MACHINE LEARNING

Introduction to ML, DL, AI and OpenVino

Session 04

Pramod Sharma
pramod.sharma@prasami.com

2

Agenda

- Why learn
- Algorithm types
- Types of machine learning
- Supervised and unsupervised learning
- Classification

3

Why “Learn” ?

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- There is no need to “learn” to calculate payroll
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (routing on a computer network)
  - Solution needs to be adapted to particular cases (user biometrics)

pra-sâmi

4

What We Talk About When We Talk About “Learning”

- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Example in retail: Customer transactions to consumer behavior:
  - People who bought “Blink” also bought “Outliers”
- Build a model that is a good and useful approximation to the data

pra-sâmi

5

Machine Learning

- "Learning is any process by which a system improves performance from experience."
  - Herbert Simon
- Definition by Tom Mitchell (1998):
  - Machine Learning is the study of algorithms that
    - improve their performance P
    - at some task T
    - with experience E.
- A well-defined learning task is given by < P, T, E>
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to
  - Solve the optimization problem
  - Representing and evaluating the model for inference

pra-sâmi

6

Algorithm types

pra-sâmi

7

ML in Short

- Tens of thousands of machine learning algorithms
  - Hundreds new every year
- Every ML algorithm has three components:
  - Representation
  - Optimization
  - Evaluation

pra-sâmi

8

Various Function Representations

- Numerical functions
  - Linear regression
  - Neural networks
  - Support vector machines
- Symbolic functions
  - Decision trees
  - Rules in propositional logic
  - Rules in first-order predicate logic
- Instance-based functions
  - Nearest-neighbor
  - Case-based
- Probabilistic Graphical Models
  - Naïve Bayes
  - Bayesian networks
  - Hidden-Markov Models (HMMs)
  - Probabilistic Context Free Grammars (PCFGs)
  - Markov networks

pra-sâmi

9

Various Search / Optimization Algorithms

- Gradient descent
  - Perceptron
  - Back-propagation
- Dynamic Programming
  - Hidden Markov Model (HMM) Learning
  - Probabilistic Context Free Grammar (PCFG) Learning
- Divide and Conquer
  - Decision tree induction
  - Rule learning
- Evolutionary Computation
  - Genetic Algorithms (GAs)
  - Genetic Programming (GP)
  - Neuro-evolution

10

Evaluation

- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Posterior probability
- Cost / Utility
- Margin
- Entropy
- K-L divergence
- and many more…

11

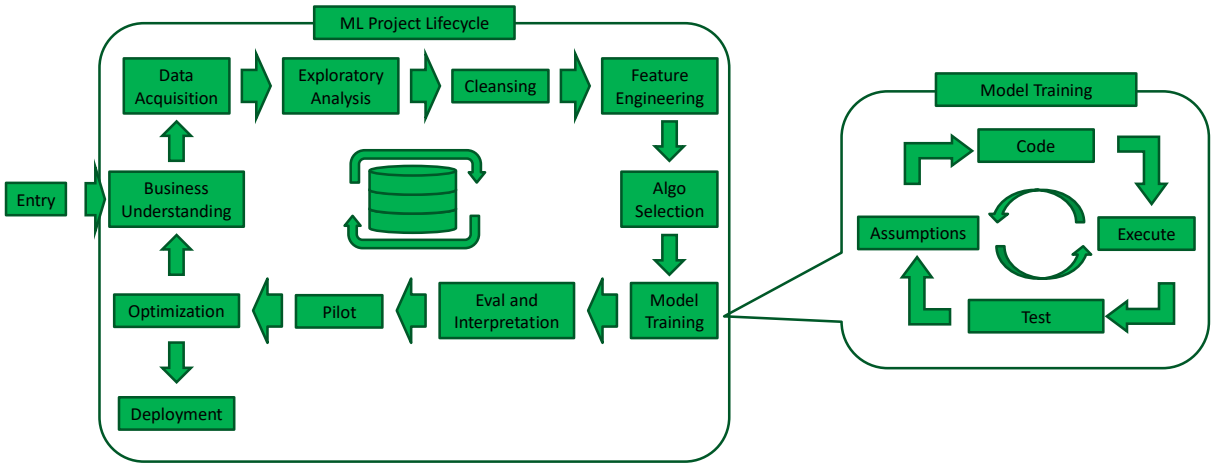
Applications

- Association
  - Basket analysis: P (*Y* | *X*) probability that somebody who buys *X* also buys *Y* where *X* and *Y* are products/services.
    - Example: *P* (chips | beer) = 0.7
- Supervised Learning
  - Given: training data + desired outputs (labels) ➔ fit a hypothesis to it
- Types of supervised learning
  - Classification
  - Regression
- Unsupervised Learning
  - Given: training data (without desired outputs); Try and determining structure in the data
  - Clustering algorithm groups data together based on data features
- Semi-supervised learning
  - Given: training data + a few desired outputs
- Reinforcement learning
  - Rewards from sequence of actions

pra-sâmi

12

Project Lifecycle



pra-sâmi

13

Common Terms

- Model
  - Collection of parameters you are trying to fit
- Data : that CSV file which is given to you
- Instance
- Feature
  - Attributes of your data that will be used in prediction
- Labels
  - Value you are trying to predict with your model
- Categorical Data
  - **Nominal** : categorically discrete data
  - **Ordinal** : quantities that have a natural ordering
  - **Interval** : intervals between each value are equally split e. g. age groups
  - **Ratio:** interval data with a natural zero point. They are ordered units that have the same difference.
- Continuous Data
- Date / time stamp

feature

features

label

Instance

	A	B	date	gender	grades	tide leve	target
0	0.623961	-2.441386	2015-01-01	male	A	3.10 m	0
1	0.262080	-0.951155	2015-01-02	female	D	3.66 m	0
2	-0.653064	0.615920	2015-01-03	male	C	3.47 m	1
3	0.526069	0.421712	2015-01-04	female	D	3.01 m	0
4	-0.472418	1.843539	2015-01-05	female	C	3.32 m	1

Continuous

Date / Timestamp

Categorical Nominal

Categorical Ordinal

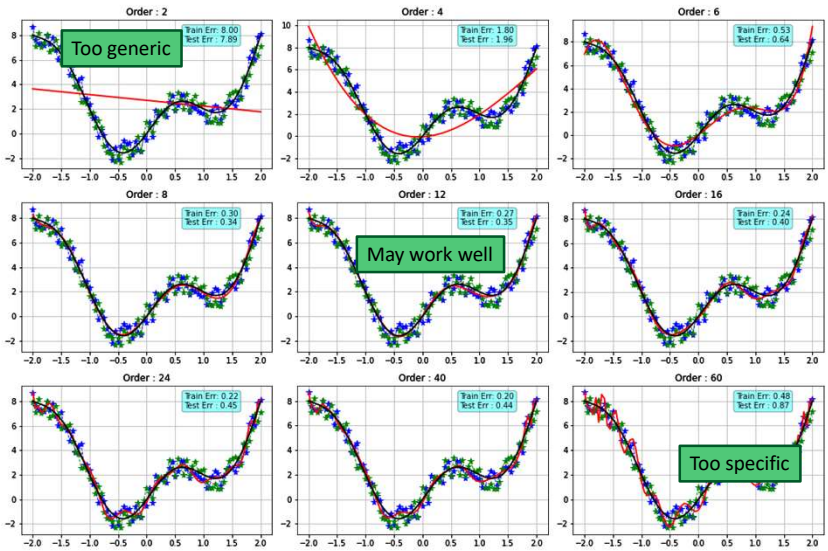
Categorical Ratio

12/23/2023

pra-sâmi

14

Under-fitting vs. Over-fitting



12/23/2023

pra-sâmi

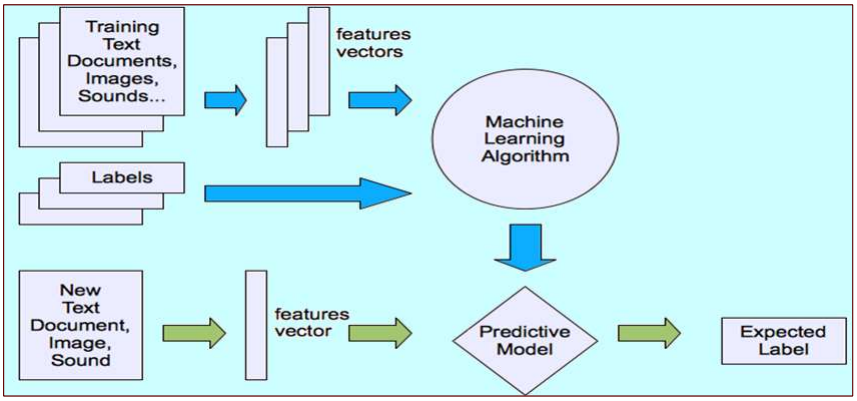
15

Supervised and Unsupervised Learning

pra-sâmi

16

Machine Learning Structure

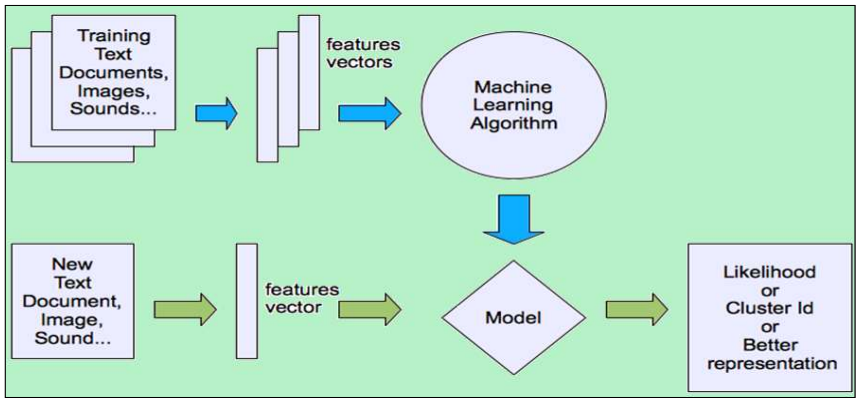


Supervised Learning

pra-sâmi

17

Machine Learning Structure

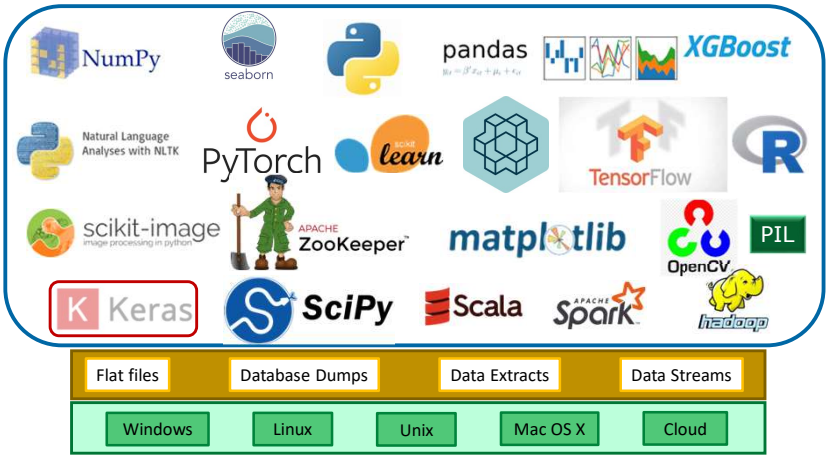


Unsupervised Learning

pra-sâmi

18

AI and ML Technology Stack...

