

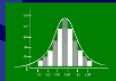
Introduction to Data Science

Winter Semester 2023/24

Oliver Ernst

TU Chemnitz, Fakultät für Mathematik, Professur Numerische Mathematik

Lecture Slides



Contents I

1 What is Data Science?

2 Learning Theory

2.1 What is Statistical Learning?

2.2 Assessing Model Accuracy

3 Linear Regression

3.1 Simple Linear Regression

3.2 Multiple Linear Regression

3.3 Other Considerations in the Regression Model

3.4 Revisiting the Marketing Data Questions

3.5 Linear Regression vs. K -Nearest Neighbors

4 Classification

4.1 Overview of Classification

4.2 Why Not Linear Regression?

4.3 Logistic Regression

4.4 Linear Discriminant Analysis

4.5 A Comparison of Classification Methods

5 Resampling Methods

Contents II

5.1 Cross Validation

5.2 The Bootstrap

6 Linear Model Selection and Regularization

6.1 Subset Selection

6.2 Shrinkage Methods

6.3 Dimension Reduction Methods

6.4 Considerations in High Dimensions

6.5 Miscellanea

7 Nonlinear Regression Models

7.1 Polynomial Regression

7.2 Step Functions

7.3 Regression Splines

7.4 Smoothing Splines

7.5 Generalized Additive Models

8 Tree-Based Methods

8.1 Decision Tree Fundamentals

8.2 Bagging, Random Forests and Boosting

8.3 More on Boosting

9 Unsupervised Learning

9.1 Principal Components Analysis

9.2 Clustering Methods

10 Neural Networks

② Learning Theory

2.1 What is Statistical Learning?

2.2 Assessing Model Accuracy

② Learning Theory

2.1 What is Statistical Learning?

2.2 Assessing Model Accuracy

Learning Theory

Example: Advertising channels

- Allocate an advertising budget for a product across the three media channels **TV**, **radio** and **newspaper**.
- Data set containing sales numbers in 200 markets (in 1000s of units) available for each medium for different advertising budget values in 200 markets.
- We will try to model the dependence of sales on advertising budgets.
- Terminology:

X_1 : TV budget
 X_2 : radio budget
 X_3 : newspaper budget

}

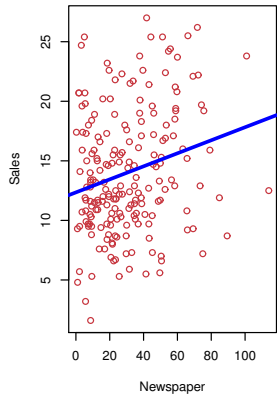
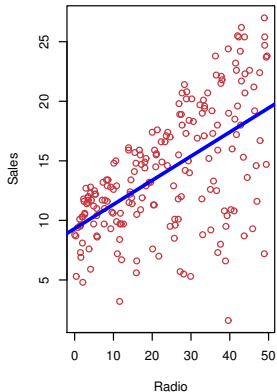
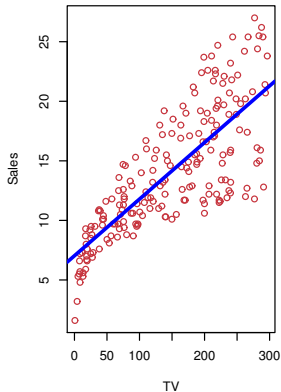
independent variables,
input variables,
predictors,
variables,
features

Y : sales

dependent variable,
output variable,
target variable,
response

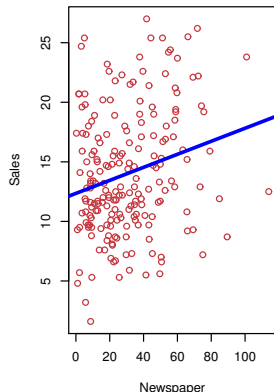
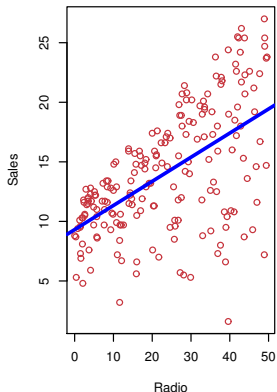
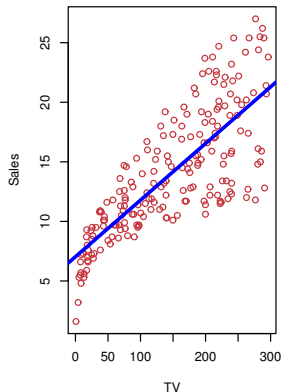
Learning Theory

Example: Advertising channels



Learning Theory

Example: Advertising channels



$$Y = f(X) + \varepsilon$$

$$\begin{aligned} X &= (X_1, \dots, X_p), & p &= \# \text{ predictors,} \\ \varepsilon &: \text{ random error term, } \mathbf{E}[\varepsilon] = 0, \\ f &: \text{ systematic information } X \text{ provides about } Y. \end{aligned} \tag{2.1}$$

Learning Theory

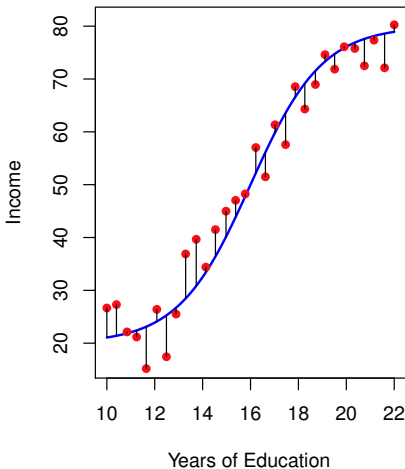
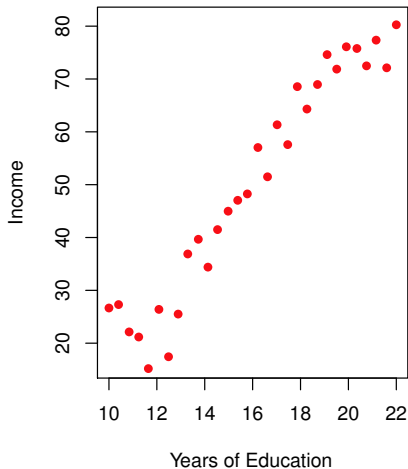
Example: Income

- Data set shows `income` against `years of education` for 30 people.
- Objective: determine function f relating `income` as response to `years of education` as predictor.
- f generally unknown, must be estimated from the data.
- Here: data simulated, so f available.
- In another data set, `income` is given with respect to two input variables: `years of education` and `seniority`.

- Statistical learning is concerned with techniques for estimating f from a data set.

