COM4509/6509

Machine Learning and Adaptive Intelligence

Lecture 1: Introduction

Mike Smith* and Matt Ellis

*m.t.smith@sheffield.ac.uk

About the Module

Textbooks

Structure

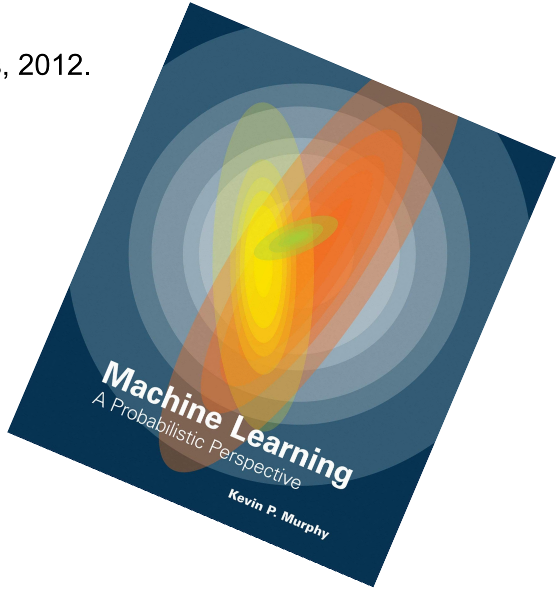
Assessment

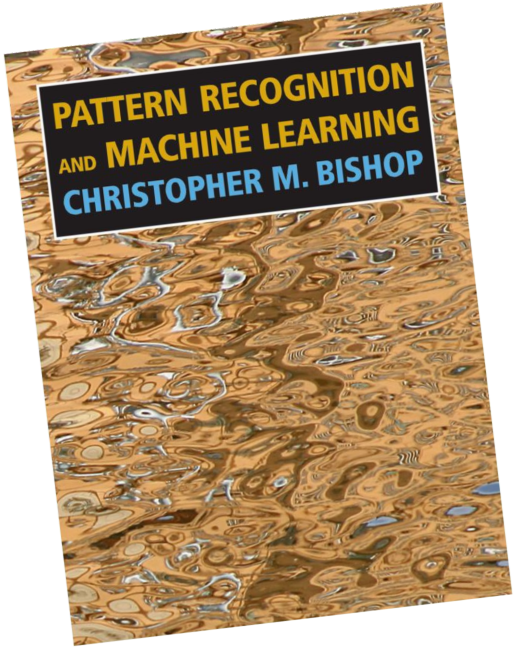
About the Module: Textbooks

Two textbooks in particular (although they are both from a while ago):

Bishop, Christopher M.
Pattern recognition and machine learning. Vol. 4. No. 4. New York: springer, 2006.

Murphy, Kevin P.
Machine learning: a probabilistic perspective. MIT press, 2012.





About the Module: Structure

Here's a rough plan. But we might modify it a little as we go…

Mike's Lectures

Session 1: Introduction to Machine Learning
Session 2: End-to-end machine learning
Session 3: Decision trees and ensemble methods
Session 4: Linear regression
Session 5: Gaussian Processes

Matt's Lectures

Session 6: Logistic regression and automatic differentiation
Session 7: Neural networks
Session 8: Unsupervised learning
Session 9: Generative models
Session 10: Advanced topics in machine learning

Reading for this week

Bishop, Christopher M. *Pattern recognition and machine learning.*

- Section 1.1 (pages 1-12) and
- Sections 1.2.1, 1.2.2 and 1.2.3.

[24 pages in total]

- Textbook available as pdf (see blackboard)

About the Module: Lectures, Labs & Assessment

- Lecture: two hours (Tuesday 10-12)
- Labs: one hour (Thursday 12-1, except week 4, when it’s Wednesday at 1pm)

Assignment

- Two parts:
    - Part 1: Performing machine learning (prediction) tasks on a dataset.
    - Part 2: Apply some of the later tools you will learn (CNNs etc).
- Release: 4th November.
- Deadline: 6th December.
- Feedback: 10th Jan.

Let’s get started

ACTIVITY
4 mins

You are approached by a clinical nephrologist:

- they want to predict 24-72 hour futures in patients on haemodialysis (3/week).
- Time series from 10,000 patients (of weight, blood pressure, etc).
- What will blood pressure be in 24-48 hours? Prob of hospitalising in 72 hours?

Discuss:

- What sort of problem are these? (regression? classification?)
- Is this machine learning or statistical learning? (what do you care about here?)
- How might you test your predictions?
- What else should you think about? (in terms of data, prediction use, etc)

What will blood pressure be in 24-48 hours?

Prob of hospitalising in 72 hours?

Let’s get started

- What sort of problem are these?

What will blood pressure be in 24-48 hours?
Prob of hospitalising in 72 hours?

Let’s get started

- What sort of problem are these?
  - **Regression**: Predicting continuous/ordinal variables
  - **Classification**: Predicting label (categorical) variables
-

What will blood pressure be in 24-48 hours?
Prob of hospitalising in 72 hours?

Let’s get started

- What sort of problem are these?
    - **Regression**: Predicting continuous/ordinal variables
    - **Classification**: Predicting label (categorical) variables
- Is this machine learning or statistical learning? (what do you care about here?)
    - Care about **prediction accuracy** more than e.g. statistical connection between variables etc?

What will blood pressure be in 24-48 hours?
Prob of hospitalising in 72 hours?

Let's get started

- What sort of problem are these?
  - **Regression**: Predicting continuous/ordinal variables
  - **Classification**: Predicting label (categorical) variables
- Is this machine learning or statistical learning? (what do you care about here?)
  - Care about **prediction accuracy** more than e.g. statistical connection between variables etc?
- How might you test your predictions?

What will blood pressure be in 24-48 hours?
Prob of hospitalising in 72 hours?

Let’s get started

- What sort of problem are these?
    - **Regression**: Predicting continuous/ordinal variables
    - **Classification**: Predicting label (categorical) variables
- Is this machine learning or statistical learning? (what do you care about here?)
    - Care about **prediction accuracy** more than e.g. statistical connection between variables etc?
- How might you test your predictions?
    - **Another dataset**? (or at least split into training and test)

What will blood pressure be in 24-48 hours?
Prob of hospitalising in 72 hours?

Let's get started

- What sort of problem are these?
  - **Regression**: Predicting continuous/ordinal variables
  - **Classification**: Predicting label (categorical) variables
- Is this machine learning or statistical learning? (what do you care about here?)
  - Care about **prediction accuracy** more than e.g. statistical connection between variables etc?
- How might you test your predictions?
  - **Another dataset**? (or at least split into training and test)
  - What are you **comparing** against (what would be a 'good' accuracy?)

What will blood pressure be in 24-48 hours?
Prob of hospitalising in 72 hours?

Let’s get started

- What sort of problem are these?
    - **Regression**: Predicting continuous/ordinal variables
    - **Classification**: Predicting label (categorical) variables
- Is this machine learning or statistical learning? (what do you care about here?)
    - Care about **prediction accuracy** more than e.g. statistical connection between variables etc?
- How might you test your predictions?
    - **Another dataset**? (or at least split into training and test)
    - What are you **comparing** against (what would be a ‘good’ accuracy?)
    - Is accuracy a good metric? (MAE, MSE, RMSE, NLPD, ROC, etc)

What will blood pressure be in 24-48 hours?
Prob of hospitalising in 72 hours?

Let’s get started

- What sort of problem are these?
  - **Regression**: Predicting continuous/ordinal variables
  - **Classification**: Predicting label (categorical) variables
- Is this machine learning or statistical learning? (what do you care about here?)
  - Care about **prediction accuracy** more than e.g. statistical connection between variables etc?
- How might you test your predictions?
  - **Another dataset**? (or at least split into training and test)
  - What are you **comparing** against (what would be a ‘good’ accuracy?)
  - Is accuracy a good metric? (MAE, MSE, RMSE, NLPD, ROC, etc)
  - Are there **subpopulations** that are significantly more likely to be misclassified?

What will blood pressure be in 24-48 hours?
Prob of hospitalising in 72 hours?

Let’s get started

- What sort of problem are these?
  - **Regression**: Predicting continuous/ordinal variables
  - **Classification**: Predicting label (categorical) variables
- Is this machine learning or statistical learning? (what do you care about here?)
  - Care about **prediction accuracy** more than e.g. statistical connection between variables etc?
- How might you test your predictions?
  - **Another dataset**? (or at least split into training and test)
  - What are you **comparing** against (what would be a ‘good’ accuracy?)
  - Is accuracy a good metric? (MAE, MSE, RMSE, NLPD, ROC, etc)
  - Are there **subpopulations** that are significantly more likely to be misclassified?
  - **Uncertainty** quantification?

What will blood pressure be in 24-48 hours?
Prob of hospitalising in 72 hours?

Let’s get started

- What sort of problem are these?
  - **Regression**: Predicting continuous/ordinal variables
  - **Classification**: Predicting label (categorical) variables
- Is this machine learning or statistical learning? (what do you care about here?)
  - Care about **prediction accuracy** more than e.g. statistical connection between variables etc?
- How might you test your predictions?
  - **Another dataset**? (or at least split into training and test)
  - What are you **comparing** against (what would be a ‘good’ accuracy?)
  - Is accuracy a good metric? (MAE, MSE, RMSE, NLPD, ROC, etc)
  - Are there **subpopulations** that are significantly more likely to be misclassified?
  - **Uncertainty** quantification?
- What else should you think about?

What will blood pressure be in 24-48 hours?
Prob of hospitalising in 72 hours?

Let’s get started

- What sort of problem are these?
  - **Regression**: Predicting continuous/ordinal variables
  - **Classification**: Predicting label (categorical) variables
- Is this machine learning or statistical learning? (what do you care about here?)
  - Care about **prediction accuracy** more than e.g. statistical connection between variables etc?
- How might you test your predictions?
  - **Another dataset**? (or at least split into training and test)
  - What are you **comparing** against (what would be a ‘good’ accuracy?)
  - Is accuracy a good metric? (MAE, MSE, RMSE, NLPD, ROC, etc)
  - Are there **subpopulations** that are significantly more likely to be misclassified?
  - **Uncertainty** quantification?
- What else should you think about?
  - **Data** quality?

What will blood pressure be in 24-48 hours?
Prob of hospitalising in 72 hours?

Let’s get started

- What sort of problem are these?
  - **Regression**: Predicting continuous/ordinal variables
  - **Classification**: Predicting label (categorical) variables
- Is this machine learning or statistical learning? (what do you care about here?)
  - Care about **prediction accuracy** more than e.g. statistical connection between variables etc?
- How might you test your predictions?
  - **Another dataset**? (or at least split into training and test)
  - What are you **comparing** against (what would be a ‘good’ accuracy?)
  - Is accuracy a good metric? (MAE, MSE, RMSE, NLPD, ROC, etc)
  - Are there **subpopulations** that are significantly more likely to be misclassified?
  - **Uncertainty** quantification?
- What else should you think about?
  - **Data** quality? What will it be **used for**?

What will blood pressure be in 24-48 hours?
Prob of hospitalising in 72 hours?

Let’s get started

- What sort of problem are these?
  - **Regression**: Predicting continuous/ordinal variables
  - **Classification**: Predicting label (categorical) variables
- Is this machine learning or statistical learning? (what do you care about here?)
  - Care about **prediction accuracy** more than e.g. statistical connection between variables etc?
- How might you test your predictions?
  - **Another dataset**? (or at least split into training and test)
  - What are you **comparing** against (what would be a ‘good’ accuracy?)
  - Is accuracy a good metric? (MAE, MSE, RMSE, NLPD, ROC, etc)
  - Are there **subpopulations** that are significantly more likely to be misclassified?
  - **Uncertainty** quantification?
- What else should you think about?
  - **Data** quality? What will it be **used for**? Future **datashift**?

What will blood pressure be in 24-48 hours?
Prob of hospitalising in 72 hours?

