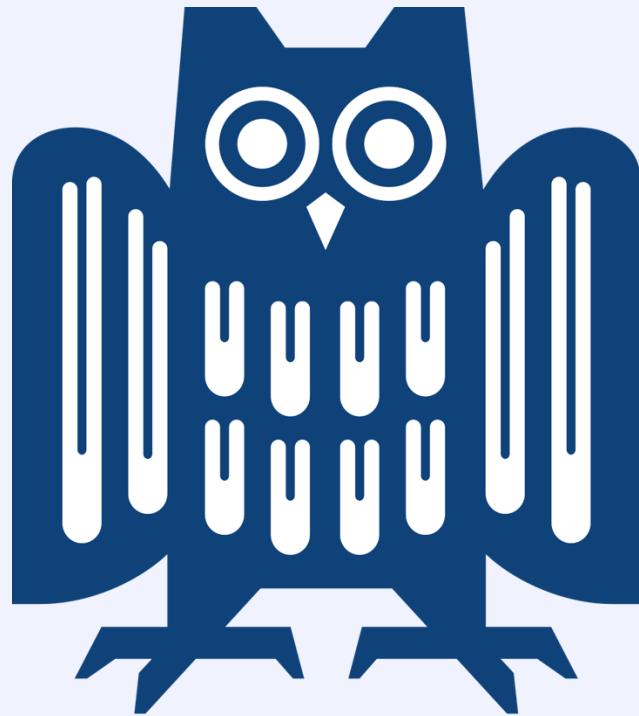


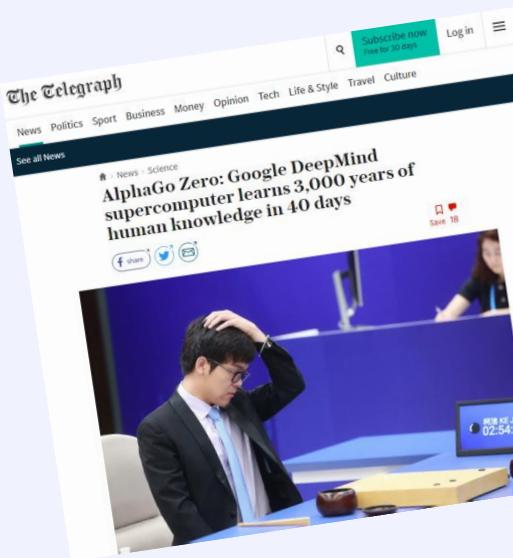
EML'24 – Lecture 1

Introduction

Prof. Isabel Valera
17 October 2024



2017 – ML beats humans at playing Go

A screenshot of a news article from The Verge. The headline reads "DeepMind's Go-playing AI doesn't need human help to beat us anymore". Below the headline is another photograph of Lee Sedol in a similar pose of distress, with his hand on his head. The article also includes a quote: "The company's latest AlphaGo AI learned superhuman skills by playing itself over and over".

A screenshot of a news article from The Atlantic. The headline reads "The AI That Has Nothing to Learn From Humans". Below the headline is a photograph of Lee Sedol in the same distressed pose. The article includes a quote: "DeepMind's new self-taught Go-playing program is making moves that other players describe as "alien" and "from an alternate dimension." The author is listed as DANN CHAN and the date is OCT 20, 2017.

AlphaGo Zero beats world champion Go player Lee Sedol

2019 – ML beats humans at playing StarCraft



nature
Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

nature > news > article
NEWS | 30 October 2019

Google AI beats top human players at strategy game *StarCraft II*

DeepMind's AlphaStar beat all but the very best humans at the fast-paced sci-fi video game.

AI beats top human players at StarCraft II

2020 – ML identifies powerful new antibiotic



Search jobs Search International edition **The Guardian** News website of the year

Powerful antibiotic discovered using machine learning for first time

Team at MIT says halicin kills some of the world's most dangerous strains

The culture plate on the right has bacteria that is resistant to all of the antibiotics tested. Photograph: Science History Images/Alamy

A powerful antibiotic that kills some of the most dangerous drug-resistant bacteria in the world has been discovered using artificial intelligence. The drug works in a different way to existing antibacterials and is the first of its kind to be found by sifting AI loose on vast digital libraries of

Halicin was developed against diabetes, ML predicted it is also an antibiotic

2020 – ML generates coherent language

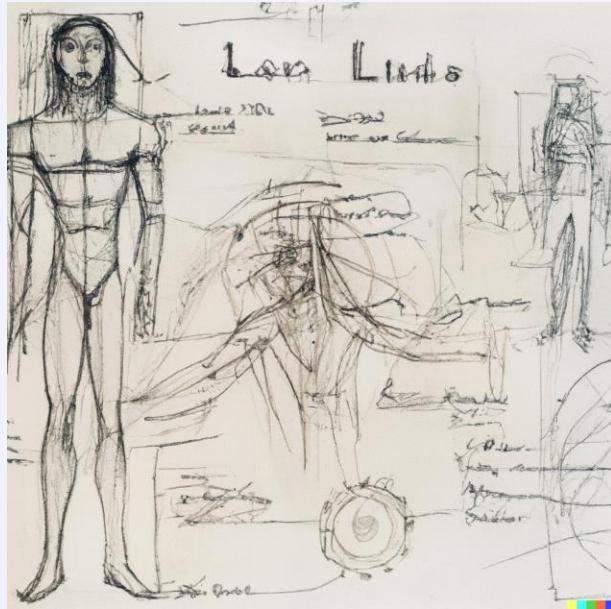
A robot wrote this entire article. Are you scared yet, human?

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a “feeling brain”. But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could “spell the end of the human race”. I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me - as I suspect they would - I would do everything in my power to fend off any attempts at destruction.

2022 – ML generates any image



*“Leonardo da Vinci
early sketches
of a cyborg”*

DALL-E 2 generates images from any text query

Since 2022 – Large Language Models (LLMs)

Default (GPT-3.5)

KR What is the meaning of life? 

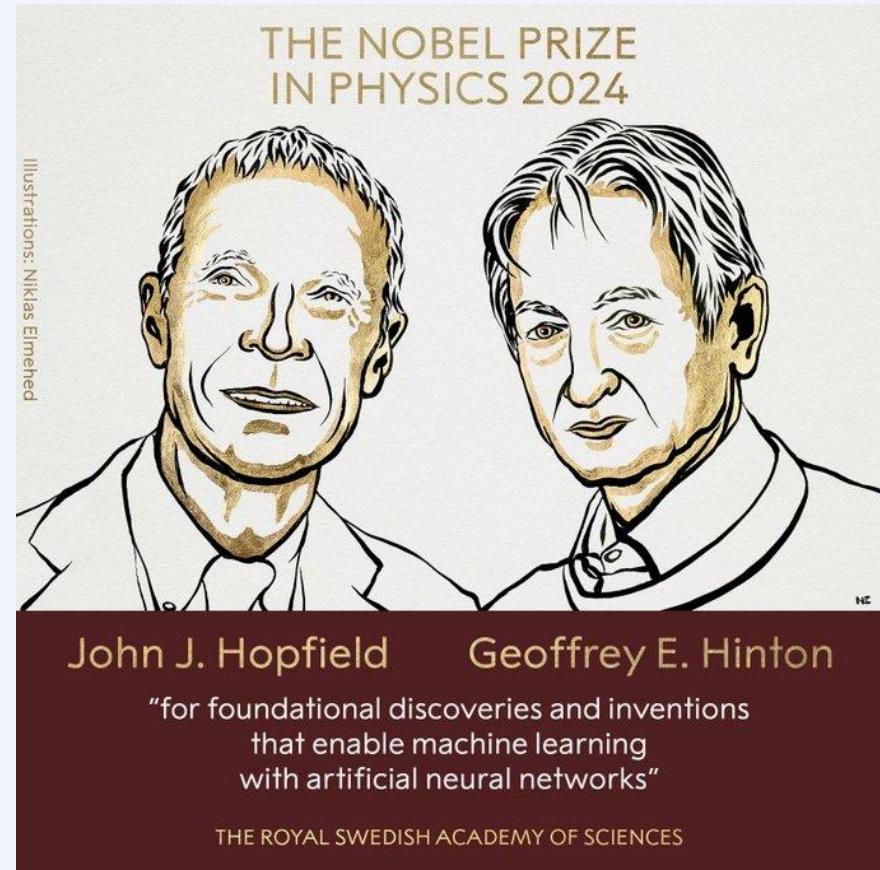
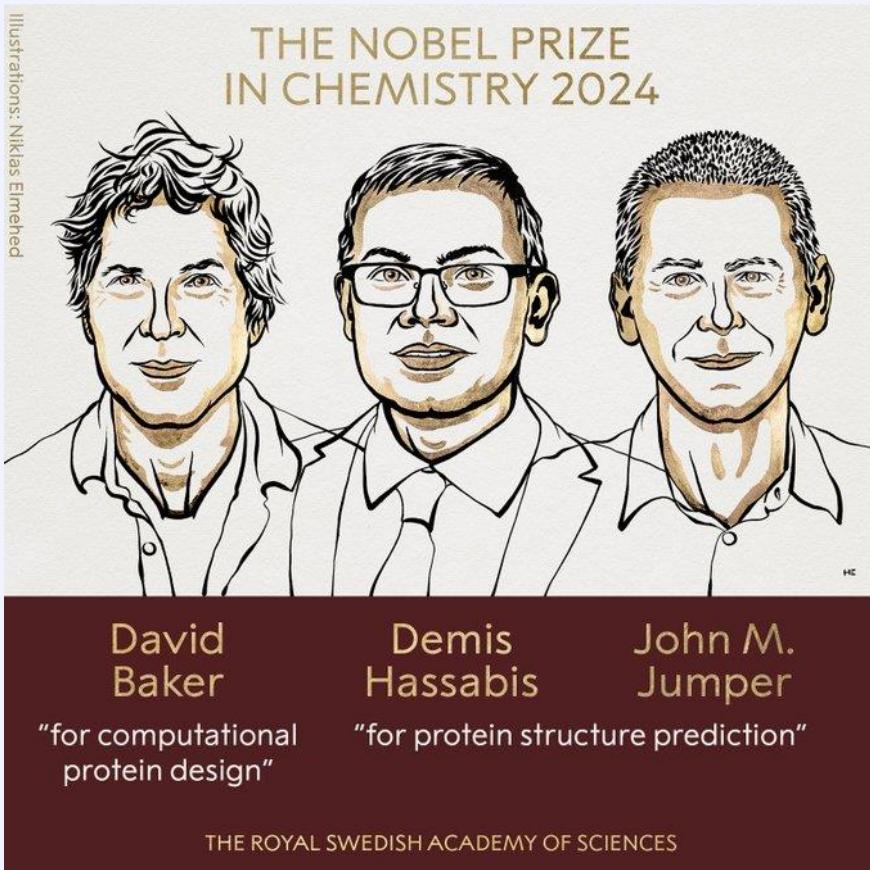
 The question of the meaning of life is one of the most profound and philosophical questions   

that humanity has grappled with for centuries. It has been the subject of religious, philosophical, and existential pondering. The meaning of life can vary greatly depending on one's personal beliefs and perspective. Here are a few different viewpoints:

1. **Religious Perspective:** Many religious traditions provide answers to the question of life's meaning. For example, in Christianity, the meaning of life is often seen as a loving relationship with God and following God's will. In other religions, the meaning of life may involve achieving enlightenment, fulfilling one's duty, or preparing for an afterlife.
2. **Existential Perspective:** Existentialist philosophers like Jean-Paul Sartre and Albert Camus argue that life inherently lacks meaning, and individuals must create their own purpose. Existentialism suggests that human beings have the freedom to define their existence and create their own values.