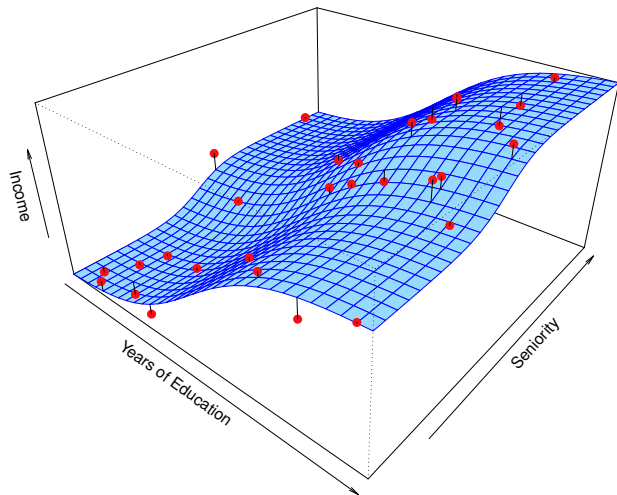
Learning Theory

Example: Income



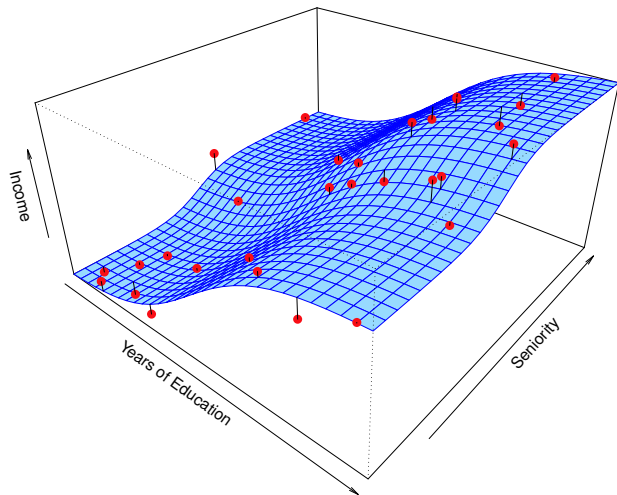
Learning Theory

Example: Income



Learning Theory

Example: Income



Two main reasons for estimating f :
prediction
and
inference.

Learning Theory

Prediction

- Suppose inputs X readily available, but outputs Y difficult to obtain.
- Since errors average out, predict Y using

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

\hat{f} : estimate for f ,
 \hat{Y} : prediction for Y .

- Precise form of \hat{f} often not important ('black box'), only that it gives good predictions.

- Suppose inputs X readily available, but outputs Y difficult to obtain.
- Since errors average out, predict Y using

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

\hat{f} : estimate for f ,
 \hat{Y} : prediction for Y .

- Precise form of \hat{f} often not important ('black box'), only that it gives good predictions.

Example:

X_1, \dots, X_p : characteristics of a patient's blood samples, measured in lab.

Y : patient's risk for severe adverse reaction to particular drug.

For obvious reasons, having an accurate estimate $\hat{Y} = \hat{f}(X)$ is preferable to evaluating $Y = f(X)$.

Learning Theory

Prediction

Accuracy of $\hat{Y} \approx Y$ depends on **reducible error** and **irreducible error**.

Learning Theory

Prediction

Accuracy of $\hat{Y} \approx Y$ depends on **reducible error** and **irreducible error**.

- **reducible error**: $f - \hat{f}$. Can be made smaller and smaller by employing increasingly sophisticated statistical learning techniques.

Accuracy of $\hat{Y} \approx Y$ depends on **reducible error** and **irreducible error**.

- **reducible error**: $f - \hat{f}$. Can be made smaller and smaller by employing increasingly sophisticated statistical learning techniques.
- **irreducible error**: ε . Present even for $f = \hat{f}$, cannot be predicted from X .

Possible sources:

- Additional variables Y may depend on but which are not observed/measured.
- Unmeasurable variation.
(E.g.: Adverse reaction may depend on manufacturing variations in drug or variations in patient's sensitivity over time.)

Accuracy of $\hat{Y} \approx Y$ depends on **reducible error** and **irreducible error**.

- **reducible error**: $f - \hat{f}$. Can be made smaller and smaller by employing increasingly sophisticated statistical learning techniques.
- **irreducible error**: ε . Present even for $f = \hat{f}$, cannot be predicted from X .
Possible sources:
 - Additional variables Y may depend on but which are not observed/measured.
 - Unmeasurable variation.
(E.g.: Adverse reaction may depend on manufacturing variations in drug or variations in patient's sensitivity over time.)
- Quantitative measure: **mean squared error** (MSE). For fixed X and \hat{f}

$$\mathbf{E} \left[(Y - \hat{Y})^2 \right] = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{reducible}} + \underbrace{\mathbf{Var} \varepsilon}_{\text{irreducible}}$$

Accuracy of $\hat{Y} \approx Y$ depends on **reducible error** and **irreducible error**.

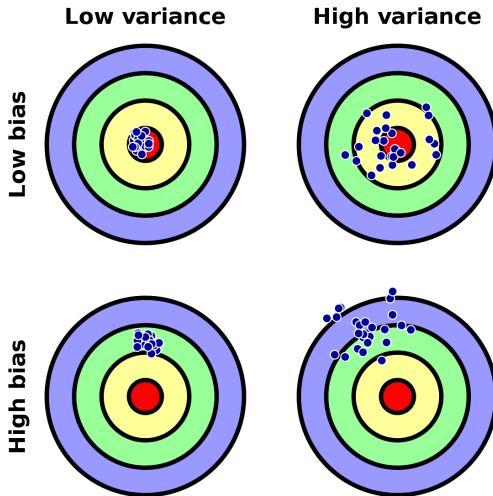
- **reducible error**: $f - \hat{f}$. Can be made smaller and smaller by employing increasingly sophisticated statistical learning techniques.
- **irreducible error**: ε . Present even for $f = \hat{f}$, cannot be predicted from X .
Possible sources:
 - Additional variables Y may depend on but which are not observed/measured.
 - Unmeasurable variation.
(E.g.: Adverse reaction may depend on manufacturing variations in drug or variations in patient's sensitivity over time.)
- Quantitative measure: **mean squared error** (MSE). For fixed X and \hat{f}

$$\mathbf{E} \left[(Y - \hat{Y})^2 \right] = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{reducible}} + \underbrace{\text{Var } \varepsilon}_{\text{irreducible}}$$

- **Note:** irreducible error always a (generally unknown) lower bound on prediction accuracy.