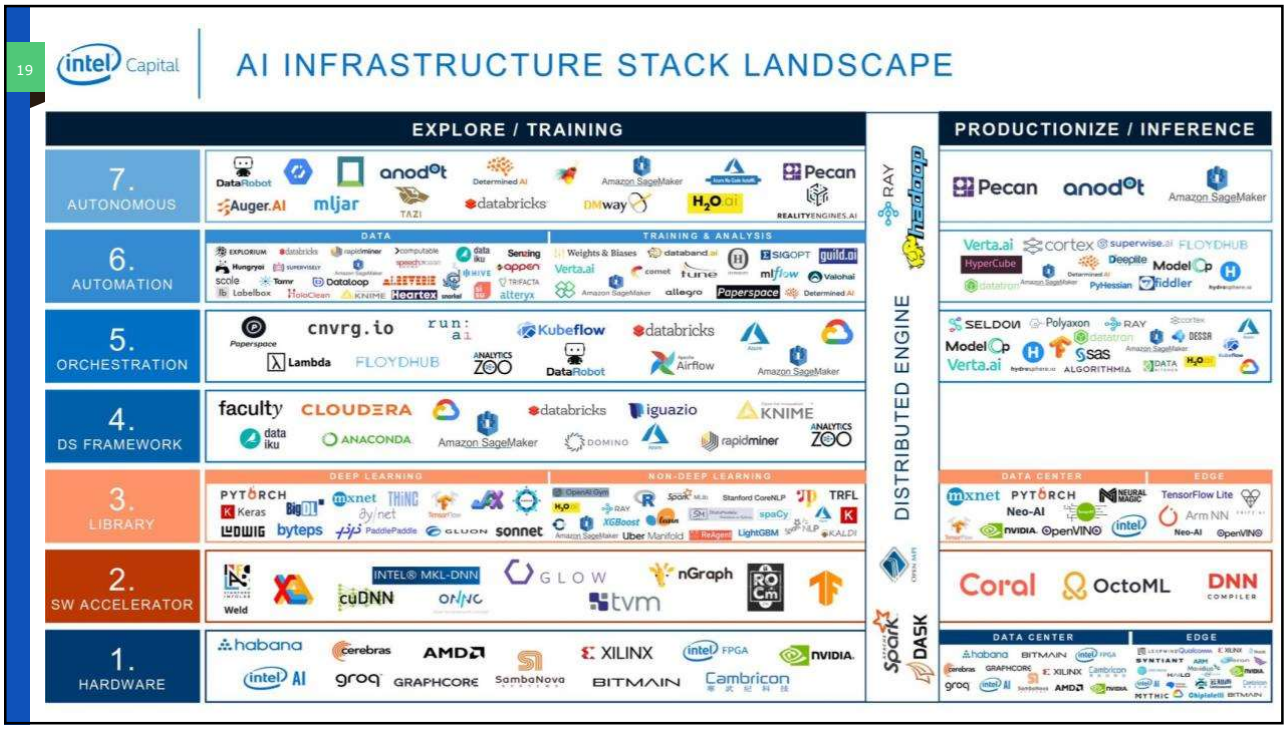


Additional

- IDE – Jupyter, Pycharm, Pydev…
- Solution design
- Implementation and execution
- Project management
- Change management

pra-sâmi

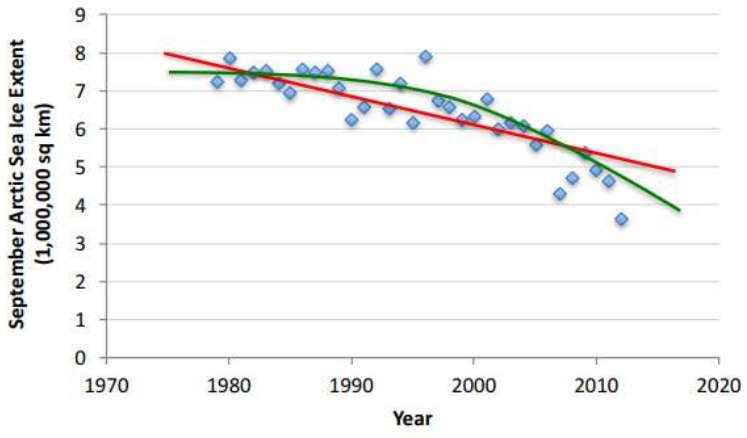
19



20

Supervised Learning: Regression

- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function f(x) to predict y given x
  - y is real-valued $\Rightarrow$ regression

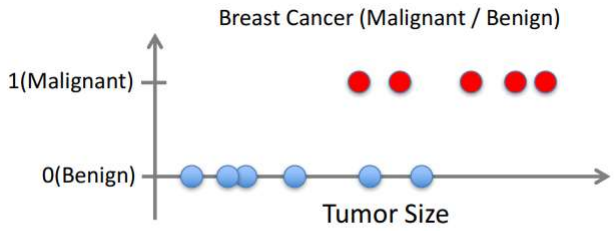


pra-sâmi

21

Supervised Learning: Classification

- Given $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- Learn a function f(x) to predict y given x
  - y is categorical ⇒ classification

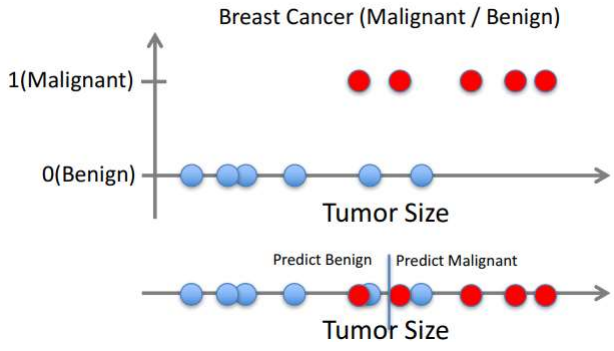


pra-sâmi

22

Supervised Learning: Classification

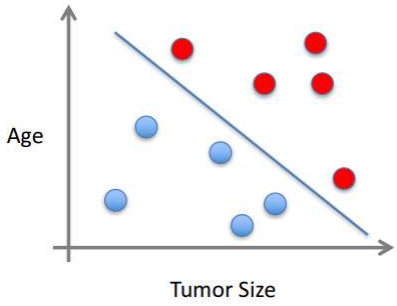
- Given $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- Learn a function f(x) to predict y given x
  - y is categorical ⇒ classification



pra-sâmi

23 Supervised Learning

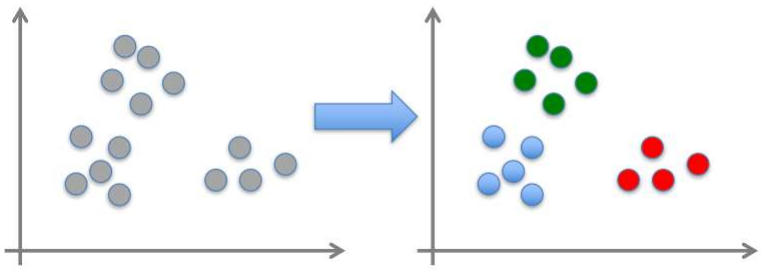
- x can be multi-dimensional
  - Each dimension corresponds to an attribute



- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape
- and other observable/ measurable parameters

24 Unsupervised Learning

- Given $x_1, x_2, \ldots, x_n$ (without labels)
- Output hidden structure behind the x's
  - e.g., clustering



25

Types of Supervised Learning

pra-sâmi

26

Classification : Definition

- Classification is a form of data analysis to extract models describing important data classes.
- Essentially, it involves dividing up objects so that each is assigned to one of a number of mutually exhaustive and exclusive categories known as classes.
  - The term “mutually exhaustive and exclusive” simply means that each object must be assigned to precisely one class
    - That is, never to more than one and never to no class at all.

pra-sâmi

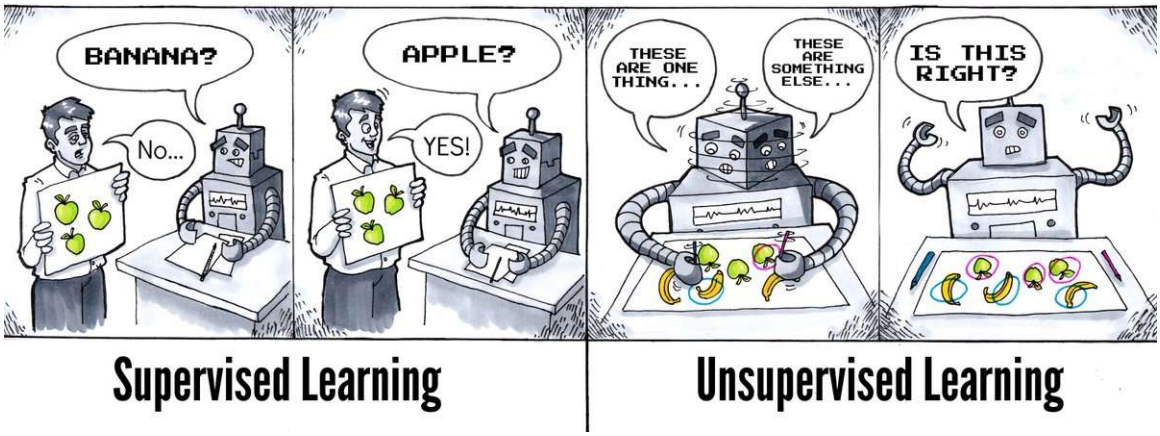
27

Classification Techniques

- Classification consists of assigning a class label to a set of new cases.
- Supervised Classification
  - The set of possible classes is known in advance.
- Unsupervised Classification
  - Set of possible classes is not known.
  - After classification we can try to assign a name to that class.
  - Unsupervised classification is called clustering.

