Introduction to Machine Learning

Data → Decisions

2 broad categories

- Unsupervised learning
- Supervised learning

Market-Basket Analysis

People who buy X also buy Y

When is such a correlation relevant?

Calculate frequently occurring subsets of items

Large number of items - N

Subsets 2^N

Unsupervised learning

Manufacture readymade clothes

What should S/M/L/XL correspond to?

Large sample of actual measurements

Find "natural" groups

Clustering

Customer segmentation

Supervised Learning

Symptoms → Diagnosis

Historical validated data

Predict outcome for current item

Cough	Pain	BP	- - - - -	TB?
Y	N	-		Y
N	Y	Y		N
⋮				⋮
				Y
Y	Y	N	- - -	?

Classification Problem

- Medical diagnosis
- Credit Card fraud
- Email spam filtering
- Loan eligibility
- ⋮

Prediction models

Regression

School uses mock exams (prelims) to predict board exam performance

$$\alpha_0 + \alpha_1 M_1 + \alpha_2 M_2 + \alpha_3 M_3 = E$$

$E \geq 45$ ✓ Threshold

Attributes: $A_1, A_2, \ldots, A_n$

Category: C – assume boolean Y/N

Training data – historical, validated classification

Assumption – uniformity of distribution of data

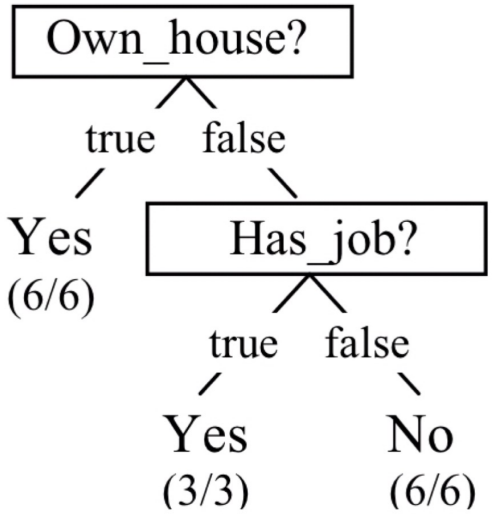
Historical Data

Measure error for given choice of α_i's

Loss function

ID	Age	Has_job	Own_house	Credit_rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**

Loan sanction

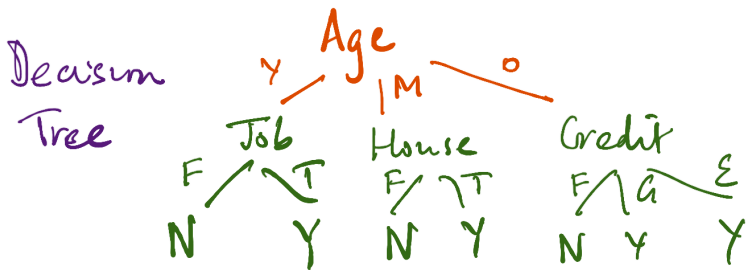


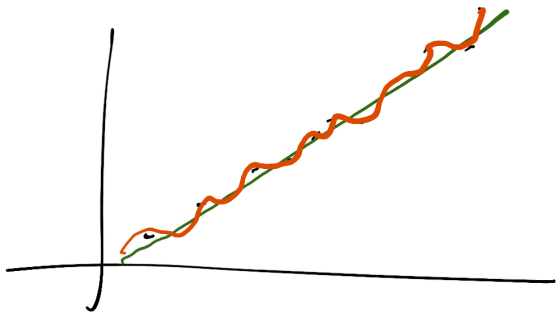
Which tree to construct?

Smaller is better

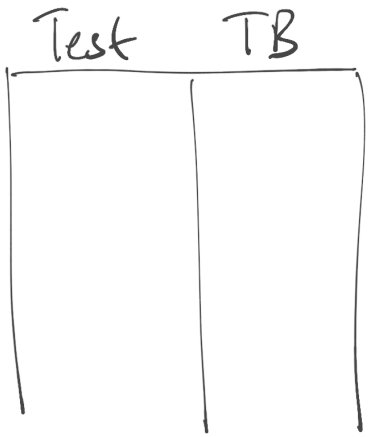
- Occam's Razor - Simplest explanation
- Explainability
- "Overfitting"

ID	Age	Has_job	Own_house	Credit_rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	~~false~~	~~good~~	**Yes**
4	young	true	~~true~~	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	~~good~~	**Yes**
9	middle	false	true	~~excellent~~	**Yes**
10	middle	false	true	~~excellent~~	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**





Probabilistic



5% have TB

Test is 90% accurate both ways

1000 rows

950 No TB — 855 N, 95 Y

50 TB — 45 Test Y, 5 Test N

Test = Y

TB = ?

$= \frac{45}{140}$

Regression

$$f = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots \alpha_k x_k$$

$$f_1 = \alpha_0 + \alpha_1 x_1 + \alpha_3 x_3$$

$$f_2 = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\vdots$$

$$f_n =$$

Bayes Rule

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(B|A) = \frac{P(A \wedge B)}{P(A)}$$

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

$P(A) = 0.14$, $P(B|A) = 0.9$, $P(B) = 0.05$

TB = Y

Test = Y

$$f = c_0 + c_1 f_1 + c_2 f_2 + \cdots c_n f_n$$