Let’s get started

- What sort of problem are these?
  - **Regression**: Predicting continuous/ordinal variables
  - **Classification**: Predicting label (categorical) variables
- Is this machine learning or statistical learning? (what do you care about here?)
  - Care about **prediction accuracy** more than e.g. statistical connection between variables etc?
- How might you test your predictions?
  - **Another dataset**? (or at least split into training and test)
  - What are you **comparing** against (what would be a ‘good’ accuracy?)
  - Is accuracy a good metric? (MAE, MSE, RMSE, NLPD, ROC, etc)
  - Are there **subpopulations** that are significantly more likely to be misclassified?
  - **Uncertainty** quantification?
- What else should you think about?
  - **Data** quality? What will it be **used for**? Future **datashift**? **Implementation**? Interpretability?

Machine Learning

We want to develop an algorithm that will make a **prediction** using a **model** and some **data**.

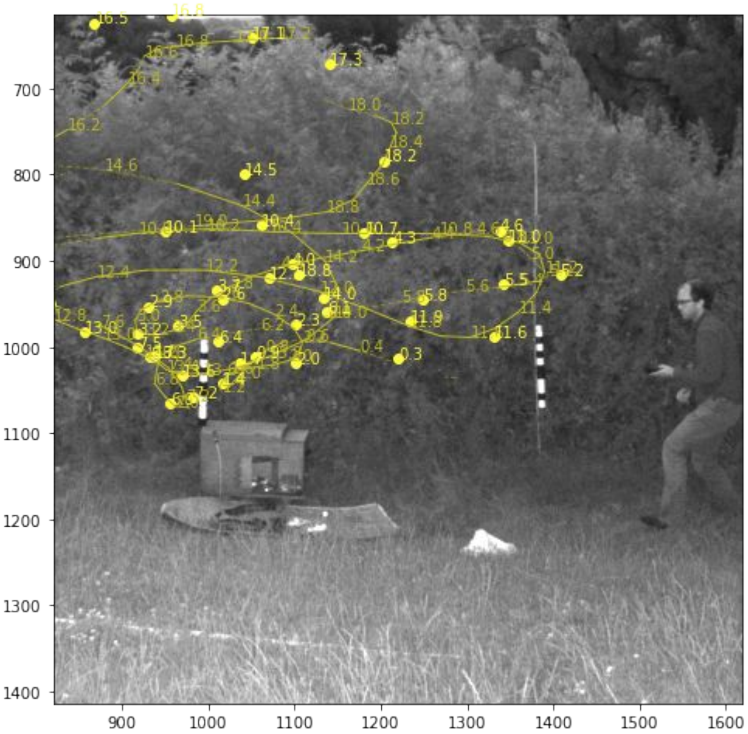
Examples:

Machine Learning

We want to develop an algorithm that will make a **prediction** using a **model** and some **data**.

Examples:

- Reconstruction of **bumblebee flight path** from photos

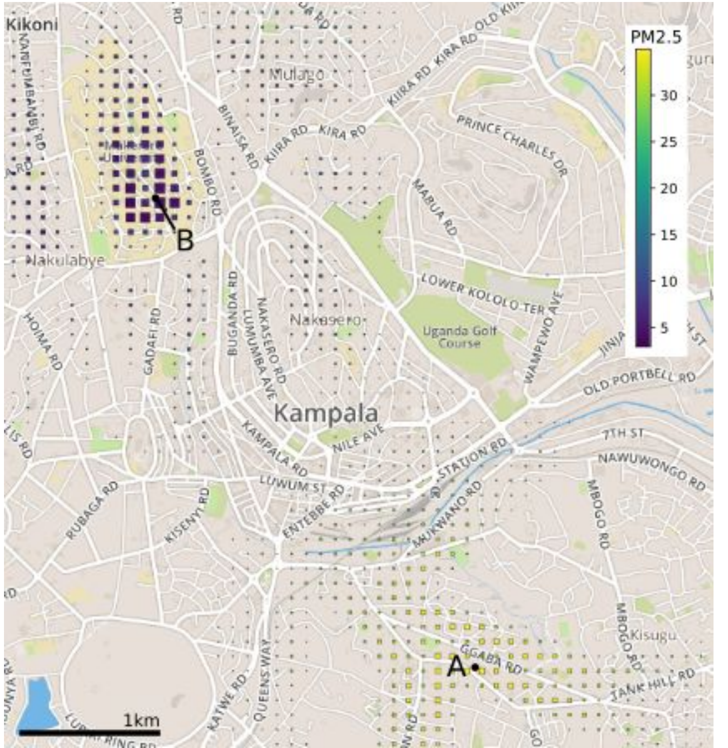


Machine Learning

We want to develop an algorithm that will make a **prediction** using a **model** and some **data**.

Examples:

- Reconstruction of bumblebee flight path from photos
- **Optimising air pollution sensor placement**

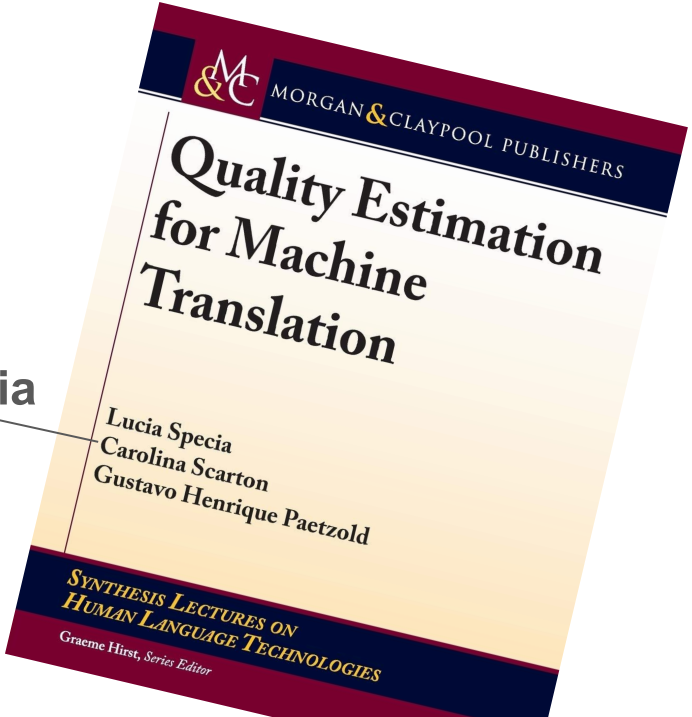


Machine Learning

We want to develop an algorithm that will make a **prediction** using a **model** and some **data**.

Examples:

- Reconstruction of bumblebee flight path from photos
- Optimising air pollution sensor placement
- **Detect/remove abusive posts on social media**



Machine Learning

We want to develop an algorithm that will make a **prediction** using a **model** and some **data**.

Examples:

- Reconstruction of bumblebee flight path from photos
- Optimising air pollution sensor placement
- Detect/remove abusive posts on social media
- **Segmenting customer groups**

Machine Learning

We want to develop an algorithm that will make a **prediction** using a **model** and some **data**.

Examples:

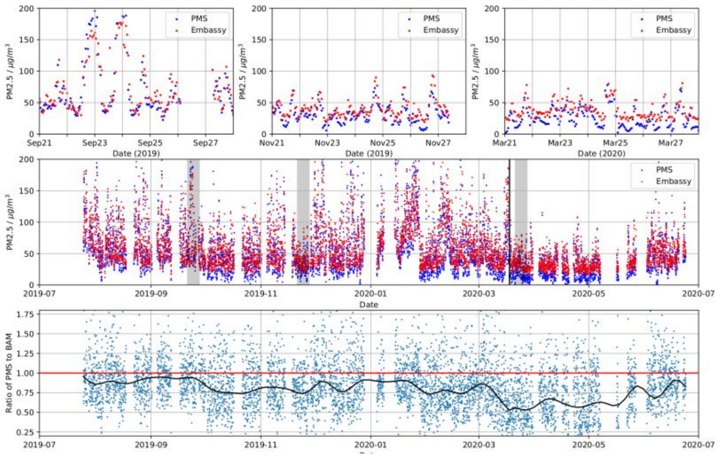
- Reconstruction of bumblebee flight path from photos
- Optimising air pollution sensor placement
- Detect/remove abusive posts on social media
- Segmenting customer groups
- **Optimising delivery distribution routing**

Machine Learning

We want to develop an algorithm that will make a **prediction** using a **model** and some **data**.

Examples:

- Reconstruction of bumblebee flight path from photos
- Optimising air pollution sensor placement
- Detect/remove abusive posts on social media
- Segmenting customer groups
- Optimising delivery distribution routing
- **Calibrating low-cost sensor networks**



Machine Learning

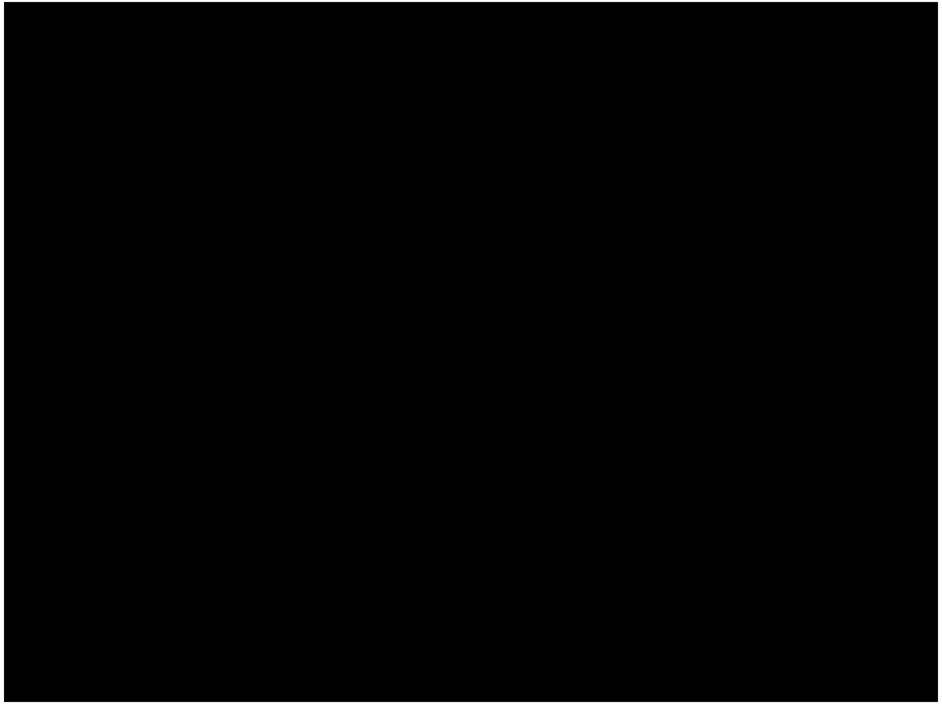
More examples:

- **Recommendation Systems**

Machine Learning

More examples:

- Recommendation Systems
- **AlphaFold**



https://alphafold.ebi.ac.uk/entry/Q8I3H7

Machine Learning

More examples:

- Recommendation Systems
- AlphaFold
- **Autonomous Driving?**

Definitions

Training set - Fit parameters, or similar

Validation set - Tune hyperparameters

Test set - Checks if it works on a held-out dataset

“The literature on machine learning often reverses the meaning of 'validation' and 'test' sets” - Ripley, 2009.

(draw fig)

Ripley, Brian D. (2009). Pattern recognition and neural networks.

