

ETC3250: Introduction

Semester 1, 2019

Professor Di Cook

Econometrics and Business Statistics

Monash University

Week 1 (b)

Outline

■ Introduction

● Learning from
data

■ Better understand or make predictions about a certain phenomenon under study

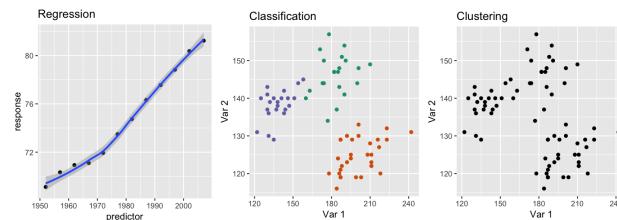
■ Construct a model of that phenomenon by finding relations between several variables

■ If phenomenon is complex or depends on a large number of variables, an analytical solution might not be available

■ However, we can collect data and learn a model that approximates the true underlying phenomenon

Outline

- Introduction
- Learning from data



$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N, \text{ where } x_i = (x_{i1}, \dots, x_{ip})^T$$

Statistical learning provides a framework for constructing models from the data.

3 / 38

Outline

- Introduction
- Learning from data
- Different learning problems

■■ Supervised learning, y_i available for all x_i

● Regression (or prediction)

● Classification

■■ Unsupervised learning, y_i unavailable for all x_i

■■ Semi-supervised learning, y_i available only for few x_i

■■ Other types of learning: reinforcement learning, online learning, active learning, etc.

Being able to identify which is the type of learning problem you have is important in practice

4 / 38

Outline

- Introduction
- Supervised learning

$$\mathcal{D} = \{(y_i, x_i)\}_{i=1}^N$$

where $(y_i, x_i) \sim P(Y, X) = P(X) \underbrace{P(Y|X)}_{\text{where } P(Y|X) \text{ means that these arise from some probability distribution. "}" \text{ means distributed as, arise from. Typically, we only are interested in } P(Y|X), \text{ the distribution of } Y \text{ conditional on } X.}$

5 / 38

Outline

- Introduction
- Supervised learning

■ $Y = (Y_1, \dots, Y_q)$: response (output) (could be multivariate, $q = 1$ for us)

■ $X = (X_1, \dots, X_p)$: set of p predictors (input)

We seek a function $h(X)$ for predicting Y given values of the input X . This function is computed using \mathcal{D} .

6 / 38

Outline

- Introduction
- Supervised learning

$\mathcal{D} = \{(y_i, x_i)\}_{i=1}^N$ where $(y_i, x_i) \sim P(Y, X)$

We are interested in minimizing the expected **out-of-sample prediction error**:

$$\text{Err}_{\text{out}}(h) = E[L(Y, h(X))]$$

where $L(y, \hat{y})$ is a non-negative real-valued **loss function**, such as $L(y, \hat{y}) = (y - \hat{y})^2$ and $L(y, \hat{y}) = I(y \neq \hat{y})$.

The goal is that the predictions from the model are accurate for future samples.

7 / 38

Outline

- Introduction
- Supervised learning
- Regression

We often assume that our data arose from a statistical model

$Y = f(X) + \varepsilon$, where f is the true unknown function, ε is the random error term with $E[\varepsilon] = 0$ and is independent of X .

■ The additive error model is a useful approximation to the truth

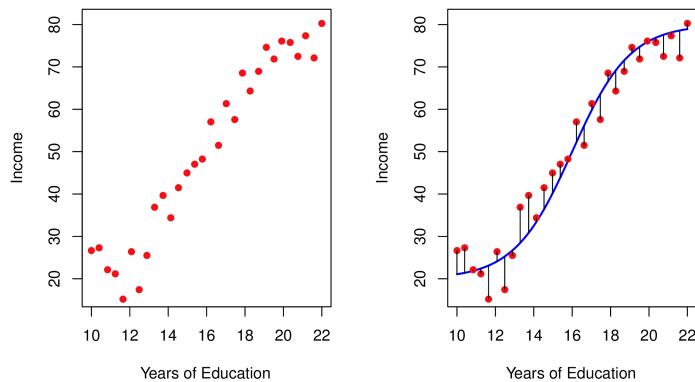
$$f(x) = E[Y|X = x]$$

■ Not a deterministic relationship: $Y \neq f(X)$

8 / 38

Outline

- [Introduction]
- [Supervised learning]
- Regression



Blue curve is $f(x)$, the true functional relationship.

