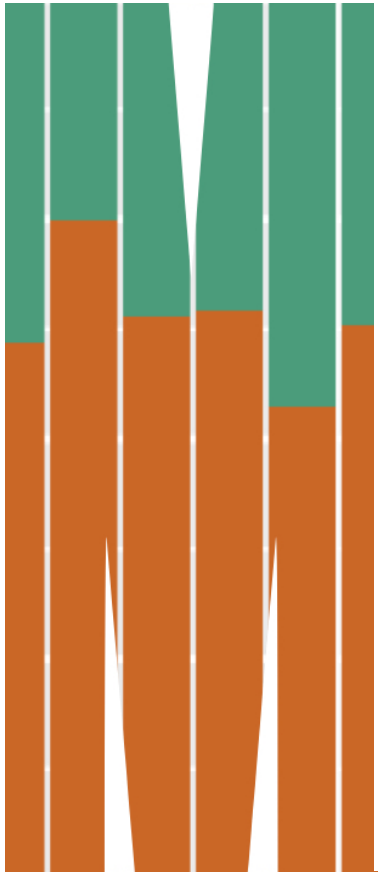


ETC3250/5250: Supervised Learning

Semester 1, 2020

Professor Di Cook

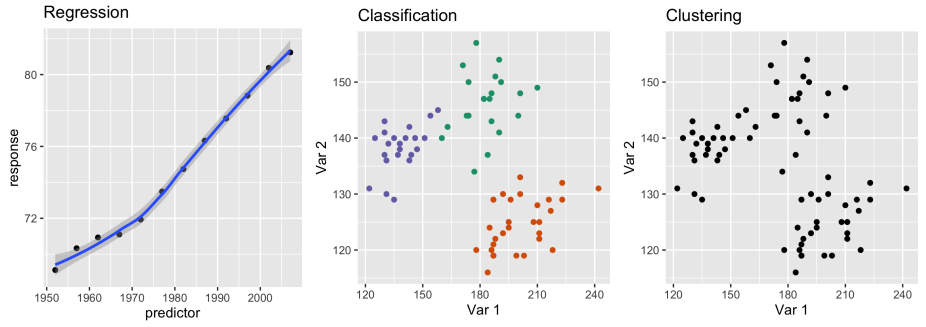
Econometrics and Business Statistics
Monash University

Week 1 (b)

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

Learning from Data



Statistical learning provides a framework for constructing models from the 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

Supervised learning

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

where $(y_i, x_i) \sim P(Y, X) = P(X)\underbrace{P(Y|X)}$ where $P(Y, X)$ means

that these arise from some probability distribution. "$\sim$" means distributed as, arise from. Typically, we only are interested in $P(Y|X)$, the distribution of Y conditional on X.

Supervised learning

- $Y = (Y_1, \ldots, Y_q)$: response (output) (could be multivariate, $q = 1$ for us)
- $X = (X_1, \ldots, X_p)$: 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 $\mathcal{D}$.

Supervised learning

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

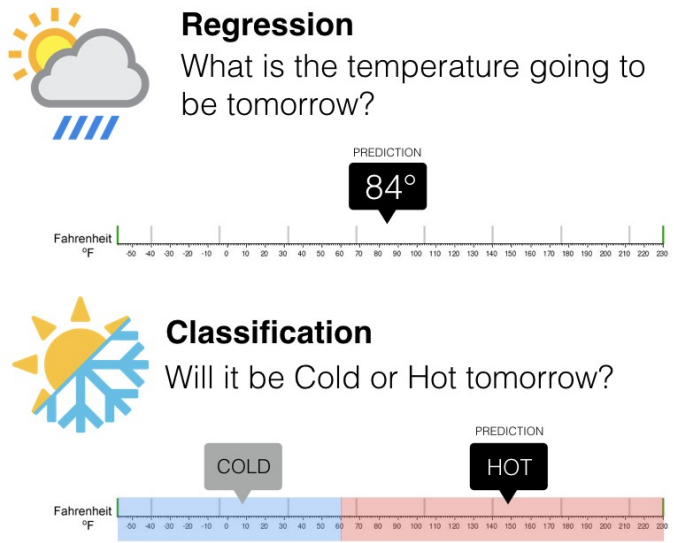
We are interested in minimizing the expected out-of-sample prediction error:

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

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.