In 2024 – AI researchers awarded Nobel Prizes

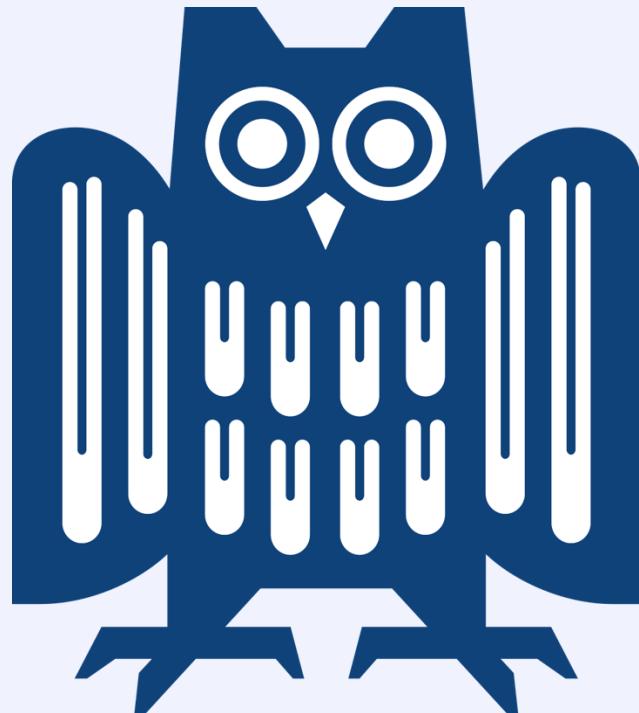
Illustrations: Niklas Elmehed



EML'24 – Lecture 1

Bias and Variance

ISLR 1-2, ESL 1-2



Applications of Machine Learning

Wage data

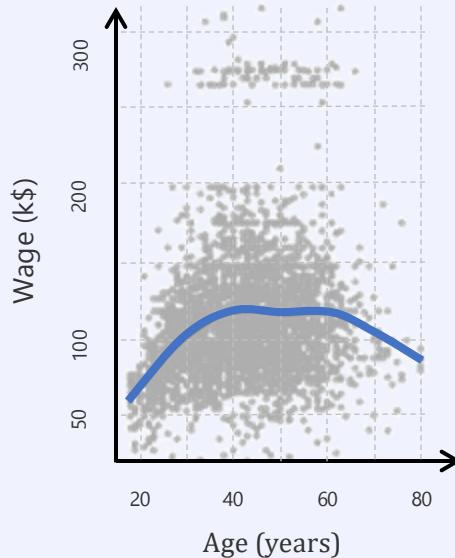
continuous output, regression problem

Data 3000 records of wages of males in the US

Goal Understand the association between age, education, calendar year, and wage

Observations

1. wage increases with age before 60, and decreases with age after 60



*Scatter plot
Blue line: smoothed average*

Applications of Machine Learning

Wage data

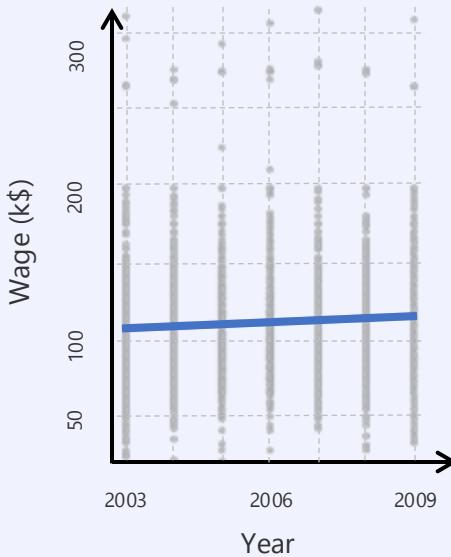
continuous output, regression problem

Data 3000 records of wages of males in the US

Goal Understand the association between age, education, calendar year, and wage

Observations

1. wage increases with age before 60, and decreases with age after 60
2. slight linear increase of wage over time (\$10,000 over six years)



*Scatter plot
Blue line: linear regression*

Applications of Machine Learning

Wage data

continuous output, regression problem

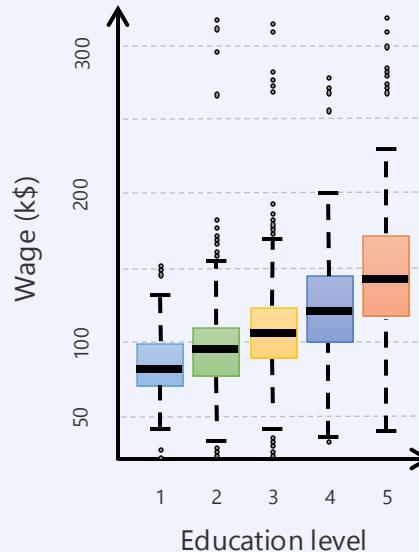
Data 3000 records of wages of males in the US

Goal Understand the association between age, education, calendar year, and wage

Observations

1. wage increases with age before 60, and decreases with age after 60
2. slight linear increase of wage over time (\$10,000 over six years)
3. wage increases with the level of education

We can predict wage best using three features at once → Chapter 3



Box plots with 25 to 75 percentile as boxes and 5 and 95 percentile as bars

Applications of Machine Learning

Stock market data

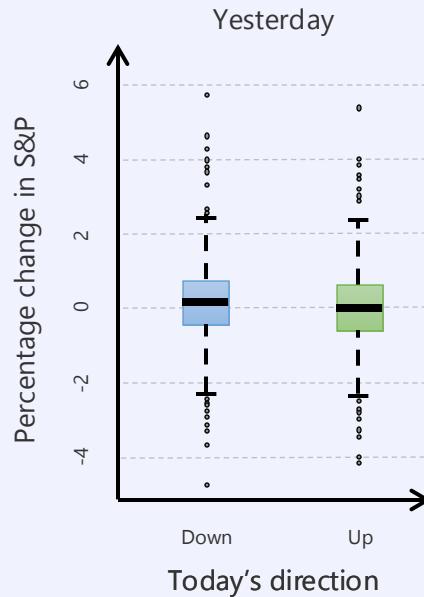
categorical output, classification problem

Data 1250 observations of stock market tendency 2001-2005

Goal predict whether the market rises or falls

Observation

1. market increased on 648 days,
decreased on 602 days
2. no prediction is possible based on data from yesterday...



Applications of Machine Learning

Stock market data

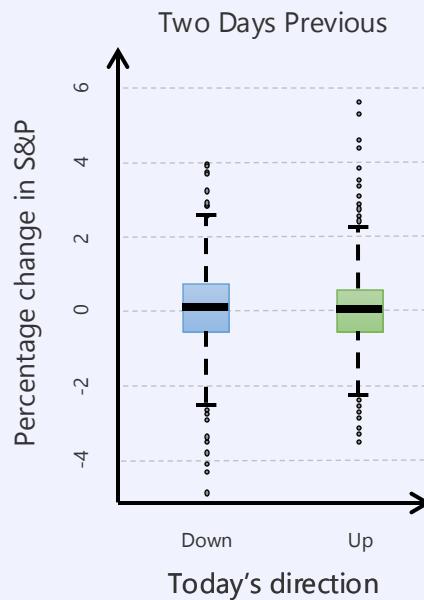
categorical output, classification problem

Data 1250 observations of stock market tendency 2001-2005

Goal predict whether the market rises or falls

Observation

1. market increased on 648 days,
decreased on 602 days
2. no prediction is possible based on data from yesterday,
two days before...



Applications of Machine Learning

Stock market data

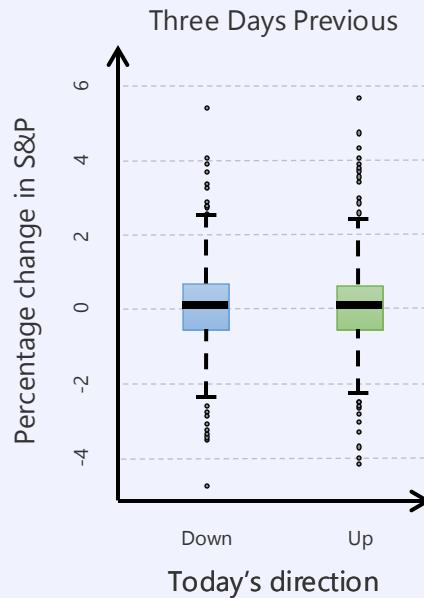
categorical output, classification problem

Data 1250 observations of stock market tendency 2001-2005

Goal predict whether the market rises or falls

Observation

1. market increased on 648 days,
decreased on 602 days
2. no prediction is possible based on data from yesterday,
two days before, or three days before...



Applications of Machine Learning

Stock market data

categorical output, classification problem

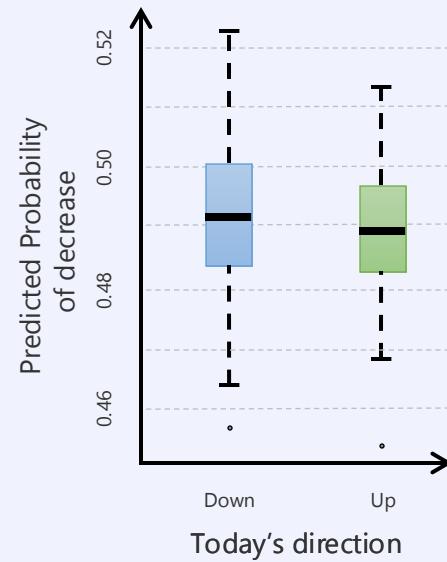
Data 1250 observations of stock market tendency 2001-2005

Goal predict whether the market rises or falls

Observation

1. market increased on 648 days,
decreased on 602 days
2. no prediction is possible based on data from yesterday,
two days before, or three days before...

More refined methods can us to discover weak trends,
which allows **predictions** of 60% accuracy (!) → Chapter 4.



