Machine Learning with Applications in R

Dr Rebecca Barter

Machine Learning, AI, Prediction, LLMs, …

What do **I** mean by Machine Learning?

Prediction problems

- Classification problems
- Regression problems
- Common ML algorithms
  Least Squares Random Forest,
  Neural Networks, XGBoost

Other things related to ML that are not covered in this course:

- Large Language Models (LLMs), such as chatGPT, BERT
- Artificial Intelligence (AI)
- Unsupervised learning, such as clustering (K-means), dimensionality reduction (PCA)

Workshop introduction

This workshop will cover

- Prediction problems
- Linear prediction algorithms
- Evaluating predictions
- Decision trees
- Random Forest
- Parameter tuning
- XGBoost
- Neural Networks

This workshop will *not* cover

This workshop will *not* cover

- Large Language Models
- Artificial Intelligence (AI)
- Natural Language Processing
- ChatBots
- Unsupervised learning (clustering, dimensionality reduction)
- Image analysis
- Text analysis

Framing Research Questions as Prediction Problems

Data-driven research questions

Real world problem: *You work at a hospital that wants to reduce the length of stay for surgical patients*

Data-driven research questions

Real world problem: *You work at a hospital that wants to reduce the length of stay for surgical patients*

Descriptive/Inferential questions (***what*** is happening?)

What is the *average* hospital stay duration for patients post-surgery?
Do patients in wards with higher nurse-to-patient ratios tend to have shorter stay lengths?

Causal questions (***why*** is it happening?)

Do higher nurse-to-patient ratios *reduce* the average length of stay for post-surgical patients?

Predictive questions (what ***will*** happen?)

Can we predict the length of stay post-surgery for future patients?

Descriptive (correlations) and **causal** relationships often be used to generate **predictions**

Predictive techniques can be used to identify **descriptive** relationships

*But **predictive** and **descriptive** relationships do **not** imply **causal** relationships*

Predicting a “response”

Question

Which patients are at high risk of readmission based on their medical and demographic characteristics?

Response variable (the thing we’re interested in)

Whether or not someone is readmitted

Predictor variables (the other information that we have)

Medical and demographic characteristics

Framing a predictive question

1. Is your question predictive? Can it be reframed as predictive?

2. What is your response variable? Is it available in your data?

3. What are your ideal predictor variables? Are they available in your data?

Framing a predictive question

"What factors influence someone's likelihood of seeking healthcare?"

1. Is your question predictive? Can it be reframed as predictive?

This is more of a causal question. Causality is complicated.

Reframe as: "Can we predict/identify which patients are likely to seek healthcare?"

Framing a predictive question

"Can we predict/identify which patients are likely to seek healthcare?"

2. What is your response variable? Is it available in your data?

What does it mean to "seek healthcare?"

What is the timeframe?

Reframe as: Can we predict which patients will have at least **one healthcare encounter** (outpatient visit, emergency department visit, or hospitalization) **within the next year**?

Framing a predictive question

"Can we predict which patients will have at least one healthcare encounter (outpatient visit, emergency department visit, or hospitalization) within the next year?"

3. What are your ideal predictor variables? Are they available in your data?

Use domain knowledge.

Sometimes this will be dictated by the data that you have.

Reframe as: "Can we predict which patients will have at least one healthcare encounter (e.g., outpatient visit, emergency department visit, or hospitalization) within the next year based on their **past medical history, demographics, and prior healthcare utilization?**"

Tips for framing predictive questions

What is the real-world task you are trying to achieve? Is your goal ethical (e.g., could it be used to exacerbate existing biases?)

Does a prediction formulation of your question address the broader research goal?

Avoid overly vague or overly narrow questions.

Can your question be answered with the data you have?

Does your data have adequate granularity (e.g., individual-level vs aggregated data)?

Does your data contain the relevant information that might be needed to predict your response?

Applying predictions in practice

For what **population** will your predictions be used in practice?

Does this population **reflect** the data that you are using to generate the predictions?

“Accurate” predictions evaluated on one population can be **biased** when applied to another population.

It is important to ***evaluate*** your predictions on data that reflects the population that they will be applied to in practice.

Examples

For the following examples

- Consider the problem and available data and formulate an appropriate **prediction problem**
- Identify a specific measurable **response** variable available in your data
- Identify a few meaningful **predictor** variables available in your data
- On what **population** will the predictions be used? Does this population reflect the data?

