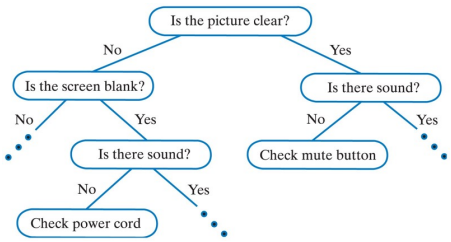
Classification and Regression Trees



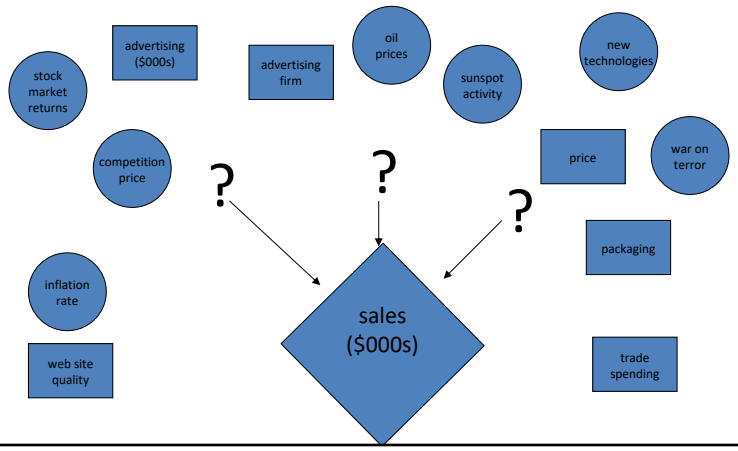
From BUS 462 (Business Analytics) at SFU
Amin Milani Fard - Fall 2024

Some slides are from Michael Brydon

1

Data Mining

- What happens when you do not have an *a priori* model?



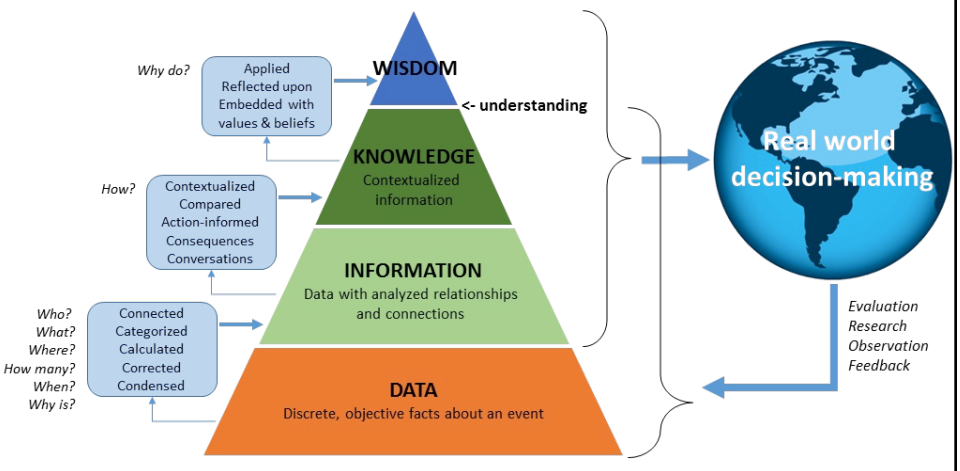
2

Data Mining

- Automatic discovery of patterns in large amounts of data
  - distinct from data warehousing; however, DW typically seen as a prerequisite to mining
- Technologies
  - linear models (e.g., multiple regression)
  - classification trees
    - tests of independence (e.g., χ^2)
    - information theoretic measures
  - neural networks

3

DIKW pyramid



4

Training Data

V1	V2	V3	V4	...	Vn	R1
1581	0.75	blue	86		bar	low
1415	0.96	blue	62		bar	low
1497	0.31	red	51		bar	high
1119	0.29	red	71		foo	high
1115	0.31	red	72		foo	high
1592	0.20	red	89		bar	high
1563	0.83	blue	70		bar	low
1989	0.54	green	59		baz	med
1743	0.75	blue	57		baz	low
1900	0.26	red	77		baz	high
1084	0.33	red	76		foo	high
1197	0.89	blue	76		foo	low
1141	0.61	green	70		foo	med
1955	0.99	blue	88		baz	low
1405	1.00	blue	75		bar	low
1127	0.87	blue	77		foo	low
1641	0.21	red	73		bar	high
1332	0.32	red	68		foo	high
1959	0.16	red	66		baz	high
1754	0.06	red	72		baz	high
1310	0.78	blue	69		foo	low
1824	0.08	red	55		baz	high

Data contains *n* independent variables

We know the value of R1 (the categorical "dependent" or "response" variable) for each row

The Data Mining Task: Specify a "cause-effect" relationship between the independent variables and the dependent variable