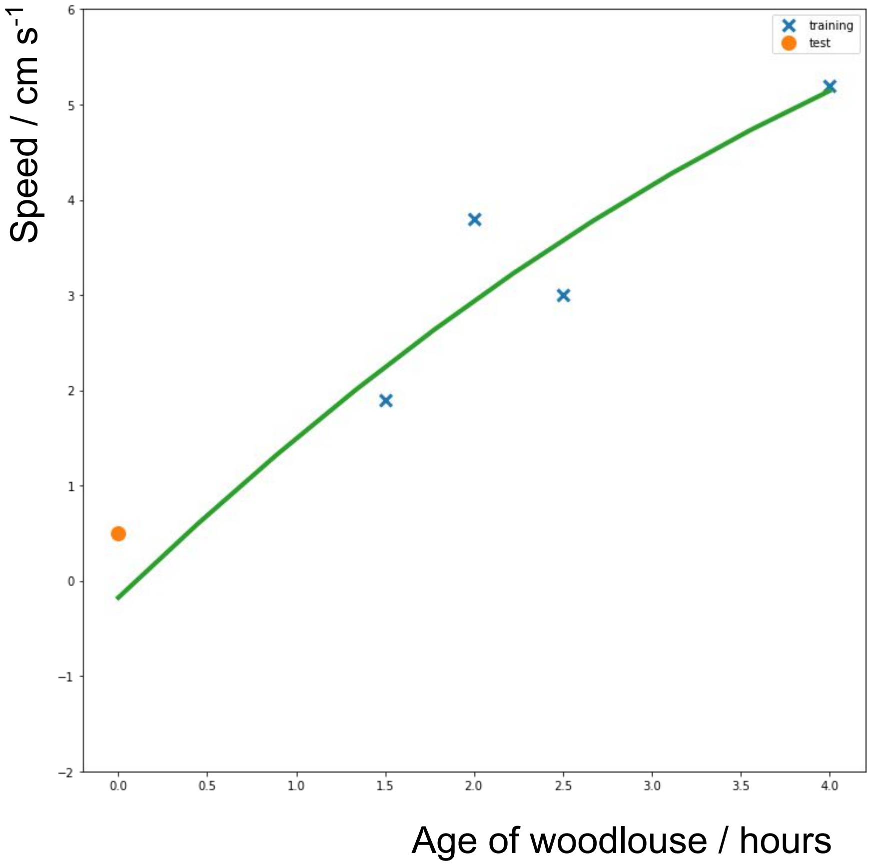
Definitions

Model: 2nd order polynomial

Data: 5, one-dimensional, points

We **train** on 4 points and use the last point as **validation**

We can repeat this, **holding out** a different point each time…



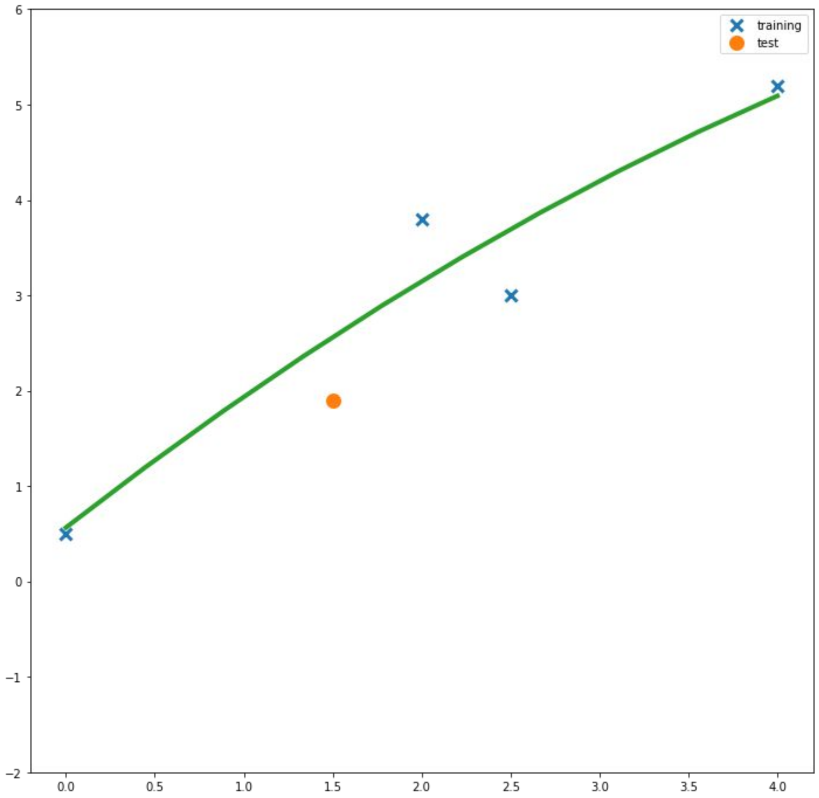
Definitions

Model: 2nd order polynomial

Data: 5, one-dimensional, points

We **train** on 4 points and use the last point as **validation**

We can repeat this, **holding out** a different point each time…



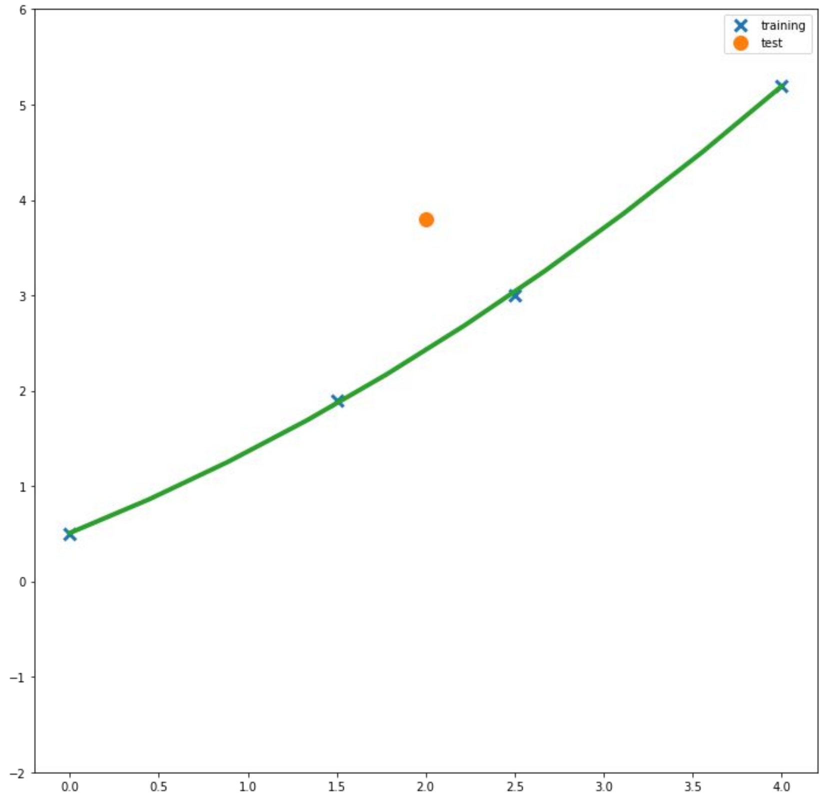
Definitions

Model: 2nd order polynomial

Data: 5, one-dimensional, points

We **train** on 4 points and use the last point as **validation**

We can repeat this, **holding out** a different point each time…



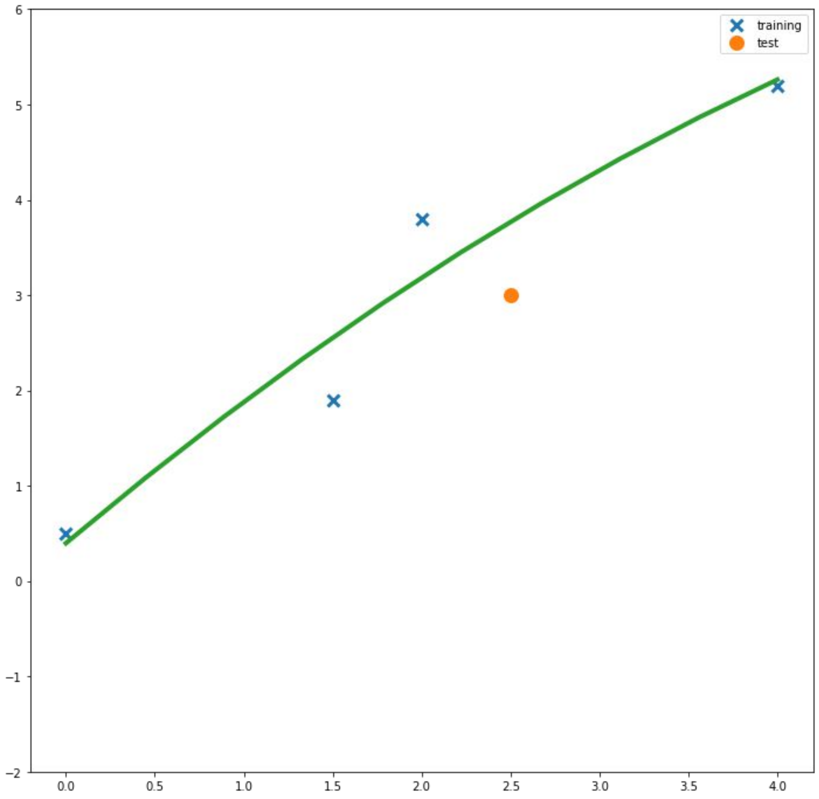
Definitions

Model: 2nd order polynomial

Data: 5, one-dimensional, points

We **train** on 4 points and use the last point as **validation**

We can repeat this, **holding out** a different point each time…



Definitions

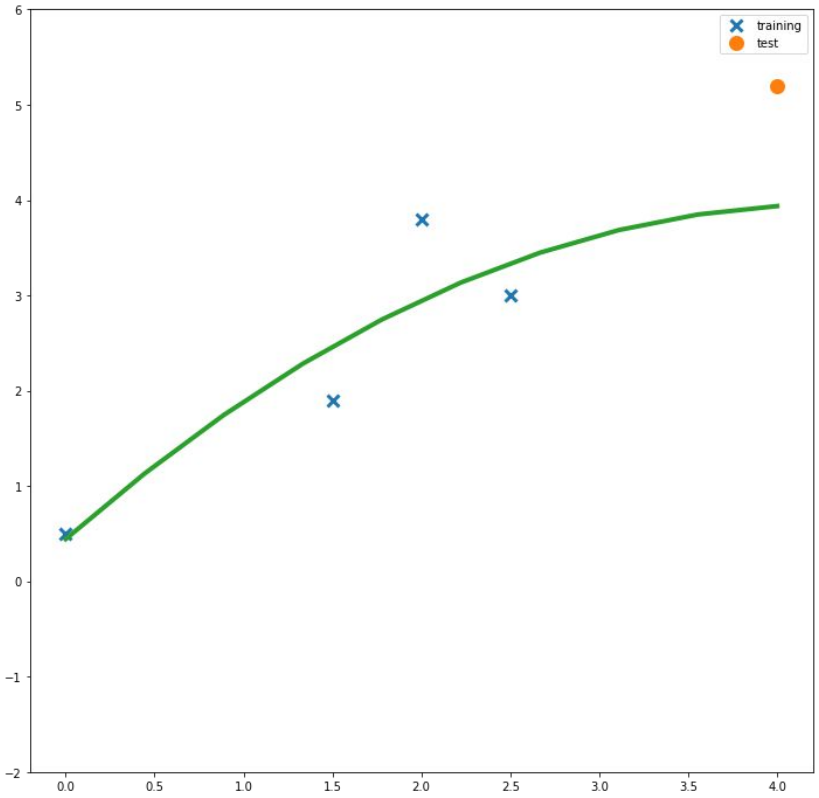
Model: 2nd order polynomial

Data: 5, one-dimensional, points

We **train** on 4 points and use the last point as **validation**

We can repeat this, **holding out** a different point each time…

This is called
  leave-one-out cross validation



Definitions

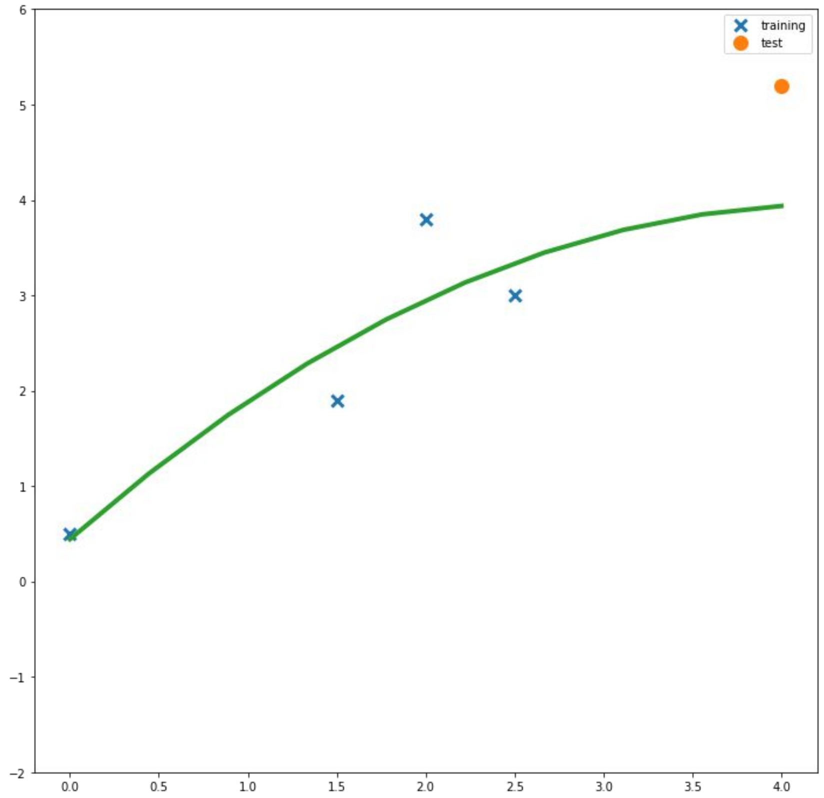
Model: 2nd order polynomial

Data: 5, one-dimensional, points

We **train** on 4 points and use the last point as **validation**

We can repeat this, **holding out** a different point each time…

This is called
leave-one-out cross validation



Alternative: k-fold cross validation.