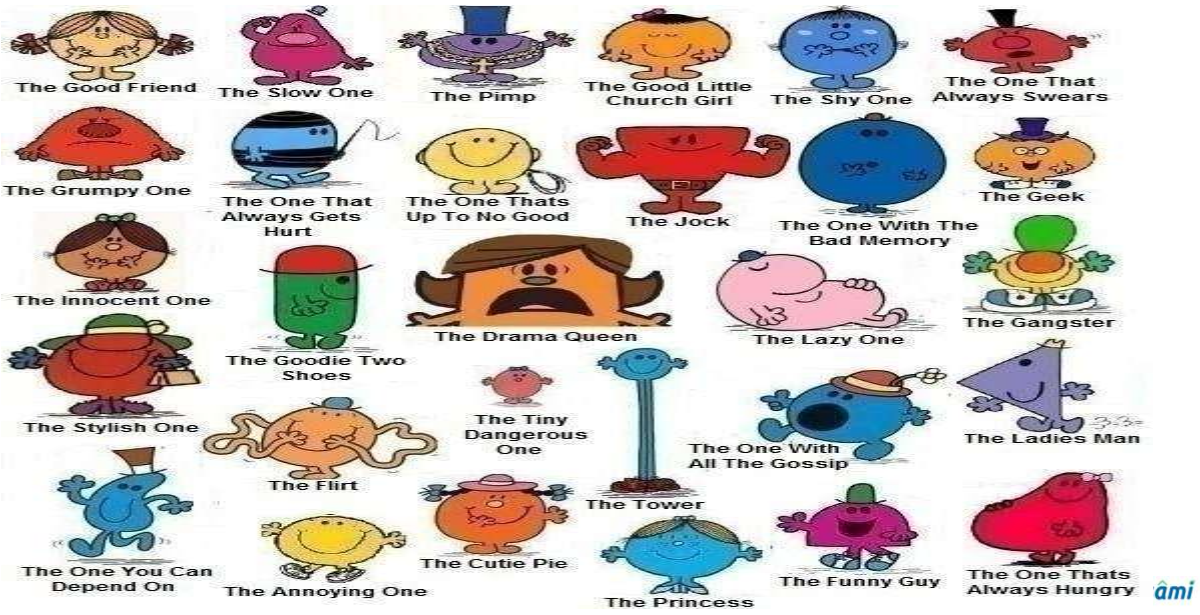
28

Supervised vs. Unsupervised



29

Supervised Classification



30

Unsupervised Classification

pra-sâmi

31

Supervised Classification Technique

- Given a collection of records (training set)
  - Each record contains a set of attributes, one of the attributes is the class.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: Previously unseen records should be assigned a class as accurately as possible.
  - Satisfy the property of "mutually exclusive and exhaustive"

32

Classification Problem

- More precisely, a classification problem can be stated as below:

Definition: Classification Problem

Given a dataset D = $\{t_1, t_2, \ldots\ldots, t_m\}$ of examples and a set of classes C = $\{c_1, c_2, \ldots\ldots, c_k\}$, the classification problem is to define a mapping $f : D \rightarrow C$,

Where each t_i is assigned to one class.

Note that example $t_i \in D$ is defined by a set of attributes $A = \{A_1, A_2, \ldots\ldots, A_n\}$.

33

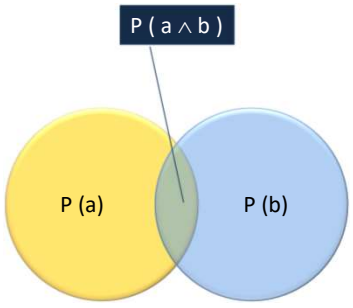
Probabilities

pra-sâmi

34

Probability Theory: Probabilities

- Some simple rules governing probabilities
- All probabilities are between 0 and 1 inclusive
  - $0 \geq P(a) \geq 1$
- If something is necessarily true it has probability 1
  - P(true) = 1; p(false) = 0
- The probability of a disjunction being true is
  - $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$
- From these laws all of probability theory can be derived.
- These axioms are sound and complete with respect to the semantics.



pra-sâmi

35

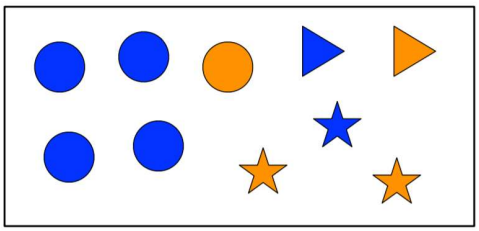
Probability Theory: Conditional Probability

- Probabilistic conditioning specifies how to revise beliefs based on new information.
- An agent builds a probabilistic model taking all background information into account.
  - This gives the prior probability
- All other information must be conditioned on
- Given evidence e (all of the information obtained), the conditional probability P(h | e) of h given e is the posterior probability of h
  - Evidence e rules out possible worlds incompatible with e.
  - $P(h \mid e) = \frac{P(h \wedge e)}{P(e)}$

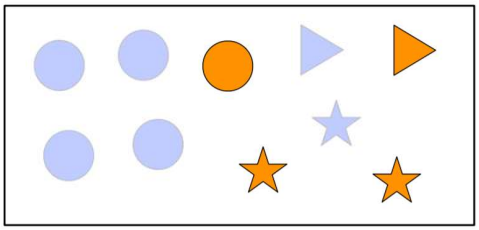
36

Probability Theory: Conditional Probability

- Possible world
- Observed color = orange



- P(Shape=circle | Color=orange) = 0.25
- P(Shape=star | Color=orange) = 0.5



37

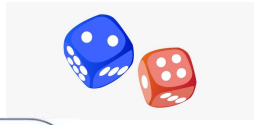
Independence

- A and B are independent iff
  - P (A | B) = P (A)
  - Or P (B | A) = P (B)
  - Or P (A, B) = P (A) . P (B)
- As we can see in case of Two dices

pra-sâmi

38

Conditional Probability



$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)}$$

$$P(\text{sum} = 12 \wedge \text{⚅}) = \frac{1}{36}$$

$$P(\text{⚅}) = \frac{1}{6}$$

$$P(\text{sum} = 12 \mid \text{⚅}) = \frac{1}{6}$$

Rolling dice is also independent of each other.

Knowing value of one dice does not change the behavior of second dice

pra-sâmi

39

Joint Porbability

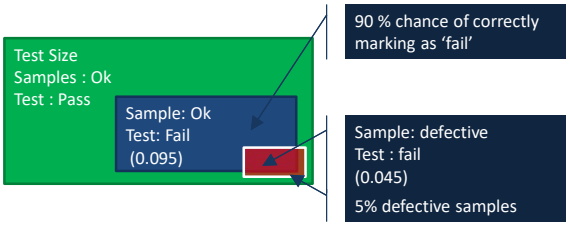
- In a factory producing mechanical parts, 5 % pieces are defective
  - P(ok) = 0.95; p(defect) 0.05
- There is a testing lab in the factory. The test is 90% accurate
  - P(correct) = 0.9; p(incorrect) = 0.1
  - There is 90% chance, that it catches failed sample – Precision
    - 10 % chance It can show failed even if the sample is ok
  - There is 90% chance, it shows pass if the sample is ok – Recall
    - 10 % chance it shows pass when the sample is not ok.
- Suppose the test result show “fail”, what is the probability that it is actually defective

pra-sâmi

40

Joint Porbability

- 5 % pieces are defective : P(ok) = 0.95; p(defect) 0.05
- The test is 90% accurate : P(correct) = 0.9; p(incorrect) = 0.1
- Joint
  - p(X) =P(defect | correct) = 0.05*0.9 = 0.045
  - p(Y) = P(ok | incorrect) = 0.95* 0.1 = 0.095
- Sum is not 1. hence the values are normalized to 1.
  - Divide both by (0.095 + 0.045) = 0.14
  - P(X) = 0.32
  - P(Y) = 0.68



pra-sâmi

41

Why is Bayes' theorem interesting?

- Often you have causal knowledge:
  - P(symptom | disease)
  - P(light is off | status of switches and switch positions)
  - P(alarm | fire)
  - P(image looks like <this> | a tree is in front of a car)
- And want to do evidential reasoning:
  - P(disease | symptom)
  - P(status of switches | light is off and switch positions)
  - P(fire | alarm)
  - P(a tree is in front of a car | image looks like <this>)

42

Exercise

- A cab was involved in a hit-and-run accident at night.
- Two cab companies, the Green and the Blue, operate in the city.
- You are given the following data:
  - 90% of the cabs in the city are Green and 10% are Blue.
- A witness identified the cab as Blue. The court tested the reliability of the witness in the circumstances that existed on the night of the accident and concluded that the witness correctly identifies each one of the two colors 75% of the time and failed 25% of the time.
- What is the probability that the cab involved in the accident was Blue?

43

Solution

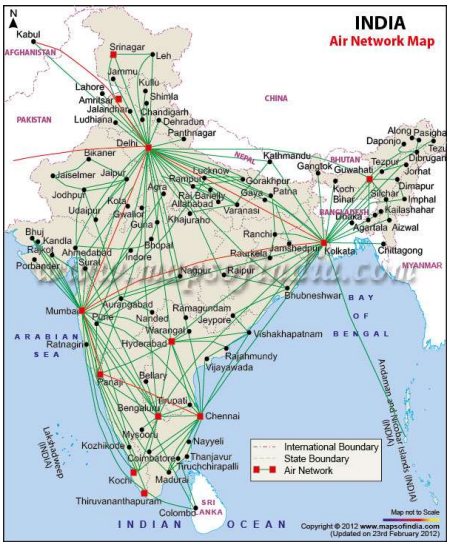
- If all taxis were blue, i.e., P(B) = 1, then obviously P(B|LB) = 1.
- Given that 9 out of 10 taxis are green
  - A taxi drawn randomly from the taxi population, we have P(B) = 0.1
- Hence
  - $P(B|LB) \propto 0.75 \times 0.1 = 0.075$
  - $P(\neg B|LB) \propto 0.25 \times 0.9 = 0.225$
- Normalizing to 1
  - $P(B|LB) = 0.075 / (0.075+0.225) = 0.25$
  - $P(\neg B|LB) = 0.225 / (0.075+0.225) = 0.75$

pra-sâmi

44

Example: Bayesian Classification - Air Traffic Data

- Let us consider a set observation recorded in a database
- Regarding the arrival of airplanes in the routes from any airport to New Delhi under certain conditions.



pra-sâmi

45

Air-Traffic Data

Days	Season	Fog	Rain	Class
Weekday	Spring	None	None	On Time
Weekday	Winter	None	Slight	On Time
Weekday	Winter	None	Slight	On Time
Holiday	Winter	High	Heavy	Late
Saturday	Summer	Normal	None	On Time
Weekday	Autumn	Normal	None	Very Late
Holiday	Summer	High	Slight	On Time
Sunday	Summer	Normal	None	On Time
Weekday	Winter	High	Heavy	Very Late
Weekday	Summer	None	Slight	On Time
Saturday	Spring	High	Heavy	Cancelled
Weekday	Summer	High	Slight	On Time
Weekday	Winter	Normal	None	Late
Weekday	Summer	High	None	On Time
Weekday	Winter	Normal	Heavy	Very Late
Saturday	Autumn	High	Slight	On Time
Weekday	Autumn	None	Heavy	On Time
Holiday	Spring	Normal	Slight	On Time
Weekday	Spring	Normal	None	On Time
Weekday	Spring	Normal	Slight	On Time

pra-sâmi

46

Air-Traffic Data

- In this dataset, there are four attributes
  - A = [Day, Season, Fog, Rain] total 20 records
- The categories of classes are:
  - C= [On Time, Late, Very Late, Cancelled]
- Given this is the knowledge of data and classes, we are to find most likely classification for any other unseen instance, for example:

Week Day	Winter	Fog High	Rain Heavy	???