(Chapter2/2.2.pdf)

9 / 38

Outline

- [Introduction]
- [Supervised learning]
- Regression

10 / 38

Outline

- Introduction
- Supervised learning
 - Regression
 - Why?

■ Prediction:

● $\hat{y}_* = \hat{f}(x_*)$ for a new observation x_*

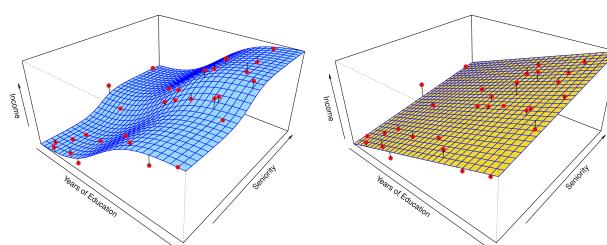
■ Inference (or explanation):

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?

11 / 38

Outline

- Introduction
- Supervised learning
 - Regression
 - Why?
 - Estimation



Linear model: $\hat{f}(\text{education}, \text{seniority}) = \hat{\beta}_0 + \hat{\beta}_1 \times \text{education} + \hat{\beta}_2 \times \text{seniority}$

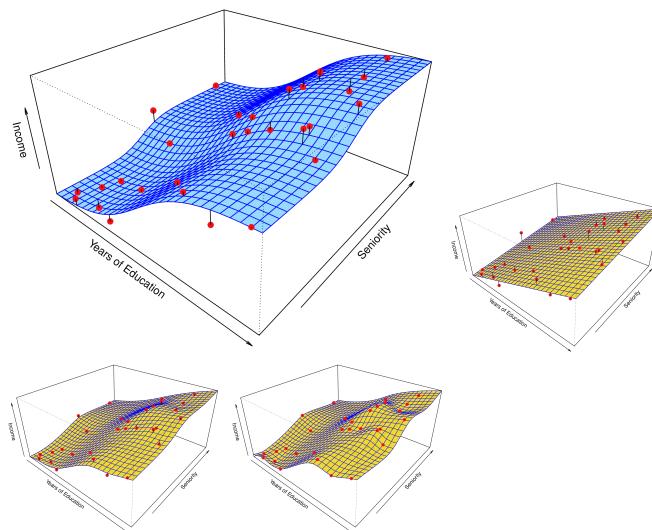
Why would we ever choose to use a **more restrictive method** instead of a **very flexible approach**?

(Chapter2/2.3.pdf, 2.4.pdf)

12 / 38

Outline

- 📊 Introduction
- 📊 Supervised learning
 - ⌚ Regression
 - ⌚ Why?
 - ⌚ Estimation



(Chapter2/2.3.pdf, 2.4.pdf, 2.5.pdf, 2.6.pdf)

13 / 38

Outline

- 📊 Introduction
- 📊 Supervised learning
 - ⌚ Regression
 - ⌚ Why?
 - ⌚ Estimation
 - ⌚ Methods

📊 Parametric methods

- ⊕ Assumption about the form of f , e.g. linear
- 😊 The problem of estimating f reduces to estimating a set of parameters
- 😊 Usually a good starting point for many learning problems
- 😢 Poor performance if linearity assumption is wrong

📊 Non-parametric methods

- 😊 No explicit assumptions about the form of f , e.g. nearest neighbours: $\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$
- 😊 High flexibility: it can potentially fit a range of shapes
- 😢 A large number of observations is required to estimate f with good accuracy

14 / 38

Outline

- Introduction
- Supervised learning
- Assessing model accuracy
- Regression

Suppose we have a regression model $y = f(x) + \varepsilon$. Estimate \hat{f} from some training data, $Tr = \{x_i, y_i\}_{i=1}^n$.