Definitions

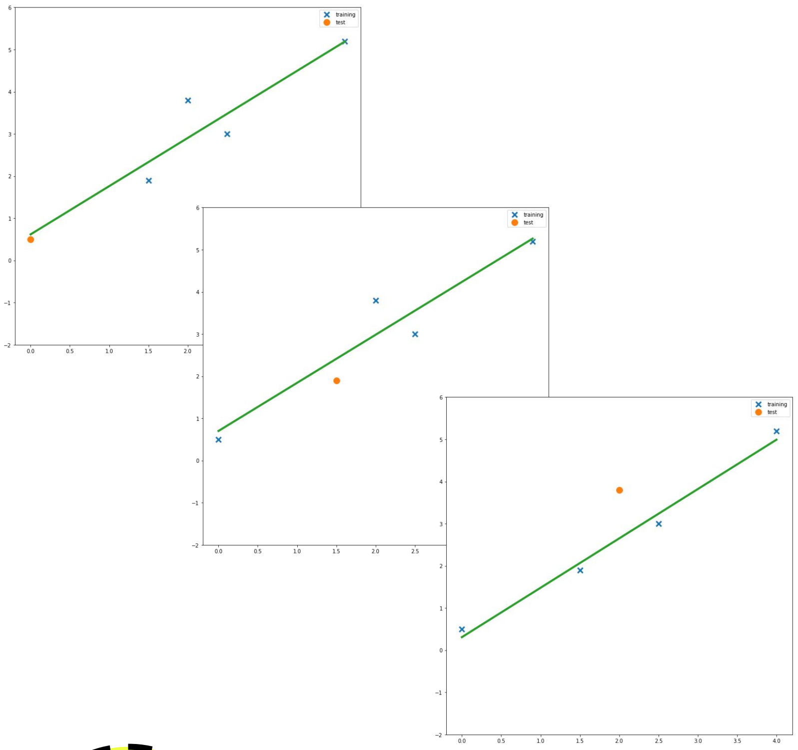
How do we decide between models?

What about a 1st-order polynomial?

We can run the cross-validation on these again, and look at some metric of error:

Order	MSE
0th	4.02
1st	0.39
2nd	1.00

The best model appears to be the 1st order model. But to properly assess this we need to use some held-out **test** data that we haven't yet looked at. Using the data we've used for training and hyperparameter/model selection will artificially inflate our estimate.



etc..

Avoid circular analysis

Definitions

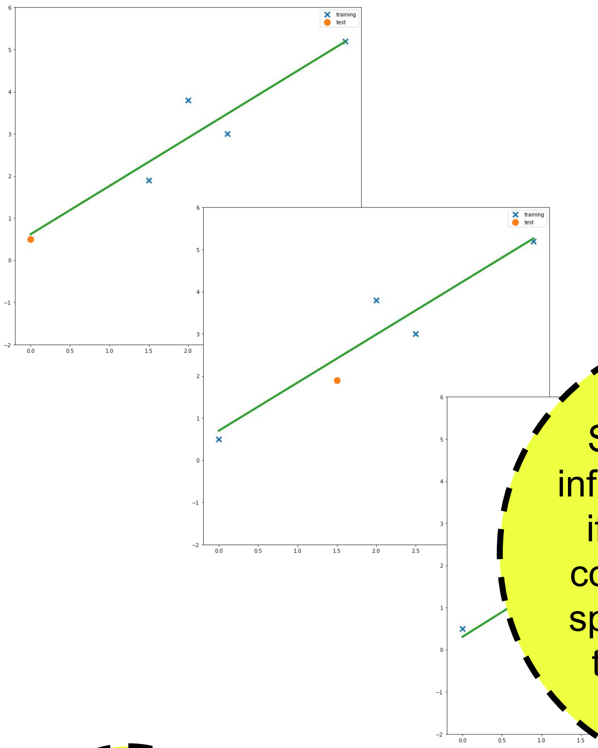
How do we decide between models?

What about a 1st-order polynomial?

We can run the cross-validation on these again, and look at some metric of error:

Order	MSE
0th	4.02
1st	0.39
2nd	1.00

The best model appears to be the 1st order model. But to properly assess this we need to use some held-out **test** data that we haven't yet looked at. Using the data we've used for training and hyperparameter/model selection will artificially inflate our estimate.



Similar accuracy inflation can happen if data points are correlated and you split them between training and test

Avoid circular analysis

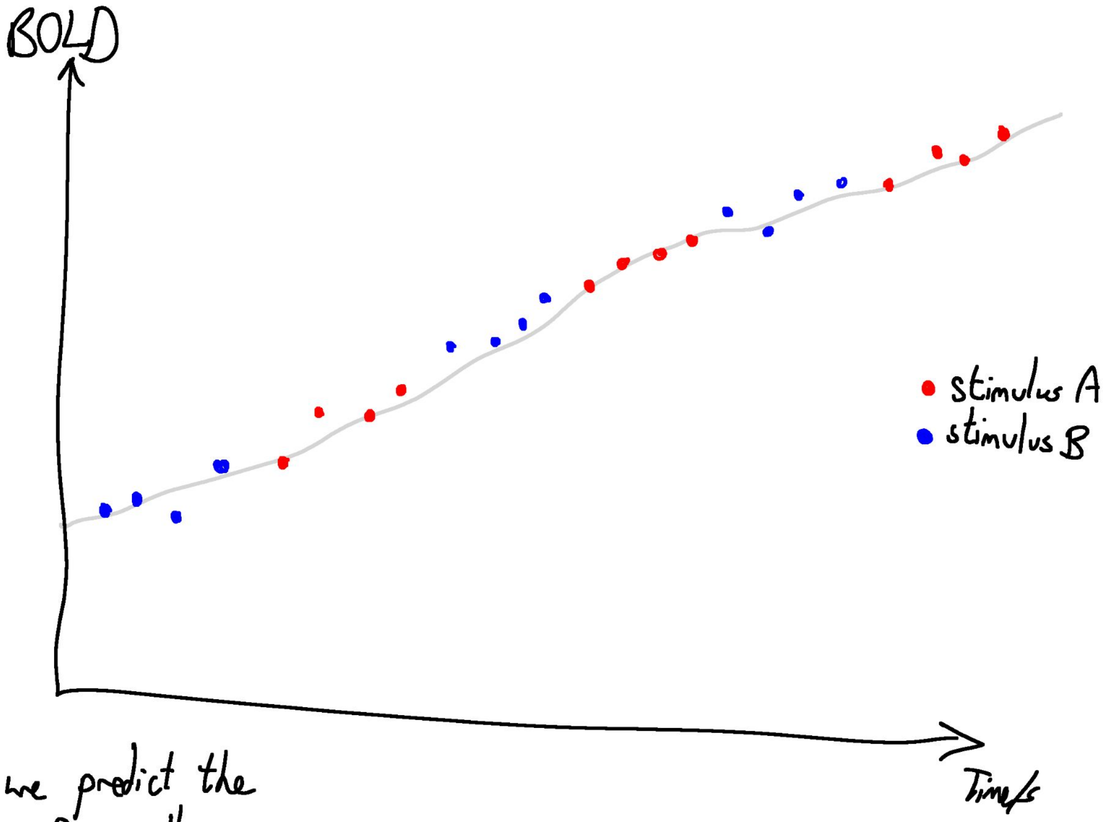
etc..

Detour…

Similar accuracy inflation can happen if data points are correlated and you split them between training and test

If the observations are correlated then the correlation must not connect between training and test sets.

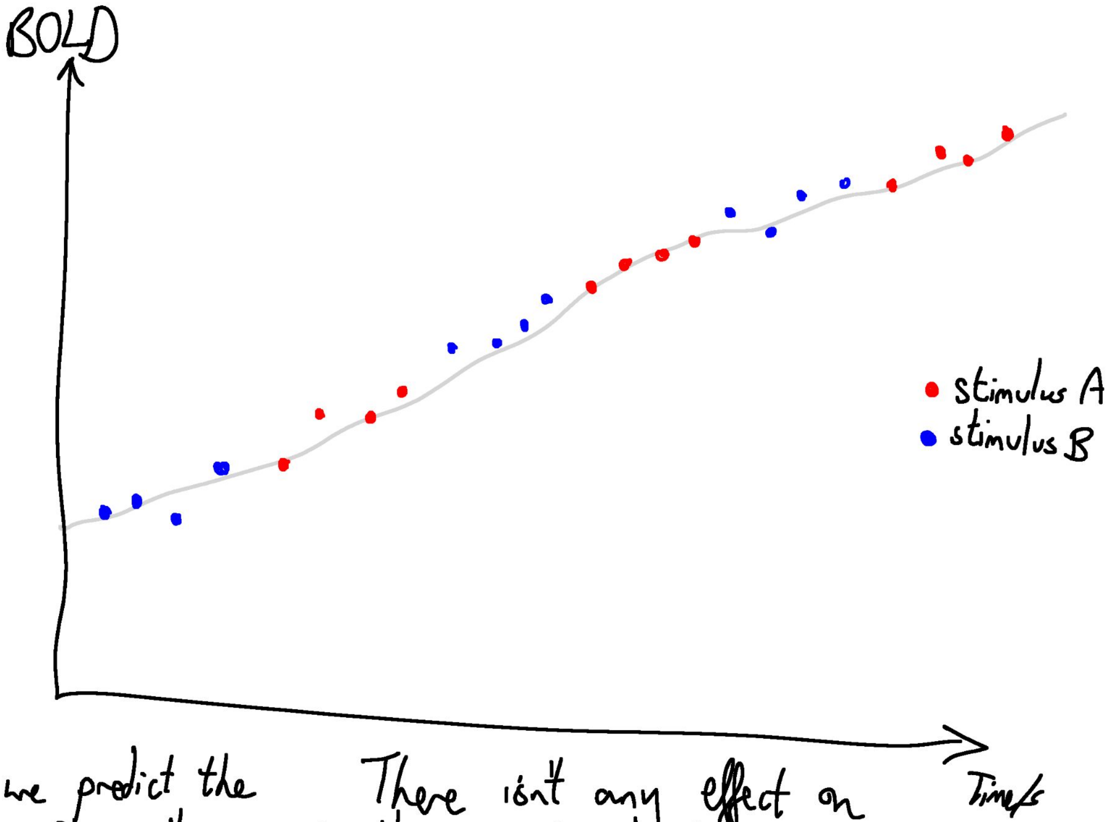
Similar accuracy inflation can happen if data points are correlated and you split them between training and test



Can we predict the stimulus from the BOLD signal?