Regression vs Classification Problems



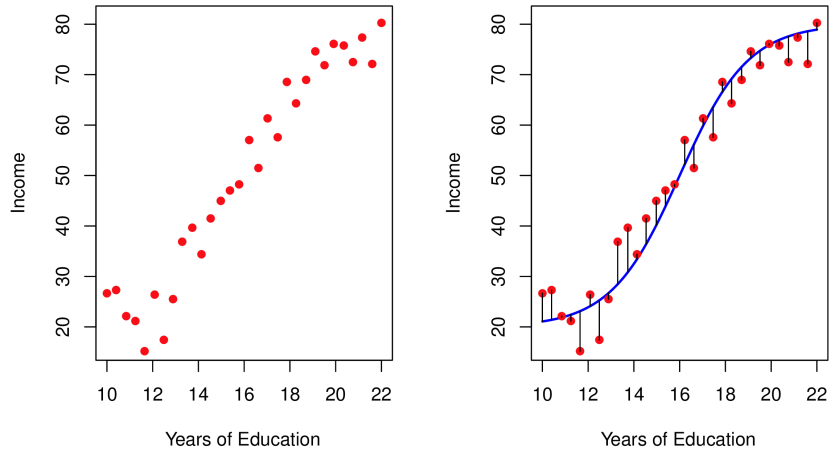

Regression

Regression

We often assume that our data arose from a statistical model

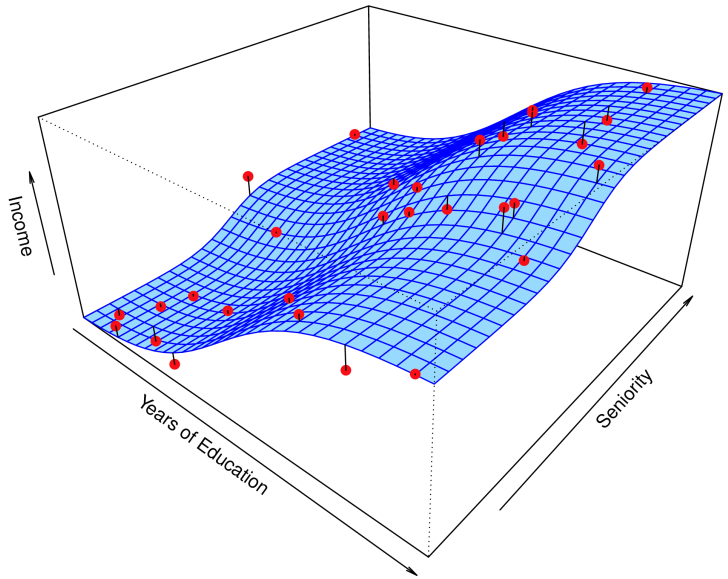
$Y = f(X) + \varepsilon$, where f is the true unknown function, ε is the random error term with $E[\varepsilon] = 0$ and is independent of X.

- The additive error model is a useful approximation to the truth
- $f(x) = E[Y|X = x]$
- Not a deterministic relationship: $Y \neq f(X)$



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

(Chapter2/2.2.pdf)



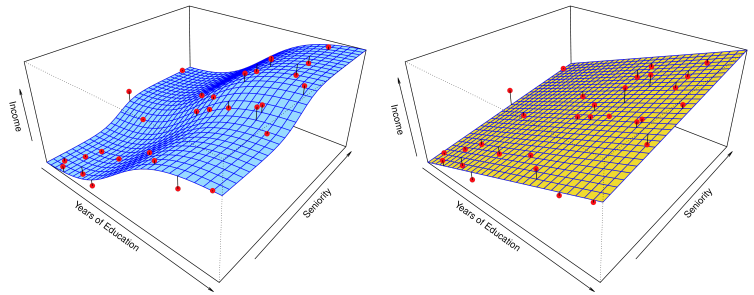
Blue surface is $f(x)$, the true functional relationship.

(Chapter2/2.3.pdf)

Goals of a regression analysis

- **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?