One common measure of accuracy is:

Training Mean Squared Error

$$MSE_{Tr} = \text{Ave}_{i \in Tr} [y_i - \hat{f}(x_i)]^2 = \frac{1}{n} \sum_{i=1}^n [(y_i - \hat{f}(x_i))^2]$$

15 / 38

Outline

- Introduction
- Supervised learning
- Assessing model accuracy
- Regression

Suppose we have a regression model $y = f(x) + \varepsilon$. Estimate \hat{f} from some training data, $Tr = \{x_i, y_i\}_{i=1}^n$.

One common measure of accuracy is: Training Mean Squared Error

$$MSE_{Tr} = \text{Ave}_{i \in Tr} [y_i - \hat{f}(x_i)]^2 = \frac{1}{n} \sum_{i=1}^n [(y_i - \hat{f}(x_i))^2]$$

Measure real accuracy using test data $Te = \{x_j, y_j\}_{j=1}^m$, Test Mean Squared Error

$$MSE_{Te} = \text{Ave}_{j \in Te} [y_j - \hat{f}(x_j)]^2 = \frac{1}{m} \sum_{j=1}^m [(y_j - \hat{f}(x_j))^2]$$

16 / 38

Outline

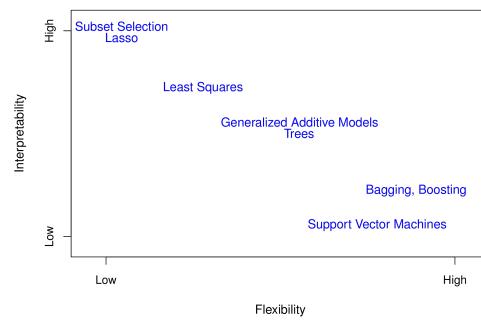
- Introduction
- Supervised learning
- Assessing model accuracy
 - Regression
 - Training vs Test MSEs

■ In general, the more **flexible** a method is, the **lower** its **training MSE** will be. i.e. it will “fit” the training data very well.
■ However, the **test MSE** may be **higher** for a more **flexible** method than for a simple approach like linear regression.
■ Flexibility also makes interpretation more difficult. There is a trade-off between **flexibility** and **model interpretability**.

17 / 38

Outline

- Introduction
- Supervised learning
 - Regression
 - Why?
 - Estimation
 - Methods
 - Interpretability vs flexibility

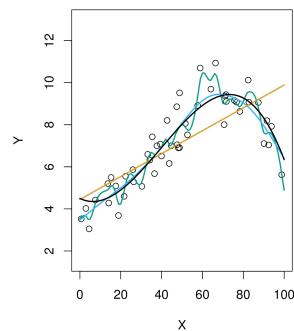


Simplistic overview of methods on the flexibility vs interpretability scale. Interpretability is when it is clear how the explanatory variable is related to the response, e.g. linear

model. Poor interpretability is often called a "black box" 18 / 38

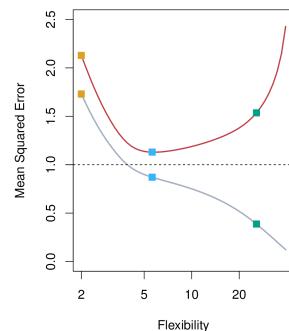
Outline

- [Introduction]
- [Supervised learning]
- [Assessing model accuracy]
- Regression
 - Training vs Test MSEs
 - Example



true curve
[Chapter2/2.9.pdf]
linear regression

Smoothing splines

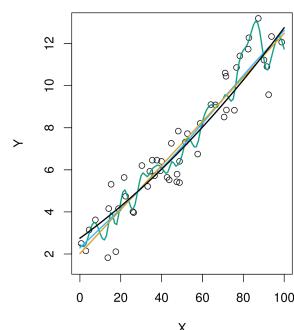


Training MSE
Test MSE

Dashed: Minimum test MSE

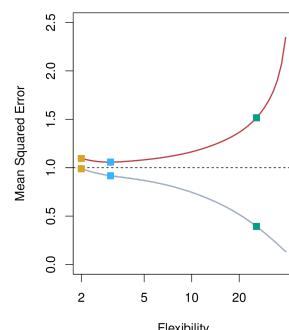
Outline

- [Introduction]
- [Supervised learning]
- [Assessing model accuracy]
- Regression
 - Training vs Test MSEs
 - Example



