Machine Learning - Lecture 1: Introduction to the Topic

Chris Thornton

January 8, 2012

What is Machine Learning?

Machine Learning (ML) is the use of data to acquire the rules for a desired behaviour.

Common tasks:

What is Machine Learning?

Machine Learning (ML) is the use of data to acquire the rules for a desired behaviour.

Common tasks:

- Using data from credit card usage, derive a rule which identifies people that represent a bad credit risk.

What is Machine Learning?

Machine Learning (ML) is the use of data to acquire the rules for a desired behaviour.

Common tasks:

- Using data from credit card usage, derive a rule which identifies people that represent a bad credit risk.
- Using data mapping visual signals to pedal/wheel movements, derive a model which allows a robot to drive a car down a motorway.

What is Machine Learning?

Machine Learning (ML) is the use of data to acquire the rules for a desired behaviour.

Common tasks:

- Using data from credit card usage, derive a rule which identifies people that represent a bad credit risk.
- Using data mapping visual signals to pedal/wheel movements, derive a model which allows a robot to drive a car down a motorway.

Is it to do with human learning?

Traditionally, ML has involved ideas about how human learning works.

But modern research is increasingly focussed on practical tasks.

What do we mean by ‘data’?

By ‘data’ we mean sets of variable values, e.g.,

What do we mean by ‘data’?

By ‘data’ we mean sets of variable values, e.g.,

- Annual rainfall in Sussex for the last twenty years;

What do we mean by ‘data’?

By ‘data’ we mean sets of variable values, e.g.,

- Annual rainfall in Sussex for the last twenty years;
- Age and salary for all members of Sussex faculty.

What do we mean by 'data'?

By 'data' we mean sets of variable values, e.g.,

- Annual rainfall in Sussex for the last twenty years;
- Age and salary for all members of Sussex faculty.
- Number of iPads sold in Brighton per week.

What do we mean by 'data'?

By 'data' we mean sets of variable values, e.g.,

- Annual rainfall in Sussex for the last twenty years;
- Age and salary for all members of Sussex faculty.
- Number of iPads sold in Brighton per week.

Datapoints

Values are organised in structures called **datapoints**.

Each datapoint combines a particular set of variables, e.g., age, salary and IQ specifically for the Informatics HoD.

Datapoints are also called **vectors** in neural-networks, and **records** in computer science.

A datapoint may also be called a **datum**.

Tabulation

Data are often presented in a tabulated form, with one datapoint per row, and one variable per column.

The relevant variable name often appears at the head of each column.

```
NAME       AGE    SALARY    IQ

smith      42     36K       130
bloggs     29     30K       140
bush       50     60K       120
...
```

A very common task in ML involves predicting one variable value from all the others.

Where this is the aim, it is usual to put the to-be-predicted variable last.

Basic data-types

Data are classified according to the number and character of variables involved.

Basic data-types

Data are classified according to the number and character of variables involved.

- Univariate, discrete: one variable with integer/symbolic values.

Basic data-types

Data are classified according to the number and character of variables involved.

- Univariate, discrete: one variable with integer/symbolic values.
- Univariate, continuous: one variable with real/continuous values.

Basic data-types

Data are classified according to the number and character of variables involved.

- Univariate, discrete: one variable with integer/symbolic values.
- Univariate, continuous: one variable with real/continuous values.
- Multivariate, discrete: more than one variable with integer/symbolic values.

Basic data-types

Data are classified according to the number and character of variables involved.

- Univariate, discrete: one variable with integer/symbolic values.
- Univariate, continuous: one variable with real/continuous values.
- Multivariate, discrete: more than one variable with integer/symbolic values.
- Multivariate, continuous: more than one variable with real/continuous values.

Basic data-types

Data are classified according to the number and character of variables involved.

- Univariate, discrete: one variable with integer/symbolic values.
- Univariate, continuous: one variable with real/continuous values.
- Multivariate, discrete: more than one variable with integer/symbolic values.
- Multivariate, continuous: more than one variable with real/continuous values.

Explicit and implicit structure

A **dataset** is a body of data, i.e., a collection of datapoints.

We will be interested in a dataset's structure.

But two meanings for 'structure'.

Explicit structure = the actual values seen in the datapoints.