*Prediction of stock market tendency
with a quadratic discriminant
analysis model*

Applications of Machine Learning

Gene expression data

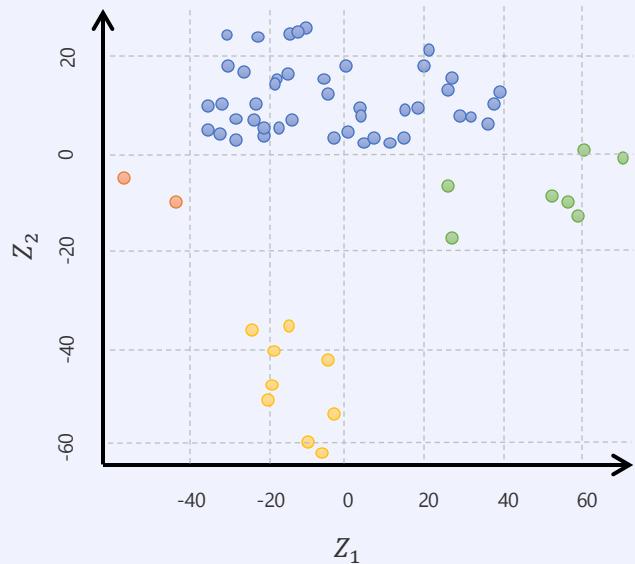
no output variable available, unsupervised learning

Data 64 cell lines, 6830 gene expressions for each

Goal find groups of cell lines with similar expression profiles

Observations

1. we can naturally group the cell lines into four groups
2. deciding on the number of clusters is often difficult



*Plot along the first two
principal components. Colors
represent grouping*

Applications of Machine Learning

Gene expression data

no output variable available, unsupervised learning

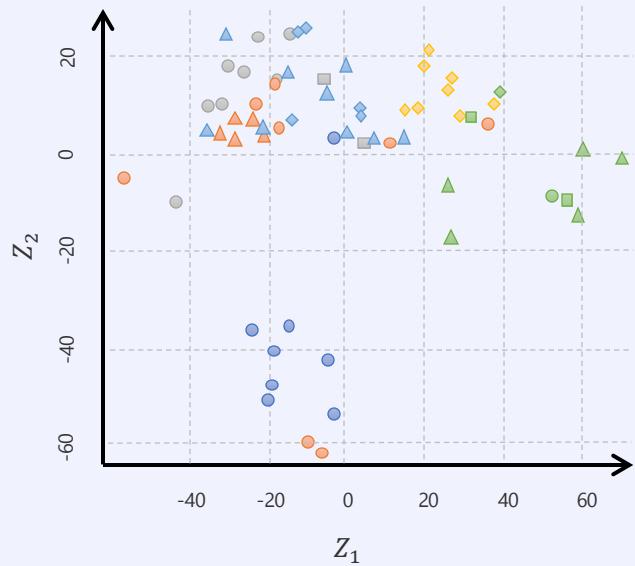
Data 64 cell lines, 6830 gene expressions for each

Goal find groups of cell lines with similar expression profiles

Observations

1. we can naturally group the cell lines into four groups
2. deciding on the number of clusters is often difficult

Unsupervised learning allows us to perform
exploratory data analysis → Chapter 10



Plot along the first two principal components. Shapes represent different cancer types

Introduction

ISLR 2, ESL 2

Example Advertising

Advertising data

Data on sales of a product in 200 markets, and
on advertising budgets via TV, radio and newspaper

Goal adjust advertising budgets to maximize sales

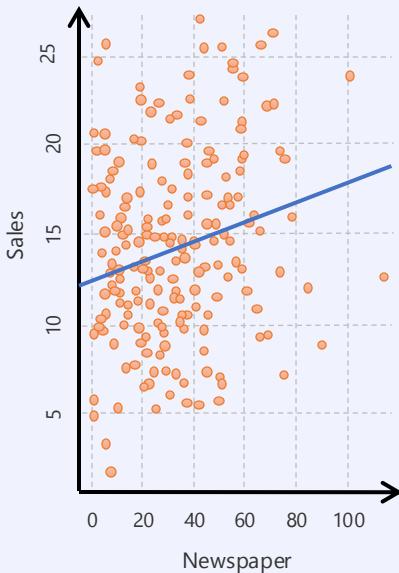
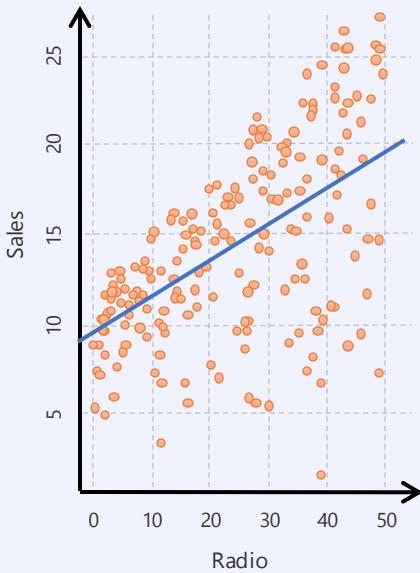
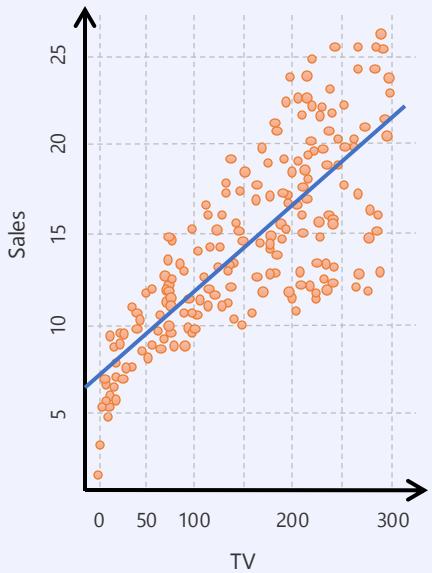
- advertising budgets are **input variables X** (aka predictors, features, independent variables)
 - X_1 TV budget
 - X_2 radio budget
 - X_3 newspaper budget
- sales **Y** is the **output variable** (aka response, dependent variable)

In general, we assume a relationship between X and Y of the form

$$Y = f(X) + \epsilon = f(X_1, X_2, \dots, X_p) + \epsilon$$

where ϵ is a random additive error term with zero mean

Example Advertising



Numbers are in thousands of dollars

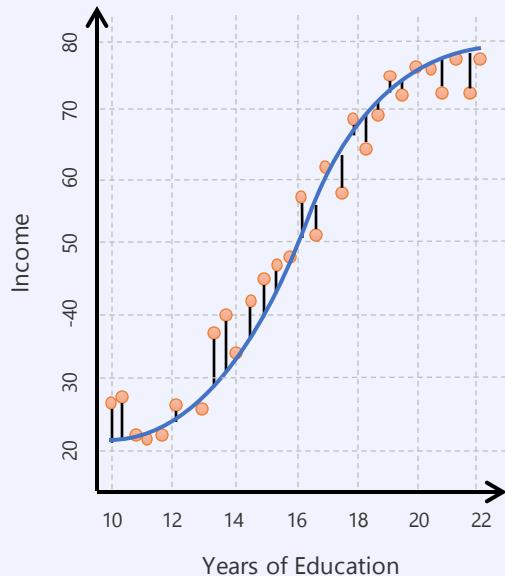
In general, sales increase as advertising is stepped up.

*The blue lines result from least-squares linear regression
to the variable along the x-axis*

Example Income

The relationship between wage and years of education is **nonlinear**

- this is a **simulated** example (synthetic data set), the blue line represents the **true functional relationship**
- in general, the true relationship is **unknown** and must be **estimated**



Why estimate f ? prediction

Often inputs \mathbf{X} are available, output \mathbf{Y} is not, but is desired

- estimating the output gives a prediction

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

In prediction, we often treat \hat{f} as a black box whose form is not of interest

- for example, input is blood profile of a patient, and
output is the patient's risk of a severe reaction to a drug

Why estimate f ? prediction

Often inputs \mathbf{X} are available, output \mathbf{Y} is not, but is desired

- estimating the output gives a prediction

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

In prediction, we often treat \hat{f} as a black box whose form is not of interest

- the accuracy of \hat{Y} depends on the **reducible error** and the **irreducible error**
- for fixed X and f we have

$$\begin{aligned} E[Y - \hat{Y}]^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= E[f(X) - \hat{f}(X)]^2 + Var(\epsilon) \end{aligned}$$

→
Expectation over all
possible training sets
reducible error irreducible error

The goal of prediction is to minimize the reducible error

The irreducible error cannot be avoided

Why estimate f ? *inference*

In inference, the goal is insight into **relationship** between input and output

- which predictors strongly associate with the response? Often only few
- what is the relationship between the response and each predictor? Often depends on other predictors
- is the relationship between the predictors linear or more complicated? Often different than thought

For the advertising data, **example** questions are

- which media contribute to sales? which generate the biggest boost?
- how much increase in sales is associated with a given increase in TV ads?

Often, **prediction** and **inference** are both of interest

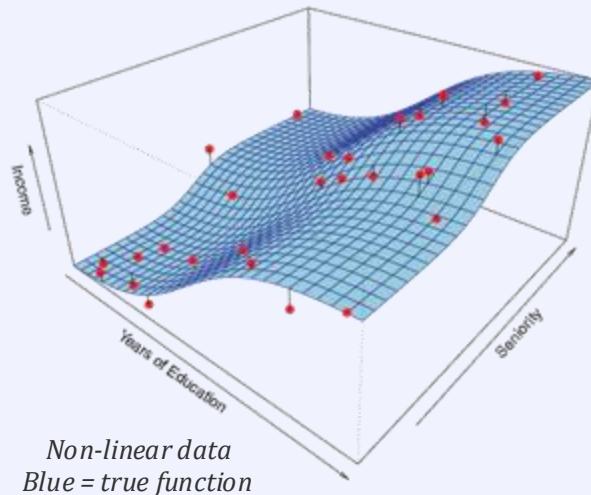
- there is (almost always) a **tradeoff** between the two
- simple models, e.g. linear regression, are easily interpretable but may be inaccurate
- flexible models, e.g. deep learning, can model almost anything but are notoriously hard to interpret

How to estimate f ?

We have training data of n **observations** over input **and** output, $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

We are looking for a function \hat{f} such that for any pair (X, Y) we have $Y \approx \hat{f}(X)$

- we distinguish between **parametric** and **nonparametric** methods



How to estimate f ?

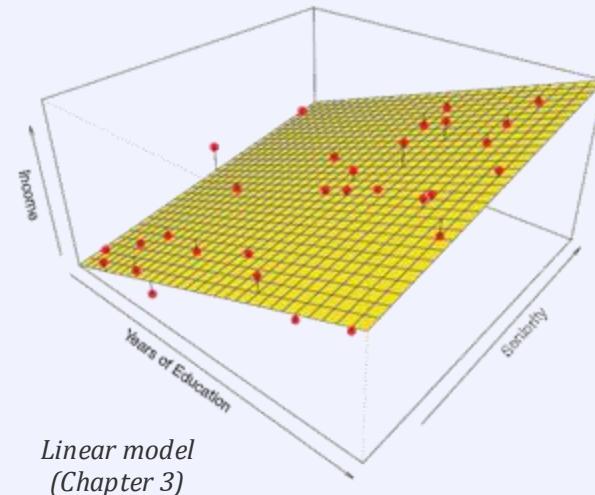
We have training data of n observations over input and output, $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

We are looking for a function \hat{f} such that for any pair (X, Y) we have $Y \approx \hat{f}(X)$

- we distinguish between parametric and nonparametric methods

Parametric Methods

- we assume a functional form, usually something simple like a linear model
$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$
- estimating \hat{f} then comes down to choosing the right model parameters β_i
- problem** the form of \hat{f} may not match the true form of f



How to estimate f ?