- Classification technique eventually to map this record to a class.

pra-sâmi

47

Naïve Bayesian Classifier

- Example: With reference to the Air Traffic Dataset mentioned earlier, let us tabulate all the posterior and prior probabilities as shown below.

Attribute		Class			
		On Time	Late	Very Late	Cancelled
Day	Weekday	9/14 = 0.64	1/2 = 0.5	3/3 = 1	0/1 = 0
	Saturday	2/14=0 .67	0/2 = 0	0/3 = 0	1/1 = 1
	Sunday	1/14 = 0.07	0/2 = 0	0/3 = 0	0/1 = 0
	Holiday	2/14 = 0.14	1/2 = 0.5	0/3 = 0	0/1 = 0
Season	Spring	4/14= 0.29	0/2 = 0	0/3 = 0	1/1 = 1
	Summer	6/14 = 0.43	0/2 = 0	0/3 = 0	0/1 = 0
	Autumn	2/14 = 0.14	0/2 = 0	1/3 = 0.33	0/1 = 0
	Winter	2/14 = 0.14	2/2 = 1.0	2/3 = 0.67	0/1 = 0

48

Naïve Bayesian Classifier

Attribute		Class			
		On Time	Late	Very Late	Cancelled
Fog	None	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
	High	4/14 = 0.29	1/2 = 0.5	1/3 = 0.33	1/1 = 1
	Normal	5/14 = 0.36	1/2 = 0.5	2/3 = 0.67	0/1 = 0
Rain	None	5/14 = 0.36	1/2 = 0.5	1/3 = 0.33	0/1 = 0
	Slight	8/14 = 0.57	0/2 = 0	0/3 = 0	0/1 = 0
	Heavy	1/14 = 0.07	1/2 = 0.5	2/3 = 0.67	1/1 = 1
Prior Probability		14/20 = 0.70	2/20 = 0.10	3/20 = 0.15	1/20 = 0.05

49

Naïve Bayesian Classifier

- Instance:

Week Day	Winter	Fog High	Rain Heavy	???

- Case1: Class = On Time
  - P(onTime) x P(weekday | onTime) x P(Winter| onTime) x P(fogHigh| onTime) x P(rainHeavy| onTime)
  - = 0.70 x 0.64 x 0.14 x 0.29 x 0.07 = 0.0013
- Case2: Class = Late
  - P(late) x P(weekday| late) x P(Winter| late) x P(fogHigh| late) x P(rainHeavy| late)
  - = 0.10 x 0.50 x 1.0 x 0.50 x 0.50 = 0.0125
- Case3: Class = Very Late
  - P(veryLate) x P(weekday| veryLate) x P(Winter| veryLate) x P(fogHigh| veryLate) x P(rainHeavy| veryLate)
  - = 0.15 x 1.0 x 0.67 x 0.33 x 0.67 = 0.0222
- Case4: Class = Cancelled
  - P(cancelled) x P(weekday| cancelled) x P(Winter| cancelled) x P(fogHigh| cancelled) x P(rainHeavy| cancelled)
  - = 0.05 x 0.0 x 0.0 x 1.0 x 1.0 = 0.0000
- Case3 is the strongest; Hence correct classification is Very Late
  - In theory, above probabilities should be normalized to 1

pra-sâmi

50

Problem with Maximum Likelihood

- In our training set too many zeros for “cancelled”
  - E.g. P(weekday| cancelled) , P(summer| cancelled), etc.
- Zero probabilities cannot be conditioned, no matter what other evidences are
  - P(cancelled) x P(weekday| cancelled) x P(Winter| cancelled) x P(fogHigh| cancelled) x P(rainHeavy| cancelled) will **always be zero** irrespective of P(cancelled), P(Winter| cancelled), P(fogHigh| cancelled), P(rainHeavy| cancelled)
- Laplace (add-1) smoothening
  - Add one to numerator and denominator
  - P(e_i| Class) = $\frac{count(e_i,Class)}{\sum_1^n(count(e_i,Class)+1)} = \frac{count(e_i,Class)+1}{\sum_1^n(count(e_i,Class))+n}$

pra-sâmi

51

Algorithm: Naïve Bayesian Classification

Input: Given a set of *k* mutually exclusive and exhaustive classes $C = \{c_1, c_2, \ldots\ldots, c_k\}$, which have prior probabilities $P(C_1)$, $P(C_2)$,..... $P(C_k)$.

There are *n*-attribute set $A = \{A_1, A_2, \ldots\ldots, A_n\}$, which for a given instance have values $A_1 = a_1$, $A_2 = a_2$,....., $A_n = a_n$

Step: For each $c_i \in C$, calculate the class condition probabilities, $i = 1,2,\ldots\ldots,k$

$$p_i = P(C_i) \times \prod_{j=1}^{n} P(A_j = a_j | C_i)$$

$$p_x = \max\{p_1, p_2, \ldots\ldots, p_k\}$$

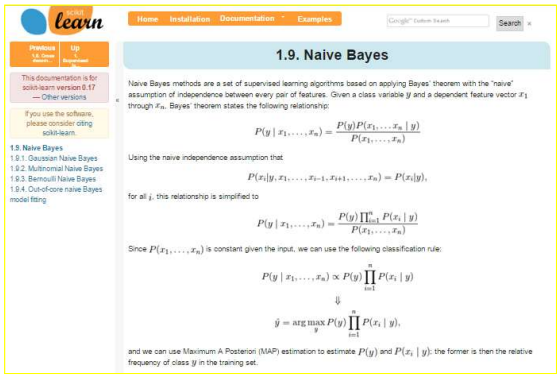
Output: C_x is the classification

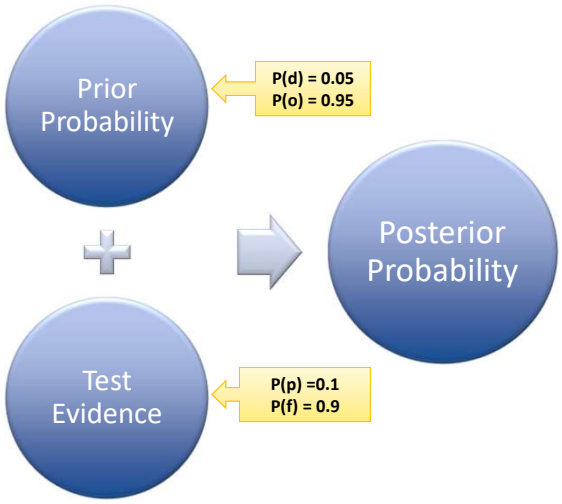
- Note: $\sum \boldsymbol{p_i} \neq \mathbf{1}$, because they are not probabilities rather proportion values (to posterior probabilities)

pra-sâmi

52

Naïve Bayes

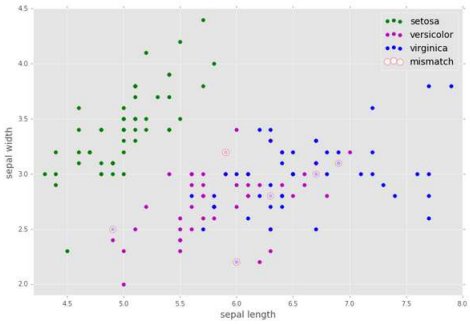




pra-sâmi

53 Analyzing using Gaussian Naïve Bayes

```
#Lets make predictions using Gaussian Naive Bayes
from sklearn.naive_bayes import GaussianNB
clf = GaussianNB()
clf.fit(features, labels)
y_pred = clf.predict(features)
```



- Read data
- Inspect features
- Import GuassianNB
- Instantiate and fit
- Make predictions

pra-sâmi

54 Naïve Bayes

# of features

Varieties of Iris

# of instances

150	4	setosa	versicolor	virginica
5.1	3.5	1.4	0.2	0
4.9	3	1.4	0.2	0
4.7	3.2	1.3	0.2	0
4.6	3.1	1.5	0.2	0
5	3.6	1.4	0.2	0
5.4	3.9	1.7	0.4	0
4.6	3.4	1.4	0.3	0
5	3.4	1.5	0.2	0
4.4	2.9	1.4	0.2	0
4.9	3.1	1.5	0.1	0
5.4	3.7	1.5	0.2	0
4.8	3.4	1.6	0.2	0
4.8	3	1.4	0.1	0
4.3	3	1.1	0.1	0

Instance

- Number of Instances: 150 (50 in each of three classes)
- Number of Attributes: 4 numeric, predictive attributes and the class
- Attribute Information:
  - sepal length in cm
  - sepal width in cm
  - petal length in cm
  - petal width in cm
  - class:
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica
- Class Distribution: 33.3% for each of 3 classes.

pra-sâmi

55

Advantages and Disadvantages

Advantages:

- Fast and space efficient
  - Look up all the probabilities with a single scan of the database
- NOT sensitive to irrelevant features...
- Handles real and discrete data well
- Handles streaming data well

Disadvantages:

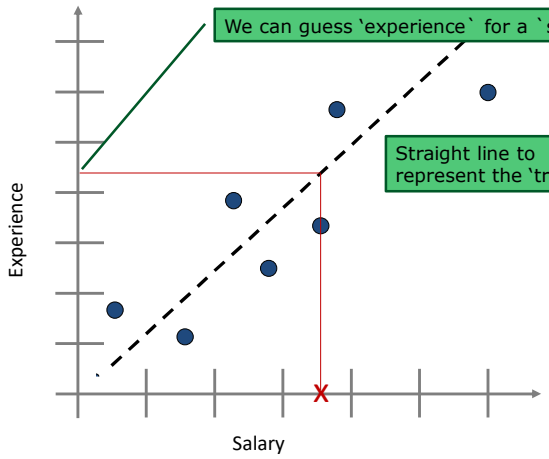
- Assumes independence of features
- It relies on all attributes being categorical
- If the data is less, then it estimates poorly

56

Linear Regression

57

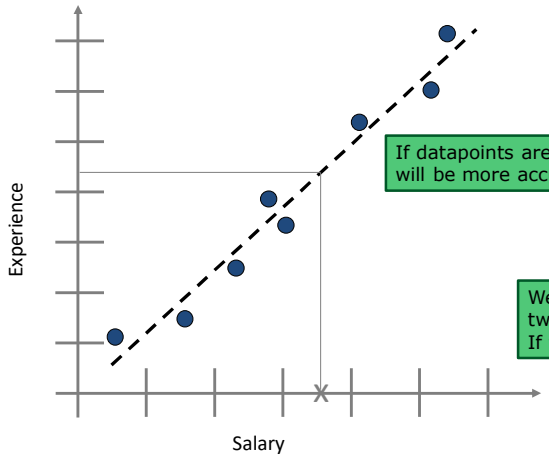
Correlation



At present, datapoints are somewhat away from the trend line.