Estimation



Linear model: $\hat{f}\left(\text{education}, \text{seniority}\right) =$

$$\hat{\beta}_0 + \hat{\beta}_1 \times \text{education} + \hat{\beta}_2 \times \text{seniority}$$

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

(Chapter2/2.3.pdf, 2.4.pdf)

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 model assumtion (such as linearity) 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

Measures of accuracy

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

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

Measures of accuracy

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

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

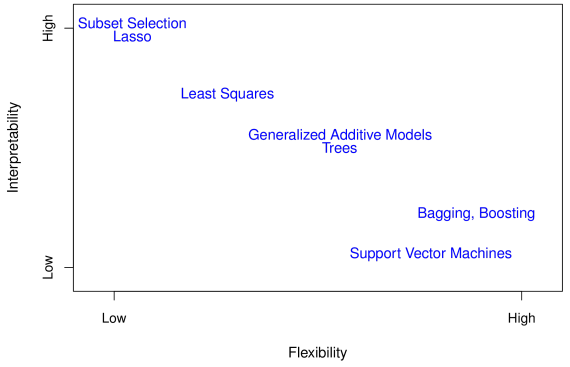
$$MSE_{Te} = \underset{j \in Te}{\text{Ave}}[y_j - \hat{f}(x_j)]^2 = \frac{1}{m}\sum_{j=1}^{m}[(y_j - \hat{f}(x_j)]^2$$

Training vs Test MSEs

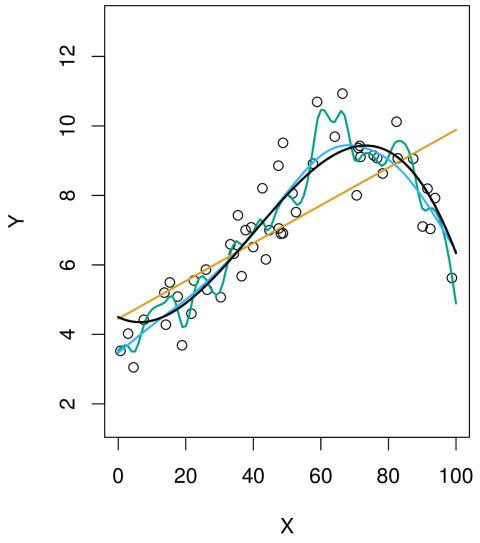
- In general, the more flexible a method is, the lower its training MSE will be. i.e. it will "fit" the training data very well.
- However, the test MSE may be higher for a more flexible method than for a simple approach like linear regression.
- Flexibility also makes interpretation more difficult.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

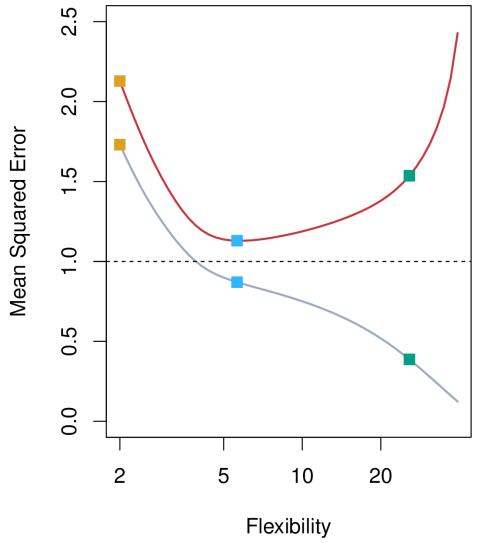
Interpretability vs Flexibility

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" method.



(Chapter2/2.7.pdf)

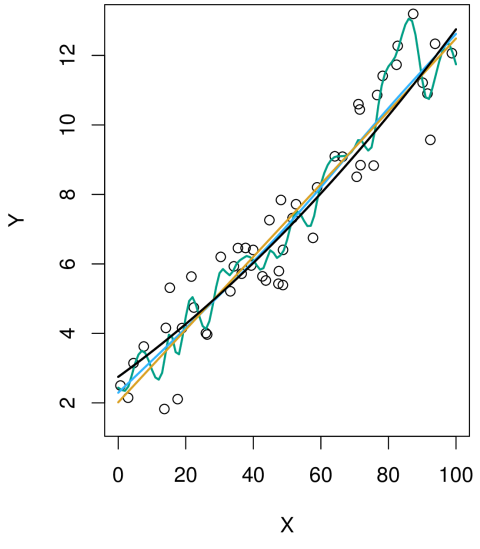




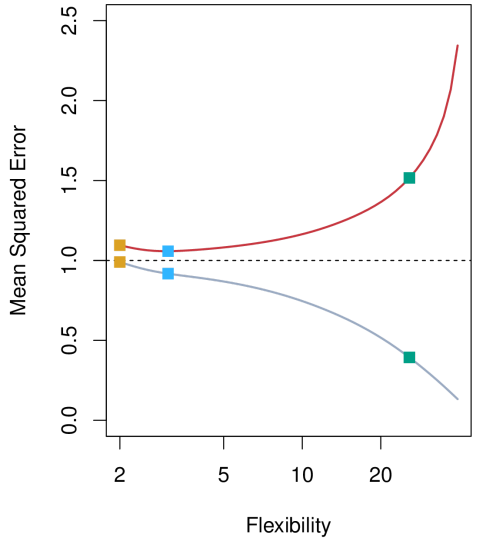
LEFT: Linear regression
Smoothing splines
True curve