If the observations are correlated then the correlation must not connect between training and test sets.

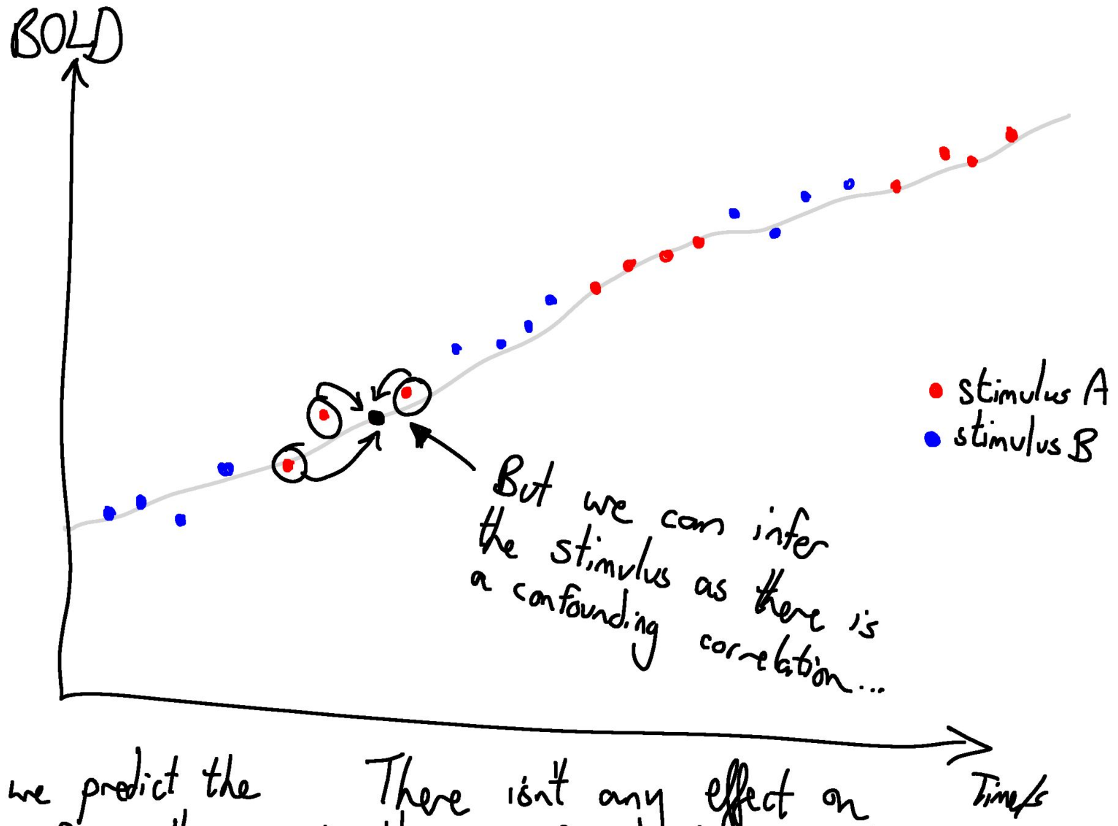
Similar accuracy inflation can happen if data points are correlated and you split them between training and test



Can we predict the stimulus from the BOLD signal? → There isn't any effect on the BOLD signal from the stimulus so our analysis shouldn't be able to...

If the observations are correlated then the correlation must not connect between training and test sets.

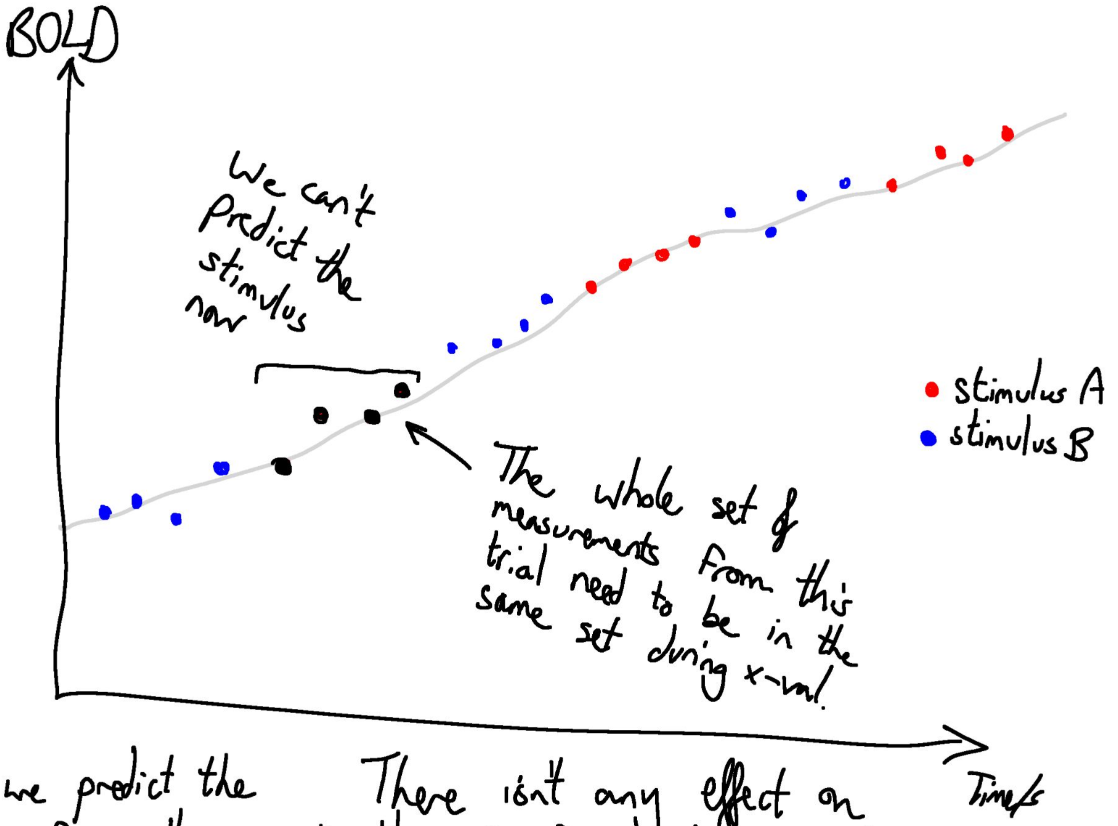
Similar accuracy inflation can happen if data points are correlated and you split them between training and test



Can we predict the stimulus from the BOLD signal? → There isn't any effect on the BOLD signal from the stimulus so our analysis shouldn't be able to...

If the observations are correlated then the correlation must not connect between training and test sets.

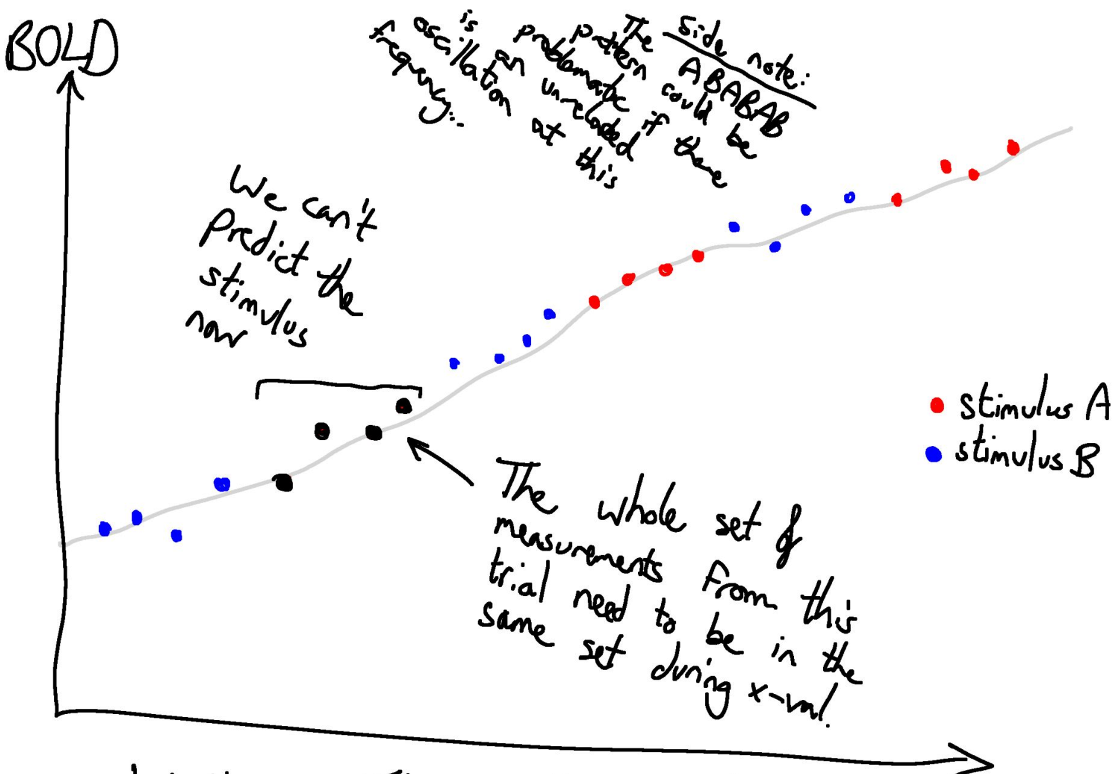
Similar accuracy inflation can happen if data points are correlated and you split them between training and test



Can we predict the stimulus from the BOLD signal? → There isn't any effect on the BOLD signal from the stimulus so our analysis shouldn't be able to...

If the observations are correlated then the correlation must not connect between training and test sets.

Similar accuracy inflation can happen if data points are correlated and you split them between training and test



Can we predict the stimulus from the BOLD signal? → There isn't any effect on the BOLD signal from the stimulus so our analysis shouldn't be able to...

If the observations are correlated then the correlation must not connect between training and test sets.

Similar accuracy inflation can happen if data points are correlated and you split them between training and test

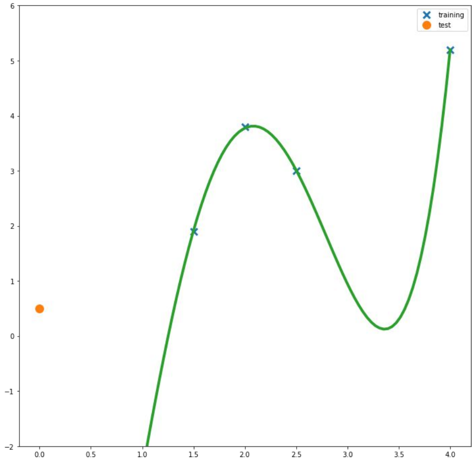
Quick Discussion

What happens at higher orders?

How well does it do,

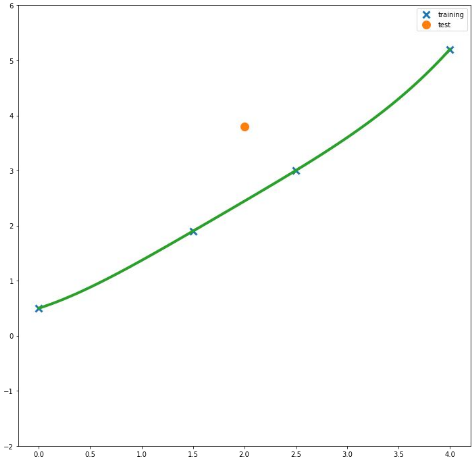
- on the training set?

- on the held-out cross validation data?



Overfitting?

3rd order polynomial



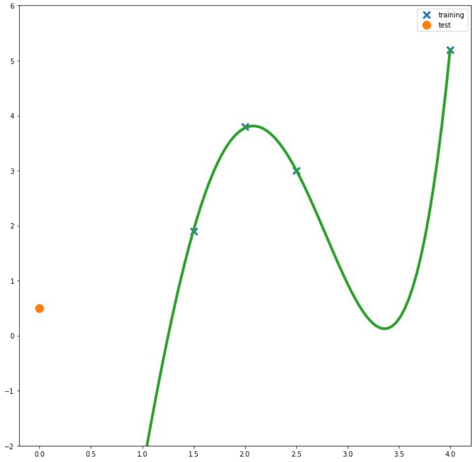
Quick Discussion

What happens at higher orders?