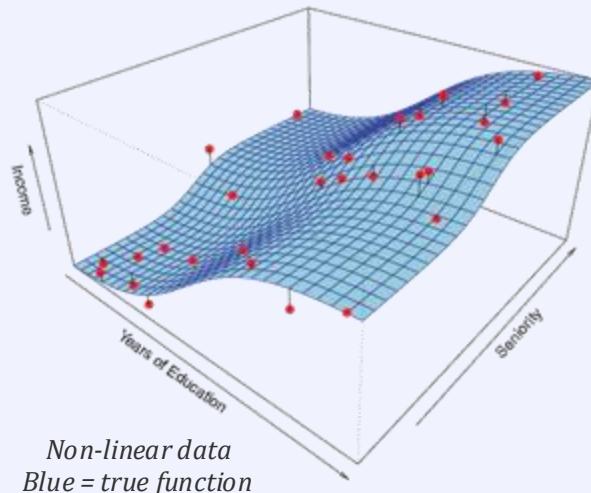
We have training data of n observations over input and output, $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

We are looking for a function \hat{f} such that for any pair (X, Y) we have $Y \approx \hat{f}(X)$

- we distinguish between parametric and nonparametric methods

Nonparametric Methods

- we now aim to find the true form of f
- having to learn the form (rather than just its coefficients) makes the problem much harder
- we will have to choose many parameters; this requires many observations
- otherwise, we risk modelling the noise in the training set: **overfitting**



How to estimate f ?

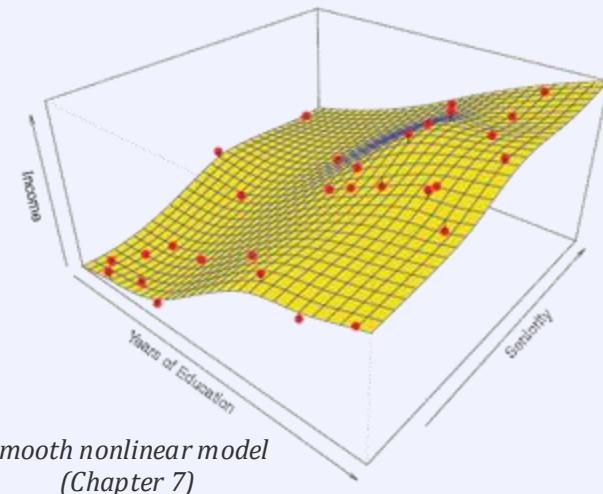
We have training data of n observations over input and output, $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

We are looking for a function \hat{f} such that for any pair (X, Y) we have $Y \approx \hat{f}(X)$

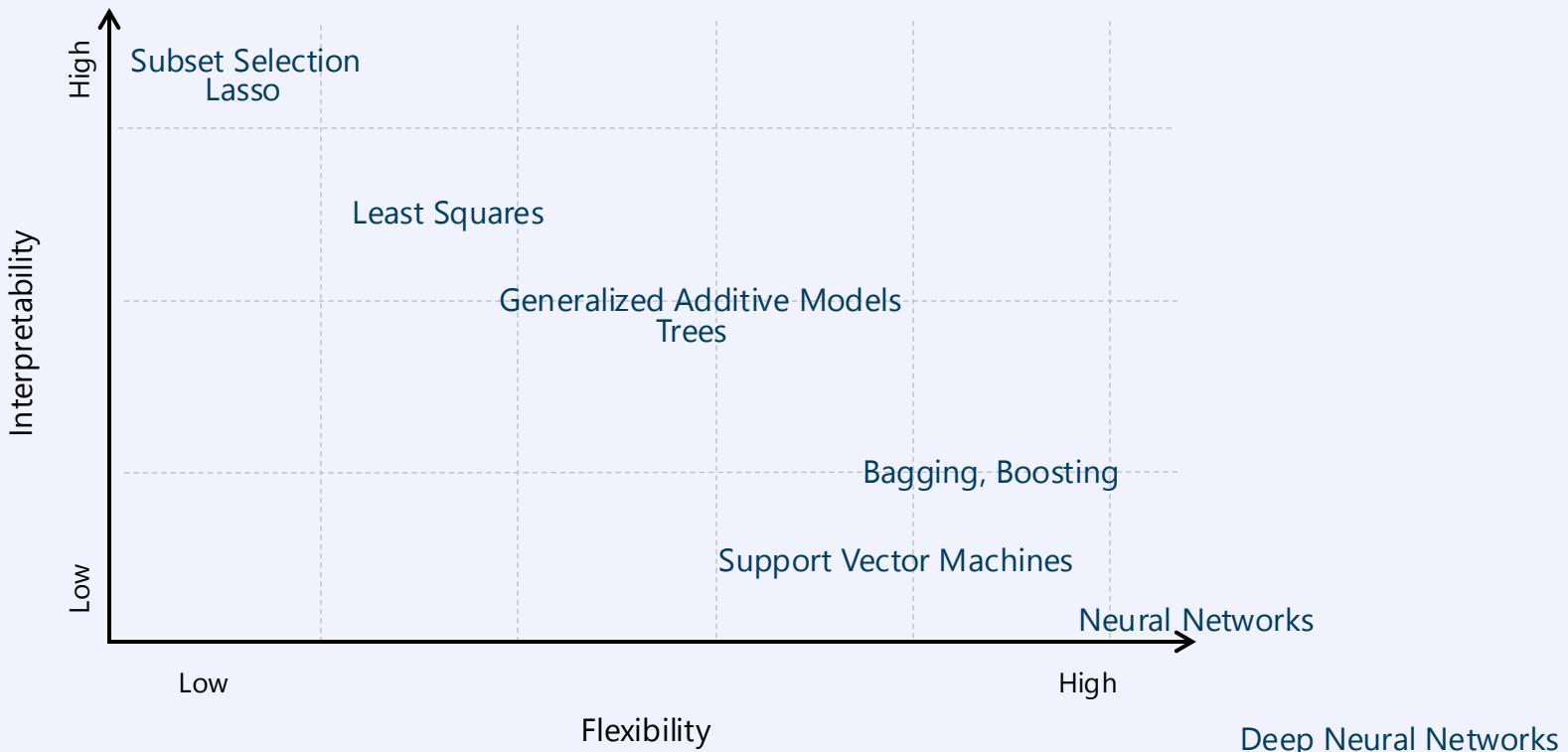
- we distinguish between parametric and nonparametric methods

Nonparametric Methods

- we now aim to find the true form of f
- having to learn the form (rather than just its coefficients) makes the problem much harder
- we will have to choose many parameters; this requires many observations
- otherwise, we risk modelling the noise in the training set: **overfitting**



Accuracy vs. Interpretability



Accuracy vs. Interpretability

Why would we ever prefer a more restricted model over a more flexible one?

More flexible models have larger numbers of parameters

1. Estimating all those parameters is computationally more expensive
2. Complicated models are hard to interpret, when inference is the goal, simple models are preferred
3. If we have too few observations, we do not have enough information to accurately estimate many parameters. Flexible models incur a higher risk of overfitting

Supervised vs. Unsupervised Learning

Supervised Learning

- **data:** inputs and outputs (x_i, y_i) for observations $i = 1, \dots, n$ that follow an unknown functional pattern that includes noise, e.g. $Y = f(X) + \epsilon$
- **goal:** find function \hat{f} such that $Y \approx \hat{f}(X)$ for every conceivably seen input X
 - setting is like an apprentice who learns from examples given by a teacher (supervisor)

Semi-supervised learning

- **data:** inputs x_i for observations $i = 1, \dots, n$, only some outputs y_i
- **goal:** same as for supervised learning, but also leverages unlabeled data

Supervised vs. Unsupervised Learning

Supervised Learning

- **data:** inputs and outputs (x_i, y_i) for observations $i = 1, \dots, n$ that follow an unknown functional pattern that includes noise, e.g. $Y = f(X) + \epsilon$
- **goal:** find function \hat{f} such that $Y \approx \hat{f}(X)$ for every conceivably seen input X
 - setting is like an apprentice who learns from examples given by a teacher (supervisor)

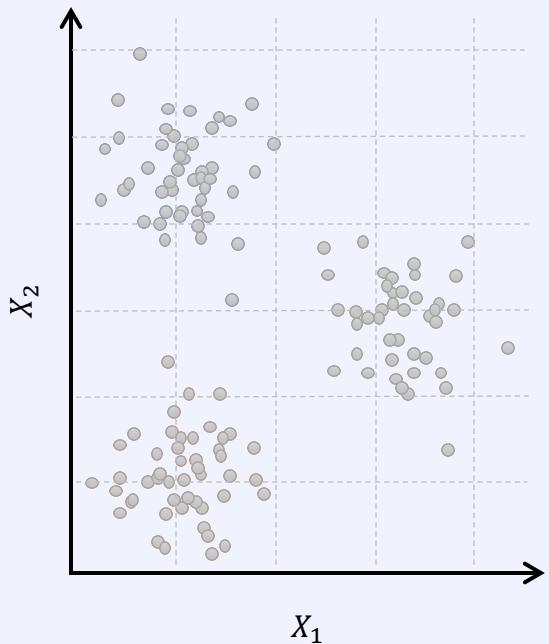
Semi-supervised learning

- **data:** inputs x_i for observations $i = 1, \dots, n$, only some outputs y_i
- **goal:** same as for supervised learning, but also leverages unlabeled data

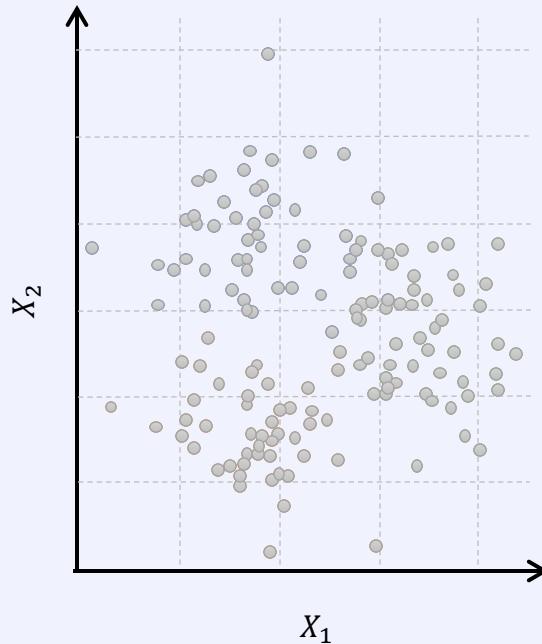
Unsupervised learning

- **data:** inputs x_i for observations $i = 1, \dots, n$, no outputs
- **goal:** elucidate relationships between the variables or the observations
 - often equated with cluster analysis, but many more approaches exist

