Predictive modelling process

A first tour

Michela Mulas

Office hours on Fridays 2 -> 4pm

https://meet.google.com/mqf-isfq-tcd

Today's goal

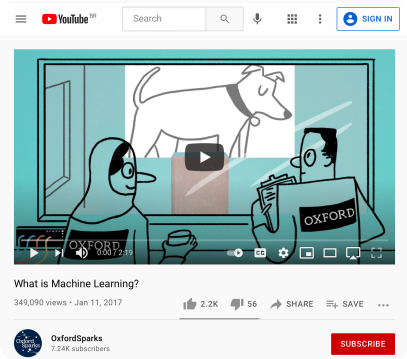
We are going to ...

- Introduce the predictive modeling process $\rightsquigarrow$ Define the terms.
- Case study: Predict fuel economy.
- Discuss HW1.

Reading list

- M. Kuhn and K. Johnson. *Applied Predictive Modeling*[1], 2014
- G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*[2], 2014

[1] Chapters 1 and 2
[2] Chapter 2



Motivations

Information has become more readily available via the internet and media and our desire to use this information to help us make decisions has intensified.

Human brain can consciously and subconsciously assemble a vast amount of data but it cannot process the even greater amount of easily obtainable, relevant information for the problem at hand.

- Websites are used to filter billions of web pages to find the most appropriate information for our queries.
- These sites use tools that take our current information, sift through data looking for patterns that are relevant to our problem, and return answers.
- The process of developing these kinds of tools has evolved throughout a number of fields such as chemistry, computer science, physics, and statistics and has been called **machine learning**, **artificial intelligence**, **pattern recognition**, **data mining**, **predictive analytics**, and **knowledge discovery**.

AI is a broad concepts that refers to the use of computers to mimic the cognitive functions of human beings in field ranging from artificial vision to robotics from video games to autonomous cars ...

Motivations

Each field approaches the problem using different perspectives and tool sets, the **ultimate objective is the same: to make an accurate prediction**.

Geisser[1] defines predictive modelling as the process by which a model is created or chosen to try to best predict the probability of an outcome.

Predictive modelling

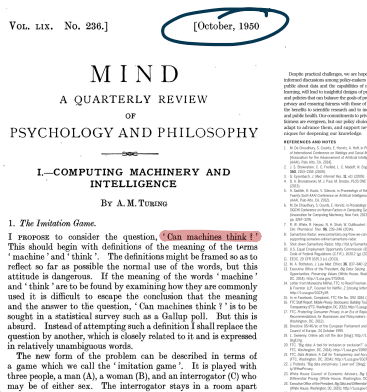
The process of developing a mathematical tool or model that generates an accurate prediction.

Examples of artificial intelligence can be found everywhere[2]:

- The Google global machine uses AI to interpret cryptic human queries.
- Credit card companies use it to track fraud.
- Netflix uses it to recommend movies to subscribers.
- Financial systems use AI to handle billions of trades.

[1] Geisser S. (1993). *Predictive Inference: An Introduction.* Chapman and Hall

[2] Levy S. (2010). *The AI Revolution is On.* Wired.



VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think ?' This should begin with definitions of the meaning of the terms 'machine ' and ' think '. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words ' machine ' and ' think ' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, ' Can machines think ? ' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the ' imitation game '. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either ' X is A and Y is B ' or ' X is B and Y is A '. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair ?

Now suppose X is actually A, then A must answer. It is A's

28 433

REVIEW

Machine learning: Trends, perspectives, and prospects

M. I. Jordan[1]* and T. M. Mitchell[2]*

Machine learning addresses the question of how to build computers that improve automatically through experience. It is one of today's most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science. Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the ongoing explosion in the availability of online data and low-cost computation. The adoption of data-intensive machine-learning methods can be found throughout science, technology and commerce, leading to more evidence-based decision-making across many walks of life, including health care, manufacturing, education, financial modeling, policing, and marketing.

Machine learning is a discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations? The study of machine learning is important both for addressing these fundamental scientific and engineering questions and for the highly practical computer software that it has produced and fielded across many applications.

Machine learning has progressed dramatically over the past two decades, from laboratory curiosity to a practical technology in widespread commercial use. Within artificial intelligence (AI), machine learning has emerged as the method of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications. Many developers of AI systems now recognize that, for many applications, it can be far easier to train a system by showing it examples of desired input-output behavior than to program it manually by anticipating the desired response for all possible inputs. The effect of machine learning has also been felt broadly across computer science and across a range of industries concerned with data-intensive issues, such as consumer services, the diagnosis of faults in complex systems, and the control of logistical chains. There has been a similarly broad range of effects across empirical sciences, from biology to cosmology to social sciences, as machine-learning methods have been developed to analyze highly voluminous experimental data in novel ways. Box 1 shows a few examples of contemporary areas of application of machine learning.

A learning problem can be defined as the problem of improving some measure of performance when executing some task, through some type of training experience. For example, in learning to detect credit-card fraud, the task is to assign a label of "fraud" or "not fraud" to any given credit-card transaction, the performance measure to be improved might be the accuracy of this fraud classifier, and the training experience might consist of a collection of historical credit-card transactions, each labeled in retrospect as fraudulent or not. Alternatively, one might define a different performance metric that assigns a higher cost to erroneously classifying a fraudulent case as "not fraud" than to incorrectly labeled "fraud." One might also define a different type of training experience—for example, including unlabeled credit-card transactions along with labeled examples.