5

Linear Regression Assumptions

- linearity of the relationship between dependent and independent variables
- independence of the errors (no serial/auto correlation)
- homoscedasticity (constant variance) of the errors
- normality of the residuals

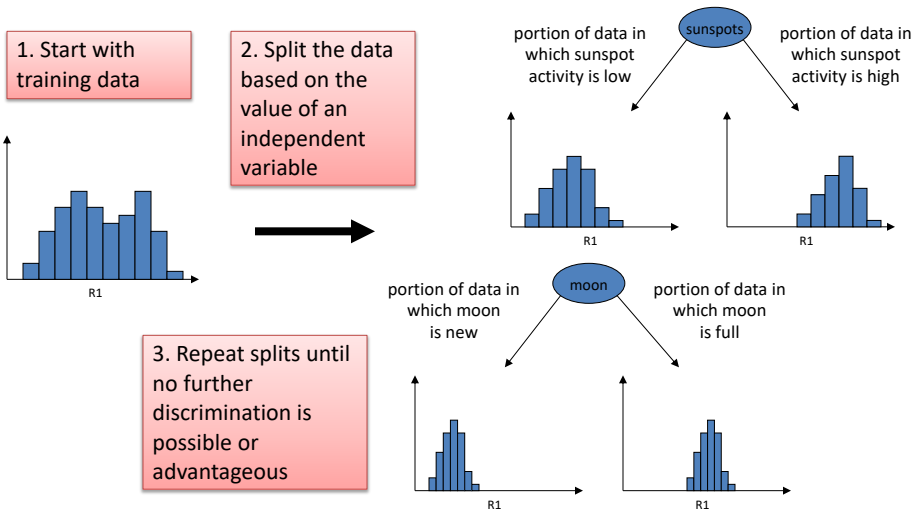
6

Classification and Regression Trees

- Non-parametric (independent of the statistical distribution of the training data)
- Can model continuous (regression trees) or categorical (classification trees) *target* variables
- Statistically selects the most useful data
- Generates classification rules that can be interpreted (by humans) and evaluated
- Is computationally rapid and can provide high quality classification results

7

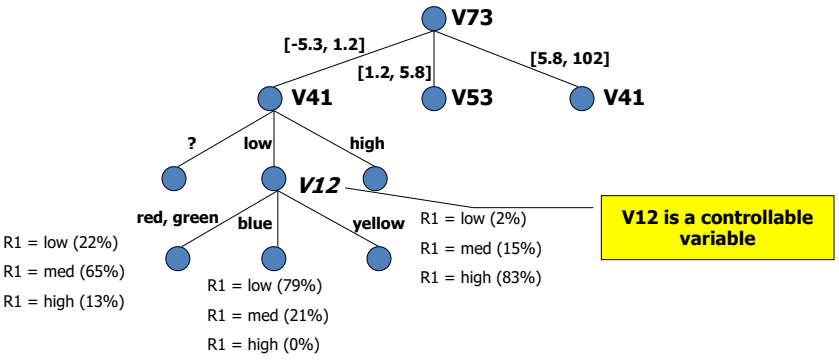
Recursive Set Partitioning



8

Data Mining Results

- Results
  - decision tree with predicted outputs at leaves
  - trade-off accuracy vs. bushiness
  - learning training data vs. over-learning training data



9

Statistical Dependence and Independence

		Preferred Brand of Shoes			
		Reebok	Nike	Adidas	
Grade in Course	A	2	3	3	8
	B	6	6	8	20
	C	2	3	4	9
		10	12	15	37

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(\text{A} \cap \text{Reebok}) = P(\text{A}) \cdot P(\text{Reebok})$$

$$= \frac{8}{37}\left(\frac{10}{37}\right) = 0.058$$

$$E(\text{A} \cap \text{Reebok}) = P(\text{A} \cap \text{Reebok})n = 0.058(37) = 2.16$$

Assumption of independence seems okay

10

Statistical Dependence and Independence

		Experience with Statistics			
		Low	Med	High	
Grade in Course	A	0	1	7	8
	B	2	10	8	20
	C	8	1	0	9
		10	12	15	37

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(\text{A} \cap \text{Low Stats}) = P(\text{A}) \cdot P(\text{Low Stats})$$

$$= \frac{8}{37}\left(\frac{10}{37}\right) = 0.058$$

$$E(\text{A} \cap \text{Low Stats}) = P(\text{A} \cap \text{Low Stats})n = 0.058(37) = 2.16$$

Assumption of independence gives a poor prediction

11

Assessing Independence

- Consider χ^2 test of independence:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