Learning Theory

Bias and variance



Inference seeks to determine how the individual predictors X_1, \dots, X_p affect the response Y . In particular, this involves more detailed knowledge about \hat{f} than simply considering it a black box.

Things to investigate:

Inference seeks to determine how the individual predictors X_1, \dots, X_p affect the response Y . In particular, this involves more detailed knowledge about \hat{f} than simply considering it a black box.

Things to investigate:

- *Identify those predictors with the strongest effect on Y .*
Can be a small subset of X_1, \dots, X_p .

Inference seeks to determine how the individual predictors X_1, \dots, X_p affect the response Y . In particular, this involves more detailed knowledge about \hat{f} than simply considering it a black box.

Things to investigate:

- *Identify those predictors with the strongest effect on Y .*
Can be a small subset of X_1, \dots, X_p .
- *Determine relationship between response and each predictor.*
Is it monotone increasing or decreasing with respect to an individual predictor? For more complex dependencies, such monotonicities can be affected by the values of the remaining predictors.

Inference seeks to determine how the individual predictors X_1, \dots, X_p affect the response Y . In particular, this involves more detailed knowledge about \hat{f} than simply considering it a black box.

Things to investigate:

- *Identify those predictors with the strongest effect on Y .*
Can be a small subset of X_1, \dots, X_p .
- *Determine relationship between response and each predictor.*
Is it monotone increasing or decreasing with respect to an individual predictor? For more complex dependencies, such monotonicities can be affected by the values of the remaining predictors.
- *Is a linear model sufficient?*
Historically, most estimation methods have produced a linear (affine) function \hat{f} . If the true dependence of Y on X is more complicated, a linear model may not be accurate enough.

Learning Theory

Prediction example: direct-marketing campaign



By Dvortygirl - Own work, CC BY-SA 3.0

- Company plans a direct-marketing campaign, wishes to identify individuals who would respond positively to a mailing.
- Response is $Y \in \{\text{positive, negative}\}$.
- Predictors X_j are demographic variables.
- Detailed relationship of response to demographic variables not of interest.
- A model which generates accurate predictions is all that is needed.

Learning Theory

Inference examples: advertising data set, purchase behavior

- In our first example (advertising across media channels TV, radio and newspaper), one may also be interested in answers to
 - Which media increase sales?
 - Of those, which has the strongest positive effect?
 - At what rate do sales increase when the TV budget is raised?

Learning Theory

Inference examples: advertising data set, purchase behavior

- In our first example (advertising across media channels TV, radio and newspaper), one may also be interested in answers to
 - Which media increase sales?
 - Of those, which has the strongest positive effect?
 - At what rate do sales increase when the TV budget is raised?
- Another example: model brand of a product chosen by a customer as a function of predictor variables **price, store location, discount levels, competitor pricing** etc.
Here detailed knowledge of how each variable affects outcome is of interest, e.g.
 - *What effect will changing the price of a product have on sales?*

Learning Theory

Inference examples: advertising data set, purchase behavior

- In our first example (advertising across media channels TV, radio and newspaper), one may also be interested in answers to
 - Which media increase sales?
 - Of those, which has the strongest positive effect?
 - At what rate do sales increase when the TV budget is raised?
- Another example: model brand of a product chosen by a customer as a function of predictor variables **price, store location, discount levels, competitor pricing** etc.
Here detailed knowledge of how each variable affects outcome is of interest, e.g.
 - *What effect will changing the price of a product have on sales?*

Learning Theory

Example: combination of prediction and inference

- There are also **mixed situations**, involving both prediction and inference: Consider value of a house depending on prediction values **size, crime rate, zoning, distance from a river/ocean, air quality, schools, income level of community, . . .**
 - *How much does an ocean view increase the value of a house?* (inference)
 - *Is this house over- or undervalued?* (prediction)
- **Note:** the two objectives may be competing. Linear models allow for easier inference, but may not be accurate enough for given prediction goal. More sophisticated (highly nonlinear) approaches may yield high prediction accuracy, but the models they produce are often difficult to interpret.

Learning Theory

Prediction techniques

Denote by

n : number of available data observations. (“**training data**”)

x_{ij} : value of j -th predictor in i -th observation

$$i = 1, \dots, n; \quad j = 1, \dots, p$$

y_i : value of response variable in i -th observation

Then training data consists of predictor-response pairs

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}, \quad \mathbf{x}_i = (x_{i1}, \dots, x_{ip}).$$

Goal is estimating function \hat{f} such that $Y \approx \hat{f}(X)$ for all observations (X, Y) .

Denote by

n : number of available data observations. (“**training data**”)

x_{ij} : value of j -th predictor in i -th observation

$$i = 1, \dots, n; \quad j = 1, \dots, p$$

y_i : value of response variable in i -th observation

Then training data consists of predictor-response pairs

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}, \quad \mathbf{x}_i = (x_{i1}, \dots, x_{ip}).$$

Goal is estimating function \hat{f} such that $Y \approx \hat{f}(X)$ for all observations (X, Y) .

Two basic approaches: **parametric** and **non-parametric**.

Two-step model-based approach

- 1 Assume specific functional form for f , popular example is the **linear model**

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \quad (2.2)$$

Estimation of function f now consists only in determining values of the $p + 1$ **parameters** $\beta_0, \beta_1, \dots, \beta_p$ (huge simplification).

- 2 **Train** or **fit** the chosen model to the data, i.e., choose parameters $\{\beta_j\}_{j=0}^p$ in order that (here for linear model (2.2))

$$f(X) \approx \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p.$$

Most common fitting technique: (ordinary) least squares, but many other techniques exist.

Problem of estimating f reduced to estimating a finite number of parameters.

Fundamental trade-off:

- Simplification comes at expense of strong restriction on type of dependence.
- For a bad choice, model cannot match the data well.
- More **flexible** models can better adapt to given data, but will generally involve more parameters to be estimated.
- Moreover, even if we are willing to fit our data extremely well with a flexible model, we may be adapting the model only to the fluctuations due to the random error contained in the data (“fitting the noise”).

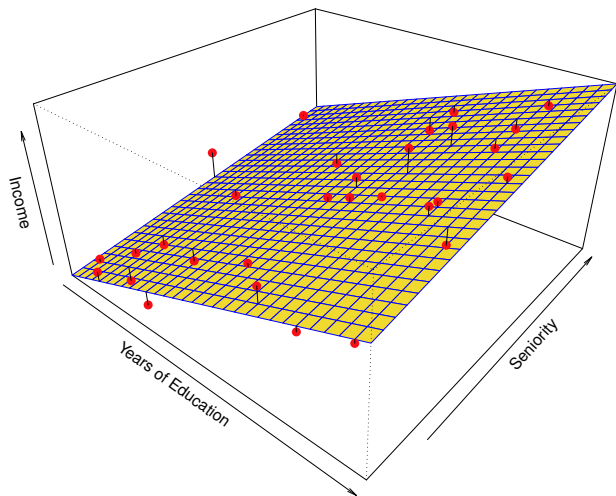
In this case, our model will not **generalize** well, i.e., have a low prediction value for new data.