Implicit structure = patterns that are seen across the values.

Example: A-level grades

Dataset containing average A-level grades for the past ten years.

Explicit structure is the year and grade values.

We also see *implicit* structure—a gradual increase in values over time.

Various ways to model this implicit structure.

We could compute the difference between all years and then average.

This might reveal that grades increase by 0.3% per year on average.

Ways of using the model

The model could then be used for

Ways of using the model

The model could then be used for

- Prediction, i.e., predict the average grade for the next year.

Ways of using the model

The model could then be used for

- Prediction, i.e., predict the average grade for the next year.
- Discounting: work out what current grades are 'worth' in terms of previous years.

Ways of using the model

The model could then be used for

- Prediction, i.e., predict the average grade for the next year.
- Discounting: work out what current grades are 'worth' in terms of previous years.

Why machine learning now?

Machine learning is an increasingly central topic in informatics.

Why machine learning now?

Machine learning is an increasingly central topic in informatics.

- With computers managing/mediating many aspects of our lives, there has been a huge increase in accumulation of electronic data.

Why machine learning now?

Machine learning is an increasingly central topic in informatics.

- With computers managing/mediating many aspects of our lives, there has been a huge increase in accumulation of electronic data.
- With computers increasingly up to the demands of complex modeling, it is getting easier to process very large datasets.

Why machine learning now?

Machine learning is an increasingly central topic in informatics.

- With computers managing/mediating many aspects of our lives, there has been a huge increase in accumulation of electronic data.
- With computers increasingly up to the demands of complex modeling, it is getting easier to process very large datasets.
- Suspicion is growing in fields such as NLP (Natural Language Processing) that approaches based on hand-coded solutions are unlikely to succeed.

Why machine learning now?

Machine learning is an increasingly central topic in informatics.

- With computers managing/mediating many aspects of our lives, there has been a huge increase in accumulation of electronic data.
- With computers increasingly up to the demands of complex modeling, it is getting easier to process very large datasets.
- Suspicion is growing in fields such as NLP (Natural Language Processing) that approaches based on hand-coded solutions are unlikely to succeed.

Real-world applications: learning consumer behaviour

Use of CCTV and automatic checkout machines in modern supermarkets enables detailed logs to be kept of purchases made, reductions on offer, counter locations etc.

These logs embody vast quantities of data and are therefore hard to analyse using traditional methods.

Machine Learning can be used to identify patterns in the data.

These may help identify potentially significant patterns of customer behaviour, enabling better management of the supermarket.

Cheese and ice cream

Modeling might reveal that increases in purchases of ice-cream tend to be accompanied by small reductions in purchases of cheese.

The supermarket could make use of this fact in manipulating sales of cheese and ice-cream.

Example: mining financial data

In this application, the data are price fluctuations and the aim is to extract regularities reflecting investment opportunities.

Modeling these patterns can reveal behavioural rules which increase profit.

For example, the discovery that sharp increases in the price of gold tends to be preceded by long periods of price stability might be the basis for an investment rule.

Fraud detection

Predicting fraudulant cases in credit-card transations

Fraud detection

Predicting fraudulant cases in credit-card transations

- Create a dataset where the values represent transactions and the attributes of account holders.

Fraud detection

Predicting fraudulant cases in credit-card transations

- Create a dataset where the values represent transactions and the attributes of account holders.
- Add a variable which records whether the transaction was fraudulent or not.

Fraud detection

Predicting fraudulant cases in credit-card transations

- Create a dataset where the values represent transactions and the attributes of account holders.
- Add a variable which records whether the transaction was fraudulent or not.
- Mine the data to find implicit structure which predicts whether a transaction is fraudulent or not.

Fraud detection

Predicting fraudulant cases in credit-card transations

- Create a dataset where the values represent transactions and the attributes of account holders.
- Add a variable which records whether the transaction was fraudulent or not.
- Mine the data to find implicit structure which predicts whether a transaction is fraudulent or not.
- Use the model to detect fraud.

Fraud detection

Predicting fraudulant cases in credit-card transations

- Create a dataset where the values represent transactions and the attributes of account holders.
- Add a variable which records whether the transaction was fraudulent or not.
- Mine the data to find implicit structure which predicts whether a transaction is fraudulent or not.
- Use the model to detect fraud.