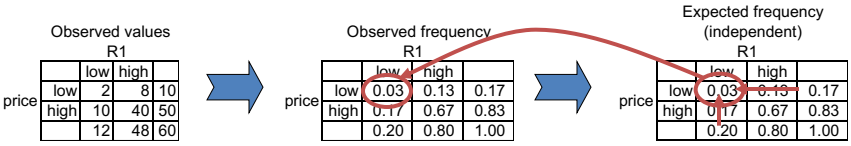
- Assuming variables are independent (expected):

$$P(A \cap B) = P(A) \times P(B)$$

- Our data provides observed joint frequency
- Key idea: variables may be *dependent* in many different ways, but they can be *independent* in only one way

12

Splitting Heuristics



Step 1. Summarize outcomes (R1 values) with respect to the values of the candidate splitting variable

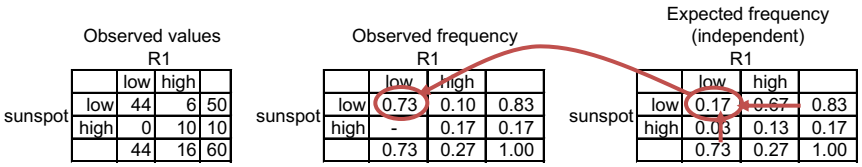
Step 2. Calculate the observed frequency of the different response/splitting variable combinations

Step 3. Compare to expected probabilities

Conclusion: R1 and the variable "price" are independent (price has no discriminatory value)

13

Try a Different Splitting Variable…



- The difference between the observed and expected (independent) frequencies is distributed according to the χ^2 distribution
- Larger χ^2 means smaller probability that R1 and the candidate splitting variable are *independent*
- Heuristic: split first on the variable with the smallest probability of being independent

14

Recursive Partitioning and Impurity

- Divide the space into rectangles such that each rectangle contains points that belong to just one class. (not always possible)
- The two most popular measures of impurity
  - the Gini index and
  - the entropy measure
- Denote the m classes of the response variable by k = 1,2,3,…,m
- p_k is the proportion of observations in a rectangle that belongs to class k

15

15

Gini Index

Gini Index for rectangle *A* containing *m* records

$$I(A) = 1 - \sum_{k=1}^{m} p_k^2$$

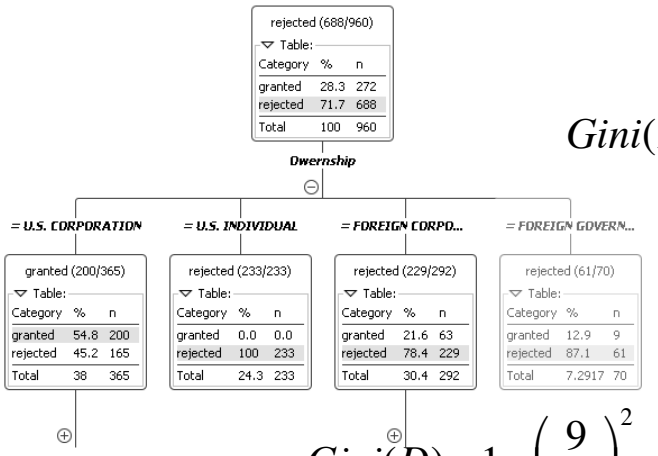
p = proportion (relative frequency) of cases in rectangle *A* that belong to class *k*

- I(A) = 0 when all cases belong to same class
- Max value when all classes are equally represented (= 0.50 in binary case)

16

16

Gini Index as a Splitting Heuristic



$$Gini(D) = 1 - \sum_{i=1}^{n} p_i^2$$

$$Gini(D) = 1 - \left(\frac{9}{70}\right)^2 - \left(\frac{6}{70}\right)^2 = 0.25$$

17

Gini Index

Best case split for a node (perfect discrimination)			
Class	Proportion (p)	p^2	Gini index
1	100%	1	
2	0%	0	
		Σp^2=1.0	1-Σp^2=**0.0**
Worst case split for a node (no discrimination)			
Class	Proportion (p)	p^2	Gini index
1	50%	0.25	
2	50%	0.25	
		Σp^2=0.5	1-Σp^2=**0.5**

18

Entropy

$$entropy\ (A) = -\sum_{k=1}^{m} p_k\ log_2(p_k)$$

p = proportion of cases (out of *m*) in rectangle *A* that belong to class *k*

- Entropy ranges between 0 (most pure) and $\log_2(m)$ (equal representation of classes)

19

19

The Overfitting (overlearning) Problem

Overfitting (aka overlearning) refers to a model that models the training data too well.