This phenomenon is called **overfitting**.

Learning Theory

Parametric model example: income data

Linear model: $\text{income} \approx \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{seniority}$.



Learning Theory

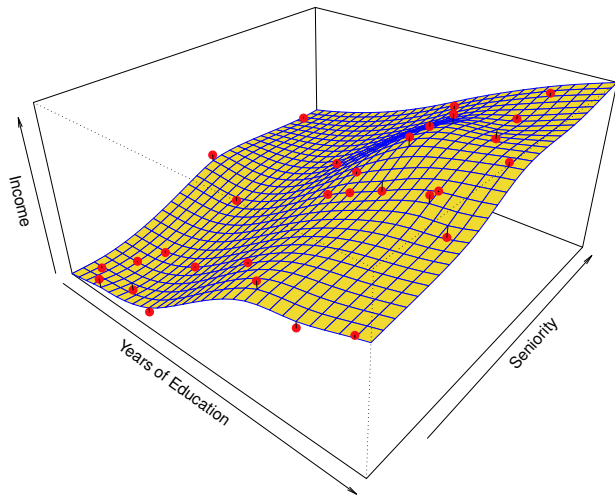
Non-parametric methods

- Non-parametric methods make no a priori assumptions on the functional form of f .
- Instead, they try to achieve as close an approximation to f as possible without being too rough or too oscillatory.
- + Bad a priori assumption can't limit approximation accuracy.
- Far more observations necessary than for parametric methods.

Learning Theory

Non-parametric model example: income data

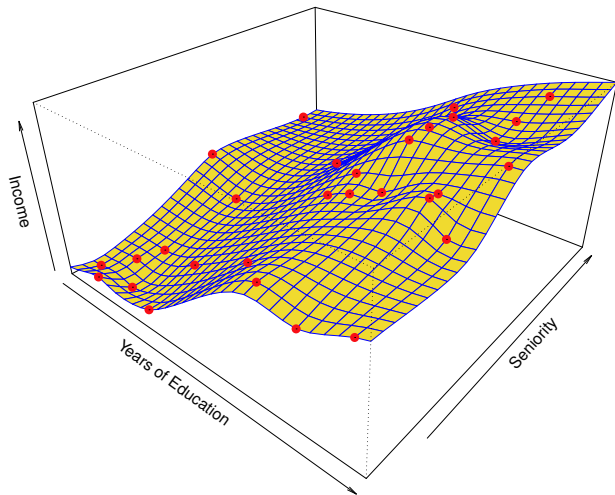
Smooth **thin-plate spline** model (later):



Learning Theory

Non-parametric model example: income data

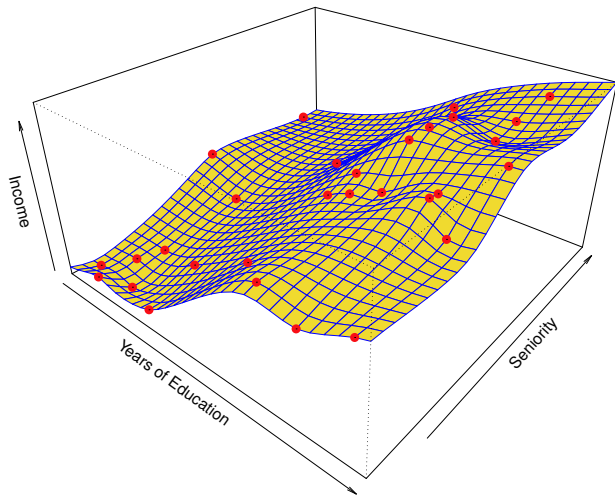
Rough **thin-plate spline** model:



Learning Theory

Non-parametric model example: income data

Rough **thin-plate spline** model:



Near perfect fit.
Are we overfitting?.

Learning Theory

Trade-off: prediction accuracy vs. model interpretability

- Less flexible/more restrictive models can only produce a small range of shapes for f . E.g.: linear regression always provides *linear* approximation to f .
- More flexible methods (e.g. thin-plate splines) offer larger variety of function shapes.

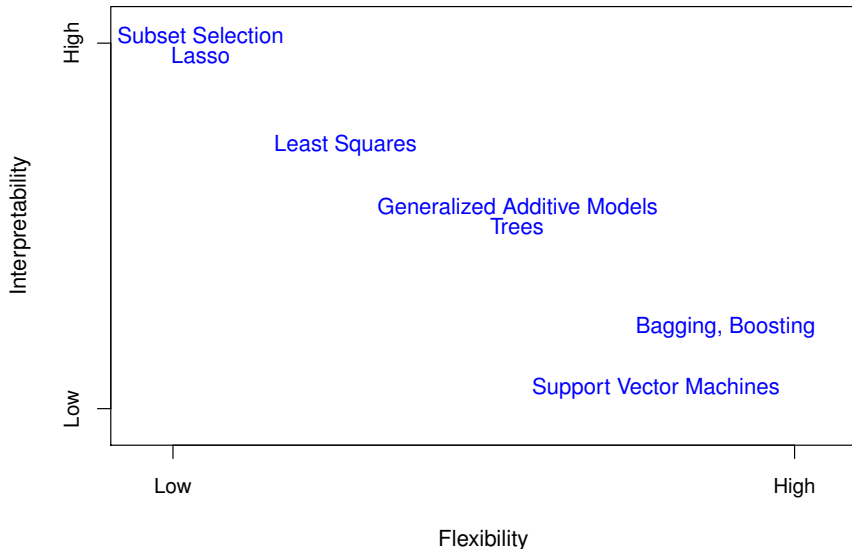
Learning Theory

Trade-off: prediction accuracy vs. model interpretability

- Less flexible/more restrictive models can only produce a small range of shapes for f . E.g.: linear regression always provides *linear* approximation to f .
- More flexible methods (e.g. thin-plate splines) offer larger variety of function shapes.
- Advantage of restrictive methods:
 - + For inference, restrictive models much more interpretable.
 - Linear least-squares easy to interpret.
 - **Lasso**: linear model, different way of selecting coefficients, sets some to zero.
More restrictive than least squares, but also more interpretable..
 - **Generalized additive models** (GAMs): extend model by certain nonlinear relationships.
More flexible, less easy to interpret.
 - **Bagging, boosting, support-vector machines**: fully nonlinear methods, very flexible, very difficult to interpret.