true curve
[Chapter2/2.9.pdf]
linear regression

Smoothing splines

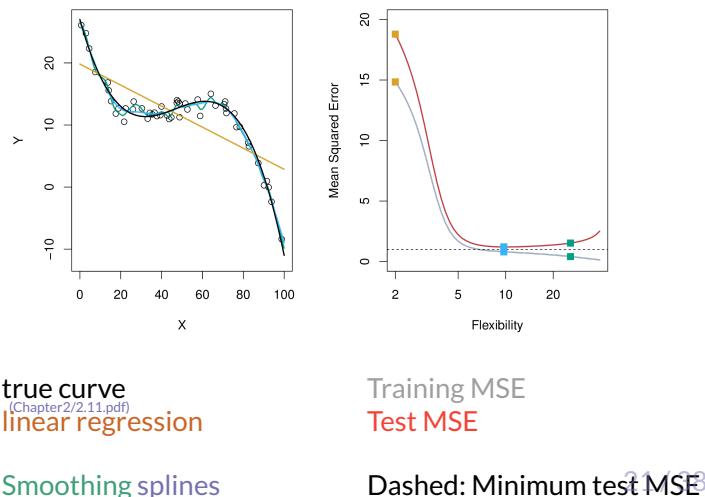


Training MSE
Test MSE

Dashed: Minimum test MSE

Outline

- [Introduction]
- [Supervised learning]
- [Assessing model accuracy]
 - Regression
 - Training vs Test MSEs
 - Example



true curve
(Chapter2/2.11.pdf)
linear regression

Smoothing splines

Training MSE
Test MSE

Dashed: Minimum test MSE

Outline

- [Introduction]
- [Supervised learning]
- [Assessing model accuracy]
 - Regression
 - Training vs Test MSEs
 - Example
 - Bias-variance tradeoff

There are two competing forces that govern the choice of learning method: **bias** and **variance**.

Bias is the error that is introduced by modeling a complicated problem by a simpler problem.

- [For example, linear regression assumes a linear relationship when few real relationships are exactly linear.]
- [In general, the **more flexible** a method is, the **less bias** it will have.]

This site has a lovely explanation, if you don't like mine.

Outline

- [Introduction]
- [Supervised learning]
- [Assessing model accuracy]
 - Regression
 - Training vs Test MSEs
 - Example
 - Bias-variance tradeoff

There are two competing forces that govern the choice of learning method: **bias** and **variance**.

Variance refers to how much your estimate would change if you had different training data.

- [In general, the more flexible a method is, the more variance it has.]
- [The size of the training data has an impact on the variance.]

22 / 38

MSE decomposition

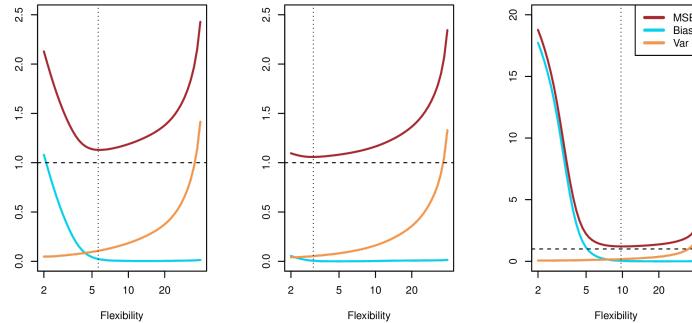
If $Y = f(x) + \varepsilon$ and $f(x) = E[Y | X = x]$, then the expected test MSE for a new y at x_0 will be equal to

$$E[(Y - \hat{f}(x_0))^2] = [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\hat{f}(x_0)) + \text{Var}(\varepsilon)$$

Test MSE = Bias² + Variance + Irreducible variance

- [The expectation averages over the variability of y as well as the variability in the training data.]
- [As the flexibility of \hat{f} increases, its variance increases and its bias decreases.]
- [Choosing the flexibility based on average test MSE amounts to a **bias-variance trade-off**]

23 / 38



squared bias, variance, $\text{Var}(\varepsilon)$ (dashed line), and test MSE for the three data sets shown earlier. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

(Chapter2/2.12.pdf)

24 / 38

Outline

The optimal MSE is obtained when

- ─ Introduction
- ─ Supervised learning
- ─ Assessing model accuracy
- ─ Regression
 - ─ Training vs Test MSEs
 - ─ Example
 - ─ Bias-variance tradeoff
 - ─ Optimal prediction

$\hat{f} = f = \mathbb{E}[Y | X = x]$.