A diverse array of machine-learning algorithms has been developed to cover the wide variety of data and problem types exhibited across different machine-learning problems (2, 3). Conceptual machine-learning algorithms can be viewed as searching through a large space of candidate programs, guided by training experience, to find a program that optimizes the performance metric. Machine-learning algorithms vary greatly in part by the way in which they represent candidate programs (e.g., decision trees, mathematical functions, and general-purpose programming languages) and in part by the way in which they search through this space of candidate programs (e.g., leveraging gradient, with well-understood convergence properties, and evolutionary search methods that evaluate successive generations of randomly mutated programs). Here, we focus on approaches that have been particularly successful to date.

Many algorithms focus on function approximation problems, in which the task is to learn a function f_{θ}, given an input transaction, output a "fraud" or "not fraud" label, and the learning problem is to improve the accuracy of this function, with experience consisting of a sample of known input-output pairs of the function. In some cases, the function is represented explicitly—for example, as a decision tree. In other cases, the function is implicit and obtained via a search process, a factorization, an optimization

[1]Department of Electrical Engineering, and Computer Sciences, Department of Statistics, University of California, Berkeley, CA, USA. [2]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA.
*Corresponding author. E-mail: jordan@cs.berkeley.edu (M.I.J.); tom.mitchell@cmu.edu (T.M.M.)

10.1126/science.aaa8415

We are talking about algorithms that use math and statistics

The idea is to develop algorithm that learn to identify specific relationships in the observed data

Motivations

Each field approaches the problem using different perspectives and tool sets, the **ultimate objective is the same: to make an accurate prediction**.

Geisser[1] defines predictive modelling as the process by which a model is created or chosen to try to best predict the probability of an outcome.

Predictive modelling

The process of developing a mathematical tool or model that generates an accurate prediction.

Examples of artificial intelligence can be found everywhere[2]:

- The Google global machine uses AI to interpret cryptic human queries.
- Credit card companies use it to track fraud.
- Netflix uses it to recommend movies to subscribers.
- Financial systems use AI to handle billions of trades.

[1] Geisser S. (1993). *Predictive Inference: An Introduction.* Chapman and Hall
[2] Levy S. (2010). *The AI Revolution is On.* Wired.

MACHINE LEARNING Vs STATISTICAL LEARNING

subfield of ARTIFICIAL INTELLIGENCE

subfield of STATISTICS

There is much overlap —> both fields focus on supervised and unsupervised learning

ML emphasis is on LARGE SCALE APPLICATIONS and PREDICTION ACCURACY

SL emphasis is on MODELS and their interpretability and precision and uncertainty

Motivations

Each field approaches the problem using different perspectives and tool sets, the **ultimate objective is the same: to make an accurate prediction**.

Geisser[1] defines predictive modelling as the process by which a model is created or chosen to try to best predict the probability of an outcome.

Predictive modelling

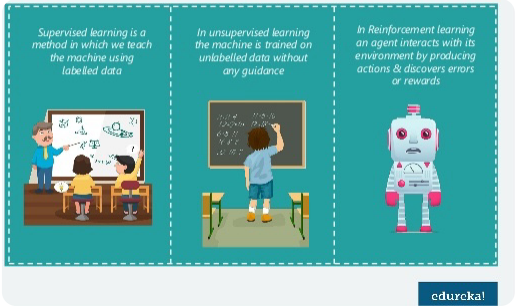
The process of developing a mathematical tool or model that generates an accurate prediction.

Examples of artificial intelligence can be found everywhere[2]:

- The Google global machine uses AI to interpret cryptic human queries.
- Credit card companies use it to track fraud.
- Netflix uses it to recommend movies to subscribers.
- Financial systems use AI to handle billions of trades.

[1] Geisser S. (1993). *Predictive Inference: An Introduction.* Chapman and Hall
[2] Levy S. (2010). *The AI Revolution is On.* Wired.

Supervised — Supervisionado
Unsupervised — Não supervisionado
Reinforcement — Por reforço



Motivations

Each field approaches the problem using different perspectives and tool sets, the **ultimate objective is the same: to make an accurate prediction**.

Geisser[1] defines predictive modelling as the process by which a model is created or chosen to try to best predict the probability of an outcome.

Predictive modelling

The process of developing a mathematical tool or model that generates an accurate prediction.

Examples of artificial intelligence can be found everywhere[2]:

- The Google global machine uses AI to interpret cryptic human queries.
- Credit card companies use it to track fraud.
- Netflix uses it to recommend movies to subscribers.
- Financial systems use AI to handle billions of trades.

[1] Geisser S. (1993). *Predictive Inference: An Introduction.* Chapman and Hall
[2] Levy S. (2010). *The AI Revolution is On.* Wired.

Focus on SUPERVISED LEARNING PROBLEM

- Y outcome measurement ~> also called dependent variable, response, target
- p vector of predictor measurements X ~> also called inputs, regressors, covariates, features, indep variables

REGRESSION PROBLEM Y is quantitative (price, pressure, concentrations)

CLASSIFICATION PROBLEM Y takes values in a finite unordered set (survive/died, digit 0–9,

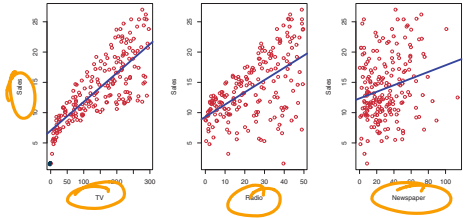
A simple example

Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.

- We get a set of data:

  Sales of that product in 200 different markets

  Advertising budgets in TV, radio, and newspaper



- Objective: Develop an accurate model that can be used to predict sales on the basis of the three media budgets.