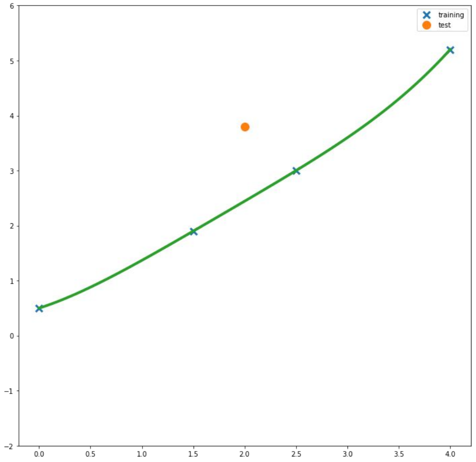
How well does it do,

- on the training set?

- on the held-out cross validation data?



Overfitting is detected if one does much better on the training data than on test data.



Quick Discussion

What happens at higher orders?

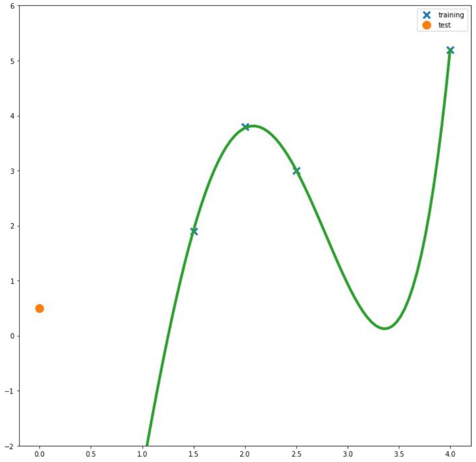
How well does it do,

- on the training set?

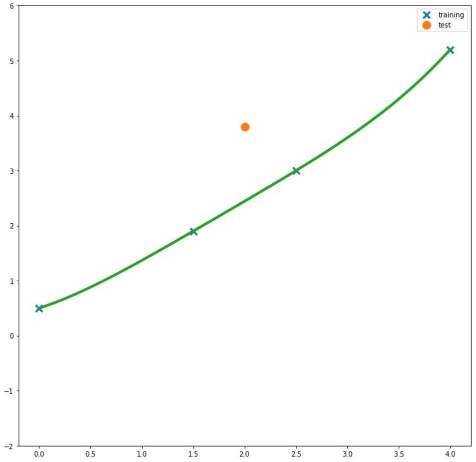
- on the held-out cross validation data?

A model's ability to predict other data (especially from an unseen dataset) is called **generalisation**.

Consider also **extrapolation** vs **interpolation**. If you want your model to extrapolate, you need to think about this during your train/test/validation split.



Overfitting is detected if one does much better on the training data than on test data.



Definitions

This has been an example of a **supervised** learning problem.

Definitions

This has been an example of a **supervised** learning problem.

- If ‘y’ is continuous [or at least ordinal]: regression

Definitions

This has been an example of a **supervised** learning problem.

- If ‘y’ is continuous [or at least ordinal]: regression
- If ‘y’ is a categorical variable: classification

Definitions

This has been an example of a **supervised** learning problem.

- If ‘y’ is continuous [or at least ordinal]: regression
- If ‘y’ is a categorical variable: classification

Unsupervised might be finding,

Definitions

This has been an example of a **supervised** learning problem.

- If ‘y’ is continuous [or at least ordinal]: regression
- If ‘y’ is a categorical variable: classification

Unsupervised might be finding,

- Similar groups [clustering]
- A probability density function [density estimation]
- A better representation [e.g. dimensionality reduction]

Definitions

This has been an example of a **supervised** learning problem.

- If ‘y’ is continuous [or at least ordinal]: regression
- If ‘y’ is a categorical variable: classification

Unsupervised might be finding,

- Similar groups [clustering]
- A probability density function [density estimation]
- A better representation [e.g. dimensionality reduction]

Other types of machine learning exist: Reinforcement learning, active learning, etc.

Objective Function

In the regression example I skipped over how we did the training. I used ‘ordinary least squares’ to find an appropriate prediction.

- Supervised learning usually involves wanting to minimise an **objective function**. The sum of squared errors is often used (we will look at why later).

Sketch out on board…
(Note 1)

Problems with Machine Learning

- Often not enough (good quality / representative) training data
- Irrelevant features
- Overfitting/Underfitting
- Failure to generalise
  - E.g. Will it work with data collected next year?
- Uncertainty quantification
  - “Don’t know” maybe should be a valid option.
- Interpretability
  - Do we trust it in safety critical systems.
- Adversarial Examples?

Take Home Messages

- Think about **what you want to use the ML algorithm for**: this will drive how you assess it.

- **Generalisation** is the ability for an algorithm to make good predictions on other similar datasets.

- You might want to select between models/hyperparameters: You can do this with validation data (e.g. using k-folds cross validation) but **you will need some held out test data** to then report the accuracy. Not doing this is circular analysis and will invalidate your entire project.

Take Home Messages

- Think about **what you want to use the ML algorithm for**: this will drive how you assess it.

- **Generalisation** is the ability for an algorithm to make good predictions on other similar datasets.

- You might want to select between models/hyperparameters: You can do this with validation data (e.g. using k-folds cross validation) but **you will need some held out test data** to then report the accuracy. Not doing this is circular analysis and will invalidate your entire project.

Activity

For each of the following decide if it's supervised/unsupervised & classification/regression and discuss the questions.

BREAK 10 minutes

Problem: Weather (rain) prediction

Your system needs to predict the rain tomorrow based on measurements from the last five days.

How would you split your data during testing/validation?

Problem: Remote sensing crop disease

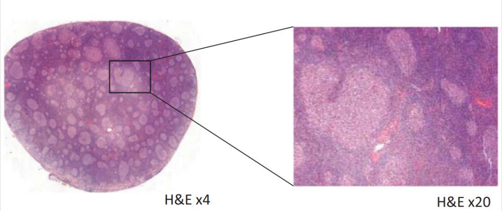
Detecting crop disease from satellite images. Ground truth also collected for some trees (to train classifier).

How can we help & assess generalisability?

Malinee, Rachane, Dimitris Stratoulias, and Narissara Nuthammachot. "Detection of oil palm disease in plantations in krabi province, thailand with high spatial resolution satellite imagery." *Agriculture* 11.3 (2021): 251.

Problem: Classifying types of follicular pathology (lymphoma vs hyperplasia)

A deep CNN can provide diagnosis.



Why is uncertainty quantification useful here? How might it be done?

Syrykh, Charlotte, et al. "Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning." *NPJ digital medicine* 3.1 (2020): 1-8.

Problem: Understand types of patient

You have time series of 3,000 patients with MND. There are probably different types and subgroups of patients in the data.

What ML approach could be used to help identify these different subgroups?

Probability

Who stole the scone?

On an isolated island, live **200 people** who really like scones.

One day, Miss McLellan’s scone was stolen.

Who did it?

Who stole the scone?

On an isolated island, live **200 people** who really like scones.

One day, Miss McLellan’s scone was stolen.

Who did it?

Who stole the scone?

On an isolated island, live **200 people** who really like scones.

One day, Miss McLellan's scone was stolen.

Who did it?

William Brown was found with jam on his jumper.

Who stole the scone?

On an isolated island, live **200 people** who really like scones.

One day, Miss McLellan's scone was stolen.

Who did it?

William Brown was found with jam on his jumper.

It seemed the case was closed…but let's compute the probability that he did steal the jam.

A new type of variable…

To help us reason about probabilities we need a new type of variable - one that doesn't just hold a single value.

Definition

Random Variable (RV)

A function that assigns a number to the outcome of an experiment.

Can be discrete (e.g. number of people; which food to have from a menu)
Or continuous (e.g. time to cycle from home to work)

We use capital letters for RVs.

We use lower case letters to denote the values they might take.

Who stole the scone?

First, before we know about the jam on his jumper… what is the probability that he did it?

Who stole the scone?

First, before we know about the jam on his jumper… what is the probability that he did it?

With 200 people on the island, if we assume everyone is equally likely to have stolen the scone… the probability it was William is,

$$P(W = true) = \frac{1}{200}$$

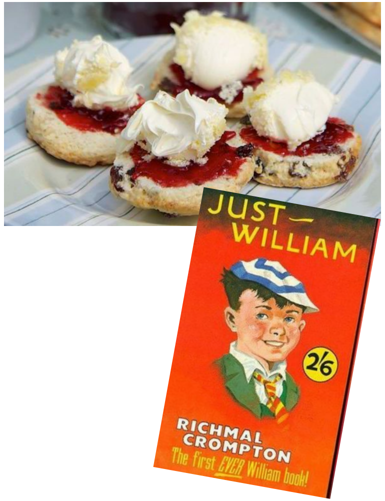
Who stole the scone?

First, before we know about the jam on his jumper… what is the probability that he did it?

With 200 people on the island, if we assume everyone is equally likely to have stolen the scone… the probability it was William is,

$$P(W = true) = \frac{1}{200}$$

P(W) is a **probability mass function**.



Who stole the scone?

First, before we know about the jam on his jumper… what is the probability that he did it?

With 200 people on the island, if we assume everyone is equally likely to have stolen the scone… the probability it was William is,

$$P(W = true) = \frac{1}{200}$$

P(W) is a **probability mass function**.

(as opposed to the probability density function that we need for continuous random variables).

Who stole the scone?

First, before we know about the jam on his jumper… what is the probability that he did it?

With 200 people on the island, if we assume everyone is equally likely to have stolen the scone… the probability it was William is,

$$P(W = true) = \frac{1}{200}$$

P(W) is a **probability mass function**.

Properties:

$$0 \leq P(X = x_i) \leq 1, \quad i = 1, ..., n$$

$$\sum_{i=1}^{n} P(X = x_i) = 1$$

(as opposed to the probability density function that we need for continuous random variables).

Who stole the scone?

Properties:

$$0 \leq P(X = x_i) \leq 1, \quad i = 1, ..., n$$

$$\sum_{i=1}^{n} P(X = x_i) = 1$$

First, before we know about the jam on his jumper… what is the probability that he did it?

With 200 people on the island, if we assume everyone is equally likely to have stolen the scone… the probability it was William is,

$$P(W = true) = \frac{1}{200}$$

In Bayesian reasoning this is known as the **prior**. It's what we believe prior to any observations.

P(W) is a **probability mass function**.

(as opposed to the probability density function that we need for continuous random variables).

Biggest confusion: Joint vs Conditional

The next thing to ask:

What is the probability of William having jam on his jumper GIVEN he stole the scone?

Biggest confusion: Joint vs Conditional

The next thing to ask:

What is the probability of William having jam on his jumper GIVEN he stole the scone?

$$P(J = True \mid W = true)$$

This is the conditional probability.

In Bayesian stats (discussed later), this is the **likelihood**. It says how likely the evidence is GIVEN our model and parameters. In this case we just have one “parameter” which is whether William is guilty.

Biggest confusion: Joint vs Conditional

The next thing to ask:

What is the probability of William having jam on his jumper GIVEN he stole the scone?

$$P(J = True \mid W = true)$$

This is the conditional probability.

Note that this is different from the **JOINT** probability:

$$P(J = true, W = true)$$

Product Rule of Probability

They are related by the **product rule of probability**:

$$P(J,W) = P(J|W)P(W)$$

Similarly,

$$P(W,J) = P(W|J)P(J)$$

It turns out that forensic science has found that scone thieves typically get jam on their jumpers 50% of the time, so P(J|W)=0.5. We learnt P(W) earlier, can we compute **P(J,W)**?