Learning Theory

Tradeoff: prediction accuracy vs. model interpretability



Learning Theory

Supervised vs. unsupervised learning

Up to now: observation pairs $(\mathbf{x}_i, y_i), i = 1, \dots, n$.

Learning Theory

Supervised vs. unsupervised learning

Up to now: observation pairs $(\mathbf{x}_i, y_i), i = 1, \dots, n$.

Seek model \hat{f} such that $Y \approx \hat{f}(X)$ for all observations. This is called **supervised learning**.

In **unsupervised learning** only predictor variables X are observed, but no associated responses Y .

Learning Theory

Supervised vs. unsupervised learning

Up to now: observation pairs $(\mathbf{x}_i, y_i), i = 1, \dots, n$.

Seek model \hat{f} such that $Y \approx \hat{f}(X)$ for all observations. This is called **supervised learning**.

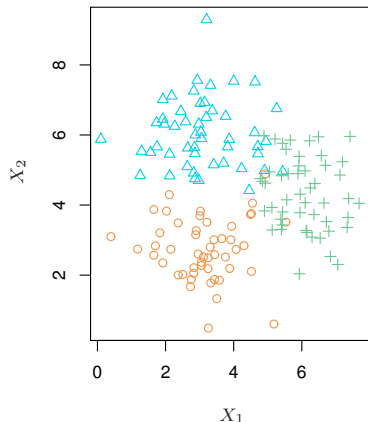
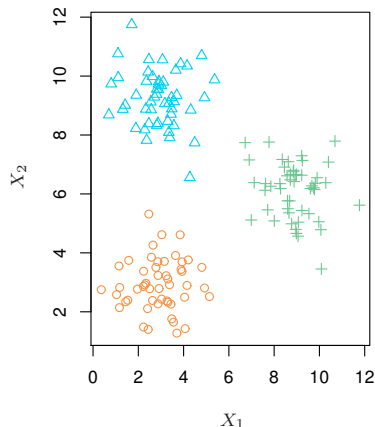
In **unsupervised learning** only predictor variables X are observed, but no associated responses Y .

- No fitting is possible (nothing to fit to); we are, in a sense, working blind.
- Less ambitious goal: discover relationships between observations, draw conclusions for predictor variables.
- **Cluster analysis** (clustering): statistical learning tool to ascertain whether observations $\{\mathbf{x}_i\}_{i=1}^n$ fall into (more or less) distinct groups.
- **Example:** market segment analysis, observe multiple characteristics of potential customers (zip code, family income, shopping habits). Possible groups: big spenders, low spenders.

In the absence of spending pattern data, clustering may reveal whether potential big spenders may be distinguished by the available data.

Learning Theory

Example: clustering



$n = 150$ observations of two variables X_1 and X_2 , each belonging to one of three groups (colored for better distinction). Left: well-separated clusters, easily identified. Right: some overlap between groups, more challenging. Some observations will likely be misclassified.

Note:

- Clustering more challenging in $p > 2$ dimensions, e.g. there are $p(p - 1)/2$ possible scatterplots to look at. Automated methods needed.
- **Semi-supervised learning**: Only $m < n$ observations come with responses. (Responses could be very expensive to obtain compared to the predictor observations).
Goal: incorporate both types of observations in an optimal way.

Learning Theory

Regression vs. classification problems

Another useful distinction is between **continuous** and **discrete** prediction and response variables.

- Continuous or **quantitative** variables – such as a person's height, age, income, the price of a house or stock – typically take on values in the real numbers.
- Discrete or **qualitative** variables – a person's gender, whether or not an event occurs, a cancer diagnosis – take on values in a in one of a finite number of different **classes** or categories.
- Problems with a quantitative response variable are typically referred to as **regression problems**, those with a qualitative response as **classification problems**.
- The distinction is not always sharp, e.g. logistic regression is used for (two-valued) qualitative responses (it estimates class probabilities).

② Learning Theory

2.1 What is Statistical Learning?

2.2 Assessing Model Accuracy

Assessing Model Accuracy

Mean squared error

Most common error metric in regression: (empirical) **mean squared error (MSE)**:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{f}(x_i) \right)^2. \quad (2.3)$$

Assessing Model Accuracy

Mean squared error

Most common error metric in regression: (empirical) **mean squared error (MSE)**:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{f}(x_i) \right)^2. \quad (2.3)$$

- When applied to training data: **training MSE**.
- More interesting (particularly for prediction): **test MSE** resulting from data not used to train (fit) the model \hat{f} .

Assessing Model Accuracy

Mean squared error

Most common error metric in regression: (empirical) **mean squared error (MSE)**:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{f}(x_i) \right)^2. \quad (2.3)$$

- When applied to training data: **training MSE**.
- More interesting (particularly for prediction): **test MSE** resulting from data not used to train (fit) the model \hat{f} .
- If a **test data** set is available in addition to the training data, different learning (fitting) methods can be compared with respect to their test MSE values.

Assessing Model Accuracy

Mean squared error

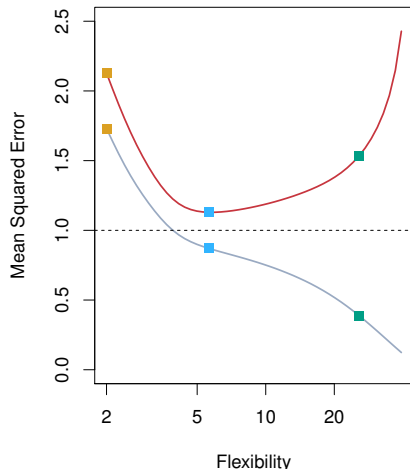
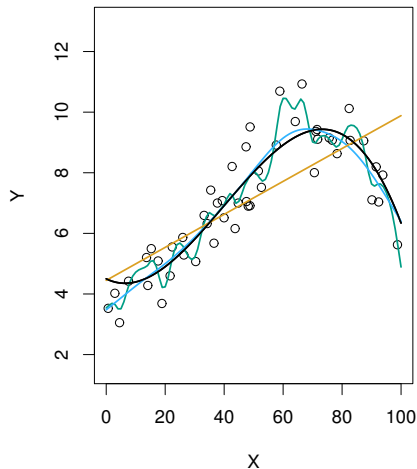
Most common error metric in regression: (empirical) **mean squared error (MSE)**:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{f}(x_i) \right)^2. \quad (2.3)$$

- When applied to training data: **training MSE**.
- More interesting (particularly for prediction): **test MSE** resulting from data not used to train (fit) the model \hat{f} .
- If a **test data** set is available in addition to the training data, different learning (fitting) methods can be compared with respect to their test MSE values.
- In the absence of a test data set, choosing a learning method based solely on the training MSE can be deceptive.