LEARNING FROM DATA

$x_1 \rightarrow y_1$
$x_2 \rightarrow y_2$

We have a set of training data $\rightarrow (x_1, y_1) \dots (x_n, y_n)$

These are observations (examples, instances)

On the basis of the training data we would like to:

→ Accurately predict unseen test cases

→ Understand which inputs affect the outcome

→ Assess the quality of our prediction and inferences

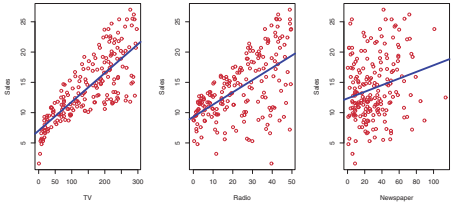
A simple example

Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.

- We get a set of data:

  Sales of that product in 200 different markets

  Advertising budgets in TV, radio, and newspaper



- Objective: Develop an accurate model that can be used to predict sales on the basis of the three media budgets.

KISS

PHILOSOPHY

- -> It is important to understand the ideas behind the various techniques, in order to know how and when to use them
- -> Understand simple methods first, in order to grasp the more sophisticated ones
- -> It is important to accurately assess the performances to know how well/badly it is working

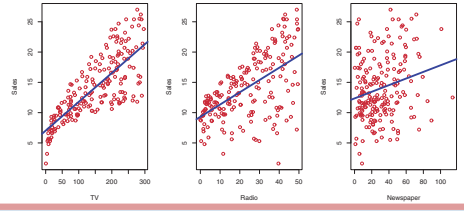
A simple example

Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.

- We get a set of data:

  Sales of that product in 200 different markets

  Advertising budgets in TV, radio, and newspaper



- Objective: Develop an accurate model that can be used to predict sales on the basis of the three media budgets.

Y -> is our output ~> SALES

$X = \begin{bmatrix} X_1 & X_2 & X_3 \end{bmatrix}$ -> is the input

TV Radio NP

$X_{1,1}$ $X_{2,1}$ $X_{3,1}$ ~> Y_1

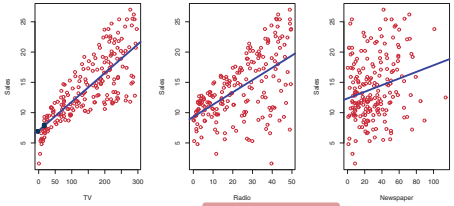
$\vdots$

$X_{1,n}$ $X_{2,n}$ $X_{3,n}$ ~> Y_n

A simple example

Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.

- We get a set of data:
  Sales is the output variable $\rightsquigarrow Y$
  Advertising budgets are the input variables $\rightsquigarrow X_1, X_2, X_3$



- Objective: Develop an accurate model $\rightsquigarrow Y = f(X) + \varepsilon$
  f is some fixed but unknown function.
  ε is a random error term (independent of X, with zero mean)

What is $f(X)$ good for ?

→ With a good f we can make predictions of Y at a new point

→ We can understand which components of X are important in explaining Y and which are irrelevant

Defining statistical learning

We observe Y (the response) and p (the predictors $X_1, X_2, \ldots, X_p$). We assume that there is some relationship between $Y = f(X) + \varepsilon$

Statistical learning refers to a set of approaches for estimating f

For prediction: In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained.

In this setting, since the error term averages to zero, we can predict Y using

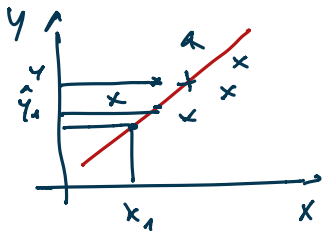
$$\underbrace{\hat{Y}}_{\text{prediction for } Y} = \underbrace{\hat{f}(X)}_{\text{estimate of } f}$$

The accuracy of $\hat{Y}$ as a prediction for Y depends on:

- Reducible error: It is our error on $\hat{f} \rightsquigarrow$ we can reduce it.
- Irreducible error: It is due to $\varepsilon \rightsquigarrow$ no matter how well we estimate f, we cannot reduce the error introduced by ε.

INFERENCE Vs PREDICTION

given a set of data you want to infer how the output is generated as function of the data



Given a new measurement you want to use an existing dataset to build a model that reliably choose the correct identifier from a set of outcomes

Defining statistical learning

We observe Y (the response) and p (the predictors $X_1, X_2, \ldots, X_p$). We assume that there is some relationship between $Y = f(X) + \varepsilon$

Statistical learning refers to a set of approaches for estimating f

For prediction: In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained.

In this setting, since the error term averages to zero, we can predict Y using

$$\underbrace{\hat{Y}}_{\text{prediction for } Y} = \underbrace{\hat{f}(X)}_{\text{estimate of } f}$$

Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = f(X)$. Assume for a moment that both f and X are fixed.

For any estimate $\hat{f}(x)$ of $f(x)$, we have

$$E\left[(Y - \hat{Y})^2 | X = x\right] = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\varepsilon)}_{\text{Irreducible}}$$

$\hat{Y}$ ~> est.
Y ~> eal

Its accuracy depends on 2 quantities:
1. REDUCIBLE ERROR (i.e. it can be reduced by using the most appropriate SL techniques)
2. IRREDUCIBLE ERROR (no matter how well we estimate f, we cannot reduce the error introduced by ε)

B2B -> EXPECTED VALUE of a random variable is the long run average of repetitions of the same experiment
H X is a generic random variable
$\text{Var}(X) = E[(X - \mu)^2]$

Defining statistical learning

We observe Y (the response) and p (the predictors $X_1, X_2, \ldots, X_p$). We assume that there is some relationship between $Y = f(X) + \varepsilon$

Statistical learning refers to a set of approaches for estimating f

For prediction: In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained.

