Introduction to Data Mining

Motivation

Lack of data a hindrance to scientific progress for centuries

- Pearson organized the collection of 1375 heights of mothers and daughters in the UK between 1893–1898

Having more of a resource usually means things are easier

- Faster CPUs, GPUs and more memory
- Higgs Boson: Tens of 10^6GB per experiment per year

With so much data we can solve any problem! **Wrong!!**

- Hard to discover meaningful patterns and regularities to exploit information contained in vast databases

Data is **not knowledge**

We are drowning in information, but starving for knowledge

Rutherford D. Roger

Learning from data

Dangerous misconception

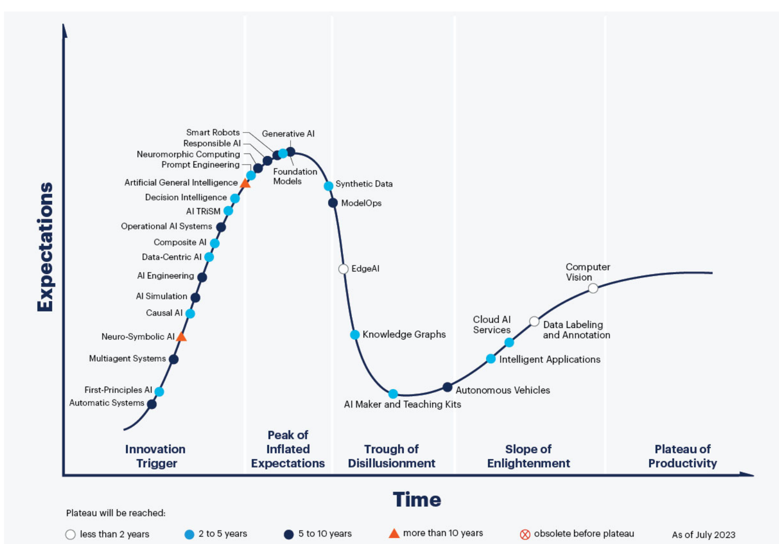
The right data mining tool will squeeze out any knowledge automatically

It is **not the tools** alone, but

1. the intelligent composition of human intuition with computational power,
2. sound background knowledge with computer-aided modeling,
3. critical reflection with convenient automatic model construction, that leads intelligent data analysis projects to success.

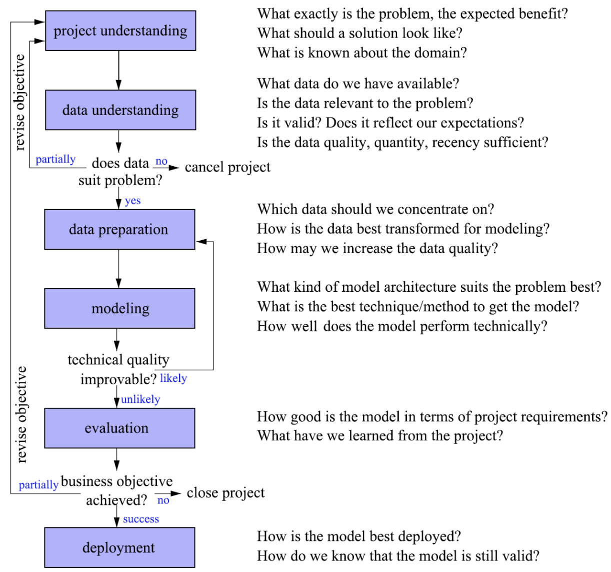
Berthold et al. (2010) *Guide to Intelligent Data Analysis*

Overinflated expectations on AI



Source: Gartner. Hypecycle for Artificial Intelligence 2023

Cross-Industry Standard Process for Data Mining



Source: Berthold et al. (2010) Guide to Intelligent Data Analysis

Types of data analysis problems

Classification

Predict the outcome with a finite number of possible results

- Is this customer credit-worthy?
- Will this customer respond to our mailing?
- Will the technical quality be acceptable?

Regression

Like classification but the value of interest is numerical

- What will sales revenue be in the next quarter?
- How much money will this customer spend?