RIGHT: Training MSE, Test MSE, Dashed: Minimum test MSE

(Chapter2/2.9.pdf)

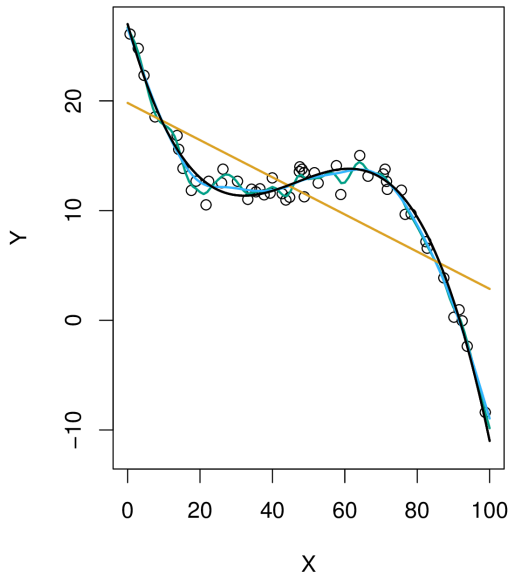


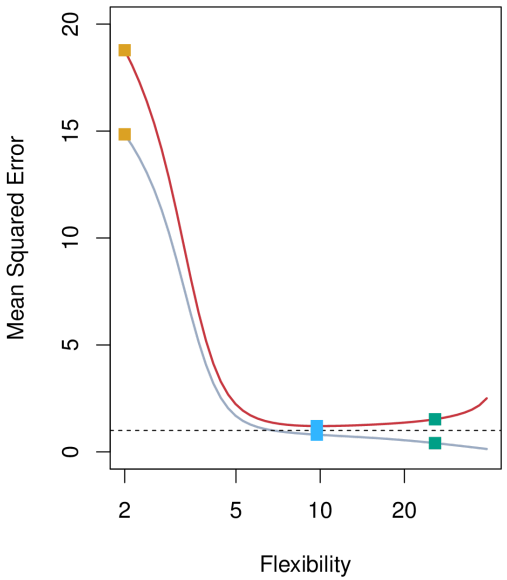
Linear regression
Smoothing splines
True curve



Training MSE, Test MSE
Dashed: Minimum test MSE

(Chapter2/2.9.pdf)





Linear regression
Smoothing splines
True curve

Training MSE, Test MSE
Dashed: Minimum test MSE

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.

Bias - variance tradeoff

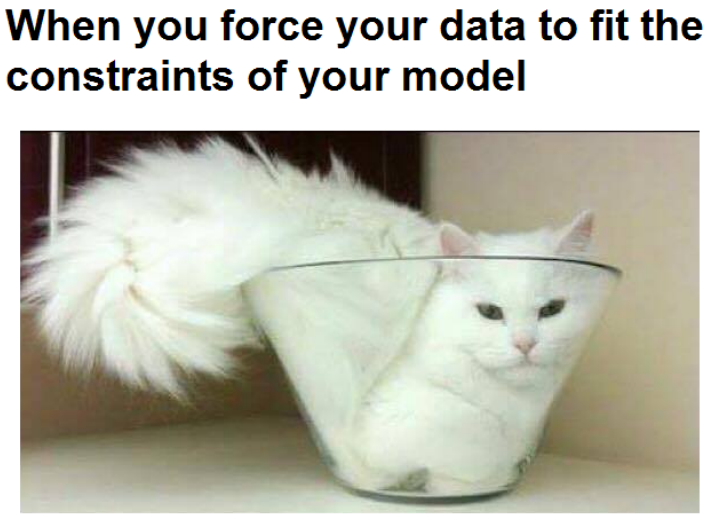
There are two competing forces that govern the choice of learning method: bias and variance.

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.