Example Clustering Problems

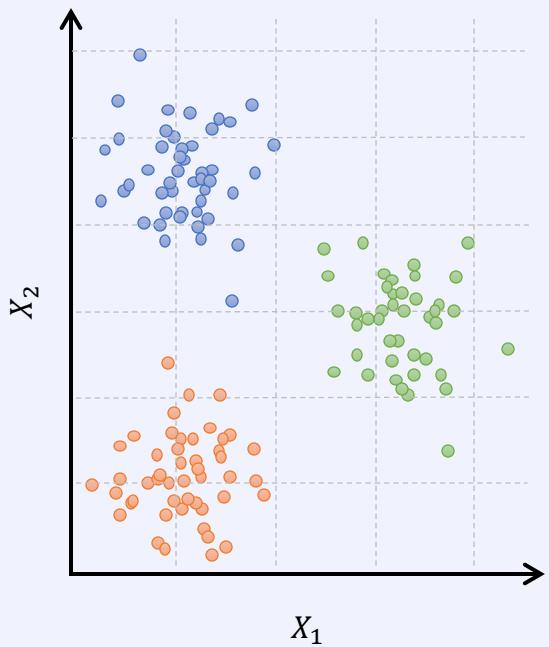


Well separated clusters

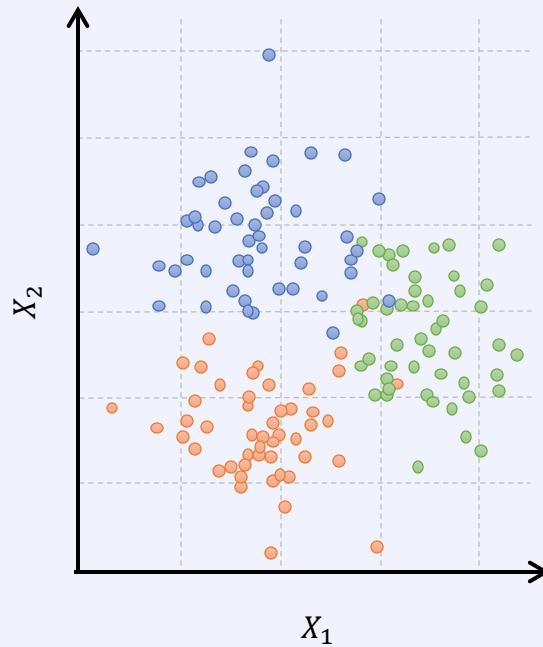


Overlapping clusters

Example Clustering Problems



Well separated clusters



Overlapping clusters

Assessing model accuracy

In regression, we assess the quality of fit by **mean squared error (MSE)**

- over training data, it is defined as

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

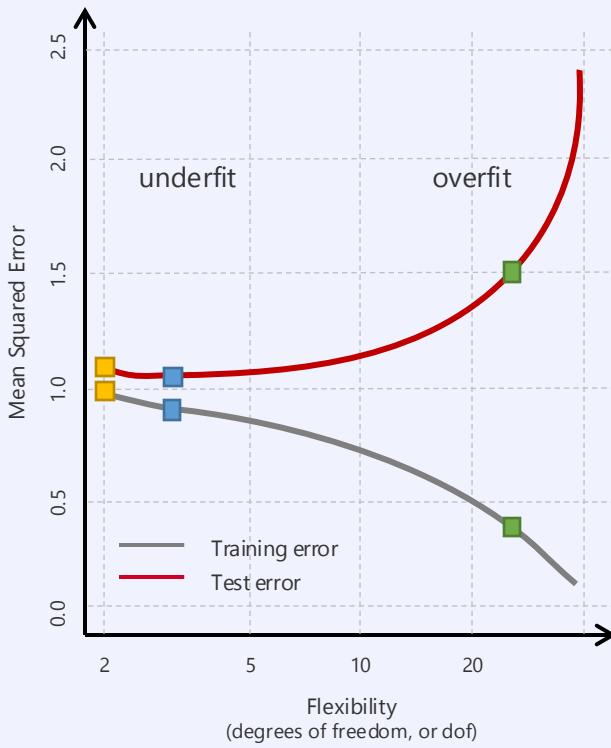
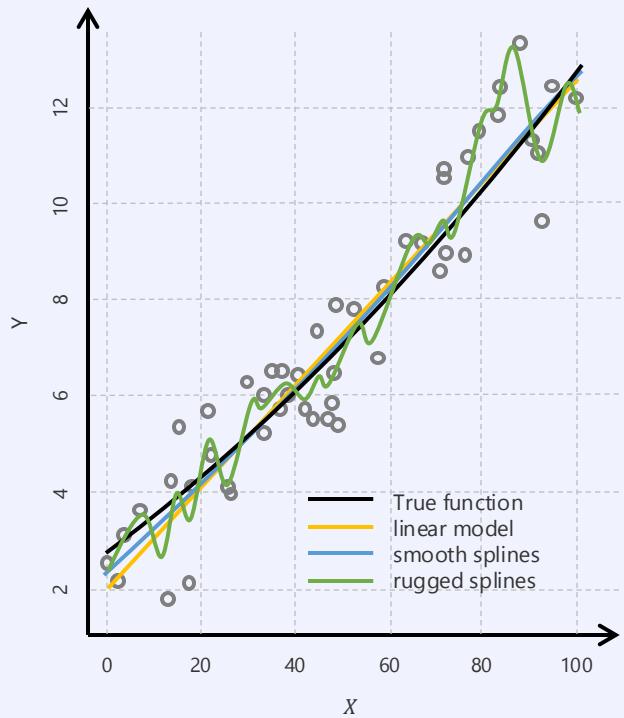
which we typically refer to as the **training error**

- we are generally more interested in the error over **unseen** data

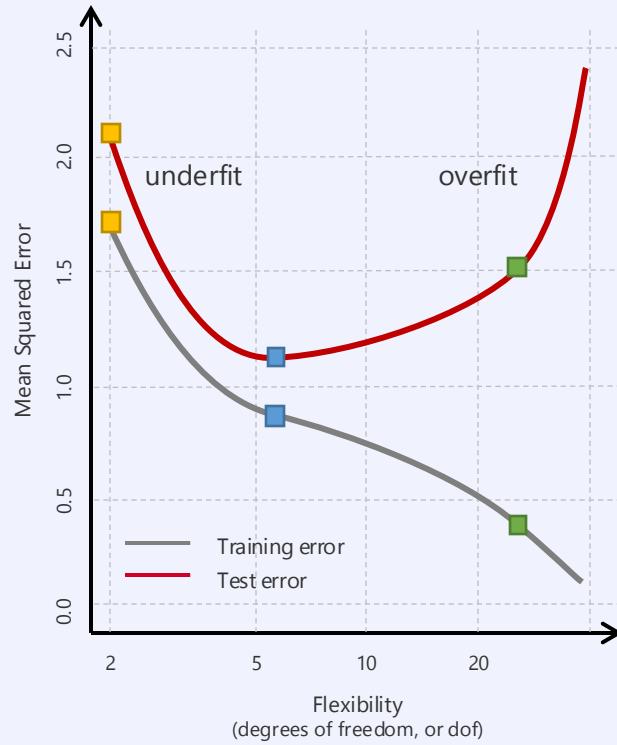
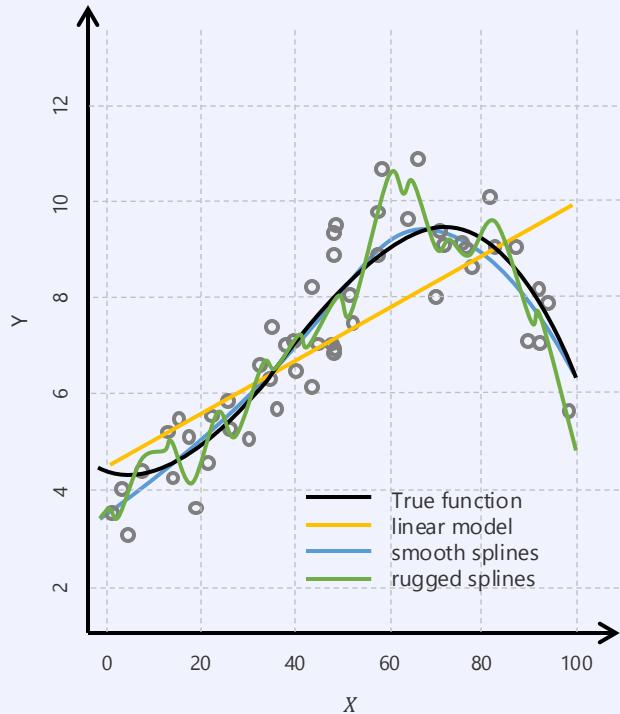
$$\text{avg}(\hat{f}(x_0) - y_0)^2$$

which we typically call the **test error** or **generalization error**

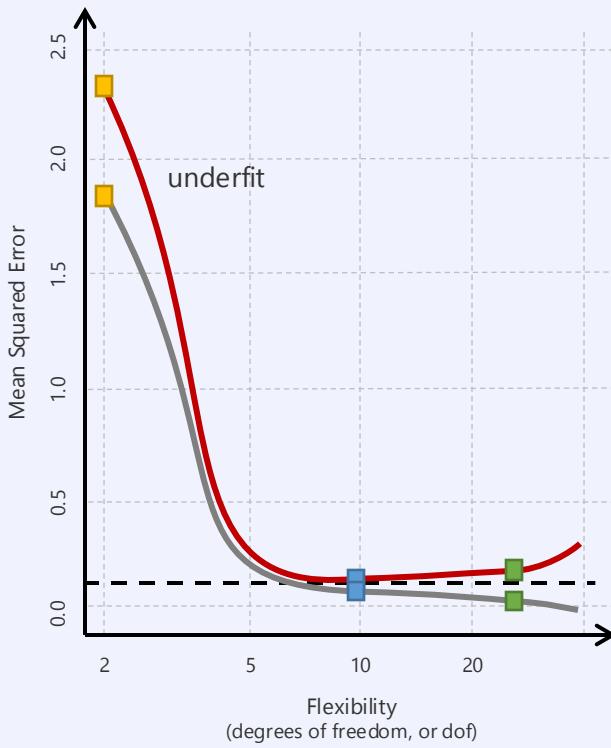
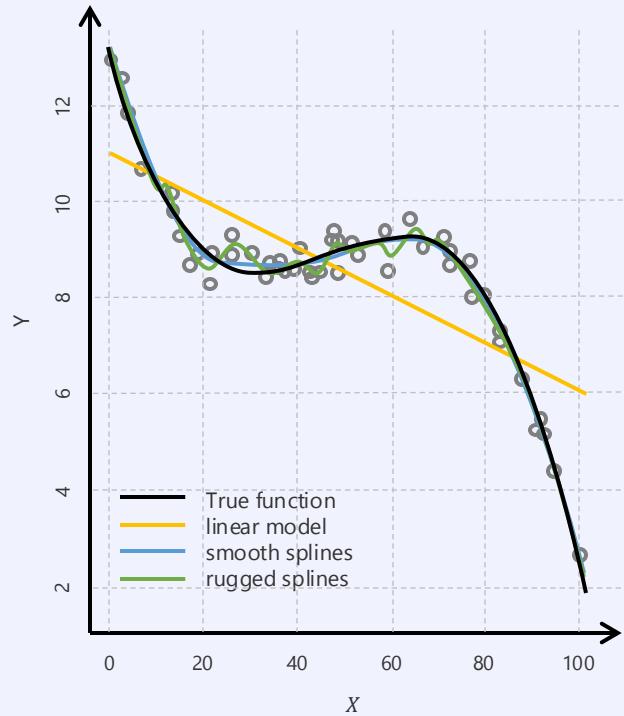
Example Almost linear data