Types of data analysis problems

Clustering/Segmentation

Summarize data by forming groups of similar cases

- Do my customers divide into different groups?

Association Analysis

Find relationships to understand interdependencies between attributes

- Which options of a mobile contract go together?
- Which products in a supermarket are sold together?

Introduction to classification

Classification problem

Information about different "objects" encoded as feature vectors X

Qualitative variable of interest Y takes (unordered) values:

- e-mail ∈ {spam, non-spam}
- debit card transaction ∈ {legitimate, fraudulent}

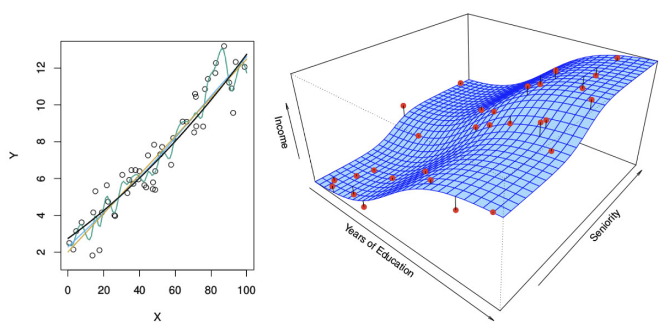
Classifier: Function $f(\cdot)$ that maps X to $P(Y|X)$

Main Goals in Classification

- Prediction
- Assess uncertainty in prediction
- Understand role of different variables / predictors

Regression versus classification

Both problems involve finding a function to predict Y from a given set of pairs of $(x_i; y_i)n_i = 1$



Source: James et al. (2021) An introduction to Statistical Learning

Regression: Y is continuous

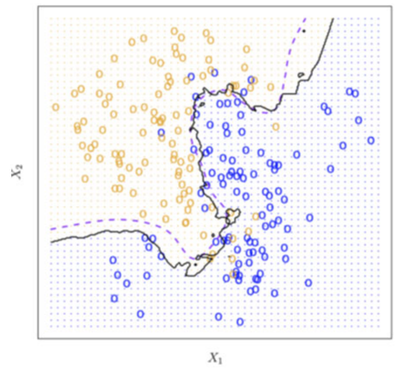
Objective: Line / surface of best fit (minimize Squared Error)

Regression versus classification

Both problems involve finding a function to predict Y from a given set of pairs of $(x_i; y_i)n_i = 1$

Classification: Y is categorical

Objective: Line / surface of best discrimination



Source: James et al. (2021) An introduction to Statistical Learning

For both, central issue is to determine the right flexibility and complexity

- For accurate predictions not only training but also on new data

Classification Data

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa

Independent Variables (Predictors, Features, Attributes, Inputs)

Can be numeric or categorical

Dependent Variable (Target, Response, Output)

Categorical

Logistic Regression

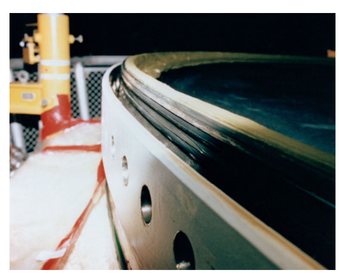
Malfunction of O-rings in the Challenger disaster

Space Shuttle Challenger launched on a cold morning (36°F) in January 1986 and exploded 73 seconds after lunch killing all 7 crew members

Caused by O-rings sealing the sections of the solid rocket booster

Need to be flexible enough to compress and expand

- Flexibility of O-rings directly related to temperature



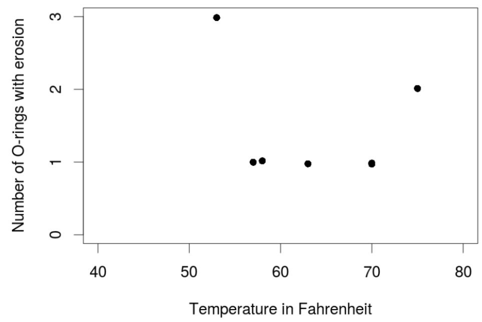
Engineers were aware that temperature on that day was particularly low and that this could affect the erosion of O-rings