Then bias=variance=0 and

MSE = irreducible variance

This is called the **oracle predictor** because it is not achievable in practice.

25 / 38

Outline

- ─ Introduction
- ─ Supervised learning
- ─ Assessing model accuracy
- ─ Classification

Here the response variable y is **qualitative**.

- ─ e.g., email is one of $\mathcal{C} = \{\text{spam, ham}\}$
- ─ e.g., voters are one of $\mathcal{C} = \{\text{Liberal, Labor, Green, National, Other}\}$

Our goals are:

1. Build a classifier $C(x)$ that assigns a class label from $\mathcal{C} = \{c_1, \dots, c_K\}$ to a future unlabeled observation x .
2. Such a classifier will divide the input space into regions \mathcal{R}_k called decision regions, one for each class, such that all points in \mathcal{R}_k are assigned to class c_k
3. Assess the uncertainty in each classification (i.e., the probability of misclassification).
4. Understand the roles of the different predictors among

$$X = (X_1, X_2, \dots, X_p)_*$$

26 / 38

Outline

- ─ Introduction
- ─ Supervised learning
- ─ Assessing model accuracy
- ─ Classification

Recall that we want to minimize the expected prediction error

$$E_{(Y,X)}[L(Y, C(X))]$$

where $L(y, \hat{y})$ is a non-negative real-valued **loss function**.

In classification, the output y is a **categorical variable**, and our loss function can be represented by a $K \times K$ matrix L , where $K = \text{card}(\mathcal{C})$. $L(k, l)$ is the cost of classifying c_k as c_l . With zero-one loss, i.e. $L(y, \hat{y}) = I(y \neq \hat{y})$ the cost is equal for mistakes between any class.

27 / 38

Outline

- Introduction
- Supervised learning
- Assessing model accuracy
 - Classification
 - Optimal classifier
 - Error

Compute \hat{C} from some training data, $T_r = \{x_i, y_i\}_1^n$.

In place of MSE, we now use the error rate (fraction of misclassifications).

Training Error Rate

$$\text{Error rate}_{Tr} = \frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{C}(x_i))$$

Measure real accuracy using test data $T_e = \{x_j, y_j\}_1^m$

Test Error Rate

$$\text{Error rate}_{Te} = \frac{1}{m} \sum_{j=1}^m I(y_j \neq \hat{C}(x_j))$$

28 / 38

Outline

- Introduction
- Supervised learning
- Assessing model accuracy
 - Classification
 - Error
 - Bayes classifier

Let $C = \{c_1, \dots, c_K\}$, and let

$$p_k(x) = P(Y = C_k \mid X = x), \quad k = 1, 2, \dots, K.$$

These are the conditional class probabilities at x .

Then the Bayes classifier at x is

$$C(x) = C_j \quad \text{if } p_j(x) = \max\{p_1(x), p_2(x), \dots, p_K(x)\}$$

This gives the minimum average test error rate.

29 / 38

Outline

$$1 - E(\max_j P(Y = C_j | X))$$

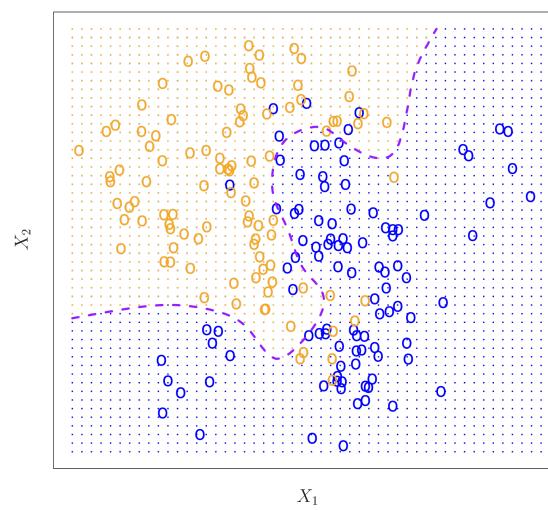
- Introduction
- Supervised learning
- Assessing model accuracy
 - Classification
 - Error
 - Bayes classifier