58

Correlation



If datapoints are closer to the trend line, our guestimates will be more accurate.

We can say, if datapoints are away from the trend line, two features have weaker relation.
If they are close together, the relationship is strong.

59

Correlation

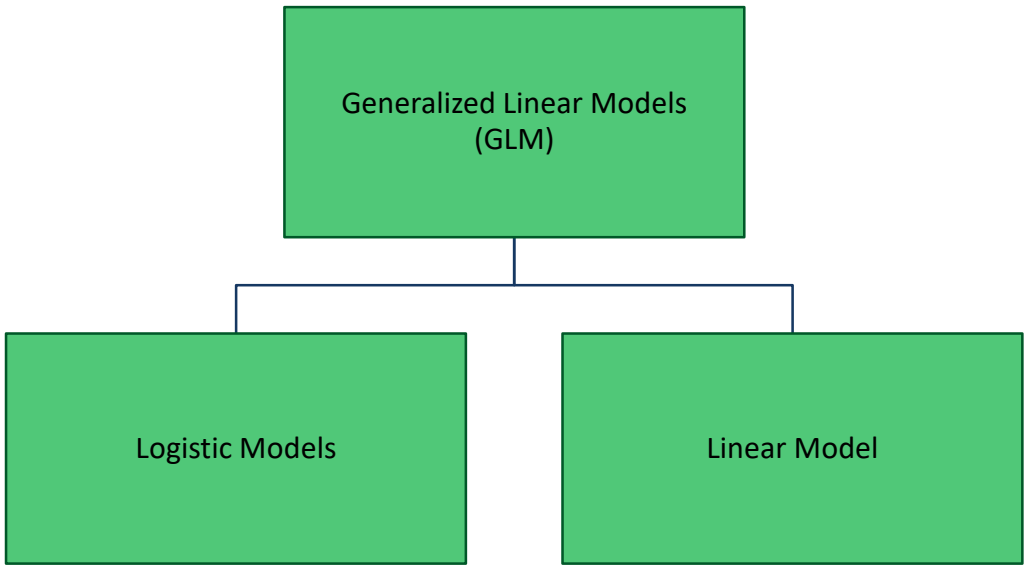
We can hope for best fit at correlation = 1
The trendline passes through each of the point.

PS: Worst case is when correlation becomes 0.

pra-sâmi

60

Generalized Linear Model

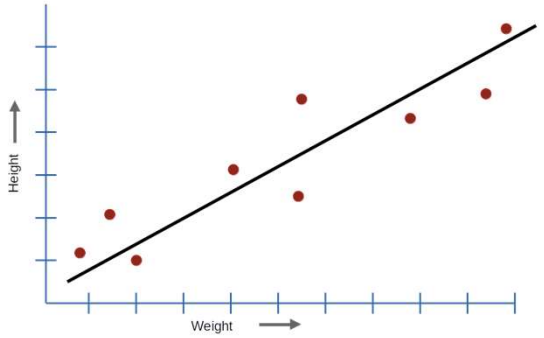


pra-sâmi

61

Fitting a line to data

- Imagine you have Height and Weight data
- How do we make prediction given a weight?
- We fit a line through the data

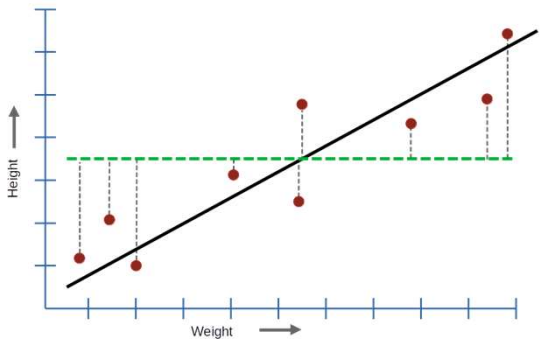


pra-sâmi

62

Residuals and squared error

- Lets assume a horizontal line at mean height
- Calculate how far away each point is from this line.
  - The distances are called residuals
- If we simply take the difference, error will cancel out
- Better approach will be to calculate sum of squared errors
  - $\sum(h_i - pred_i)^2$

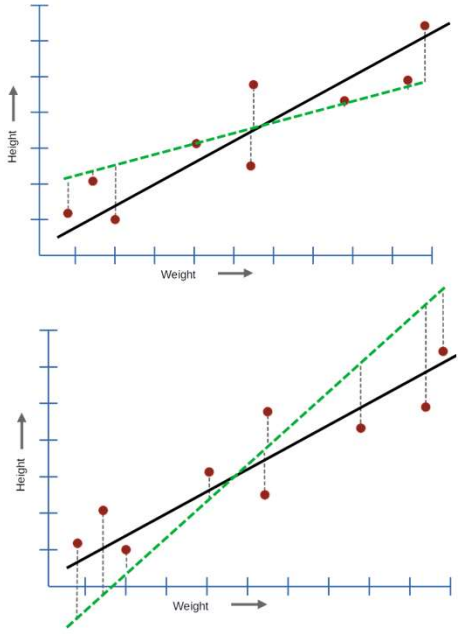


pra-sâmi

63

Least Squared Error

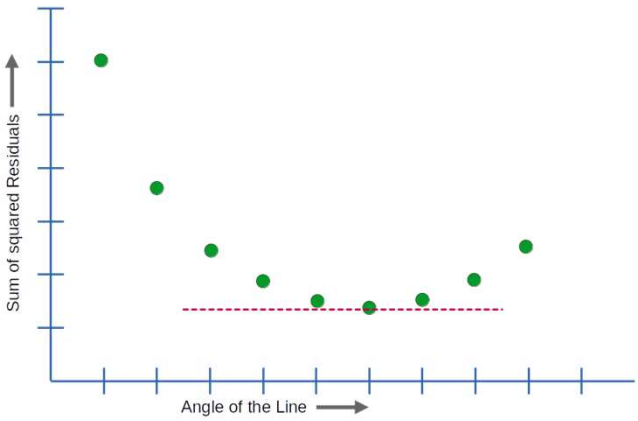
- If we rotate the line counter clock wise, errors will reduce
- If we keep rotating further the errors will start increasing
- Lets plot residual error wrt rotation



pra-sâmi

64

Least Squared Error

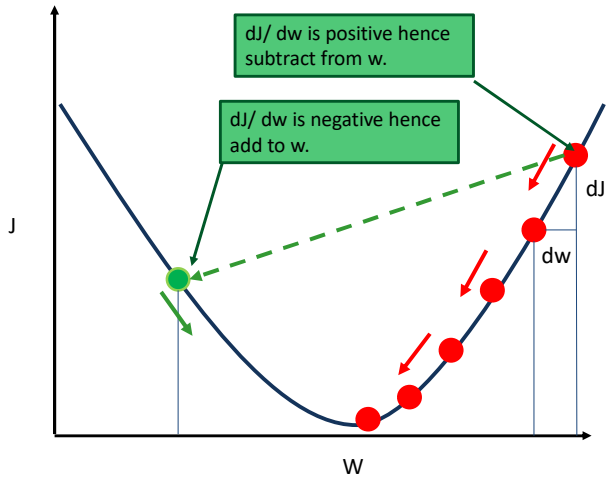


For one particular rotation, the sum of squared errors will be minimum and that's the rotation we are looking for.

pra-sâmi

65

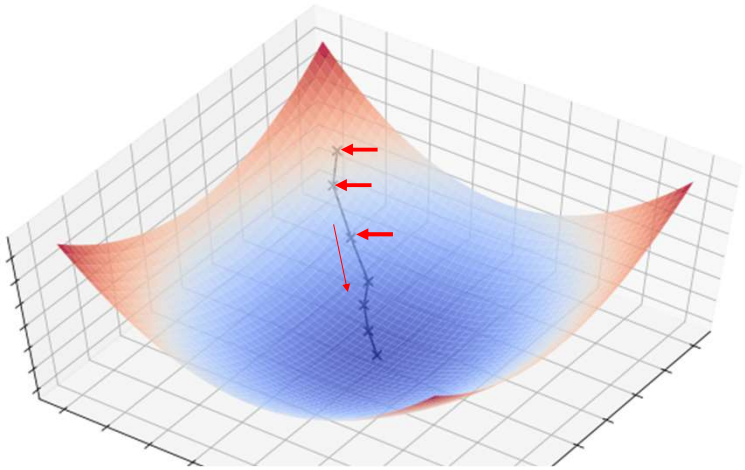
Gradient Descent

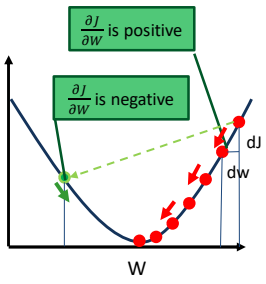


pra-sâmi

66

Solve for Optimization



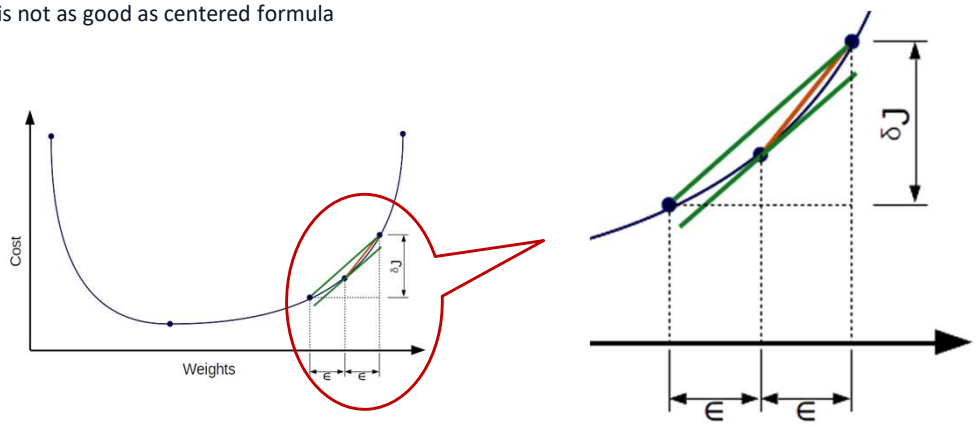


pra-sâmi

67

Calculation of Derivative

- Use the centered formula
  - The formula you may have seen for the finite difference approximation when evaluating the numerical gradient is not as good as centered formula

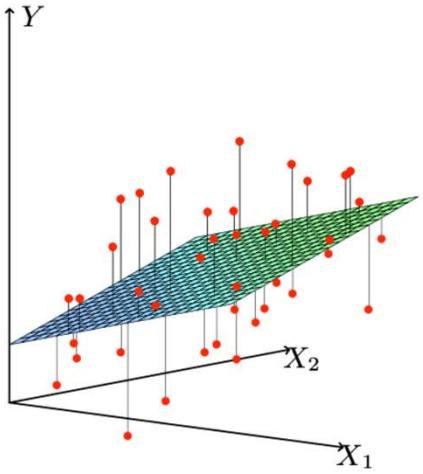


pra-sâmi

68

Linear Regression with Multiple Features

- Imagine that we have one more feature
  - Zodiac Sign
- All our calculations will be same
  - This time we will be fitting a plane instead of line
- If Zodiac sign has no bearing on Height
  - Probably slope in that direction will be zero
- Note: adding extra parameter will not make predictions any worst!