In this setting, since the error term averages to zero, we can predict Y using

$$\underbrace{\hat{Y}}_{\text{prediction for } Y} = \underbrace{\hat{f}(X)}_{\text{estimate of } f}$$

Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = f(X)$. Assume for a moment that both f and X are fixed.

For any estimate $\hat{f}(x)$ of $f(x)$, we have

$$E\left[(Y - \hat{Y})^2 | X = x\right] = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\varepsilon)}_{\text{Irreducible}}$$

We have $Y = f(x) + \varepsilon$ $\quad \hat{Y} = \hat{f}(x)$

$$E\left[(Y - \hat{Y})^2\right] = E\left[(f(x) + \varepsilon - \hat{f}(x))^2\right] =$$

$$= E\left[\left(f(x) - \hat{f}(x)\right)^2 - 2\varepsilon\left(f(x) - \hat{f}(x)\right) + \varepsilon^2\right] =$$

(linearity of E) $\quad = E\left[\left(f(x) - \hat{f}(x)\right)^2\right] - 2E\left[\varepsilon\left(f(x) - \hat{f}(x)\right)\right] + E(\varepsilon^2) =$

(f and $\hat{f}$ are constant) $\quad = \left(f(x) - \hat{f}(x)\right)^2 - 2\left(f(x) - \hat{f}(x)\right)E(\varepsilon) + E(\varepsilon^2)$

ε has 0 mean $\quad = \left(f(x) - \hat{f}(x)\right)^2 + E(\varepsilon^2) =$

($E(\varepsilon^2) = \text{Var}(\varepsilon^2)$) $\quad = \left(f(x) - \hat{f}(x)\right)^2 + \text{Var}(\varepsilon)$

Defining statistical learning

We observe Y (the response) and p (the predictors $X_1, X_2, \ldots, X_p$). We assume that there is some relationship between $Y = f(X) + \varepsilon$

Statistical learning refers to a set of approaches for estimating f

For inference: We want to understand how Y changes as a function of $X_1, \ldots, X_p$. $\hat{f}$ cannot be treated as a black box: We need to know its exact form.

Which predictors are associated with the response?

- ⇝ Identify the few important predictors among a large set of possible variables.

What is the relationship between the response and each predictor?

- ⇝ Identify the relationship between the response and a given predictor that may also depend on the values of the other predictors.

Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

- ⇝ Identify if the true relationship is more complicated than linear.

Defining statistical learning

We observe Y (the response) and p (the predictors $X_1, X_2, \ldots, X_p$). We assume that there is some relationship between $Y = f(X) + \varepsilon$

Statistical learning refers to a set of approaches for estimating f

We can use many linear and non-linear approaches for estimating f. These methods generally share certain characteristics.

- We will always assume that we have observed a set of n different data points.

  We use these observations to train, or teach, our method how to estimate f.

  Our goal is to apply a statistical learning method to the training data in order to estimate the unknown function f.

Defining statistical learning

We observe Y (the response) and p (the predictors $X_1, X_2, \ldots, X_p$). We assume that there is some relationship between $Y = f(X) + \varepsilon$

Statistical learning refers to a set of approaches for estimating f

Most statistical learning methods for this task can be characterized as:

- Parametric methods reduce the problem of estimating f down to one of estimating a set of parameters.
- Non-parametric methods do not make explicit assumptions about the functional form of f. They seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.

Key ingredients

Since many scientific domains have contributed to this field, there are synonyms:

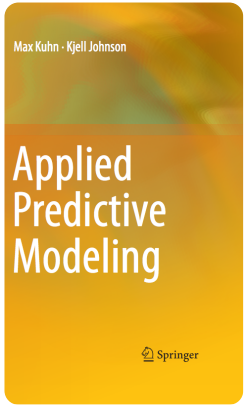
- Sample, data point, observation, or instance: refer to a single, independent unit of data, such as a customer, patient, or compound.
  - ⇝ Sample can also refer to a subset of data points (e.g. training set sample).
- Training set consists of the data used to develop models while the test or validation sets are used solely for evaluating the performance of a final set of candidate models.
- Predictors, independent variables, attributes, or descriptors are the data used as input for the prediction equation.
- Outcome, dependent variable, target, class, or response refer to the outcome event or quantity that is being predicted.

Key ingredients

Since many scientific domains have contributed to this field, there are synonyms:

- Continuous data have natural, numeric scales.
  - ⇝ Blood pressure, the cost of an item, or the number of bathrooms, ...

    In the last case, the counts cannot be a fractional number, but is still treated as continuous data.
- Categorical data, also known as nominal, attribute, or discrete data, take on specific values that have no scale.
  - ⇝ Credit status (“good” or “bad”) or color (“red”, “blue”,etc.) are examples of these data.
- Model building, model training, and parameter estimation all refer to the process of using data to determine values of model equations.



Key ingredients

For an effective predictive model we need:

- Intuition and deep knowledge of the problem context:
  - ⇝ Vital for driving decisions about model development.
- Relevant data:
  - ⇝ The whole process begins with data.
- Versatile computational toolbox:
  - ⇝ Including techniques for data pre-processing and visualization;
  - ⇝ Suite of modeling tools for handling a range of possible scenarios.

Example data sets and typical data scenarios

We can have very diverse problems as well as the characteristics of the collected data.

Grant applications

- This data set was published as a context data set: `http://www.kaggle.com`

  Historical database of 8707 University of Melbourne grant applications from 2009 and 2010 with 249 predictors.

