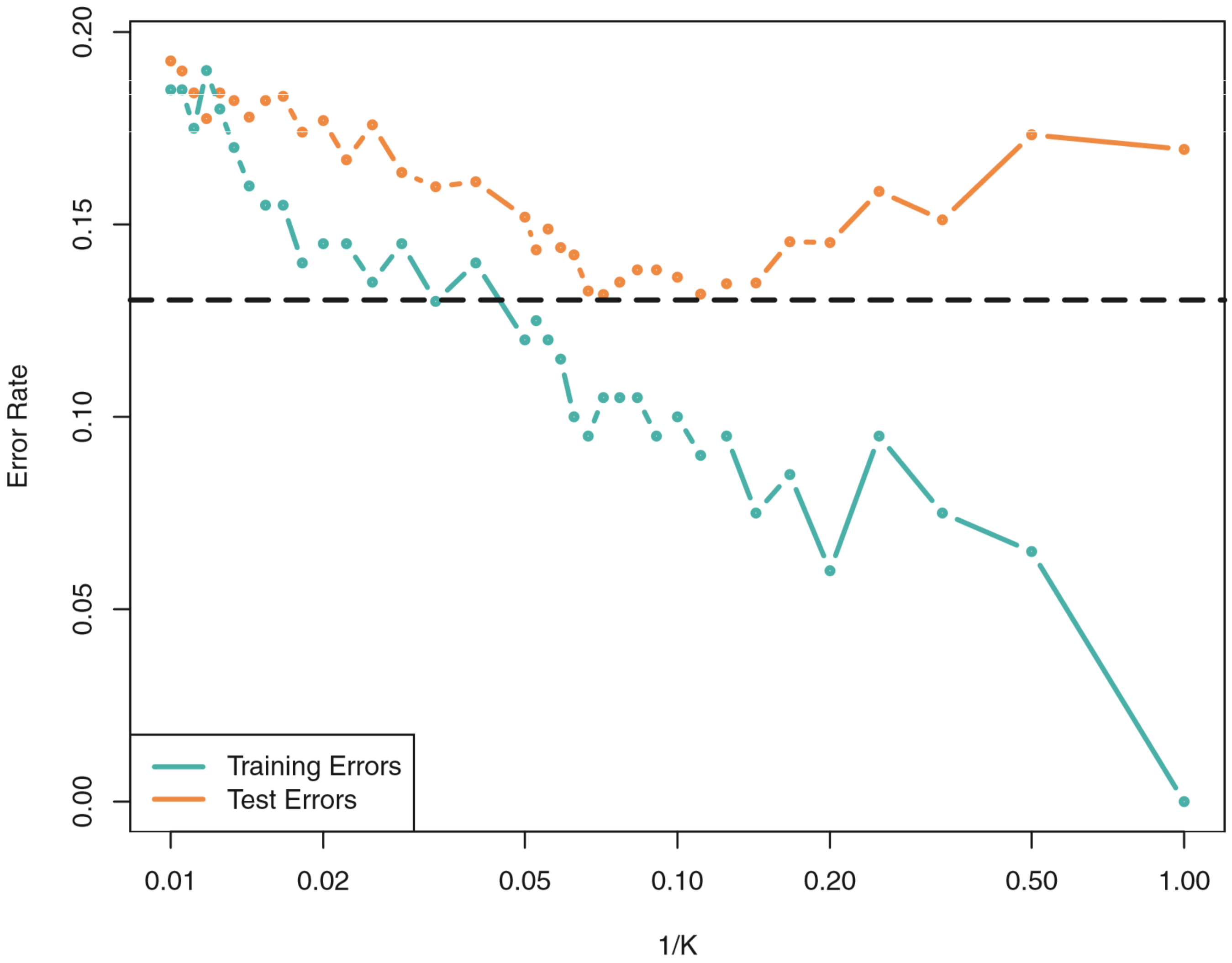
KNN with K=10 versus Bayes Decision Boundary



KNN with K=1 versus K=10



Error versus Complexity for KNN



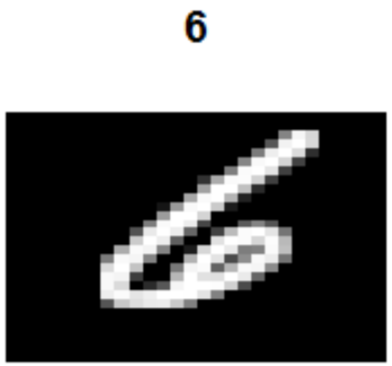
What’s Left to Talk About?

- Lab
  - Install R from https://cran.r-project.org/
  - Execute the commands from the Lab in Section 2.3 of the textbook
  - Use the following R command to install the “ISLR” package:

    install.packages(“ISLR”)
    # choose “USA (WA) [https]” for the mirror
- Homework
  - Submit your response for Assignment #1 to the http://canvas.uw.edu site
    a. Please include a brief note about …
       1. your education
       2. your current job
       3. how you would like to use knowledge acquired through this certificate program
    b. Answer question #2 from the exercises in Section 2.4 (page 52)
    c. Answer question #9 from the exercises in Section 2.4 (page 56)
    d. https://kaggle.com/join/ml210_mnist

KNN Example

```
> set.seed(2^17 - 1)
> start.time = Sys.time()
>
> trn_X = read.csv("C:/Data/mnist/trn_X.csv", header = F)
> trn_y = scan("C:/Data/mnist/trn_y.txt")
Read 60000 items
> tst_X = read.csv("C:/Data/mnist/tst_X.csv", header = F)
>
> rotate = function(X) t(apply(X, 2, rev))
> windows(height = 3, width = 3)
> i = sample.int(nrow(trn_X), size = 1)
> image(rotate(matrix(as.numeric(trn_X[i,]), nrow = 28, byrow = T)),
+         col = gray.colors(256, 0, 1),
+         main = trn_y[i], axes = F)
>
> library(class)
> subset = sample(1:nrow(trn_X), 0.25 * nrow(trn_X))
> predictions = knn(trn_X[subset,], tst_X, factor(trn_y[subset]), k = 1)
> output = data.frame(Id = 1:length(predictions), Prediction = predictions)
> write.csv(output, "C:/Data/mnist/predictions.csv", quote=F, row.names = F)
> Sys.time() - start.time
Time difference of 14.55124 mins
```



Agenda