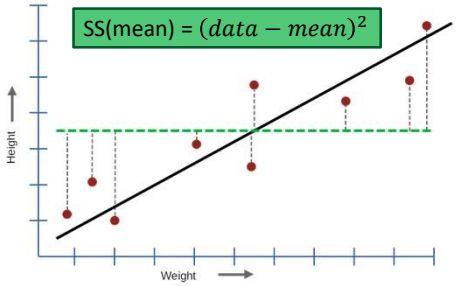


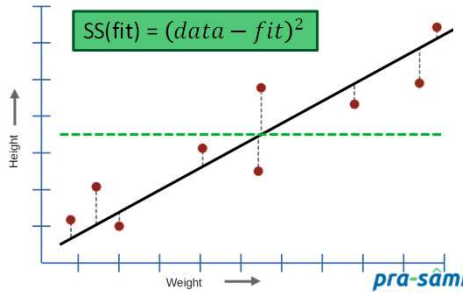
pra-sâmi

69

How Good is Fitted Line? - Calculate R^2

- Straight line has two parameters
  - Slope and intercept
- Calculating R^2
  - Sum of squared residuals around mean = SS(mean)
    - Variation around mean = SS(mean) /n
  - Sum of squared residuals around fit = SS(fit)
    - Variation around fit = SS(fit) /n
- R^2 tells us how much of variation in Height can be explained by Weight
  - $R^2 = \frac{Var(mean) - Var(fit)}{Var(mean)}$ or
  - $R^2 = \frac{SS(mean) - (fit)}{SS(mean)}$ in other words
  - $R^2 = \frac{Variation\ in\ Height\ explained\ by\ Weight}{Overall\ variation\ in\ Height}$





pra-sâmi

70

p-Value

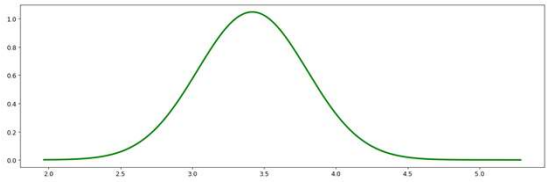
- Remember $R^2 = \frac{Variation\ in\ Height\ explained\ by\ Weight}{Overall\ variation\ in\ Height}$
- Calculate 'F'
  - $F = \frac{Variation\ in\ Height\ explained\ by\ Weight}{Variation\ in\ Height\ \mathbf{not}\ explained\ by\ weight}$
  - The residuals left after fitting the line are '**not**' explained by weight
- $F = \frac{\frac{[SS(mean) - SS(fit)]}{p_{fit} - p_{mean}}}{\frac{SS(fit)}{n - p_{fit}}}$ (p represents number of parameters)
  - p_{fit}= 2 as it has two parameters and p_{mean}=1 as it has only one parameter
  - n is number of datapoints

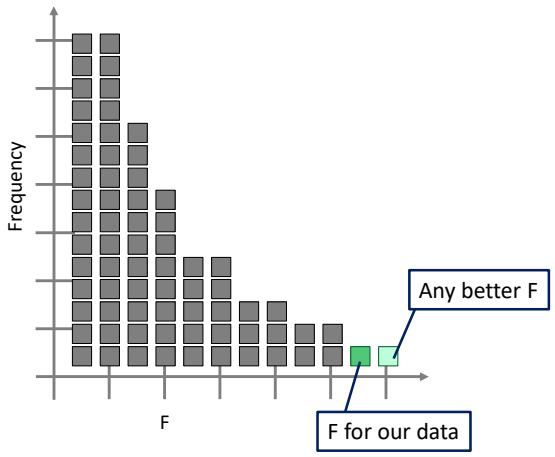
pra-sâmi

71

p-Value

- Keep generating weight data at random and keep measuring F,
  - The frequency distribution will be exponential
- p-Value is ratio of frequency of F for current distribution plus any other distributions which can produce higher than F to all possible F values
  - We need P-value to be small



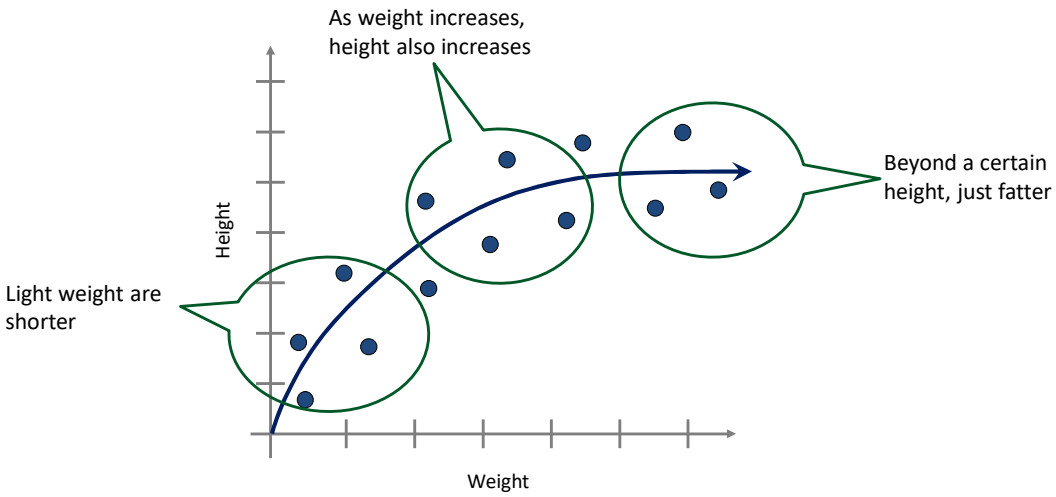


72

Bias vs. Variance

73

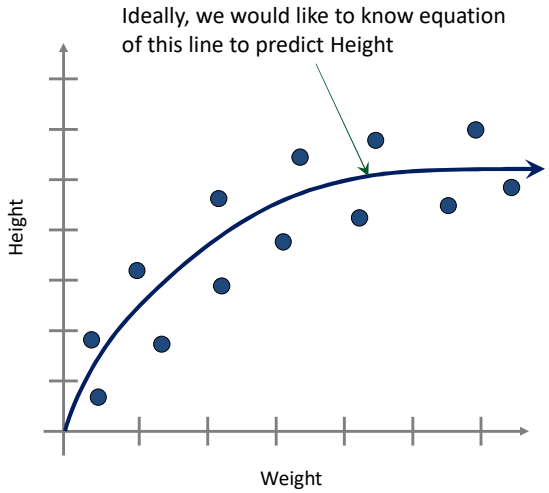
The Data and True Model



pra-sâmi

74

The Data and True Model

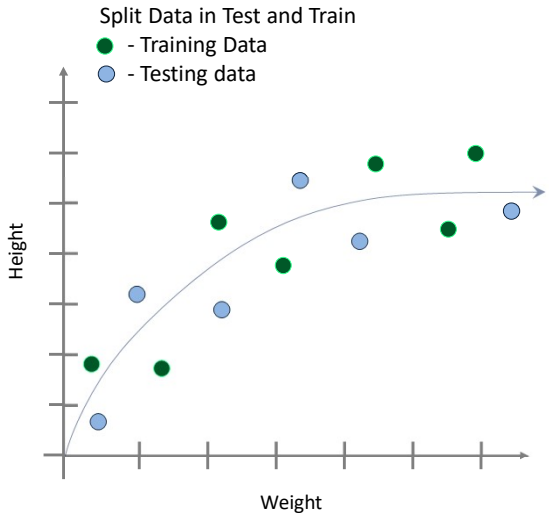


In this case lets
use ML to make
predictions

pra-sâmi

75

Train-Test Split



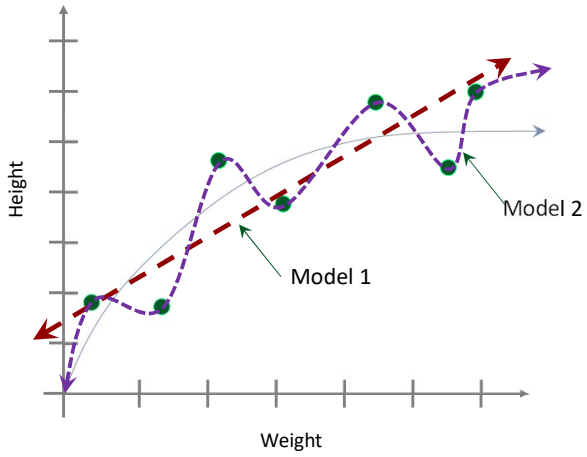
pra-sâmi

76

Fitting a Model

Removing the Testing Data

We can fit two models to the data.

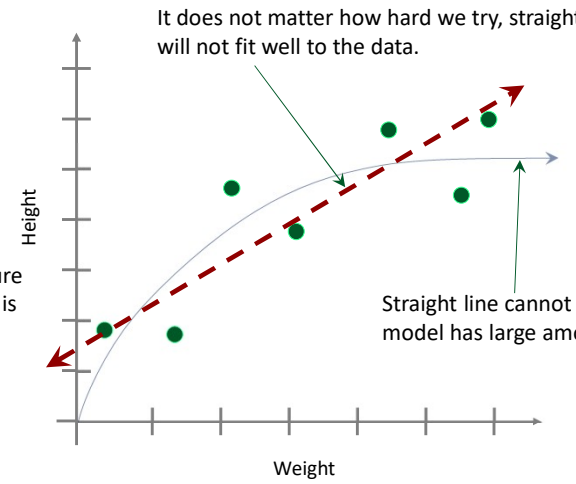


pra-sâmi

77

Fitting a Simple Model

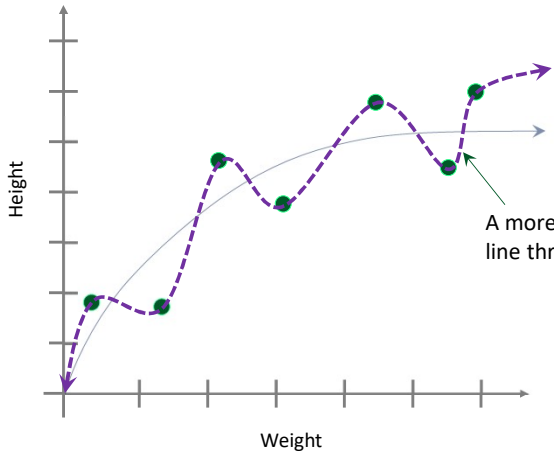
Inability of ML algorithm to capture true relationship in Training Data is called **Bias**