  Grant status: “unsuccessful” or “successful” (with 46% successful).

  Australian grant success rates are less than 25%: the dataset is not representative of the Australian rates.

  Predictors include Sponsor ID, Grant Category, Grant Value Range, Research Field and Department.

  Data are continuous, count and categorical.

  Many predictor missing values (83%).

  The samples are not independent: the same grant writers occurred multiple times.

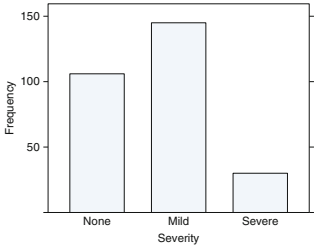
- Objective: Develop a predictive model for the probability of successful application.

Example data sets and typical data scenarios

We can have very diverse problems as well as the characteristics of the collected data.

Hepatic injury

- Data set from the pharmaceutical industry.



- 281 unique compounds, 376 predictors measured or computed for each.
- Categorical response: “does not cause injury”, “mild injury”, “severe injury”.
- Highly unbalanced, common in pharmaceutical data.
- Measurements from 184 biological screens and 192 chemical feature predictors.

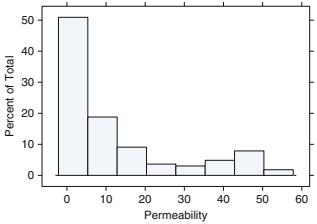
- Objective: Develop a model for predicting compounds’ probability of causing hepatic injury (i.e., liver damage).

Example data sets and typical data scenarios

We can have very diverse problems as well as the characteristics of the collected data.

Permeability

- Data set from the pharmaceutical industry.



- 165 unique compounds.
- For each 1107 molecular fingerprints (binary sequence of numbers that represents the presence or absence of a specific molecular substructure).
- The response is highly skewed, the predictors are sparse (15.5% are present).
- Attempt to potentially reduce the need for the assay.

- Objective: Develop a model for predicting compounds' permeability (the measure of a molecule's ability to cross a membrane).

Example data sets and typical data scenarios

We can have very diverse problems as well as the characteristics of the collected data.

Chemical Manufacturing Process

- Data set from a chemical process industry for producing pharmaceuticals.

  177 samples of biological material with 57 measured characteristics.

  12 of the biological starting material and 45 of the manufacturing process.

  Process variables included temperature, drying time, washing time, and concentrations of by-products at various steps.

  Some of the process measurements can be controlled, while others are observed.

  Predictors are continuous, count, categorical; some are correlated

  Some contain missing values.

  Samples are not independent (sets of samples come from the same batch of biological starting material).

- Objective: Develop a model to predict percent yield of the manufacturing process.

Example data sets and typical data scenarios

We can have very diverse problems as well as the characteristics of the collected data.

Fraudulent Financial Statements

- Data set from public data sources, such as U.S. Securities and Exchange Commission documents.

  Fanning and Cogger[1] sample non-fraudulent companies for important factors (e.g., company size and industry type).

  150 data points were used to train models and the 54 to evaluate them.

  The analysis started with an unidentified number of predictors derived from key areas, such as executive turnover rates, litigation, and debt structure.

  20 predictors were used in the models.

- Objective: Predict management fraud for publicly traded companies.

[1] Fanning K, Cogger K (1998). *Neural Network Detection of Management Fraud Using Published Financial Data.* Int. J of Intell Sys in Accounting, Finance & Management, 7(1), 21-41.

Comparing the data

Data characteristic	Data set: Music genre	Grant applications	Hepatic injury	Fraud detection	Permeability	Chemical manufacturing
Dimensions						
# Samples	12,495	8,707	281	204	165	177
# Predictors	191	249	376	20	1,107	57
Response characteristics						
Categorical or continuous	Categorical	Categorical	Categorical	Categorical	Continuous	Continuous
Balanced/symmetric		×		×		×
Unbalanced/skewed	×		×	×		
Independent			×		×	
Predictor Characteristics						
Continuous	×	×	×	×		×
Count	×	×	×			×
Categorical		×	×	×	×	×
Correlated/associated	×	×	×	×	×	×
Different scales	×	×	×	×		×
Missing values		×				×
Sparse					×	

Check the data set available in the Applied Predictive modeling package

http://faculty.marshall.usc.edu/gareth-james/ISL/Chapter%202%20Lab.txt

for an introduction to the basic commands

Case Study: Predicting Fuel Economy

Consider a simple example that illustrates the broad concepts of model building.

The data set of different estimates of fuel economy for passenger cars and trucks is provided by: `fueleconomy.gov`

From the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy and the U.S. Environmental Protection Agency.

Various characteristics are recorded for each vehicle such as engine displacement or number of cylinders.

Laboratory measurements are made for the city and highway miles per gallon (MPG) of the car.

Objective: Building a model to predict the MPG for a new car line by using

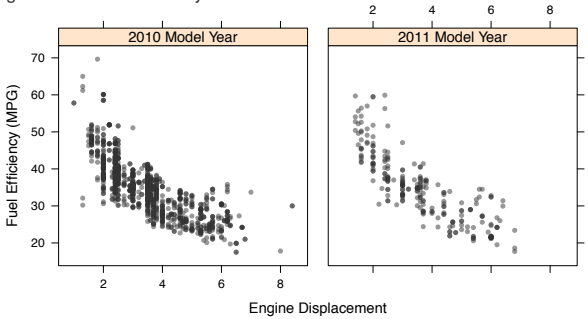
- A single predictor: Engine displacement (the volume inside the engine cylinders).
- A single response: Unadjusted highway MPG for 2010-2011 model year cars.