Assessing Model Accuracy

Example: smoothing spline models



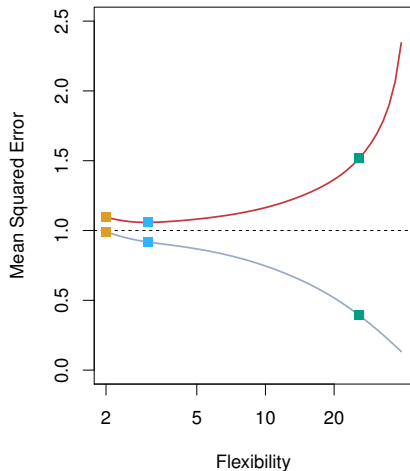
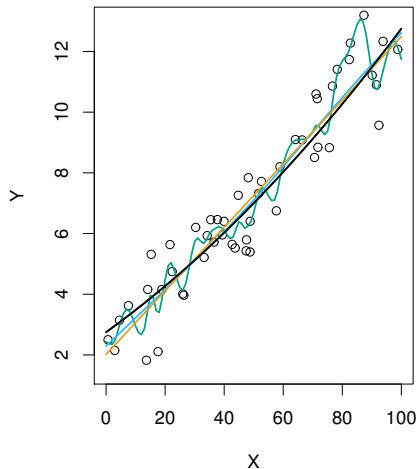
Left: Observations from model (2.1), true f in black, estimates in orange, blue, green.

Right: Average MSE for training data (gray), test data (red) vs. flexibility parameter.

Horizontal dashed line denotes $\text{Var } \varepsilon$.

Assessing Model Accuracy

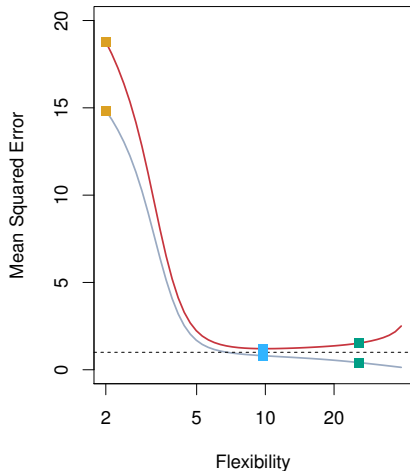
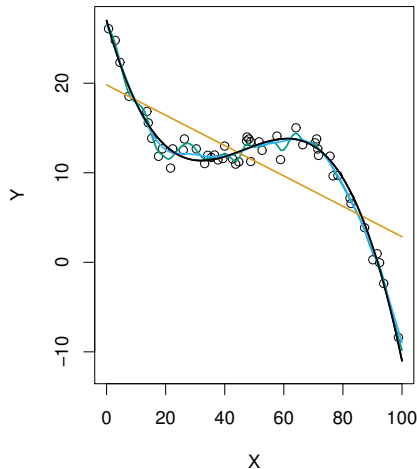
Example: smoothing spline models



Same plots as on previous figure, but with a true model that is nearly linear. Initial estimate (with few degrees of freedom) already quite accurate.

Assessing Model Accuracy

Example: smoothing spline models



Another such plot, now the true model is highly nonlinear. Maximal accuracy for training and test data not attained until model contains many degrees of freedom.

Assessing Model Accuracy

Example: smoothing spline models

Recap:

- Monotone decrease of training MSE as model becomes more flexible (more degrees of freedom (DoF) and can more flexibly follow data variation.
- Typically test MSE curve U-shaped, rises again once overfitting sets in.
- This is a fundamental property of statistical learning, regardless of data set and regardless of statistical technique being used.
- Interpretation: in overfitting, estimate is finding patterns (signal variation) where there are none.
James et al.: “When we overfit the training data, the test MSE will be very large because the supposed patterns that the method found in the training data simply don’t exist in the test data.”
- Overfitting: less flexible model would have yielded smaller test MSE.

Note: Many estimation methods based on **minimizing** MSE by adjusting the DoF in the method, hence training MSE almost always less than test MSE.

Assessing Model Accuracy

Apophenia

The tendency to misclassify random events as systematic or, more generally, to see patterns where there are none, is common to human experience and known as **Apophenia**.

- It is believed to be an advantage in the process of natural selection.
- It encourages conspiracy theories.
- It is used to explain the **gambler's fallacy** in probability theory.
- In his bestselling book **Thinking Fast and Slow**, the famous behavioral economist Kahneman calls this phenomenon the “law of small numbers”.



Jesus and Mary in an orange.

Source: anorak.co.uk

Assessing Model Accuracy

Trade-off: bias vs. variance

We now take into account the sampling variation of the data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ used to construct the approximation $\hat{f} \approx f$.

- We would like to determine the expectation $\mathbf{E} \left[(y_0 - \hat{f}(\mathbf{x}_0))^2 \right]$ for a new pair of observations (\mathbf{x}_0, y_0) which was not used in the construction of the estimate \hat{f} .
- Note that, for a fixed value of \mathbf{x}_0 , the estimate $\hat{f}(\mathbf{x}_0)$ will fluctuate depending on the specific selection of (randomly) sampled observations $\{(\mathbf{x}_i, y_i)_{i=1}^n\}$. This statistical variation is called its **sampling distribution**.
- We denote the **bias** of the estimate $\hat{f}(\mathbf{x}_0)$ by $\text{Bias } \hat{f}(\mathbf{x}_0) := f(\mathbf{x}_0) - \mathbf{E} \left[\hat{f}(\mathbf{x}_0) \right]$ and by $\mathbf{Var } \hat{f}(\mathbf{x}_0)$ its **variance**.