pra-sâmi

78

Fitting a Complex Model

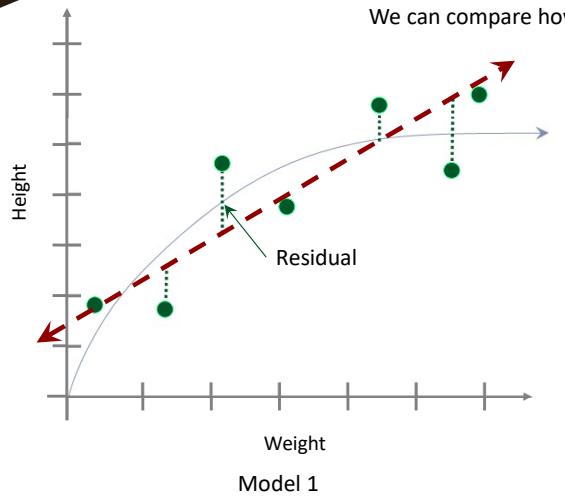


pra-sâmi

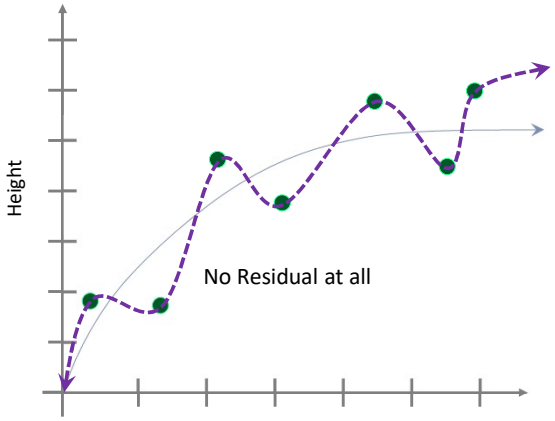
79

Comparison

We can compare how well two models are fitting **Training Data**



Model 1



Model 2

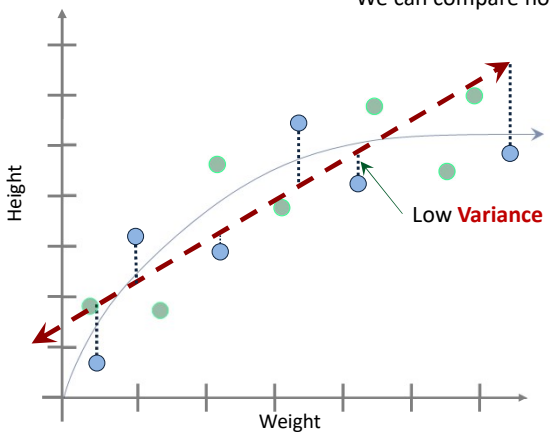
Complex model does a good job at Training!

pra-sâmi

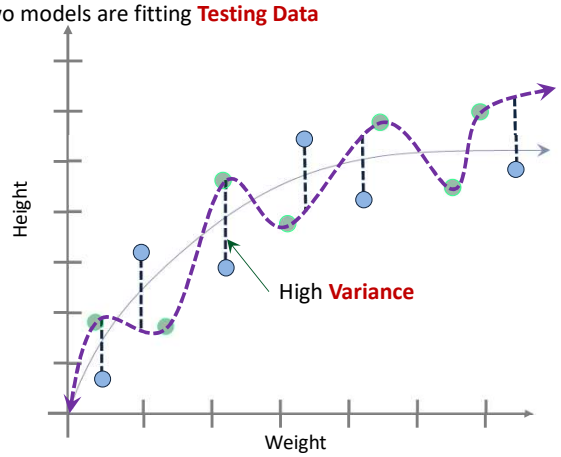
80

Comparison

We can compare how well two models are fitting **Testing Data**



Model 1



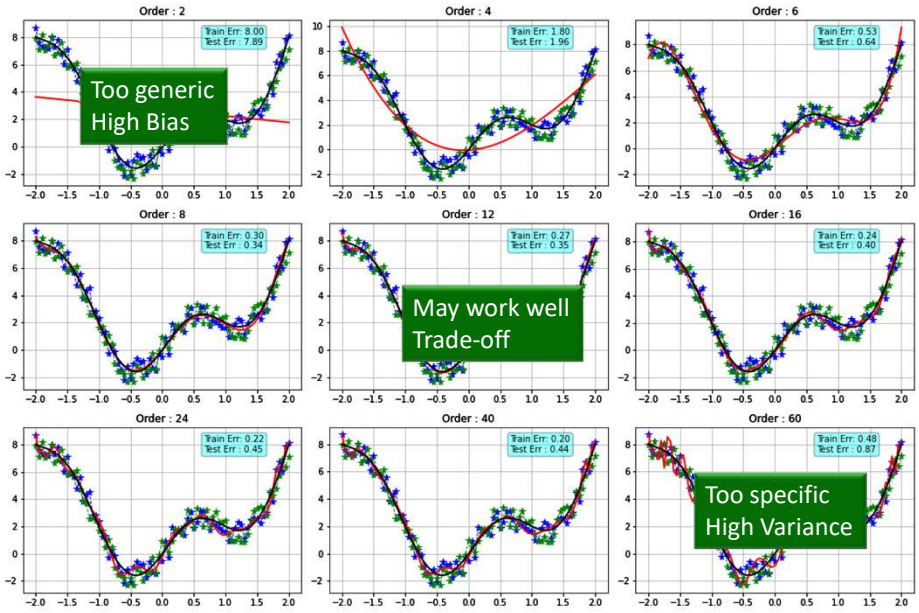
Model 2

Simple model does similar work at Testing data
However, performance of Complex model is worst!

pra-sâmi

81

Under-fitting vs. Over-fitting

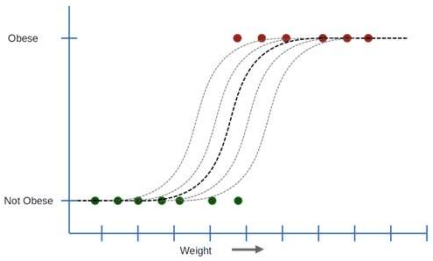


pra-sâmi

82

Odds and Odds Ratio

- Odds of being Obese = 6/7
  - 6 obese person compared to 7 non-obese person
- Odds Ratio is "Ratio of Odds" = $\frac{6/7}{3/5}$
  - Taking log helps in scaling to similar values
- Given an Obese person, odds that person is vegetarian = $\frac{23}{117}$
- Given an non-obese person, odds that person is vegetarian = $\frac{6}{210}$
- Odds Ratio = $\frac{23/117}{6/210}$
- Odds Ratio is telling us odds are 6.88 times that someone obese will also be non-vegetarian (log odd ratio is 1.93)



		Vegetarians		
		Yes	No	
Obese	Yes	23	117	140
	No	6	210	216
Total		29	327	

pra-sâmi

83

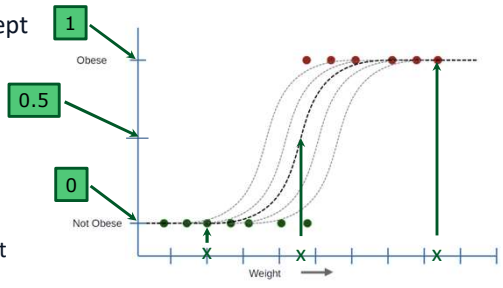
Logistic Regression

pra-sâmi

84

Logistic Regression

- Logistic regression is similar to linear regression, except
  - It's a classification task
  - Predicts class, rather than continuous value
  - It fits 'S' shaped logistic function curve
- Logistic function curve goes from 0 to 1
  - Curve tells you if the mouse is obese based on its weight
- The manner in which this line is fitted is also different
  - Linear Regression → Least squares
  - Logistic Regression → Maximum Likelihood

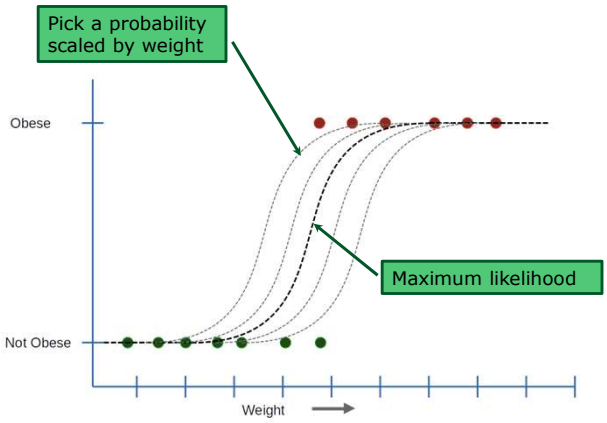


pra-sâmi

85

Maximum Likelihood

- Unlike linear regression, logistic regression fits an 'S' shaped "**Logistic Function**"
  - Curve is for probability that a person is obese based on its weight
- Calculate the likelihood of the observed data - obese or not
- Multiply likelihood of all data
  - Likelihood of the data given this distribution
- Keep shifting the line and calculate likelihood
- Look for the distribution resulting in maximum likelihood



pra-sâmi

86

Coefficients

- At present Y axis varies from 0 to 1
- Lets transform this axis to range from -∞ to +∞
  - From likelihood scale to log(odds of obesity) = $\log(\frac{p}{1-p})$

#	p	$Log_e(\frac{p}{1-p})$
1	0.5	0
2	0.731	1
3	0.88	2
4	0.95	3
5	1	∞

- With log (odds) on y-axis 'S' curve becomes straight line

- At present Y axis varies from 0 to 1
- Lets transform this axis to range from -∞ to +∞
  - From likelihood scale to log(odds of obesity) = $\log(\frac{p}{1-p})$

- In short it varies from -∞ to -∞
- With log (odds) on y-axis 'S' curve becomes straight line

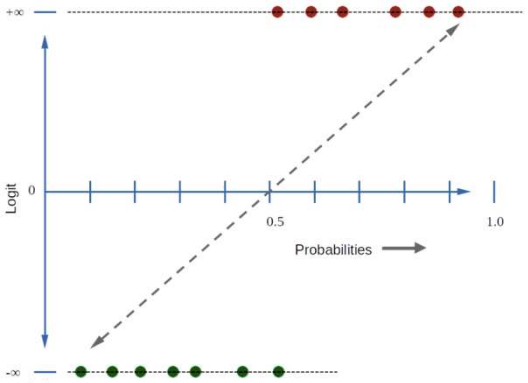
pra-sâmi

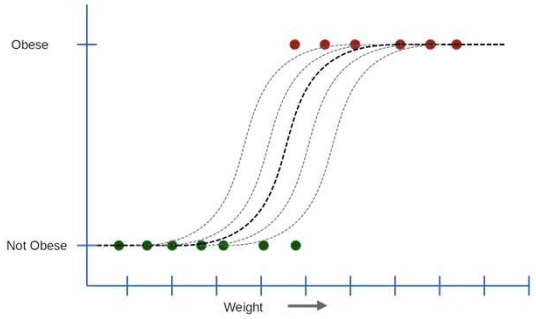
87

Coefficients

With log (odds) on y-axis 'S' curve becomes straight line

- Coefficients are presented in terms of log (odds).
- Coefficients are similar to Linear Regression (y = -4.78 + 0.89 * Weight)





pra-sâmi

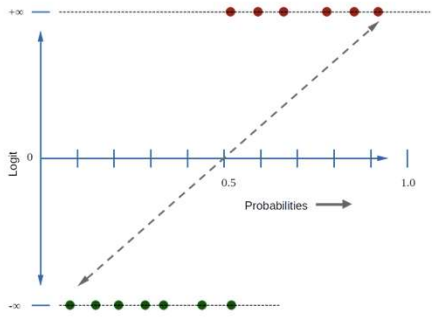
88

Fitting the line

- Use maximum likelihood to find best fitting line
- Project each of the data point onto candidate line
  - Candidate log(odd) value
- Transform the candidate log(odd) to candidate probability
  - p = $\frac{e^{\log(odd)}}{1-e^{\log(odd)}}$