Question: Who is at increased risk of hypertension in the US?

Available data: Your university's hospital EHR database

Examples

For the following examples

- Consider the problem and available data and formulate an appropriate **prediction problem**
- Identify a specific measurable **response** variable available in your data
- Identify a few meaningful **predictor** variables available in your data
- On what **population** will the predictions be used? Does this population reflect the data?

Question: Do older women in the US with pre-existing diabetes have worse pregnancy outcomes?

Available data: Your university's hospital EHR database

Examples

For the following examples

- Consider the problem and available data and formulate an appropriate **prediction problem**
- Identify a specific measurable **response** variable available in your data
- Identify a few meaningful **predictor** variables available in your data
- On what **population** will the predictions be used? Does this population reflect the data?

Question: Can we use genetic information to identify patients in the US over 60 with increased risk of Alzheimer's disease?

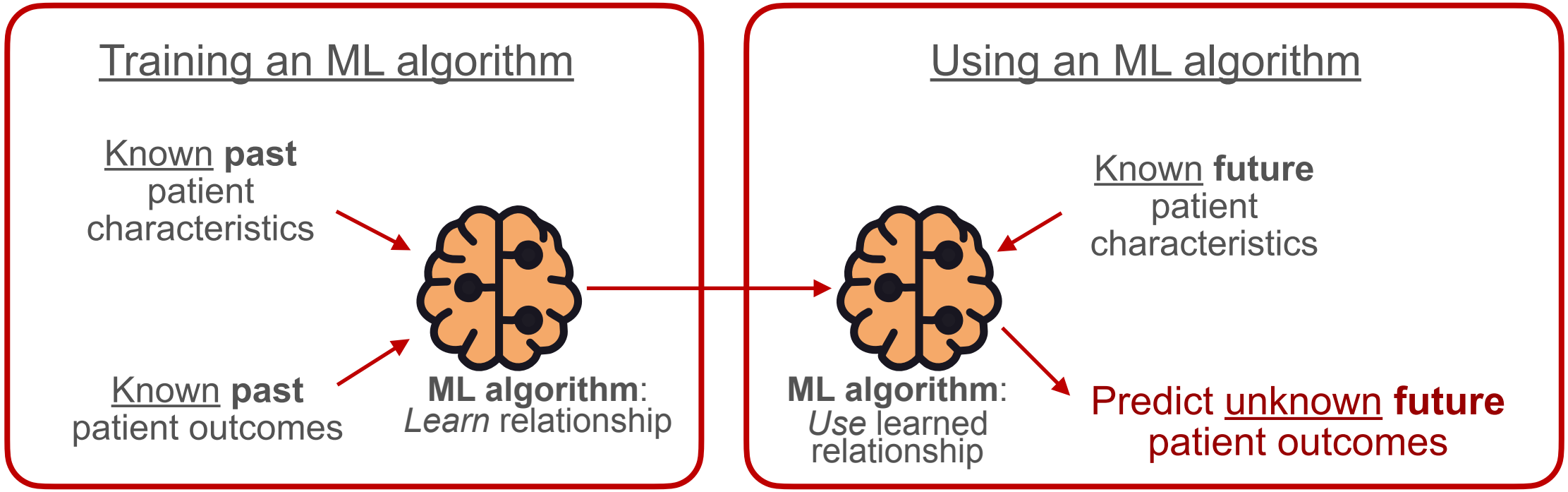
Available data: The UK Biobank data containing genetic, demographic and other medical information for over 500,000 adults from the UK aged 40-69

Introduction to Machine Learning Algorithms

Prediction and ML Algorithms

Prediction problems involve using patterns/relationships learned from past examples to make educated guesses about similar future or unseen situations.

A **ML algorithm** uses the data from past examples to learn a "pattern" that can be generalized to generate predictions for similar future or unseen situations.



Continuous responses

Most prediction problems can be categorized as either **continuous (regression)** or **binary (classification)** response prediction problems

Continuous response (regression) prediction problems involve predicting a numeric value.

Examples

Predicting patient **recovery time after surgery** (days) to provide recovery timelines and plan post-operative care

Possible response value: 1, 14, 2, 100, …

Predicting **kidney function (eGFR value)** in patients with chronic kidney disease to determine treatment plan

Possible response value: 95, 52, 25, 71, …

Binary responses

Most prediction problems can be categorized as either **continuous (regression)** or **binary (classification)** response prediction problems

Binary response (classification) prediction problems involve predicting a binary categorical value.