Theorem 2.1

In the notation introduced above, assuming independence of noise and sampling distributions, we have

$$\mathbf{E}_{X, \epsilon} \left[(y_0 - \hat{f}(\mathbf{x}_0))^2 \right] = \mathbf{Var } \hat{f}(\mathbf{x}_0) + [\text{Bias } \hat{f}(\mathbf{x}_0)]^2 + \mathbf{Var } \epsilon.$$

Assessing Model Accuracy

Trade-off: bias vs variance

Expected test MSE for new value x_0 of test data:

$$\mathbf{E} \left[(y_0 - \hat{f}(x_0))^2 \right] = \mathbf{Var} \hat{f}(x_0) + [\mathbf{Bias} \hat{f}(x_0)]^2 + \mathbf{Var} \varepsilon. \quad (2.4)$$

Assessing Model Accuracy

Trade-off: bias vs variance

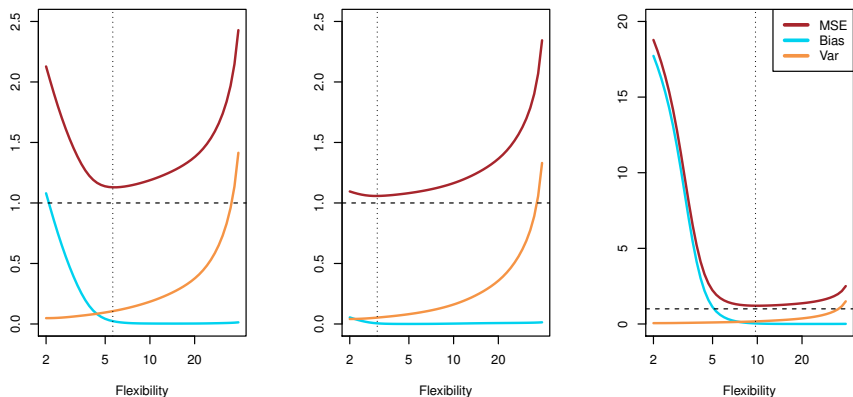
Expected test MSE for new value x_0 of test data:

$$\mathbf{E} \left[(y_0 - \hat{f}(x_0))^2 \right] = \mathbf{Var} \hat{f}(x_0) + [\mathbf{Bias} \hat{f}(x_0)]^2 + \mathbf{Var} \varepsilon. \quad (2.4)$$

- $\mathbf{E} \left[(y_0 - \hat{f}(x_0))^2 \right]$ is the expected test MSE with respect to the distribution of the predictor variable X , i.e., the average test MSE we would obtain by repeatedly estimating f using a large number of training sets, and testing each at x_0 .
- (2.4) implies that a good statistical learning method needs to achieve both low bias and low variance.
- **Variance**: amount by which \hat{f} would change if estimated using a different training data set. A method with high variance is sensitive to small changes in the data set.
- **Bias**: error introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model (systematic error). More flexible methods have lower bias.

Assessing Model Accuracy

Trade-off: bias vs variance



Bias-variance decomposition for last 3 examples. Horizontal dashed line: $\text{Var } \epsilon$. Flexibility level achieving minimal test MSE varies due to different rates of change in bias and variance.

Assessing Model Accuracy

Bias vs. variance for classification

For classification, i.e., in case the response variable Y is qualitative (discrete), replace MSE with training **error rate**:

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{y_i \neq \hat{y}_i\}} \quad (2.5)$$

expressing the fraction of incorrect classifications, where

\hat{y}_i : predicted class label for i -th observation using \hat{f} ,

$$\mathbf{1}_{\{y_i \neq \hat{y}_i\}} = \begin{cases} 1 & y_i \neq \hat{y}_i, \\ 0 & y_i = \hat{y}_i, \end{cases} \quad (\text{indicator variable}).$$

Assessing Model Accuracy

Bias vs. variance for classification

For classification, i.e., in case the response variable Y is qualitative (discrete), replace MSE with training **error rate**:

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{y_i \neq \hat{y}_i\}} \quad (2.5)$$

expressing the fraction of incorrect classifications, where

$$\hat{y}_i : \text{predicted class label for } i\text{-th observation using } \hat{f},$$
$$\mathbf{1}_{\{y_i \neq \hat{y}_i\}} = \begin{cases} 1 & y_i \neq \hat{y}_i, \\ 0 & y_i = \hat{y}_i, \end{cases} \quad (\text{indicator variable}).$$

As in regression setting, of more interest than training error rate (2.5) is **test error rate**, which averages classification errors $\mathbf{1}_{\{y_0 \neq \hat{y}_0\}}$ over a test set of observations (\mathbf{x}_0, y_0) with classification prediction \hat{y}_0 for predictor variable \mathbf{x}_0 .

Assessing Model Accuracy

Classification: Bayes classifier

Expectation of test rate error minimized by the **Bayes classifier**: assign to test observation with predictor vector \mathbf{x}_0 the class j for which conditional probability

$$\mathbf{P}(Y = j|X = \mathbf{x}_0)$$

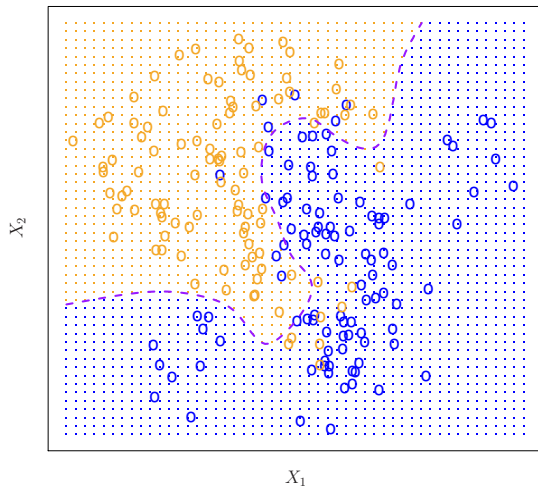
is maximized over all j .

Special case: two-class problem, i.e., $Y \in \{1, 2\}$; predict

$$\hat{y}_0 = \begin{cases} 1 & \text{if } \mathbf{P}\{Y = 1|X = \mathbf{x}_0\} > 0.5 \\ 2 & \text{otherwise.} \end{cases}$$

Assessing Model Accuracy

Example: Bayes classifier, 2 classes



Predictors: X_1, X_2

Response: $Y \in \{\text{orange, blue}\}$

Observations: circles

(simulated data)

Orange shading:

$$\mathbf{P}\{Y = \text{orange}|X\} > 0.5$$

Blue shading

$$\mathbf{P}\{Y = \text{orange}|X\} < 0.5$$

Dashed line: **Bayes decision boundary** defined by

$$\mathbf{P}\{Y = \text{orange}|X\} = 0.5$$

Assessing Model Accuracy

Bayes error rate

- Bayes classifier produces lowest possible test error rate, the **Bayes error rate**.
- By definition, error rate at $X = \mathbf{x}_0$ is

$$1 - \max_j \mathbf{P}\{Y = j|X\}.$$

- Overall Bayes error rate:

$$\mathbf{E} \left[1 - \max_j \mathbf{P}\{Y = j|X\} \right] = 1 - \mathbf{E} \left[\max_j \mathbf{P}\{Y = j|X\} \right],$$

expectation with respect to distribution of X .

- Previous example: Bayes error rate is 0.1304, positive since some observations on wrong side of decision boundary, hence $\max_j \mathbf{P}\{Y = j|X = \mathbf{x}_0\} < 1$ for some \mathbf{x}_0 .
- Bayes error rate analogous to irreducible error.