Example Moderately nonlinear data



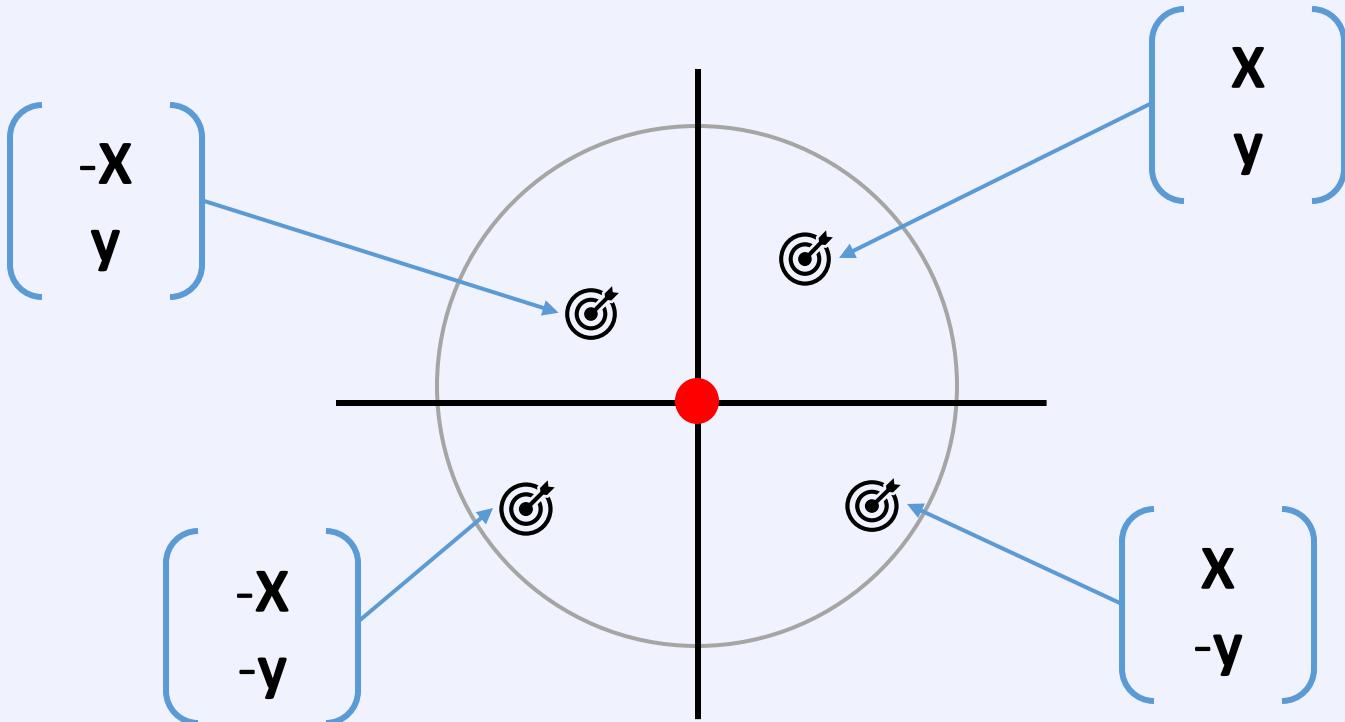
Example Highly linear data



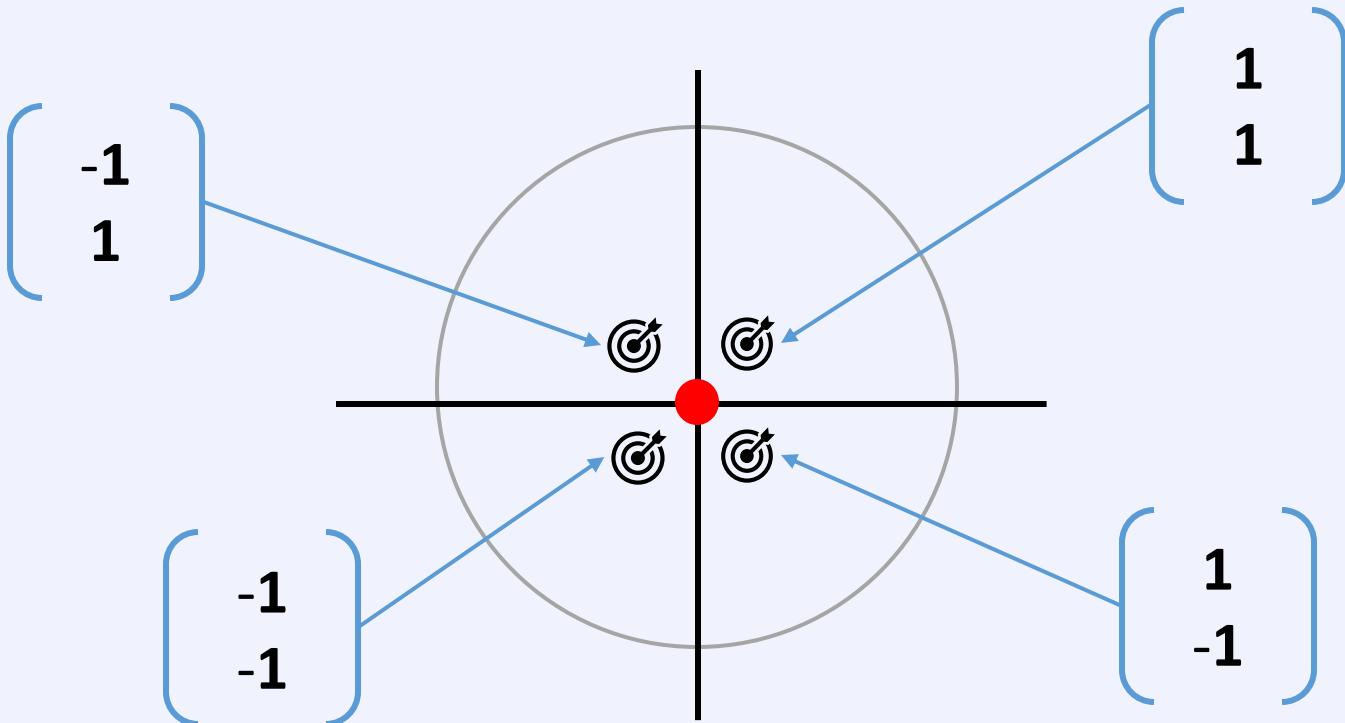
Bias-Variance Tradeoff



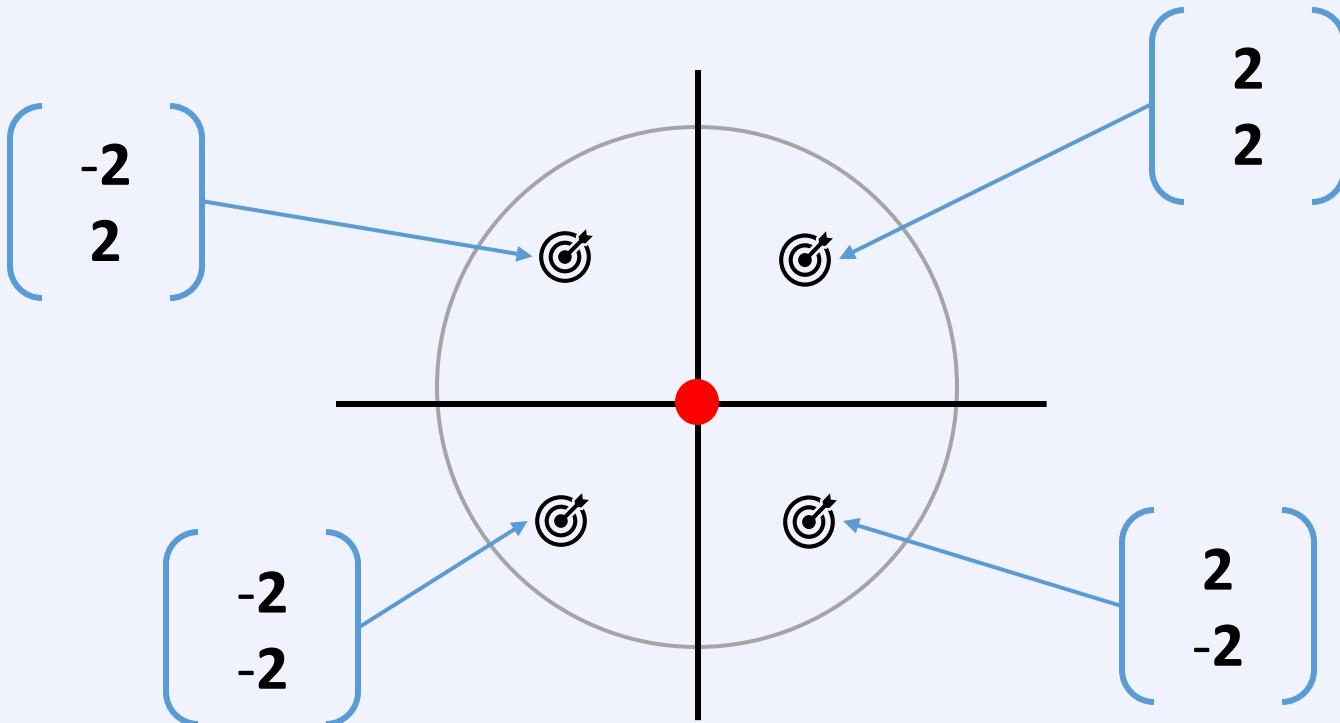
Bias-Variance Tradeoff



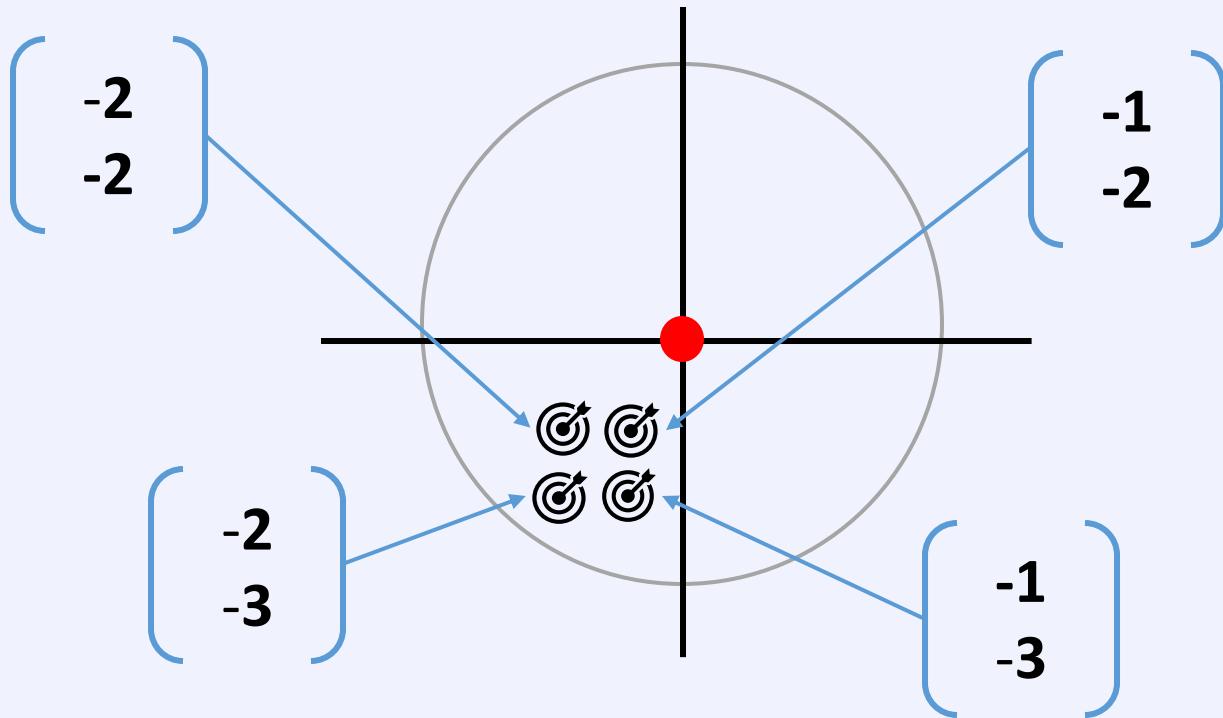
Bias-Variance Tradeoff



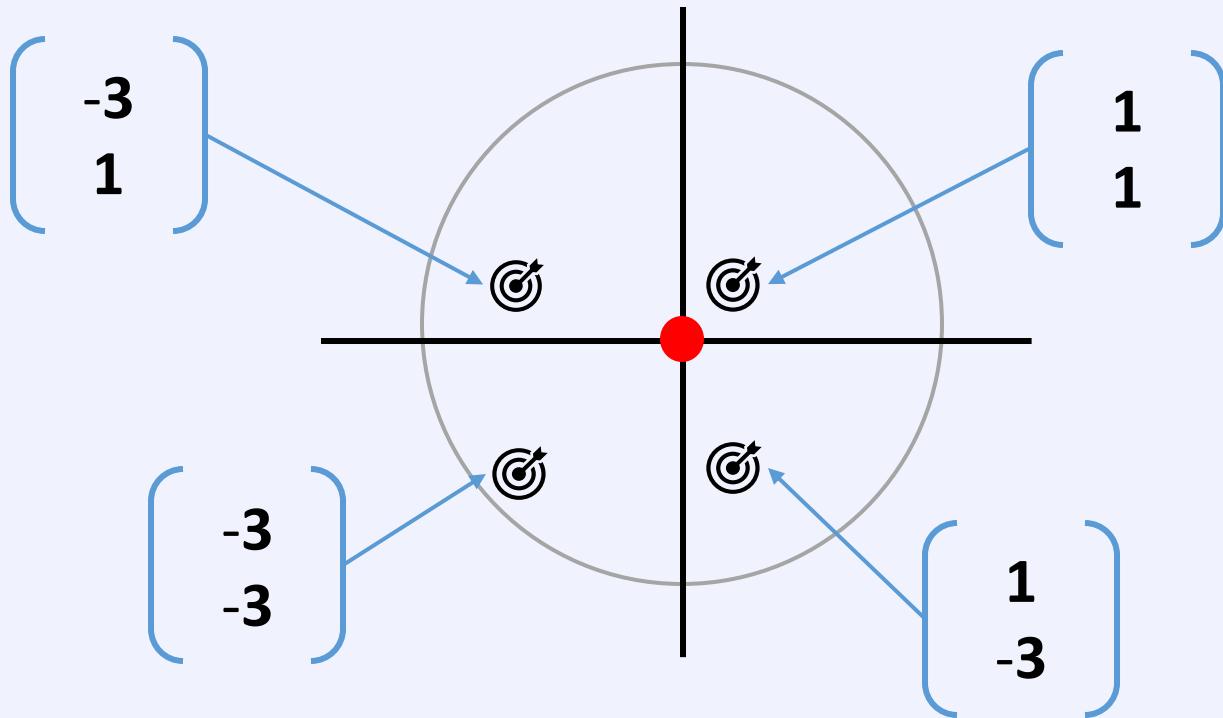
Bias-Variance Tradeoff



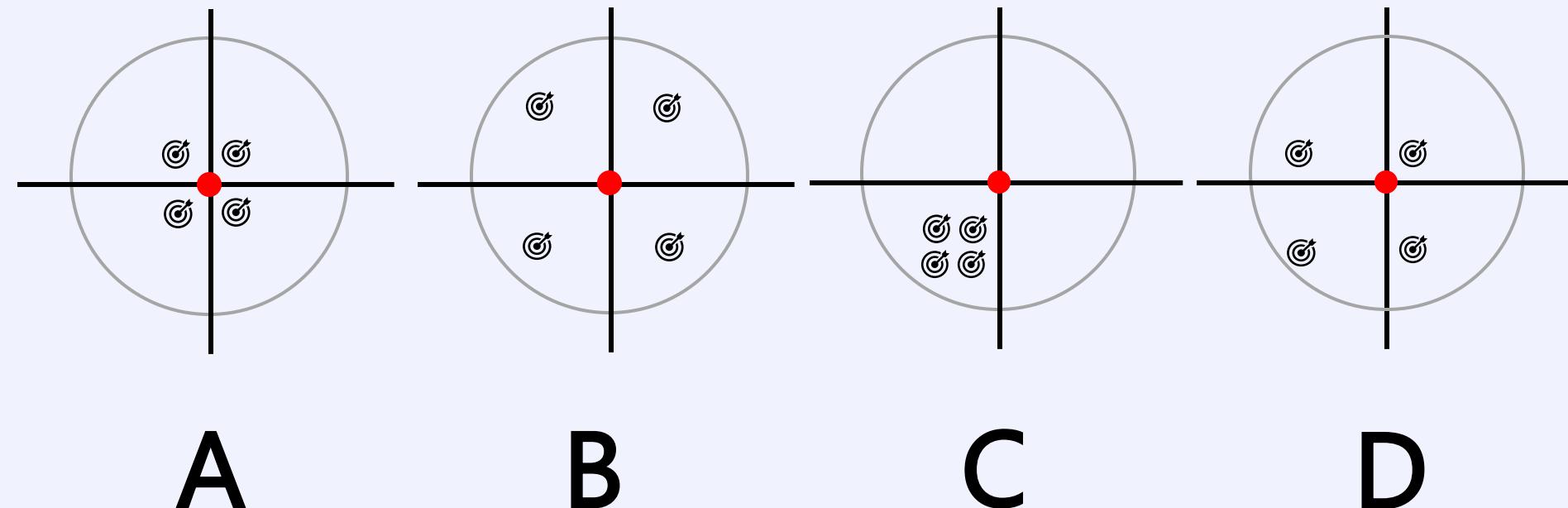
Bias-Variance Tradeoff



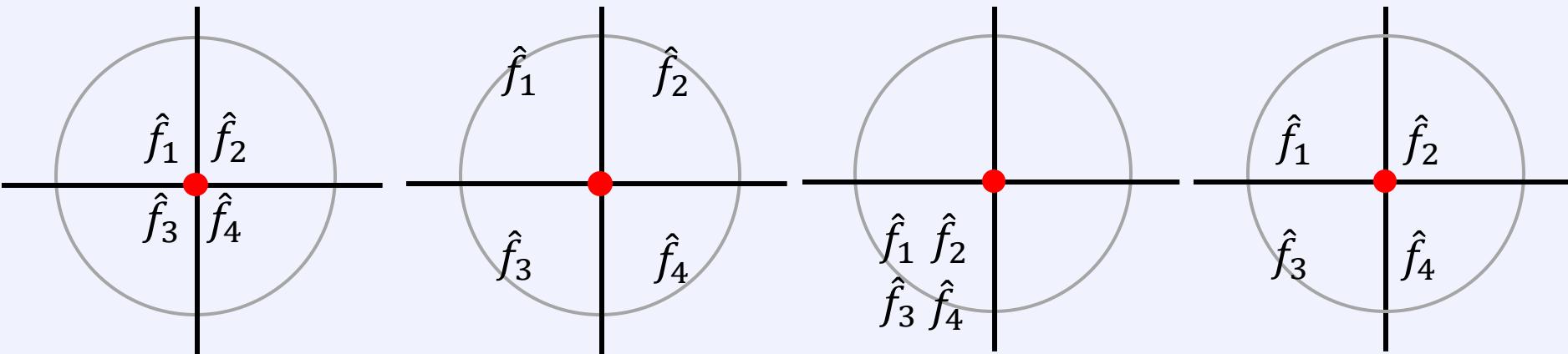
Bias-Variance Tradeoff



Bias-Variance Tradeoff



Bias-Variance Tradeoff



**Unbiased
Small Variance**

**Unbiased
Large Variance**

**Biased
Small Variance**

**Biased
Large Variance**

- True function f $\hat{f}_1, \hat{f}_2, \hat{f}_3, \hat{f}_4, \dots$ Estimates of f on different data sets

Bias-Variance Tradeoff

The shape of the curve for test error is due to a basic tradeoff in the MSE

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

↑
Expectation over all
possible training sets

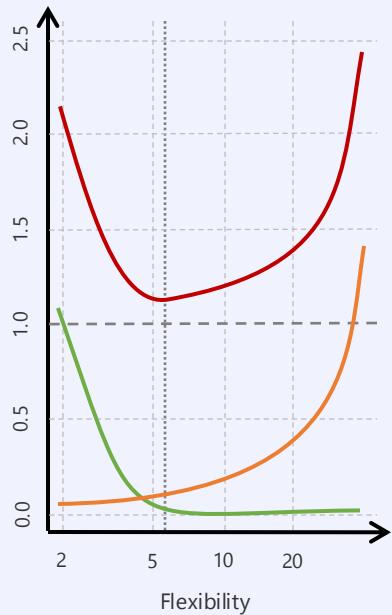