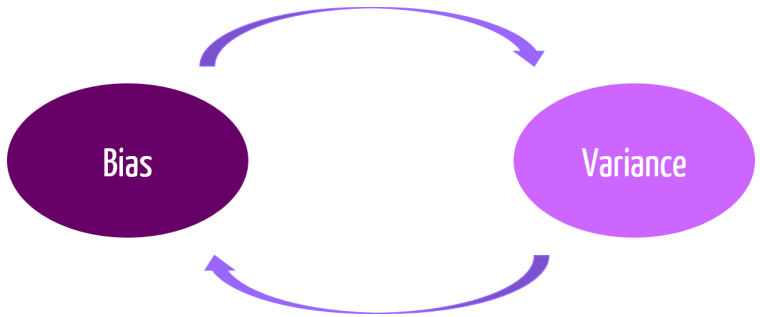
High variance (over fitted) models can be problematic...

... as well as high bias (under fitted) models!



The Bias Variance Tradeoff

As you may have guessed, there is a trade off between increasing variance (flexibility) and decreasing bias (simplicity) and vice versa.



Decomposing the MSE can give us a mathematical intuition as to why this occurs.

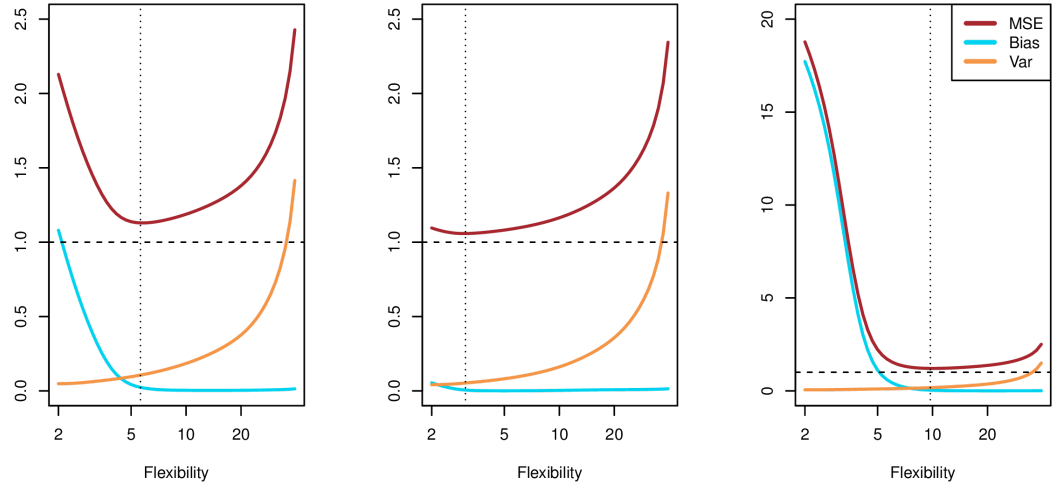
MSE decomposition

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

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

Test MSE = Bias ^{2} + 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 the bias-variance trade-off**



Squared bias, variance, Var(ε) (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)

Optimal Prediction

The optimal MSE is obtained when

$$\hat{f} = f = \mathrm{E}[Y \mid X = x].$$

Then bias=variance=0 and

$$\mathrm{MSE} = \text{irreducible variance}$$

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


Classification

Classification

Here the response variable Y is qualitative.

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

Classification - goals

Our goals are:

1. Build a classifier $C(x)$ that assigns a class label from $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ 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 $\mathcal{C}_k$
3. Assess the uncertainty in each classification (i.e., the probability of misclassification).
4. Understand the roles of the different predictors among $X = (X_1, X_2, \ldots, X_p)$.

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.

Error rates

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

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 $Te = \{x_j, y_j\}_1^m$

Test Error Rate

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

Bayes Classifier

Let $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, and let

$$p_k(x) = \mathrm{P}(Y = C_k \mid X = x), \qquad k = 1, 2, \ldots, K.$$

These are the conditional class probabilities at x.

Then the Bayes classifier at x is

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

This gives the minimum average test error rate.