Examples

Predicting if a **birth is high-risk** to help determine the safest delivery plan

Possible response value: Yes or no

Predicting if a patient is at high risk of **developing breast cancer** to determine if preventative measures are required

Possible response value: Yes or no

Machine Learning algorithms

Common ML algorithm	Classification or regression?
Linear regression	Regression only
Logistic regression	Classification only
Random Forest	Both regression and classification
XGBoost	Both regression and classification
Neural Networks	Both regression and classification

Labeled versus Unlabeled data

Labeled data

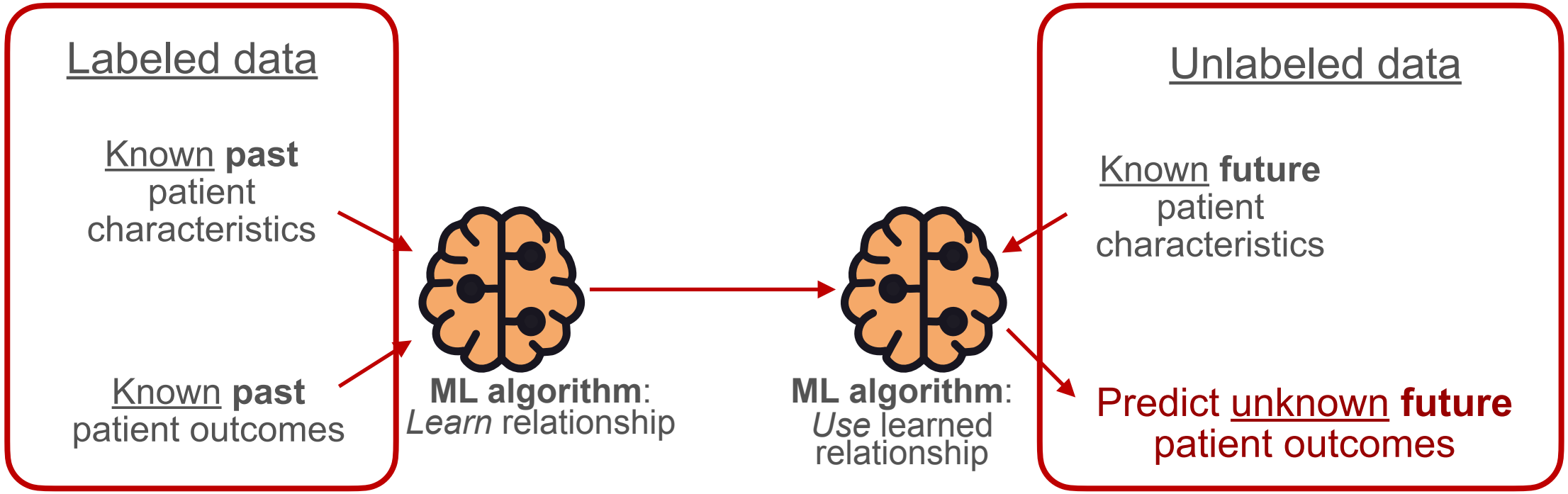
Response/outcome (thing we want to predict) is **known/observed**

Additional related features **known/observed**

Unlabeled data

Response/outcome (thing we want to predict) is **not known/observed**

Additional related features **known/observed**



Labeled versus Unlabeled data

Outcome of interest: Amount spent on healthcare in next year

Unlabeled data

Age	Sex	Weight	Diabetes
54	M	132	N
76	F	155	Y
49	M	166	Y
39	F	129	N
47	M	177	N
70	F	192	N

Labeled versus Unlabeled data

Outcome of interest: Amount spent on healthcare in next year

Labeled data

Age	Sex	Weight	Diabetes	Healthare expenses
54	M	132	N	4,300
76	F	155	Y	3,467
49	M	166	Y	103
39	F	129	N	844
47	M	177	N	6,591
70	F	192	N	8,089

All features must be observed prior to the outcome AND desired time of prediction

The ML process

Formulate prediction problem	**Problem set-up**
Collect labelled data Clean/pre-process data Split data into training and test sets	**Data preparation**
Train ML algorithm(s) using training data Evaluate ML algorithm(s) using labelled test data Select final ML algorithm	**ML training/evaluation**
Deploy/use ML algorithm on unlabelled data	**ML deployment**