Case Study: Predicting Fuel Economy

Understand the data.

It can most easily be done through a graph.

- We have just one predictor and one response $\rightsquigarrow$ use a scatter plot.
- Left: Contains all the 2010 data.
- Right: Shows the data only for new 2011 vehicles.

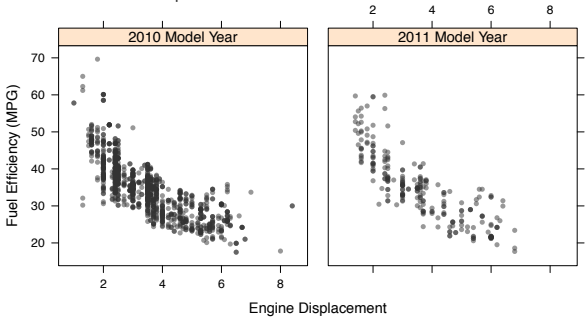


Case Study: Predicting Fuel Economy

Understand the data.

It can most easily be done through a graph.

- As engine displacement increases, the fuel efficiency drops regardless of year.
- The relationship is somewhat linear but does exhibit some curvature towards the extreme ends of the displacement axis.

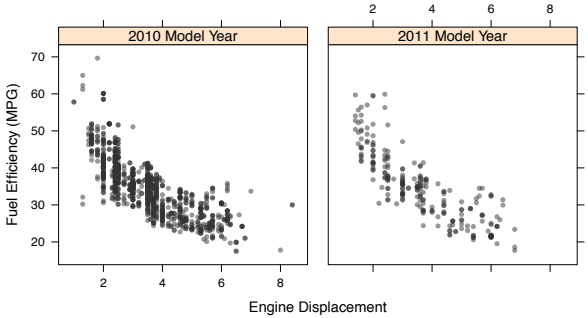


Case Study: Predicting Fuel Economy

Understand the data.

What if we had more than one predictor?

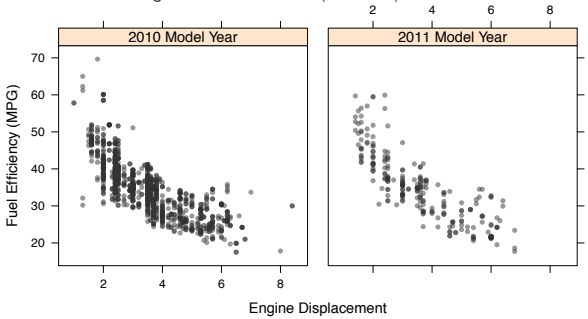
- Need to further understand characteristics of the predictors and their relationships.
- These characteristics may suggest important and necessary pre-processing steps that must be taken prior to building a model.



Case Study: Predicting Fuel Economy

Build and evaluate a model on the data.

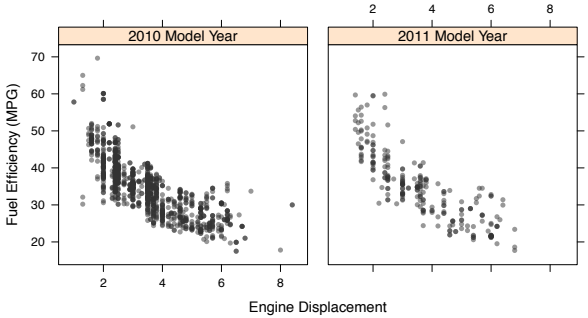
- Standard approach: Take a random sample of the data for model building and use the rest to understand model performance.
- Build the models using the 2010 data (1107 cars): training set
- Test the models using the new 2011 data (245 cars): test or validation set.



Case Study: Predicting Fuel Economy

Build and evaluate a model on the data.

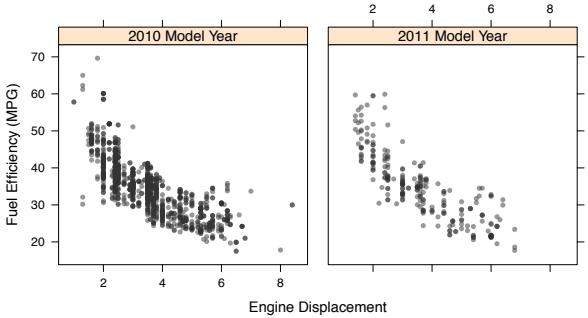
- To evaluate the performance we cannot use the same data used for building.
- Re-predict the training set data: potential to produce overly optimistic estimates.
- Alternative approach use resampling: different subversions of the training data set are used to fit the model.



Case Study: Predicting Fuel Economy

Measure the performance of the model.

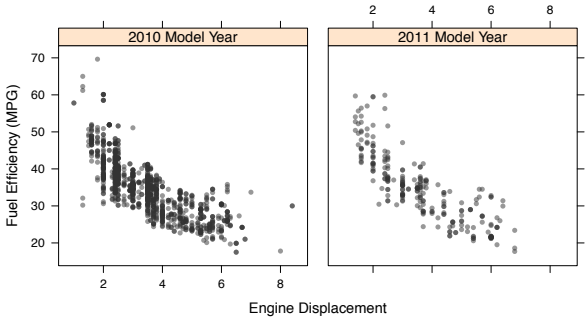
- For regression problems, the residuals are important sources of information.
- Computed as observed value minus the predicted value ($y_i - \hat{y}_i$).
- The root mean squared error is commonly used to evaluate models.
- RMSE is interpreted as how far, on average, the residuals are from zero.



Case Study: Predicting Fuel Economy

Define the relationship between the predictor and outcome.