No data available for temperatures as low as 36°F

- Lowest temperature measured was 53°F

Looking at data

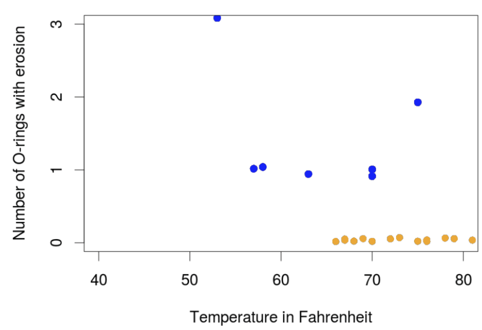
Team recognized lack of data and decided to look at all cases where there had been signs of O-ring distress



Pattern between temperature and O-ring problems?

Anything striking in this plot?

Looking at all data

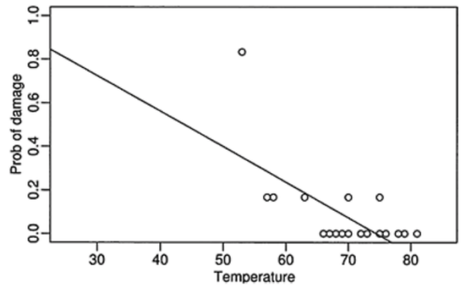


Looking at all data there is a clear pattern of higher temperature being associated with lower chance of O-ring problem

We can see that only in 3 out of 19 flights with **temp > 65°F** there were any O-rings with problems while in all launches with **temp < 65°F** at least one O-ring exhibited fault

Modelling probability of damage

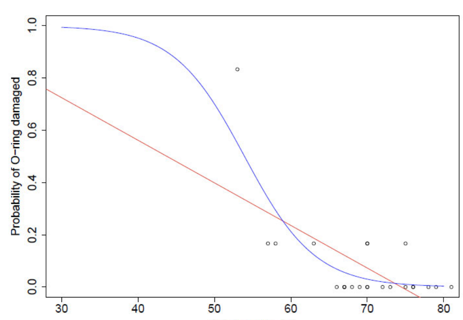
$$P(\text{damage=1 }|\text{Temp}) = \beta_0 + \beta_1\text{Temp} + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$



Standard linear regression model **not suitable**

- Probabilities can be < 0, and > 1. Truncation is unreasonable
- If probability of failure is function of temperature, then the number of damaged O-rings is Binomially distributed
- $\varepsilon \sim N(0, \sigma^2)$ is untenable for Binomially distributed data with few observations
- Variance of binomial distribution $n_1 P_1(1 - P_1)$ not constant

Modelling probability of damage



Logistic regression fit to data more sensible

Monotone function

- Probability increases as $(\beta_0 + \beta_1 x)$ increases

Performance measures

Misclassification rate & Accuracy

Truth Table / Confusion Matrix

- Rows represent true class; Columns predicted class
- Each entry specifies how many objects from a given class are classified into the class of the corresponding column

		Predicted Class 1	Predicted Class 0
True Class	**1**	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = \frac{\text{Correctly Classified}}{\text{Number of examples}} = \frac{TP+TN}{TP+TN+FN+FP}$$

$$\text{Misclassification rate} = \frac{\text{Incorrectly Classified}}{\text{Number of examples}} = \frac{FP+FN}{TP+TN+FN+FP}$$

Recall: Truth table depends on threshold used!

Limitations of Accuracy

Class Imbalance Problem: Vast majority of cases belong to one class

- Direct Marketing, Credit Scoring, Fraud Detection, Medical Diagnosis

Example: Credit card default dataset considered

- default="No": 9667 (customer repaid debt in time)
- default="Yes": 333 (customer failed to repay debt in time)

Naive Classifier:

- Predicts all observations to belong to Majority Class (here default="No")
- Achieves accuracy of 0.9667