[the probability of William having jam on his jumper AND William being the thief]

$$P(J,W) = P(J|W)P(W)$$

1/2

$$P(J, W) = P(J|W)P(W)$$

$$P(J,W) = \overset{1/2}{P(J|W)}\,\overset{1/200}{P(W)}$$

$$\overset{1/400}{P(J,W)} = \overset{1/2}{P(J|W)}\,\overset{1/200}{P(W)}$$

Product Rule of Probability

As a quick exercise, let’s use the product rule to solve a few puzzles:

We have two enclosures at a zoo.

Product Rule of Probability

As a quick exercise, let's use the product rule to solve a few puzzles:

We have two enclosures at a zoo.

Enclosure A has two lions and three tigers.

Enclosure B has six lions and two tigers.

Product Rule of Probability

As a quick exercise, let’s use the product rule to solve a few puzzles:

We have two enclosures at a zoo.

Enclosure A has two lions and three tigers.

Enclosure B has six lions and two tigers.

If we pick a random animal from enclosure A, what is the probability that it is a lion?

P(Animal=lion | Enclosure=a) =

Product Rule of Probability

As a quick exercise, let's use the product rule to solve a few puzzles:

We have two enclosures at a zoo.

Enclosure A has two lions and three tigers.

Enclosure B has six lions and two tigers.

If we pick a random animal from enclosure A, what is the probability that it is a lion?

P(Animal=lion | Enclosure=a) = 2/5

Product Rule of Probability

How do we find probabilities in the first place?

We can compute a probability from data by repeating an experiment several times.

For example if we didn’t know what was in the enclosure, we could take an animal out at random, make a note of it, and put it back, and repeat…

$$P(X = x_i) \approx \frac{n_{X=x_i}}{N}$$

As a quick exer … ew puzzles:

We have two e…

Enclosure A ha…

Enclosure B ha…

If we pick a ran…
what is the prob…

P(Animal=lion | Enclosure=a) = 2/5

Product Rule of Probability

As a quick exercise, let's use the product rule to solve a few puzzles:

We have two enclosures at a zoo.

Enclosure A has two lions and three tigers.

Enclosure B has six lions and two tigers.

An animal has escaped! If we assume the two enclosures are equally likely to fail, what is the probability that the escaped animal is a tiger **and** from enclosure B?

P(Animal=tiger, Enclosure=b)

Product Rule of Probability

As a quick exercise, let's use the product rule to solve a few puzzles:

We have two enclosures at a zoo.

Enclosure A has two lions and three tigers.

Enclosure B has six lions and two tigers.

An animal has escaped! If we assume the two enclosures are equally likely to fail, what is the probability that the escaped animal is a tiger **and** from enclosure B?

P(Animal=tiger, Enclosure=b)

P(Animal=tiger | Enclosure = b) * P(Enclosure = b)

Product Rule of Probability

As a quick exercise, let's use the product rule to solve a few puzzles:

We have two enclosures at a zoo.

Enclosure A has two lions and three tigers.

Enclosure B has six lions and two tigers.

An animal has escaped! If we assume the two enclosures are equally likely to fail, what is the probability that the escaped animal is a tiger **and** from enclosure B?

P(Animal=tiger, Enclosure=b)

P(Animal=tiger | Enclosure = b) * P(Enclosure = b) = $\frac{1}{4}$ * $\frac{1}{2}$

Product Rule of Probability

As a quick exercise, let's use the product rule to solve a few puzzles:

We have two enclosures at a zoo.

Enclosure A has two lions and three tigers.

Enclosure B has six lions and two tigers.

An animal has escaped! If we assume the two enclosures are equally likely to fail, what is the probability that the escaped animal is a tiger **and** from enclosure B?

P(Animal=tiger, Enclosure=b)

P(Animal=tiger | Enclosure = b) * P(Enclosure = b) = $\frac{1}{4}$ * $\frac{1}{2}$ = $\frac{1}{8}$

Bayes' Theorem

Bayes' theorem can be easily derived using the product rule.

We had two ways of writing down the joint probability of W=true AND J=true:

$$P(W, J) = P(W|J)P(J)$$

$$P(J, W) = P(J|W)P(W)$$

Note that the left hand sides are equal. P(W,J)=P(J,W).

Bayes' Theorem

Bayes' theorem can be easily derived using the product rule.

We had two ways of writing down the joint probability of W=true AND J=true:

$$P(W, J) = P(W|J)P(J)$$

$$P(J, W) = P(J|W)P(W)$$

Note that the left hand sides are equal. P(W,J)=P(J,W).

Bayes' Theorem

Bayes' theorem can be easily derived using the product rule.

We had two ways of writing down the joint probability of W=true AND J=true:

$$P(W, J) = P(W|J)P(J)$$

$$P(J, W) = P(J|W)P(W)$$

Note that the left hand sides are equal. P(W,J)=P(J,W). This lets us equate the right hand sides: $P(W|J)P(J) = P(J|W)P(W)$

Bayes' Theorem

We can then just divide through by P(J):

$$P(W|J)P(J) = P(J|W)P(W)$$

To give us Bayes' Theorem:

$$P(W|J) = \frac{P(J|W)P(W)}{P(J)}$$

We can now use this to find out the probability that William stole the jam…

Bayes' Theorem

We know:
P(J|W) = 1/2 - the **likelihood**:
probability of Jam being on William's jumper GIVEN he stole the jam.
P(W) = 1/200 - the **prior**:
probability (before we've observed anything) that William was the thief.

$$P(W|J) = \frac{P(J|W)P(W)}{P(J)}$$

Bayes' Theorem

We know:
P(J|W) = 1/2 - the **likelihood**:
probability of Jam being on William's jumper GIVEN he stole the jam.
P(W) = 1/200 - the **prior**:
probability (before we've observed anything) that William was the thief.

We also need P(J) - the probability of William having jam on his jumper **anyway**.

$$P(W|J) = \frac{P(J|W)P(W)}{P(J)}$$

Marginalisation

So we need P(J) - the probability of William having jam on his jumper **anyway**.

This leads us to the last tool we need: **Marginalisation**.

It's called this as if you write your probabilities in a table the sum is in the margin.

We know:

- the joint probability of William having jam on his jumper AND being the thief
- the probability of him having jam on his jumper and NOT being a thief

we can add these together to get the probability of having jam on his jumper:

$$P(J) = P(J,W) + P(J,\neg W)$$

Remember we can estimate probabilities by sampling…

We next need $P(J|\neg W)$.

We look back at the last two months and found he had jam on him on 6 days of the last 60, so we could assume that **$P(J|\neg W) = 0.1$**.

*Remembering there are 200 people on the island, what is $P(\neg W)$? And using the product rule what is **$P(J,\neg W)$**?*

ACTIVITY

Finally: Compute, using Bayes' Theorem, $P(W \mid J)$.

Answer on board
Note 2.

Bayes' Theorem

Remember the tiger that escaped earlier…we need to work out which enclosure it came from to stop other animals escaping…

We know a tiger escaped, *so what is the probability that it escaped from enclosure A (vs B)*?

Reminder:
- Enclosure A has two lions and three tigers.
- Enclosure B has six lions and two tigers.
- We assume *a priori* that the enclosures are equally likely to have failed.

ACTIVITY

Answer on board

Bayes' Theorem

Remember the tiger that escaped earlier…we need to work out which enclosure it came from to stop other animals escaping…

We know a tiger escaped, *so what is the probability that it escaped from enclosure A (vs B)*?

Reminder:
- Enclosure A has two lions and three tigers.
- Enclosure B has six lions and two tigers.
- We assume *a priori* that the enclosures are equally likely to have failed.

Answer: P(E=a|A=t) = about 70%

ACTIVITY

Note 2b

Expectation

An escaped lion will eat 2 people, while an escaped tiger will eat 10.

The expected value of a function, g( ),
of a discrete random variable X, is:

$$E[g(X)] = \sum_{i=1}^{n} g(x_i) P(X = x_i)$$

ACTIVITY

If we don't know which enclosure has failed / which animal escaped.
How many people might we expect to get eaten?
Previously we computed P(A=tiger)=34/80=42.5% so P(A=lion)=46/80=57.5%

Answer on board

Expectation

Two common expected values or statistical moments are the **mean** and the **variance**.

$$\mu_X = E[X] = \sum_{i=1}^{n} x_i P(X = x_i)$$

$$\sigma_X^2 = E[(X - \mu_X)^2] = \sum_{i=1}^{n} (x_i - \mu_X)^2 P(X = x_i) = E[X^2] - \mu_X^2$$

Board: derive last equality.
See note 3

Expectation

Realistically, there is some uncertainty about the number of people eaten by lions and tigers, if they escape.

[compute mean number eaten on the board]

ACTIVITY

What's the mean number eaten by an escaped tiger?

Lion

Number Eaten	Probability
0	0.4
1	0.5
2	0.1

Tiger

Number Eaten	Probability
0	0.1
1	0.3
2	0.2
3	0.2
4	0.1
5	0.1

Expectation

Realistically, there is some uncertainty about the number of people eaten by lions and tigers, if they escape.

[compute mean number eaten on the board]

0.7

ACTIVITY

What's the mean number eaten by an escaped tiger?

2.2

Lion

Number Eaten	Probability
0	0.4
1	0.5
2	0.1

Tiger

Number Eaten	Probability
0	0.1
1	0.3
2	0.2
3	0.2
4	0.1
5	0.1

Expectation

Realistically, there is some uncertainty about the number of people eaten by lions and tigers, if they escape.

[compute mean number eaten on the board]

0.7

ACTIVITY

What's the mean number eaten by an escaped tiger?

2.2

What's the variance of both?

Lion

Number Eaten	Probability
0	0.4
1	0.5
2	0.1

Tiger

Number Eaten	Probability
0	0.1
1	0.3
2	0.2
3	0.2
4	0.1
5	0.1

Continuous Random Variables

We use a **probability density function** for continuous random variables.

Properties of a pdf

1. $p_X(x) \geq 0.$
2. $\int_{-\infty}^{\infty} p_X(x)\mathrm{d}x = 1.$
3. $P(X \leq a) = \int_{-\infty}^{a} p_X(x)\mathrm{d}x.$
4. $P(a \leq X \leq b) = \int_{a}^{b} p_X(x)\mathrm{d}x.$

Note that p_x **can** be more than one.

Continuous Random Variables

With multiple variables we might have a **joint probability density function**:

Properties of a joint pdf

1. $p_{X,Y}(x,y) \geq 0$.
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{X,Y}(x,y) \mathrm{d}x \mathrm{d}y = 1$.
3. $P(X \leq a, Y \leq c) = \int_{-\infty}^{a} \int_{-\infty}^{c} p_{X,Y}(x,y) \mathrm{d}x \mathrm{d}y$.
4. $P(a \leq X \leq b, c \leq Y \leq d) = \int_{a}^{b} \int_{c}^{d} p_{X,Y}(x,y) \mathrm{d}x \mathrm{d}y$.

The marginalisation rule, expectations, etc - all still work, but we use an integral instead of a summation.

Independence and Conditional Independence

If two variables are **independent** then the joint probability (or joint probability density) of the two of them is equal to the product of the two probabilities (or probability densities):

$$P(A, B) = P(A)P(B)$$
$$P(A|B) = P(A)$$

$$A \perp\!\!\!\perp B$$

Conditionally independent variables are independent when a third variable is fixed.

$$P(A, B|C) = P(A|C)P(B|C)$$
$$P(A|B, C) = P(A|C)$$

$$A \perp\!\!\!\perp B \mid C$$

Independence

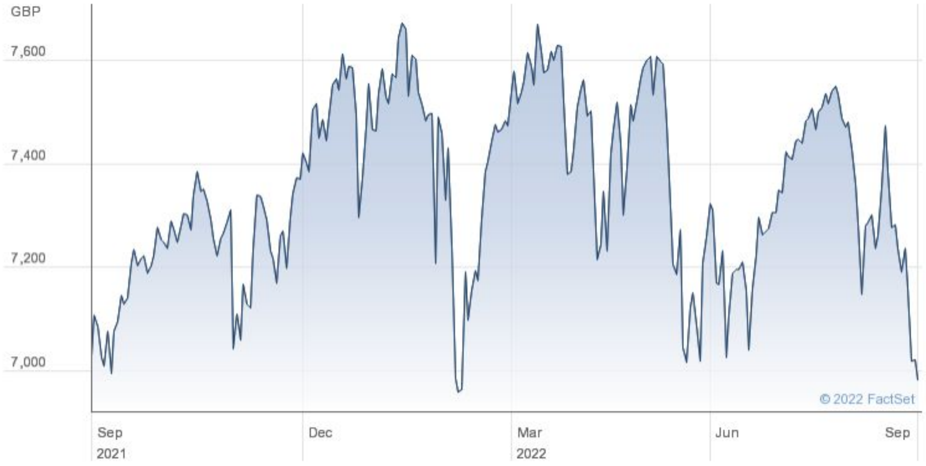
We might be interested in knowing (or assuming) that two random variables are dependent or independent: I.e. whether knowing one of them will tell you something about the value of the other. Here are some examples, can you answer the last two?