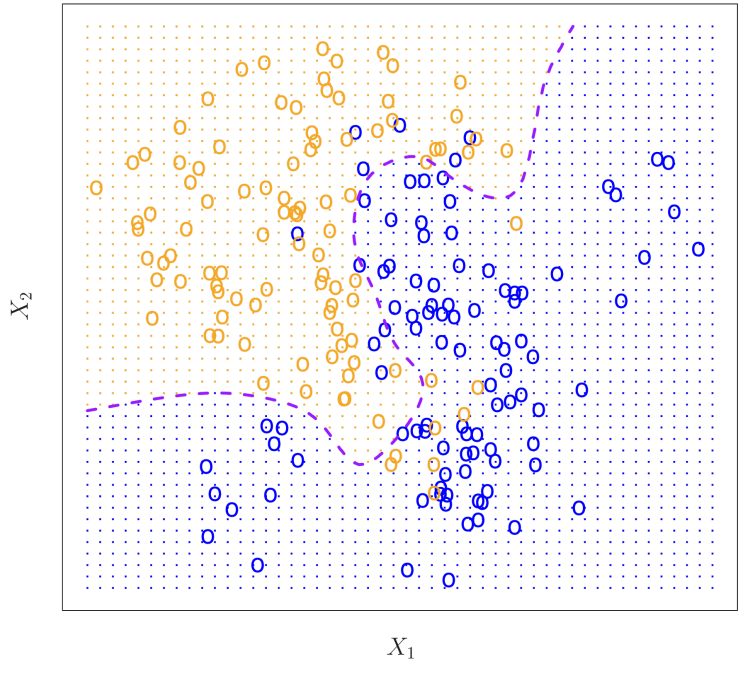
Bayes Classifier

$$1 - \mathrm{E}\left(\max_j \mathrm{P}(Y = C_j|X)\right)$$

- 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!!

Bayes Classifier

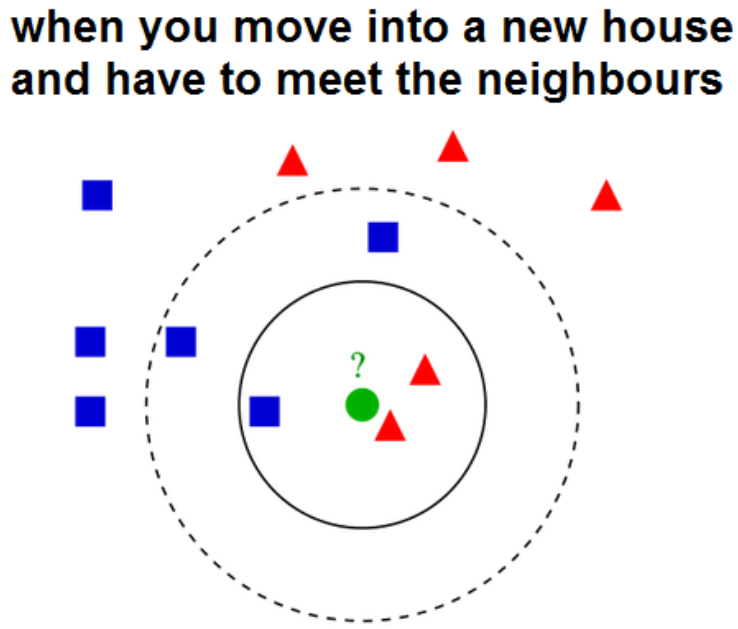


K Nearest Neighbours (KNN)

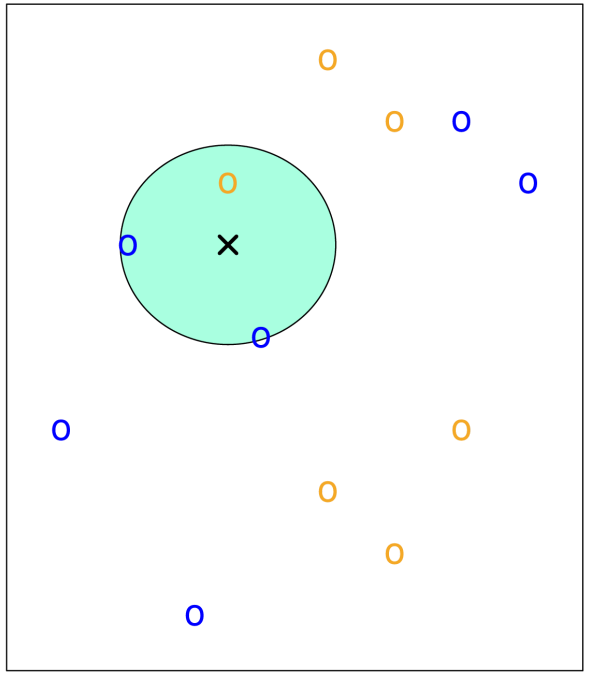
One of the simplest classifiers. Given a test observation x_0,