- The Bayes error rate is the lowest possible error rate that could be achieved if we knew exactly the true probability distribution of the data.
- It is analogous to the irreducible error in regression.
- On test data, no classifier can get lower error rates than the Bayes error rate.
- In reality, the Bayes error rate is not known exactly.

30 / 38

Outline

- Introduction
- Supervised learning
- Assessing model accuracy
 - Classification
 - Error
 - Bayes classifier



Outline

- Introduction
- Supervised learning
- Assessing model accuracy
 - Classification
 - Error
 - Bayes classifier
 - kNN

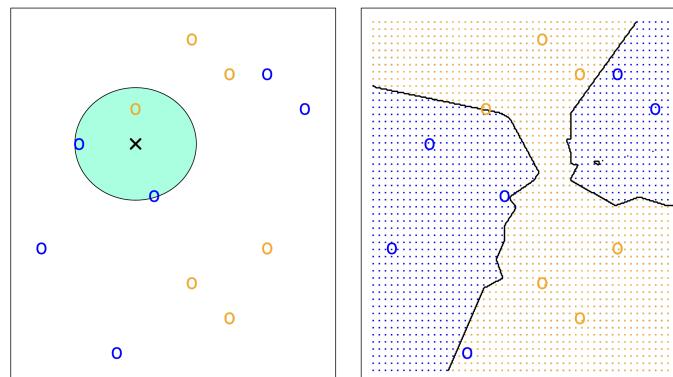
One of the simplest classifiers. Given a test observation x_0 ,

- Find the K nearest points to x_0 in the training data: \mathcal{N}_0 .
- Estimate conditional probabilities
 $= P(Y = C_j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = C_j)$.
- Classify x_0 to class with largest probability.

32 / 38

Outline

- Introduction
- Supervised learning
- Assessing model accuracy
- Classification
 - Error
 - Bayes classifier
 - kNN

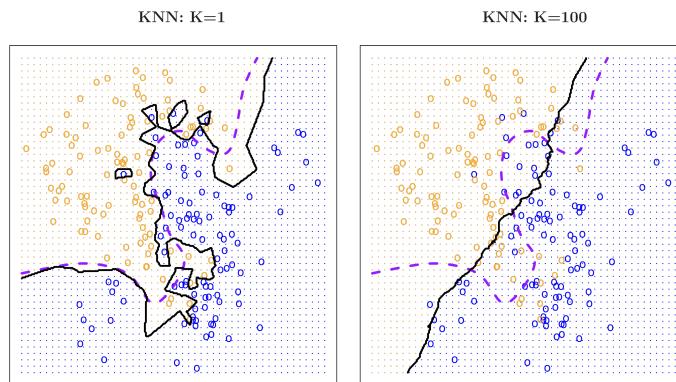


(Chapter2/2.14.pdf)

33 / 38

Outline

- [Introduction]
- [Supervised learning]
- [Assessing model accuracy]
 - Classification
 - Error
 - Bayes classifier
 - kNN



(Chapter2/2.16.pdf)

34 / 38

Outline

- [Introduction]
- [Supervised learning]
- [Assessing model accuracy]
 - Classification
 - Error
 - Bayes classifier
 - kNN

KNN: K=10

35 / 38

Outline

- 📊 Introduction
- 📊 Supervised learning
- 📊 Assessing model accuracy
 - Classification
 - Error
 - Bayes classifier
 - kNN

0.20

36 / 38

A fundamental picture

37 / 38



Made by a human with a computer

Slides at <https://monba.dicook.org>.

Code and data at
https://github.com/dicook/Business_Analytics.

Created using R Markdown with flair by [xaringan](#), and
[kunoichi](#) (female ninja) style.



This work is licensed under a Creative Commons Attribution-
ShareAlike 4.0 International License.

38 / 38