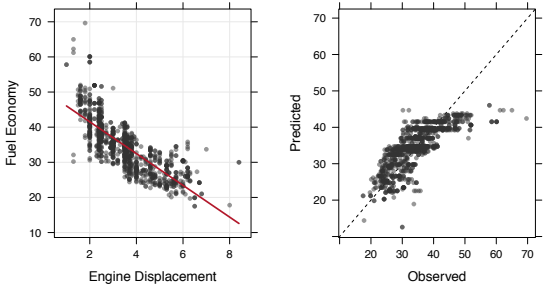
- The modeler will try various techniques to mathematically define this relationship.
- Training set used to estimate the various values needed by the model equations.
- Test set used only when a few strong candidate models have been finalized.



Case Study: Predicting Fuel Economy

First attempt: linear regression model.

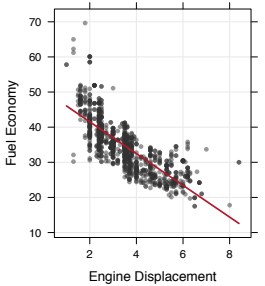
- The predicted MPG is a basic slope and intercept model.
- With the training data, we estimate the model using the least squares method.
- Left: A linear model fit defined by the estimated slope and intercept.
- Right: Observed and predicted MPG.

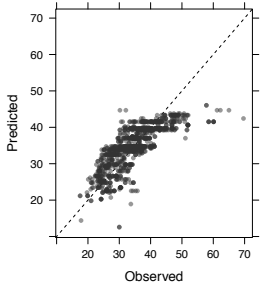


Case Study: Predicting Fuel Economy

First attempt: linear regression model.

- The model misses some of the patterns in the data.
- Under-predicting fuel `economy` when the `displacement` is < than 2L or > 6L.
- We resample the data and estimate a RMSE=4.6 MPG.

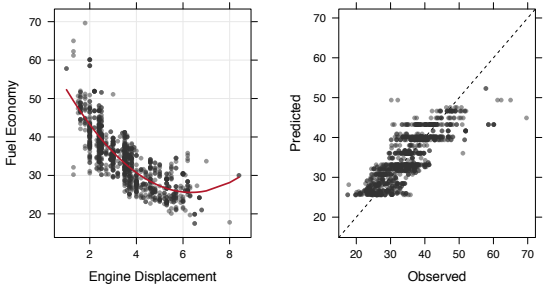




Case Study: Predicting Fuel Economy

Improve the model: Introduce some nonlinearity.

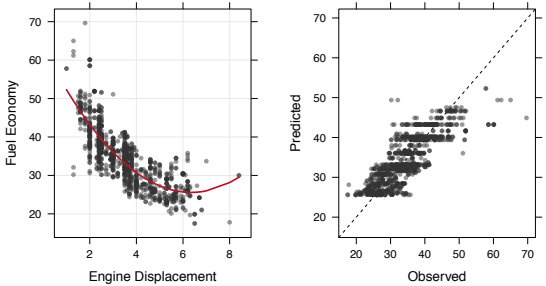
- The most basic approach is to add complexity.
- As for example, by adding a squared term:
  $\texttt{economy} = 63.2 - 11.9 \times \texttt{displacement} + 0.94 \times \texttt{displacement}^2$.
- This is a quadratic model $\rightsquigarrow$ It includes a squared term.



Case Study: Predicting Fuel Economy

Improve the model: Introduce some nonlinearity.

- The quadratic term improves the model fit (RMSE = 4.2 MPG). 4.6
- Drawback: it can perform poorly on the extremes of the predictor.
- The model appears to be bending upwards unrealistically. Predicting new vehicles with large displacement values may produce significantly inaccurate results.



Case Study: Predicting Fuel Economy

Improve the model: Multivariate Adaptive Regression Spline[1].

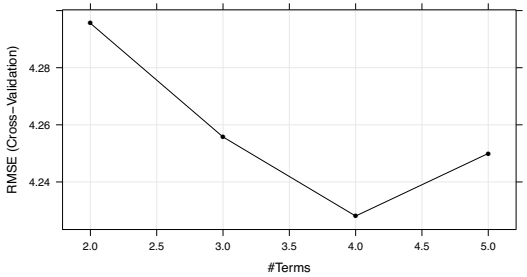
- With a single predictor, MARS can fit separate linear regression lines for different ranges of engine displacement.
- The slopes and intercepts are estimated for this model, as well as the number and size of the separate regions for the linear models.
- Unlike the linear regression models, MARS has a tuning parameter which cannot be directly estimated from the data.
- There is no analytical equation that can be used to determine how many segments should be used to model the data.
- We can try different values and use resampling to determine the appropriate one.
- Once the value is found, a final MARS model would be fit using all the training set data and used for prediction.

[1] Friedman J. (1991). *Multivariate Adaptive Regression Splines*. The Annals of Statistics, 19(1), 1-141.

Case Study: Predicting Fuel Economy

Improve the model: Multivariate Adaptive Regression Spline.

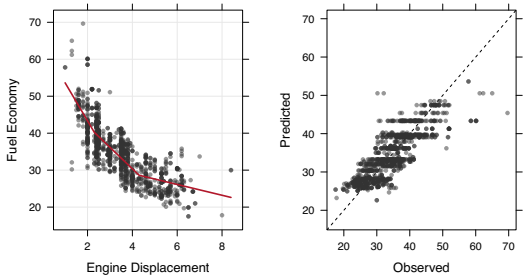
- For a single predictor, MARS can allow for up to five model terms (similar to the previous slopes and intercepts).
- The lowest RMSE value is associated with four terms, although the scale of change in the RMSE values indicates that there is some insensitivity to this tuning parameter.