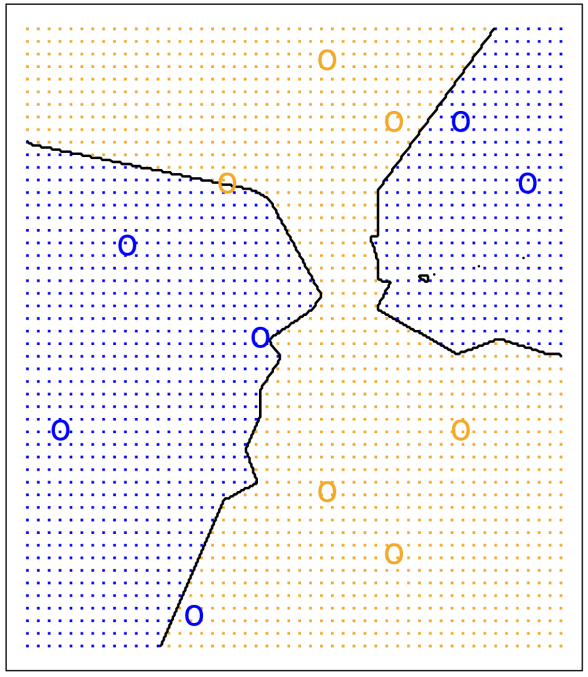
- Find the K nearest points to x_0 in the training data: $\mathcal{N}_0$.
- Estimate conditional probabilities $= P(Y = C_j \mid X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = C_j)$.
- Classify x_0 to class with largest probability.

KNN

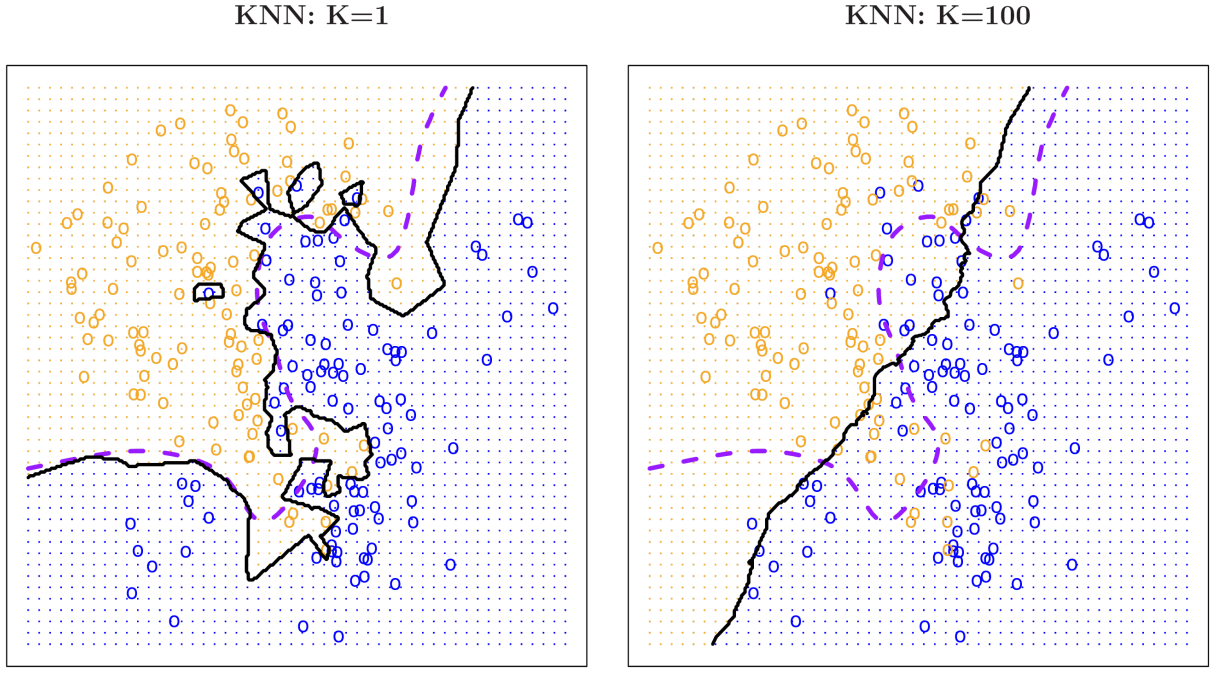

when you move into a new house
and have to meet the neighbours
?