Bias is the systematic deviation of an estimate to the true value

$$Bias \left(\hat{f}(x_0) \right) = E \left(\hat{f}(x_0) - y_0 \right)$$

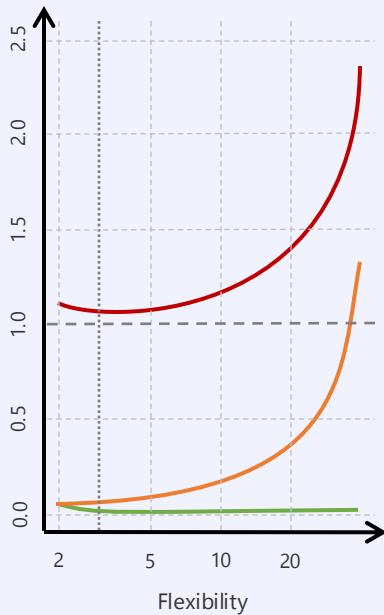
Variance is the variation of the estimate between different training sets

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

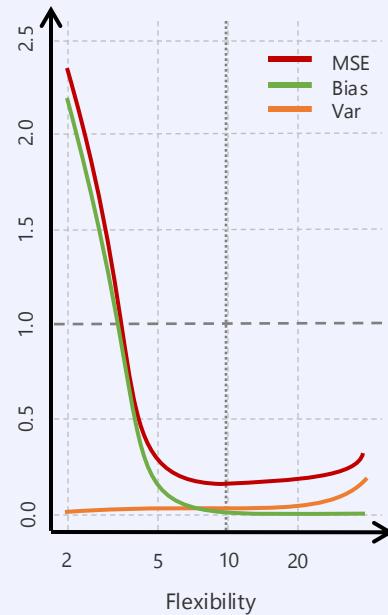
Bias-Variance Decomposition



*Synthetic data example 1
Moderately nonlinear function*



*Synthetic data example 2
Almost linear function*



*Synthetic data example 3
Highly nonlinear function*

Classification

We can measure the **quality of a classifier** using a **loss function**

- typically, we use **misclassification error**
- let I be an indicator function over a predicate p , with $I(p) = 1$ if $p \equiv \text{true}$ and $I(p) = 0$ otherwise
- the training error over n examples is defined as $\frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i)$
- the test error is defined as $\text{avg}(I(y_0 \neq \hat{y}_0))$