Variable A	**Variable B**	**Dependent / Independent?**
Child’s age	Child’s height	Dependent
Number on dice	Price of coffee	Independent
Height of tide	Size of dinner	Independent
Having a stroke	Mode (Cycle/Walk/Bus/Drive) used to get to work	[answer]
Annual Crop yield	Annual Rainfall	[answer]

Dependence is not the same as Correlation

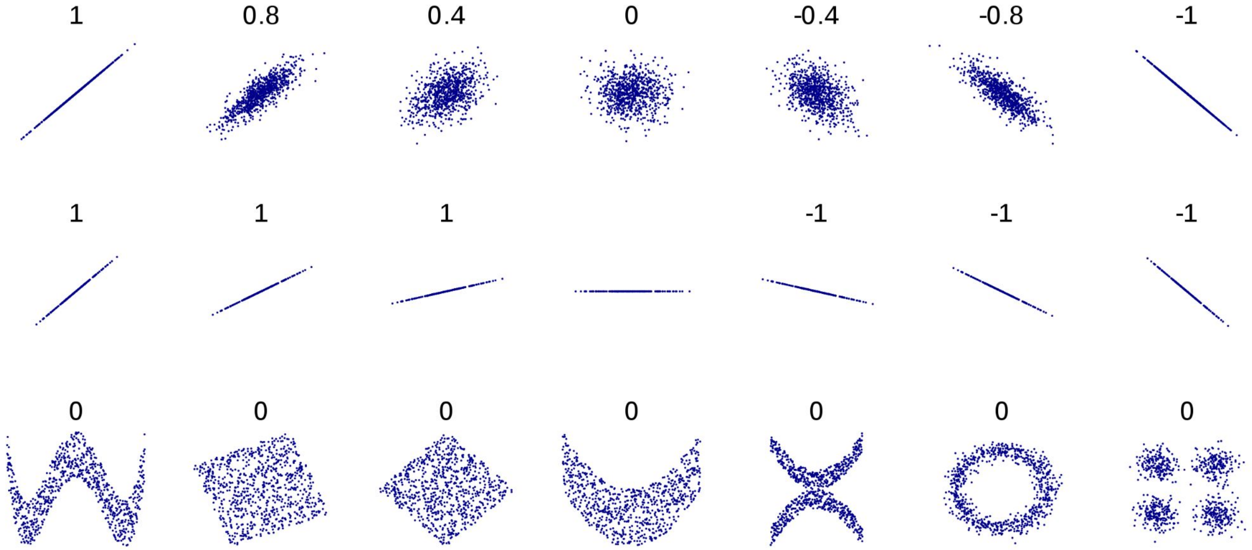
Two variables can be uncorrelated but still dependent!

E.g. FSTE100. Over a year it is mostly **uncorrelated** with time, but has lots of **dependence** on time.



Dependence is not the same as Correlation

Two variables can be uncorrelated but still dependent!



Conditional Independence $A \perp\!\!\!\perp B \mid C$

Conditionally independent variables are independent when a third variable is fixed.

$$P(A, B|C) = P(A|C)P(B|C)$$

$$P(A|B, C) = P(A|C)$$

Example:

“The probability of my car’s gearbox failing is conditionally independent of the probability of its alternator failing GIVEN its age”

The gearbox and alternator are both more likely to fail as a car gets older, but we’re saying that GIVEN a specific age, the probability of the two is independent.

~~Conditional~~ Independence

In this dataset, is favourite drink independent of bedtime?

Favourite drink	**Bedtime**	**Age**
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Beer	after 9pm	Adult
Beer	after 9pm	Adult
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	after 9pm	Child
Coke	before 9pm	Adult

~~Conditional~~ Independence

In this dataset, is favourite drink independent of bedtime?

Favourite drink	Bedtime	Age
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Beer	after 9pm	Adult
Beer	after 9pm	Adult
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	after 9pm	Child
Coke	before 9pm	Adult

If independent: P(D)P(B) = P(D,B)
P(D=milk) = 4/10 = 40%.
P(B=before9pm) = 7/10 = 70%.
P(D=milk, B=before9pm) = 4/10 = 40%.
P(D=milk) P(B=before9pm)= 4/10 * 7/10 = 28% a long way from 40%, so these are not independent.

We ideally need to consider other values, but we'll just look at milk…

Conditional Independence

In this dataset, is favourite drink CONDITIONALLY independent of bedtime GIVEN Age?

Favourite drink	**Bedtime**	**Age**
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Beer	after 9pm	Adult
Beer	after 9pm	Adult
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	after 9pm	Child
Coke	before 9pm	Adult

Conditional Independence

In this dataset, is favourite drink CONDITIONALLY independent of bedtime GIVEN Age?

Favourite drink	Bedtime	Age
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Beer	after 9pm	Adult
Beer	after 9pm	Adult
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	after 9pm	Child
Coke	before 9pm	Adult

If conditionally independent: $P(D|A)P(B|A) = P(D,B|A)$

$P(D=milk|Age=child) = 4/7$

$P(B=before9pm|Age=child) = 6/7$

Conditional Independence

In this dataset, is favourite drink CONDITIONALLY independent of bedtime GIVEN Age?

Favourite drink	Bedtime	Age
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Beer	after 9pm	Adult
Beer	after 9pm	Adult
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	after 9pm	Child
Coke	before 9pm	Adult

If conditionally independent: $P(D|A)P(B|A) = P(D,B|A)$

$P(D=milk|Age=child) = 4/7$

$P(B=before9pm|Age=child) = 6/7$

$P(D=milk, B=before9pm \mid A=child) = 3/7 = 42\%$

Conditional Independence

In this dataset, is favourite drink CONDITIONALLY independent of bedtime GIVEN Age?

Favourite drink	Bedtime	Age
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Beer	after 9pm	Adult
Beer	after 9pm	Adult
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	after 9pm	Child
Coke	before 9pm	Adult

If conditionally independent: P(D|A)P(B|A) = P(D,B|A)

P(D=milk|Age=child) =4/7

P(B=before9pm|Age=child) = 6/7

P(D=milk, B=before9pm | A=child) = 3/7 = 42%

P(D=milk | A=child) * P(B=before9pm | A=child) = 4/7 * 6/7 = 49%

Similar, so maybe CONDITIONALLY independent?

Conditional Independence

In this dataset, is favourite drink CONDITIONALLY independent of bedtime GIVEN Age?

Favourite drink	Bedtime	Age
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Beer	after 9pm	Adult
Beer	after 9pm	Adult
Milk	before 9pm	Child
Apple juice	before 9pm	Child
Milk	after 9pm	Child
Coke	before 9pm	Adult

We don't often try to infer (conditional) independence from data, but instead make independence assumptions to build our model.
We'll learn about this more in Lab 1, when we look at Naive Bayes.

If conditionally independent: P(D|A)P(B|A) = P(D,B|A)

P(D=milk|Age=child) =4/7

P(B=before9pm|Age=child) = 6/7

P(D=milk, B=before9pm | A=child) = 3/7 = 42%

P(D=milk | A=child) * P(B=before9pm | A=child) = 4/7 * 6/7 = 49%

Similar, so maybe CONDITIONALLY independent?

Estimating moments

Finally:

- We might want to estimate moments (especially if we have a continuous random variable) from data.

An estimator for μ_X is given as

$$\widehat{\mu}_X = \frac{1}{N}\sum_{k=1}^{N} x_k.$$

An estimator for σ_X^2 is given as

$$\widehat{\sigma^2}_X = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \widehat{\mu}_X)^2.$$

Take Home Messages

- The product rule for probabilities: P(A,B) = P(A|B) P(B)

- Can use this to derive Bayes' Theorem.
    - P(A|B)P(B)=P(B|A)P(A) & divide by P(B).

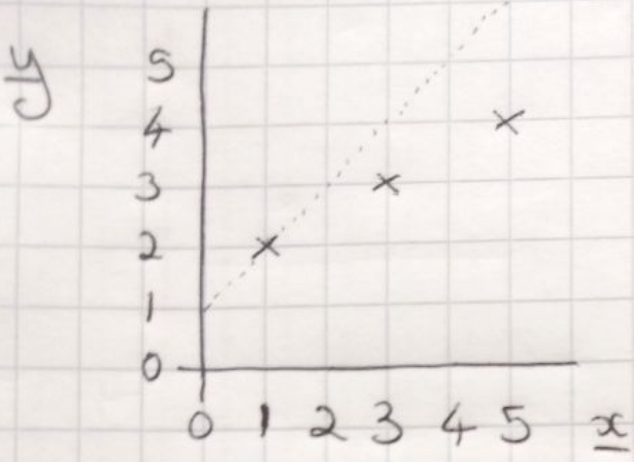
- An expectation is the sum (or integral) of a function over X multiplied by the probability (density) at each X.

$$E[g(X)] = \sum_{i=1}^{n} g(x_i) P(X = x_i)$$

Derivation of least squares linear regression

- I’ll cover this in more detail in Lecture 4
- There’s a missing ‘2’ in the derivative!
- We’ll do this properly later.
    - I just wanted to introduce some of the later topics now to give you an idea…
- So don’t worry about looking through this now, focus on the probability content.

$\underline{w} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \rightarrow \quad y = 1x + 1$



In matrix notation:

$$\underline{y} = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix}$$

We need a cost function that says how good or bad a fit is.

We compare each value in $\underline{y}$ to our prediction: $X\underline{w}$

Eg if $w = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$:

$$\underline{y} - X\underline{w} = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} - \begin{pmatrix} 1+1 \\ 1+3 \\ 1+5 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \\ -2 \end{pmatrix}$$

We can square these numbers and sum them:

$$\begin{pmatrix} 0 & -1 & -2 \end{pmatrix} \begin{pmatrix} 0 \\ -1 \\ -2 \end{pmatrix} = \begin{array}{r} \begin{pmatrix} 0^2 \\ -1^2 \\ -2^2 \end{pmatrix} \\ \hline + \quad \underline{\underline{5}} \end{array}$$

We can write this whole thing as:

$$\text{cost} = (\underline{y} - X\underline{w})^T (\underline{y} - X\underline{w})$$

$$\text{cost} = (\underline{y} - X\underline{\omega})^T (\underline{y} - X\underline{\omega})$$

We want to minimise the cost.

DIFFERENTIATE & SET TO ZERO!

$$\frac{d\,\text{cost}}{d\omega} = X^T(\underline{y} - X\underline{\omega}) = 0$$

$$\text{cost} = (\underline{y} - X\underline{w})^T (\underline{y} - X\underline{w})$$

We want to minimise the cost.

DIFFERENTIATE & SET TO ZERO!

$$\frac{d\,\text{cost}}{d\omega} = X^T(\underline{y} - X\underline{\omega}) = 0$$

$$X^T y - X^T X \underline{w} = 0$$

$$X^T y = X^T X \underline{w}$$

$$(X^T X)^{-1} X^T y = \cancel{(X^T X)^{-1}}\,\cancel{X^T X}\, w$$

$$(X^T X)^{-1} X^T \underline{y}$$

$$\left[\begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix}^T \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix}\right]^{-1} \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix}^T \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$$