Case Study: Predicting Fuel Economy

Improve the model: Multivariate Adaptive Regression Spline.

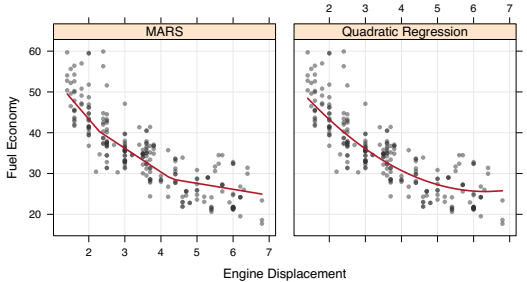
- After fitting the final MARS model with four terms, the training set fit is shown below where several linear segments were predicted.



Case Study: Predicting Fuel Economy

Compare the models on the test set

- Both models fit very similarly.
- For the test set: $\text{RMSE}_{\text{quadratic}} = 4.72$ MPG and the $\text{RMES}_{\text{MARS}} = 4.69$ MPG.
- Either model would be appropriate for the prediction of new car lines.



Case Study: Predicting Fuel Economy

Considerations on the model building process: Data Splitting

How allocate data to model building and evaluating performance?

- The primary interest was to predict the fuel economy of *new* vehicles, which is not the same population as the data used to build the model.
- This means that, to some degree, we are testing how well the model extrapolates to a different population.
- If we were interested in predicting from the same population of vehicles (i.e., interpolation), taking a simple random sample of the data would be more appropriate.
- How the training and test sets are determined reflects how the model will be applied.

Case Study: Predicting Fuel Economy

Considerations on the model building process: Data Splitting

How much data should be allocated to the training and test sets?

- Generally it depends on the situation.
- If the pool of data is small, the data splitting decisions can be critical.
  - A small test would have limited utility as a judge of performance.
  - A sole reliance on resampling techniques might be more effective.
- Large data sets reduce the criticality of these decisions.

Case Study: Predicting Fuel Economy

Considerations on the model building process: Predictor data.

The fuel economy example has revolved around one of many predictors: the engine displacement.

- The original data contain many other factors: number of cylinders, the type of transmission, and the manufacturer.
- An earnest attempt to predict the fuel economy would examine as many predictors as possible to improve performance.
- Using more predictors, it is likely that the RMSE for the new model cars can be driven down further.
- Some investigation into the data can also help.
  - ⇝ For example, none of the models were effective at predicting fuel economy when the engine displacement was small. Inclusion of predictors that target these types of vehicles would help improve performance.

Feature selection, the process of determining the minimum set of relevant predictors needed by the model, is a common way to approach the problem.

Case Study: Predicting Fuel Economy

Considerations on the model building process: Estimating performance.

We used two techniques to determine the effectiveness of the model.

1. Quantitative assessments of statistics (i.e., the RMSE) using resampling help the user understand how each technique would perform on new data.
2. Simple visualizations (e.g., plotting the observed and predicted values) to discover areas of the data where the model does particularly good or bad.

This type of qualitative information is critical for improving models and is lost when the model is gauged only on summary statistics.

Case Study: Predicting Fuel Economy

Considerations on the model building process: Evaluating several models.

Three different models were evaluated.

- The No Free Lunch Theorem[1] argues that, without having substantive information about the modeling problem, there is no single model that will always do better than any other model.
- A strong case can be made to try a wide variety of techniques, then determine which model to focus on.

In the fuel economy example, a simple plot of the data shows that there is a nonlinear relationship between the outcome and the predictor.

Given this knowledge, we might exclude linear models from consideration, but there is still a wide variety of techniques to evaluate.

One might say that “model X is always the best performing model” but, for these data, a simple quadratic model is extremely competitive.

[1] Wolpert D (1996). “The Lack of a priori Distinctions Between Learning Algorithms”. Neural Computation, 8(7), 1341–1390 (`http://www.no-free-lunch.org`)

Case Study: Predicting Fuel Economy

Considerations on the model building process: Model selection.

At some point in the process, a specific model must be chosen.

This example demonstrated two types of model selection:

- Between models: The linear regression model did not fit well and was dropped.
- Within models: For MARS, the tuning parameter was chosen using cross-validation.

In either case, we relied on cross-validation and the test set to produce quantitative assessments of the models to help us make the choice.

Because we focused on a single predictor, which will not often be the case, we also made visualisations of the model fit to help inform us.

At the end of the process, the MARS and quadratic models appear to give equivalent performance. However, knowing that the quadratic model might not do well for vehicles with very large displacements, our intuition might tell us to favor the MARS model.

Next...

Applied computational intelligence

Homework 1

For this exercise set, choose either Alternative 1 or Alternative 2, below. Regardless of your choice, your submission must comply with the guidelines at the end of this document.

Data selection

You are give the possibility to choose one set of data attached to the HW assignment:

- Alternative 1 - Gapminder: The dataset contains an excerpt of the Gapminder[1] data on life expectancy, GDP per capita, and population by country. The attached package contains two main datasets: (i) `gapminder`: 12 rows for each country (1952, 1957,. . ., 2007) and it is a subset of (ii) `gapminder_unfiltered`: more lightly filtered and therefore about twice as many rows[2].
- Alternative 2 - Your choice: Assuming you have at your disposal a set of data of your own interest. The dataset should comprise of a certain number of observations, each observation consists of a certain number of predictors (the predictors should be numerical, not categorical) and corresponding class label.

DL Jxe6
2021

Where to find data?

kaggle ?

...