Assessing Model Accuracy

K -nearest neighbors

- Bayes classifier not realizable, since based on unknown conditional distribution, represents unattainable reference value.
- **K-nearest neighbors (KNN)** classifier: classify based on *estimate* of conditional distribution.
- Given $X = \mathbf{x}_0$, denote by \mathcal{N}_0 the $K \in \mathbb{N}$ training set points (geometrically) closest to \mathbf{x}_0 and estimate

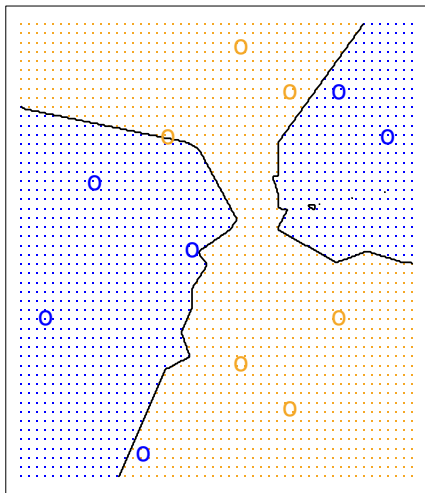
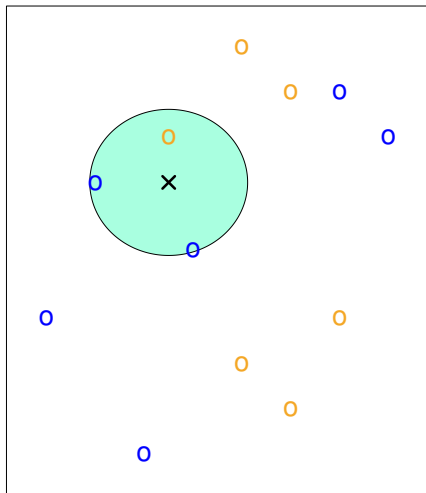
$$\mathbf{P}\{Y = j | X = \mathbf{x}_0\} \approx \frac{1}{K} \sum_{i \in \mathcal{N}_0} \mathbf{1}_{\{y_i = j\}},$$

i.e., by fraction of K nearest neighbors belonging to class j .

- Now proceed as in Bayes estimate with this approximate conditional distribution.

Assessing Model Accuracy

Example: KNN, $K=3$

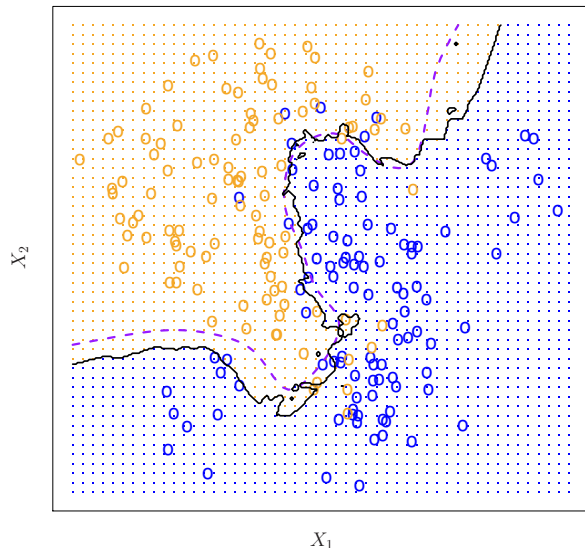


Left: x : x_0 ; green circle: \mathcal{N}_0

Right: KNN applied to all shaded points, resulting decision boundary

Assessing Model Accuracy

Example: KNN, $K=10$



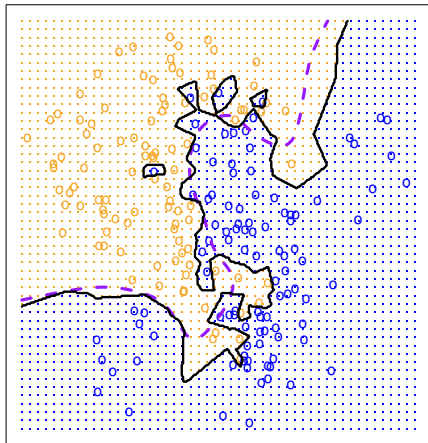
KNN decision boundary for $K = 10$ applied to data set from Slide 66.

Test error rates:
Bayes: 0.1304
KNN: 0.1363.

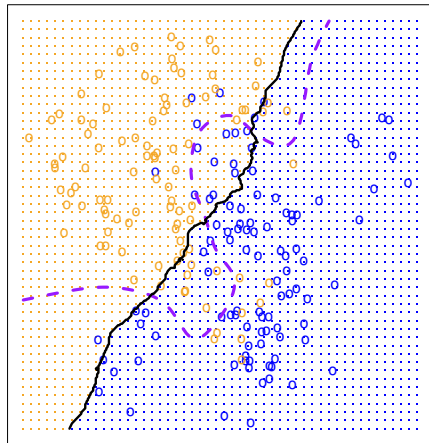
Assessing Model Accuracy

Example: KNN, $K=1,100$

KNN applied to data set from Slide 66:



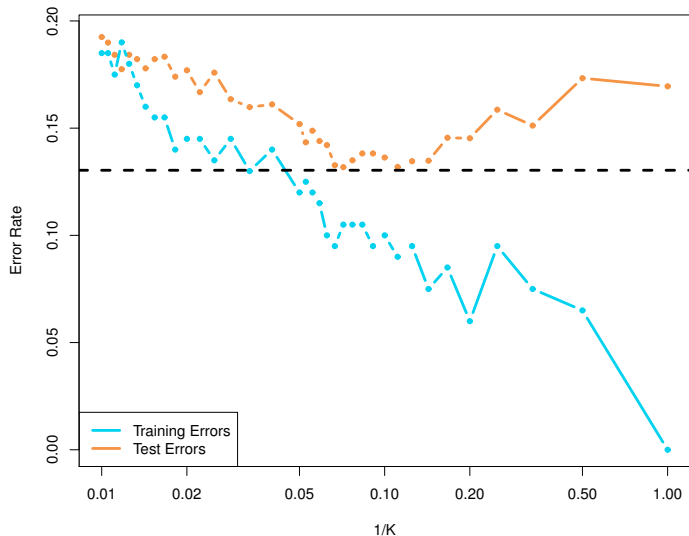
Left: $K = 1$, high variance; test error rate 0.1695.



Right: $K = 100$, high bias; test error rate 0.1925.

Assessing Model Accuracy

Example: KNN error rates against K



Training and test errors of KNN classification for same data plotted against $1/K$.