KNN

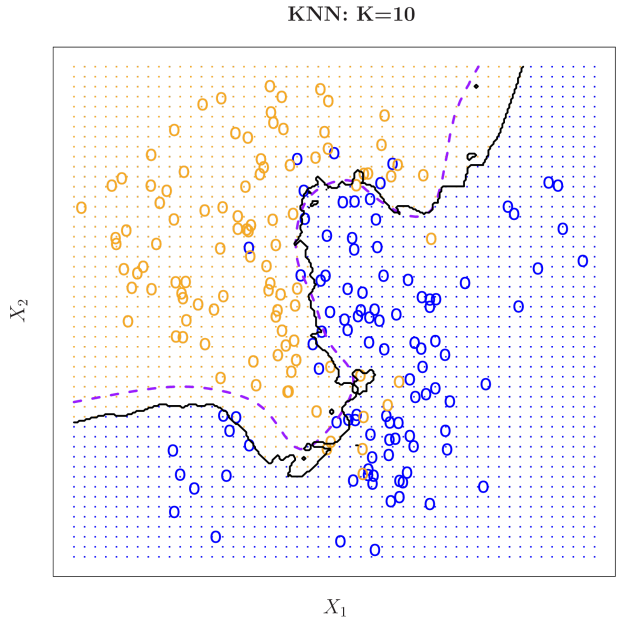




KNN

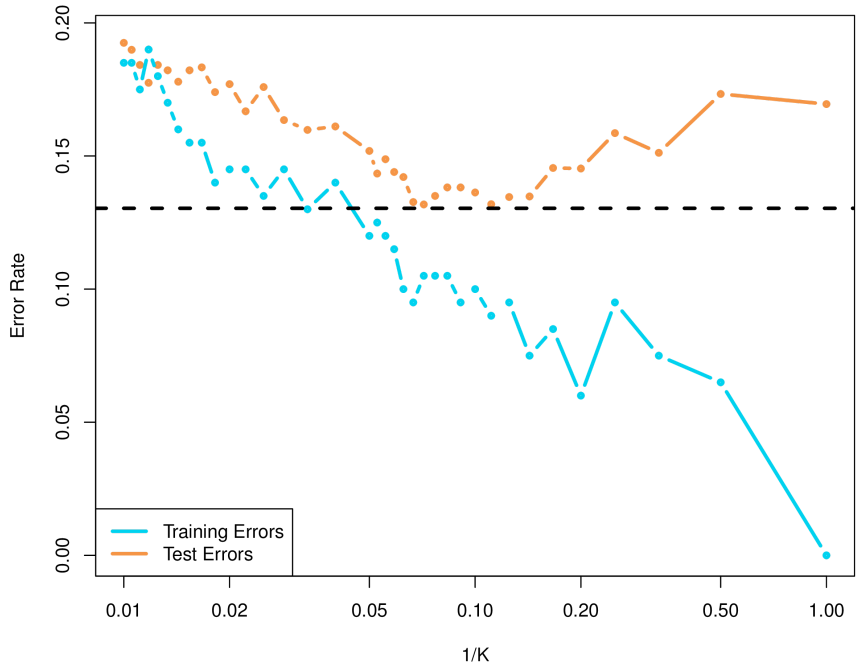


KNN



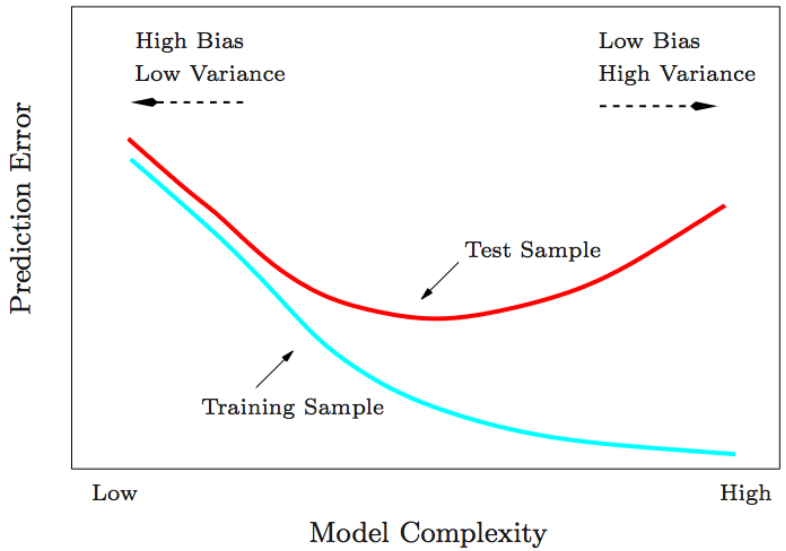
(Chapter2/2.15.pdf)

KNN



(Chapter2/2.17.pdf)

A fundamental picture



👨‍💻 Made by a human with a computer

Slides at https://iml.numbat.space.

Code and data at https://github.com/numbats/iml.

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