Detecting instances of the rare class is like finding a needle in a haystack

Alternative Measures of Performance

		Predicted	
		1	0
Actual	1	TP	FN
	0	FP	TN

$$\textbf{Sensitivity} \text{ or } \textbf{TPR} = \frac{TP}{TP + FN}$$

$$\textbf{Specificity} \text{ or } \textbf{TNR} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Sensitivity: Minimize misclassification of Class 1 records (aka Recall)

- 100 people with COVID of which 42 test positive = 42% sensitivity

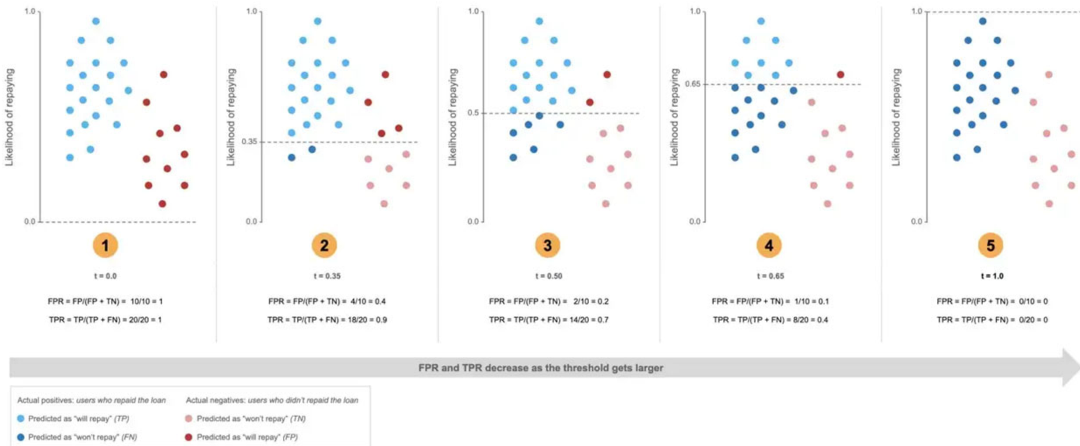
Specificity: Minimize misclassification of Class 0 records

- 100 people with Non-COVID of which 90 test negative = 90% specificity

Precision: Minimize misclassification of records predicted to be in Class 1

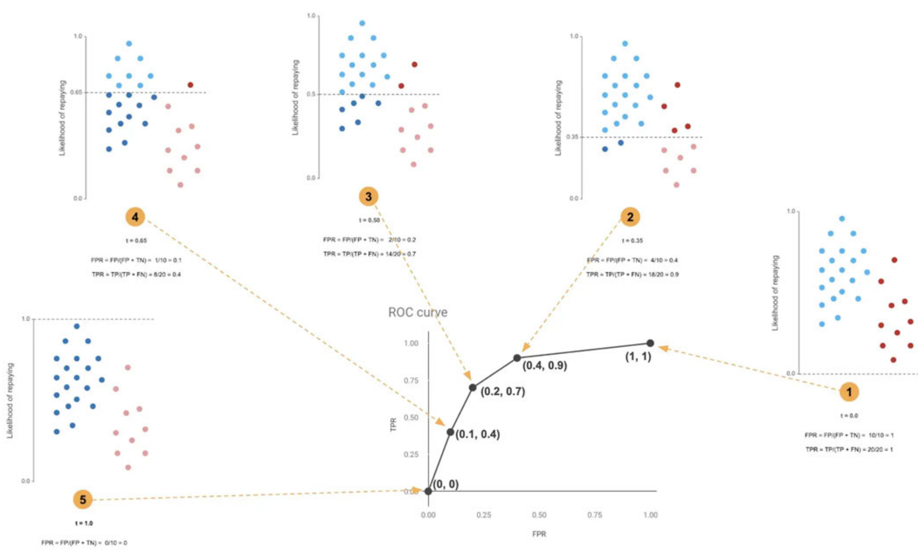
- 43 people with COVID tested positive, and 10 people w/o COVID tested positive gives us a precision of 81%

Interplay between TPR and FPR



Source: Towardsdatascience, https://bit.ly/3YgwyaE

Receiver Operating Characteristic (ROC)



Source: Towardsdatascience, https://bit.ly/3YgwyaE