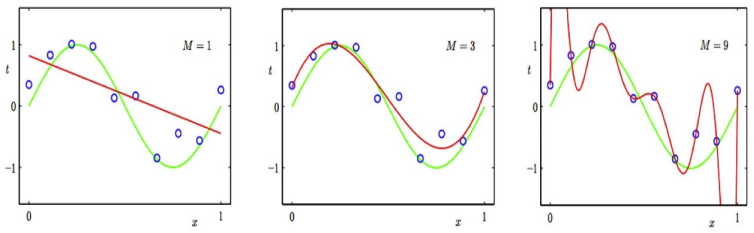
It happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

20

20

The Overfitting Problem

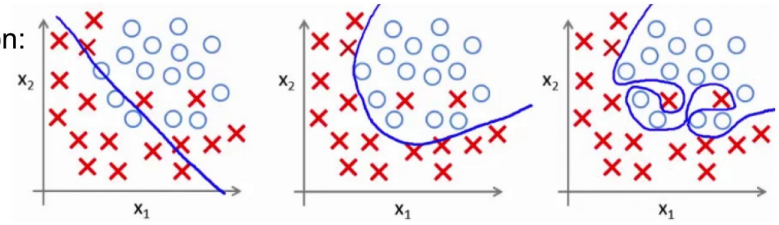
Regression:



predictor too inflexible: cannot capture pattern

predictor too flexible: fits noise in the data

Classification:



21

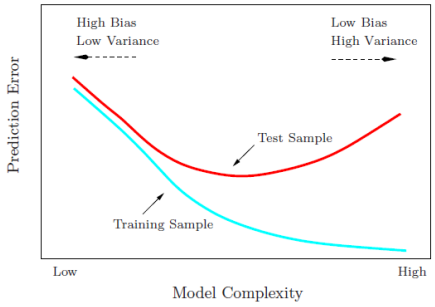
Bias and Variance

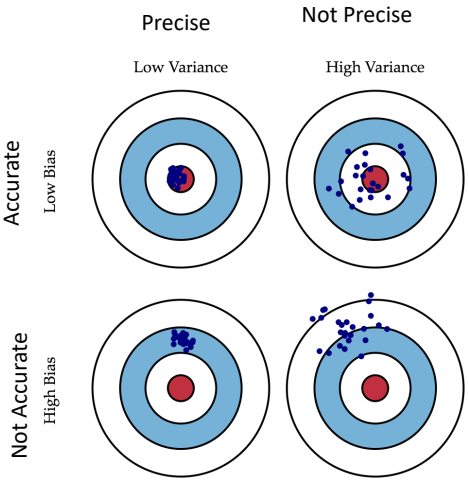
- Bias: The amount of assumptions the model makes. If it has many assumptions then it has high bias, meaning underfitting.
- Variance: The sensitivity of the model to the data it was trained on. High sensitivity leads to high variance, meaning overfitting.

22

22

Bias and Variance





Overfitting: low bias but high variance

Underfitting: high bias but low variance

23

23

Dealing with High Bias and Variance

Problem	**High Bias** (Training performance is low)	**High Variance** (Validation performance is low)
Causes	**Underfitting**	**Overfitting**
Solutions	• Train more • Increase model complexity • Try a different learning model	• Introduce more data • Decrease model complexity (e.g. use regularization) • Try a different learning model

24

24

Bias and Variance

- The ***bias*** is an error from erroneous assumptions in the **learning algorithm**. **High bias** can cause an algorithm to miss the relevant relations between features and target outputs (**underfitting**).
- The ***variance*** is an error from **sensitivity to small fluctuations** in the training set. **High variance** can cause an algorithm to model the random **noise** in the training data, rather than the intended outputs (**overfitting**).

25

25

Avoiding Overfitting

- Natural end of process is 100% purity in each leaf.
- Too many rectangles implies too many splits
- Solutions
  - Stopping Tree Growth => Splitting stops when purity improvement is not statistically significant (using chi-square) - CHAID
  - Pruning the Tree => Use validation error or cost complexity to choose the best tree at that stage

26

26

Response Variable: Two Cases

CASE 1: Categorical response (high, med, low)

- "classification tree"
- leaf nodes provide class membership probabilities

CASE 2: Interval (continuous) response

- "regression tree"
- leaf nodes provide mean and standard deviation

27

Regression Trees for Prediction

- Procedure similar to classification tree
- Many splits attempted, choose the one that minimizes impurity (sum of squared deviations from leaf mean)
- Prediction is computed as the average of numerical target variable in the rectangle (in CT it is majority vote)
- Performance measured by RMSE (root mean squared error)

28

28

Classification Trees vs. Regression Trees

- Splitting Criteria:
  - Gini, Entropy, Twoing
- Goodness of fit measure:
  - misclassification rates
- Prior probabilities and misclassification costs
  - available as model "tuning parameters"

- Splitting Criterion:
  - sum of squared errors
- Goodness of fit:
  - sum of squared errors
- No priors or misclassification costs…
  - … just let it run

29