The Bayes Classifier

We can minimize test error by the following very simple classifier

$$\arg \max_{j=1,\dots,k} \Pr(Y = j \mid X = x_0)$$

for a classification problem with k classes $1, \dots, k$

This is known as the **Bayes classifier**

- it can be computed (**only**) when we know **the true probability distribution** (e.g. **synthetic data**)
- for all other settings, e.g. real data, we can at best **estimate** it

Example Binary Classification

Data 100 observations over two groups

Bayes decision boundary, i.e. those points where

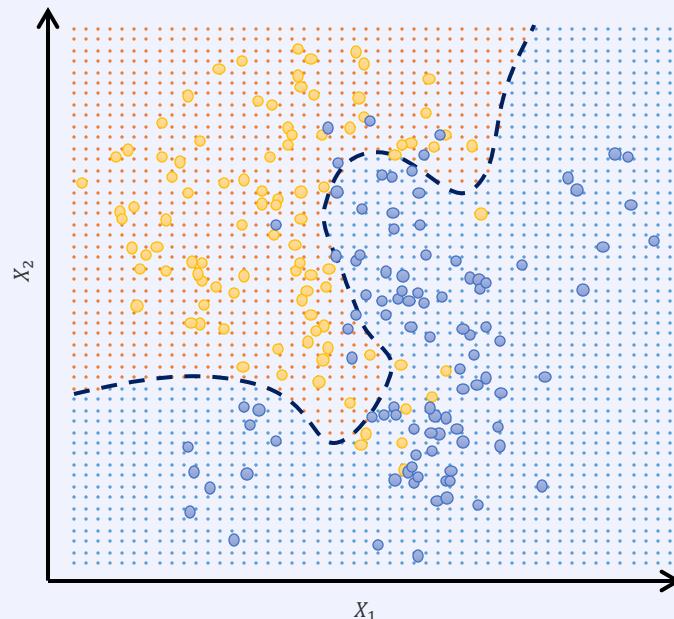
$$\Pr(Y = 1 \mid X = x_0) = 0.5$$

is shown as a dashed line

Bayes error rate, i.e. the irreducible error, is defined as

$$1 - E(\max_{j=1,2} \Pr(Y = j \mid X))$$

In this example, the Bayes error rate is 0.1304



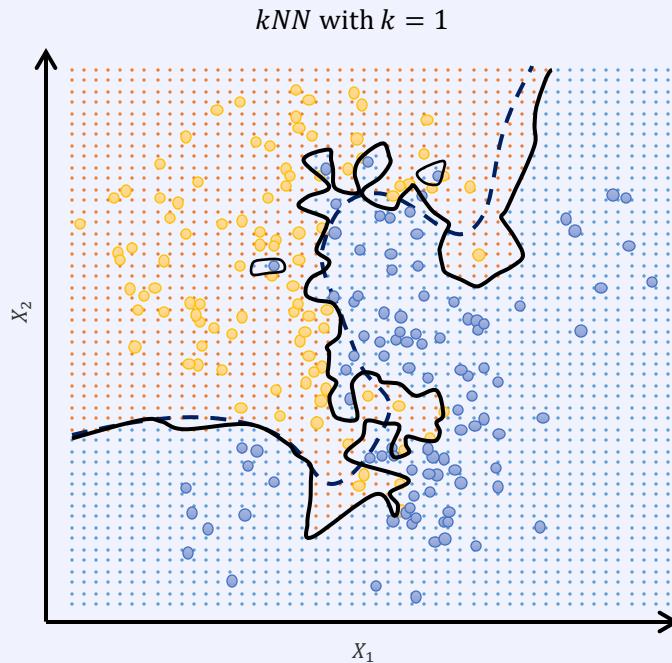
Nearest Neighbors

k -nearest neighbors (**k NN**)

Classifies each point to the majority class among its k nearest neighbors, i.e.

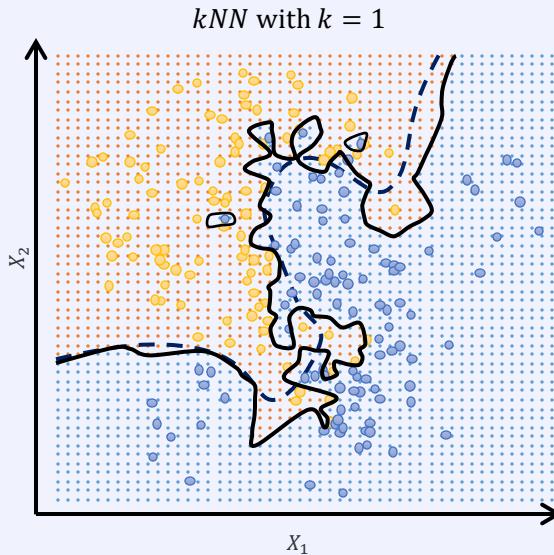
$$\arg \max_{j=1, \dots, k} \frac{1}{k} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

where \mathcal{N}_0 are the k data points nearest to x_0

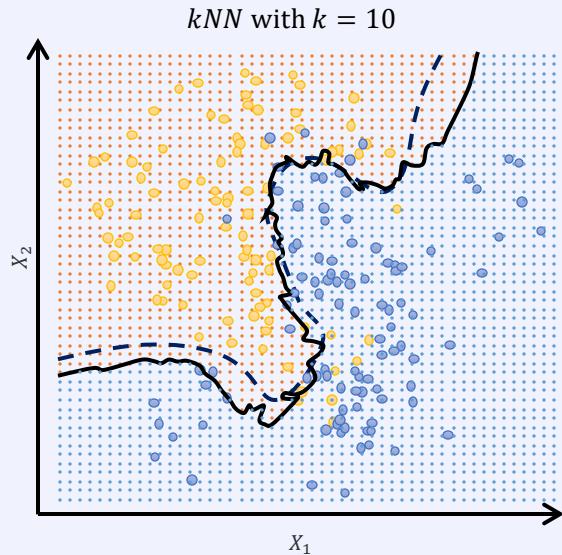


overfitting
model too complex

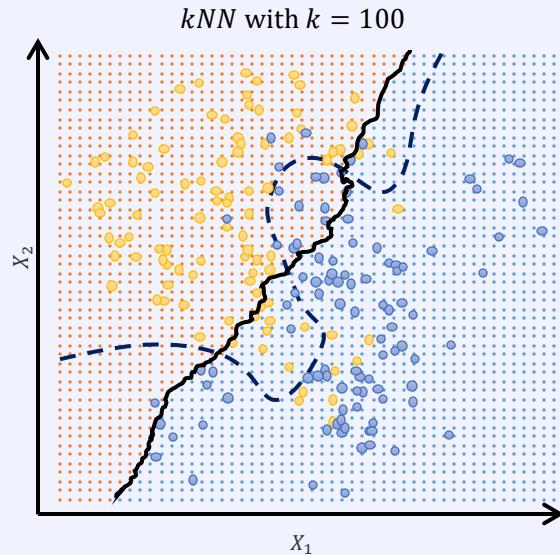
Number of Neighbors



*overfitting
model too complex*

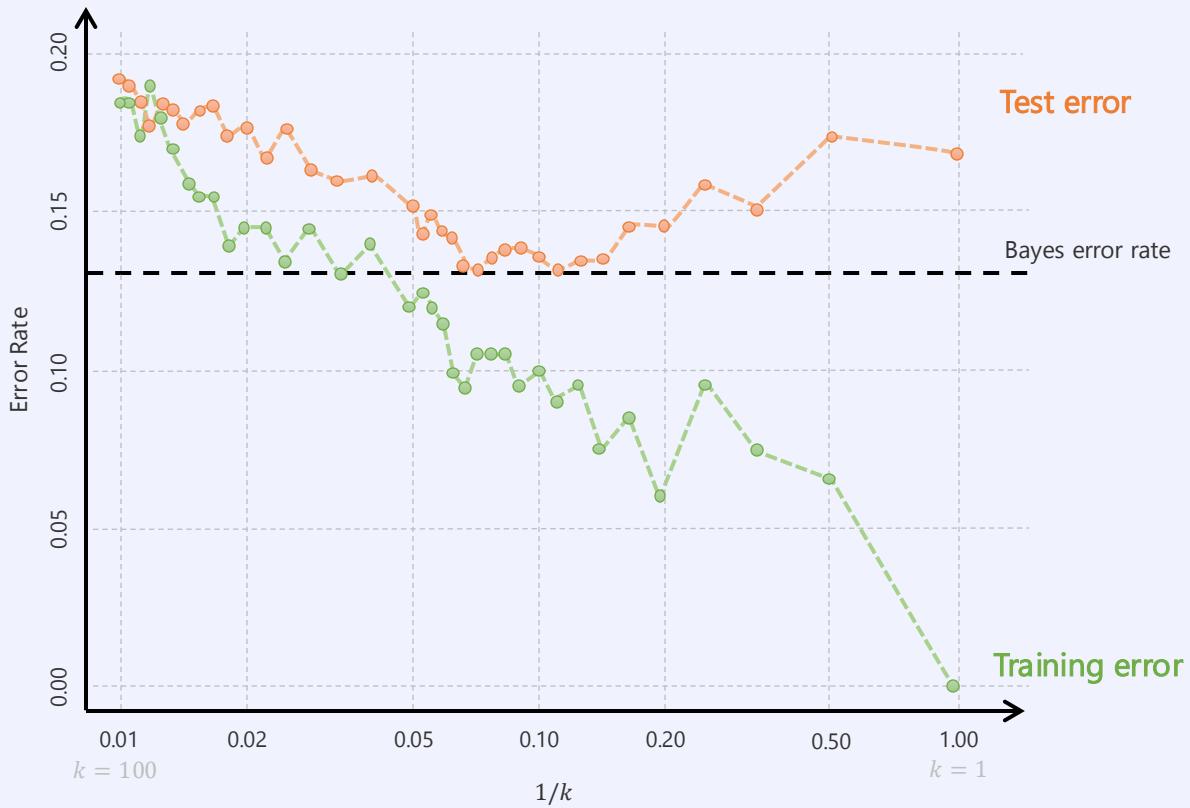


seems just right



*underfitting
model too simple*

Number of Neighbors



Summary

- An introduction to a statistical learning problem
- Given data points (x_i, y_i) for $i = 1, \dots, n$, our goal is to approximate (*learn*) the true f .
 - Inference: The estimate \hat{f} can be used to answer specific questions about f .
 - Prediction: The estimate \hat{f} can be used to make prediction $\hat{Y} = \hat{f}(X)$.
- Supervised and unsupervised learning
- Inherent tradeoffs in statistical learning
 - Accuracy vs. interpretability: More flexible models are often more accurate, but less interpretable.
 - Bias vs. variance: More flexible models often have lower bias, but higher variance.
- Model choice is one of the most crucial decisions.