Summary

Summary

- Machine Learning involves identifying and representing patterns in data, for purposes of obtaining a desired behaviour.

Summary

- Machine Learning involves identifying and representing patterns in data, for purposes of obtaining a desired behaviour.
- Data expressed in terms of variables and datapoints.

Summary

- Machine Learning involves identifying and representing patterns in data, for purposes of obtaining a desired behaviour.
- Data expressed in terms of variables and datapoints.
- Tabulation conventions.

Summary

- Machine Learning involves identifying and representing patterns in data, for purposes of obtaining a desired behaviour.
- Data expressed in terms of variables and datapoints.
- Tabulation conventions.
- Univariate v. multivarite, discrete v. continuous

Summary

- Machine Learning involves identifying and representing patterns in data, for purposes of obtaining a desired behaviour.
- Data expressed in terms of variables and datapoints.
- Tabulation conventions.
- Univariate v. multivarite, discrete v. continuous
- Explicit v. implicit structure.

Summary

- Machine Learning involves identifying and representing patterns in data, for purposes of obtaining a desired behaviour.
- Data expressed in terms of variables and datapoints.
- Tabulation conventions.
- Univariate v. multivarite, discrete v. continuous
- Explicit v. implicit structure.
- ML involves modeling implicit structure on the basis of explicit structure.

Summary

- Machine Learning involves identifying and representing patterns in data, for purposes of obtaining a desired behaviour.
- Data expressed in terms of variables and datapoints.
- Tabulation conventions.
- Univariate v. multivarite, discrete v. continuous
- Explicit v. implicit structure.
- ML involves modeling implicit structure on the basis of explicit structure.

Questions

Questions

- If a supermarket wants to increase its sales of frozen pizzas, what data should it aim to collect?

Questions

- If a supermarket wants to increase its sales of frozen pizzas, what data should it aim to collect?
- In univariate discrete data, how many values would we expect to find in each datapoint?

Questions

- If a supermarket wants to increase its sales of frozen pizzas, what data should it aim to collect?
- In univariate discrete data, how many values would we expect to find in each datapoint?
- How many data should we expect to find in a multivariate dataset?

Questions

- If a supermarket wants to increase its sales of frozen pizzas, what data should it aim to collect?
- In univariate discrete data, how many values would we expect to find in each datapoint?
- How many data should we expect to find in a multivariate dataset?
- How many variables are involved in the specification of multivariate data?

Questions

- If a supermarket wants to increase its sales of frozen pizzas, what data should it aim to collect?
- In univariate discrete data, how many values would we expect to find in each datapoint?
- How many data should we expect to find in a multivariate dataset?
- How many variables are involved in the specification of multivariate data?
- When tabulating data, how is the number of columns determined?

Questions

- If a supermarket wants to increase its sales of frozen pizzas, what data should it aim to collect?
- In univariate discrete data, how many values would we expect to find in each datapoint?
- How many data should we expect to find in a multivariate dataset?
- How many variables are involved in the specification of multivariate data?
- When tabulating data, how is the number of columns determined?
- In the domain of politics, give one example of a continuous variable and one example of a discrete variable.

Questions

- If a supermarket wants to increase its sales of frozen pizzas, what data should it aim to collect?
- In univariate discrete data, how many values would we expect to find in each datapoint?
- How many data should we expect to find in a multivariate dataset?
- How many variables are involved in the specification of multivariate data?
- When tabulating data, how is the number of columns determined?
- In the domain of politics, give one example of a continuous variable and one example of a discrete variable.
- Newspapers sometimes rank universities in terms of numbers of applicants. What is the explicit structure of the data? Suggest some possible forms of implicit structure.

Questions

- If a supermarket wants to increase its sales of frozen pizzas, what data should it aim to collect?
- In univariate discrete data, how many values would we expect to find in each datapoint?
- How many data should we expect to find in a multivariate dataset?
- How many variables are involved in the specification of multivariate data?
- When tabulating data, how is the number of columns determined?
- In the domain of politics, give one example of a continuous variable and one example of a discrete variable.
- Newspapers sometimes rank universities in terms of numbers of applicants. What is the explicit structure of the data? Suggest some possible forms of implicit structure.