Also Called 'Sigmoid' Function

- Plot them back on S curve and observe values on y-axis (likelihood)
  - These are estimated values
  - Probabilities of Not Obese is 1-p



pra-sâmi

89

Overall Likelihood

- Likelihood of data
  - p_o^1 x p_o^2 x p_o^3 x … p_o^n x (1-p_{no}^1) x (1- p_{no}^2) x (1- p_{no}^3) x … x (1–p_{no}^m)
  - Note : o – obese; no – not obese
- Its better if we calculate log of likelihood
  - log(p_o^1) + log(p_o^2) + log(p_o^3) + … + log(p_o^n) + log(1-p_{no}^1) + log(1- p_{no}^2) + log(1- p_{no}^3) + … + log(1- p_{no}^m)
  - = -2.85 (say)
- Rotate the line and calculate again.
- Finally we find the line that maximizes the likelihood

- Likelihood of data
  - p_o^1 x p_o^2 x p_o^3 x … p_o^n x (1-p_{no}^1) x (1- p_{no}^2) x (1- p_{no}^3) x … x (1-p_{no}^m)
  - Note : o – obese; no – not obese
- Its better if we calculate log of likelihood
  - log(p_o^1) + log(p_o^2) + log(p_o^3) + … + log(p_o^n) + log(1-p_{no}^1) + log(1- p_{no}^2) + log(1- p_{no}^3) + … + log(1- p_{no}^m)
  - = -2.85 (say)
- Rotate the line and calculate again.
- Finally we find the line that maximizes the likelihood

pra-sâmi

90

McFadden's Pseudo R^2

- Log(likelihood of data given 'S' curve)
  - log(p_o^1) + log(p_o^2) + log(p_o^3) + … + log(p_o^n) + log(1-p_{no}^1) + log(1- p_{no}^2) + log(1- p_{no}^3) + … + log(1- p_{no}^m)
  - Call this LL(fit); substitute for SS(fit)
  - We need an estimate for something analogous to SS(mean) ➔ without using weight
- That's simple ➔ number of sample marked as obese divided by total number of samples
  - = $\frac{Nu \quad obese}{Total}$
  - LL(overall) = log($p^{overall}$) + log($p^{overall}$) + log($p^{overall}$) + … + log($p^{overall}$) + log(1-$p^{overall}$) + log(1- $p^{overall}$) + log(1-$p^{overall}$) + … + log(1-$p^{overall}$)
- For sample size of 9 with 5 obese records; $p^{overall}$ = 5/9 = 0.56
  - LL(overall) = log(0.56) + log(0.56) + log(0.56) + … + log(0.56) + log(1-0.56) + log(1-0.56) + log(1-0.56) + … + log(1-0.56)
- $R^2 = \frac{LL(overall) - LL(fit)}{LL(overall)}$

pra-sâmi

91

Log (likelihood)

- Log(likelihood) =
  - $\log(p_o^1) + \log(p_o^2) + \log(p_o^3) + \ldots + \log(p_o^n) +$
    $\log(1-p_{no}^1) + \log(1-p_{no}^2) + \log(1-p_{no}^3) + \ldots + \log(1-p_{no}^m)$

Simple log where ground truth is 1

Two different treatment

Log(1-p) where ground truth is 0

- What if we replace
  - $\log(p_o^1) = y_o^1 * \log(p_o^1) + (1 - y_o^1) * \log(1 - p_o^1)$
    $\log(1-p_{no}^1) = y_{no}^1 * \log(1-p_{no}^1) + (1-y_{no}^1) * \log(1-p_{no}^1)$

Note: ground truth y_o^1 for obese is 1

Note: : ground truth y_{no}^1 for not obese 0

- Both have similar form, hence in general
  - Log(likelihood) = $\sum y^i * \log(p^i) + (1 - y^i) * \log(1 - p^i)$

pra-sâmi

92

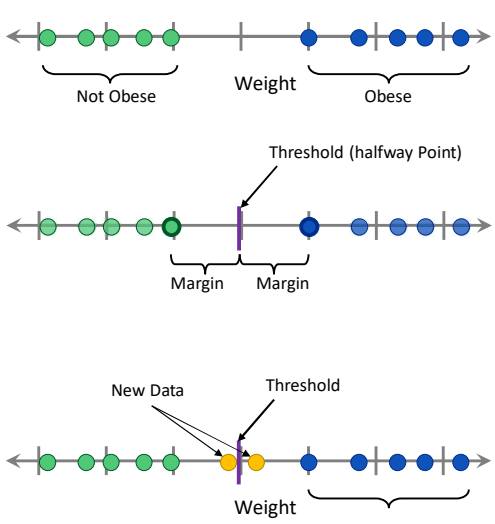
Support Vector Machines

pra-sâmi

93

Maximum Margin Classifier

- Given a classification data of obese and not obese
- We can pick a threshold somewhere at mid point to separate between `obese` and `not obese`
  - Shortest Distance between threshold and datapoints is called 'margin'
  - If we move the threshold to the left then Margin will be smaller of the two
- When new data comes base on threshold we can pick its class
- This is referred as 'Maximum Margin Classifier'

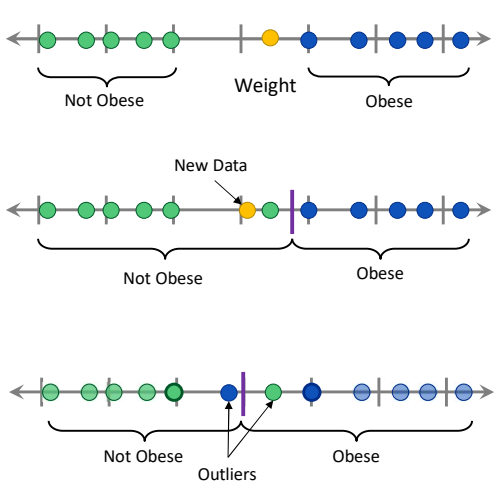


pra-sâmi

94

Outliers are a Challenge

- What if we have observation close to Obese group?
  - But it is not obese - Higher muscle mass!
  - Max Margin Classifier will be close to Obese Observations
  - New observation as shown, will be marked as Not Obese, it is quite far from Not obese
- Max. Margin is very sensitive to Outliers!
- Keep the Threshold as per majority and allow some misclassifications
  - Bias-Variance Tradeoff – use Soft Margin
  - Which is right value for Margin – Cross Validation

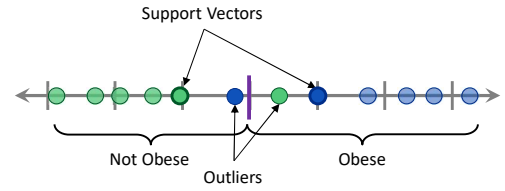


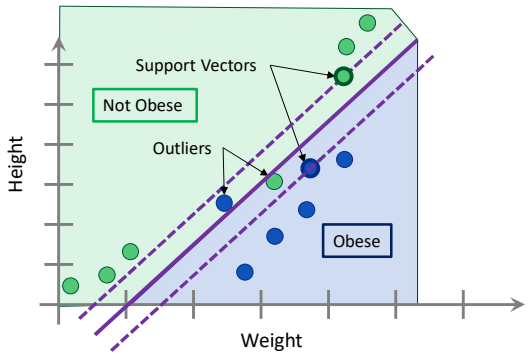
pra-sâmi

95

Support Vector Classifier

- Threshold is selected so that it balances between Bias and Variance
- What's good value of Soft Margin
  - Cross validation helps us best classification
- If we also have weight data
  - The data would be 2 Dimensional
  - Support vector classifier is a line
- Similarly, for 3-D data; classifiers will be a plane
  - All these classifiers are called hyperplane



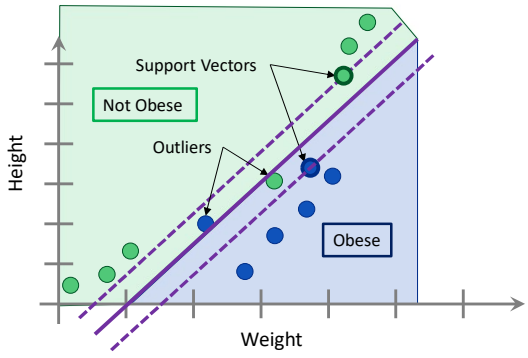


pra-sâmi

96

Support Vector Classifiers

- Assume that, we have same data but no Heights
  - Data would look like figure at the bottom
  - All Obese are concentrated in the center with non-obese on both side
  - No matter where we put classifier error will be high
- Support Vector Classifiers will not be able to handle such data
- Welcome **Support Vector Machines**

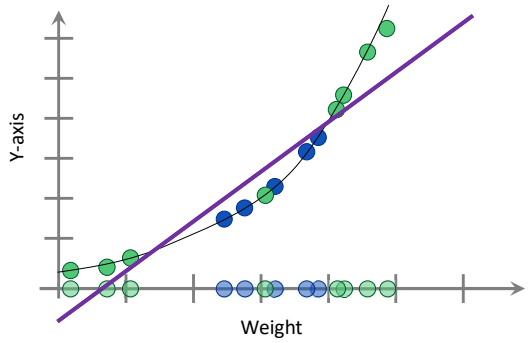


pra-sâmi

97

Intuition behind Support Vector Machines

- Add a y-axis
- X-axis will remain weights which we have observed
  - Y-axis coordinate → $(\text{Weight})^2$
- Not we can separate the data into respective classes easily



pra-sâmi

98

Support Vector Machines

- Start in some low dimensions
- Move the data into higher dimensions
- Find a Classifier that separates the datapoints
- Support vector machines use Kernel Functions to find Support Vector Classifiers
- Example:
  - Suppose we use polynomial Kernel with degree d = $(a\ x\ b\ + r\ )^d$
  - = $a\ x\ b\ +\ a^2 x\ b^2 + \frac{1}{4}$ For r = ½ and d =2; a and b are two different observations.
  - = $(a, a^2, \frac{1}{2})$. $(b,\ b^2, \frac{1}{4})$ { a dot product}
  - First term is base axis, second term is second axis and so and so forth

pra-sâmi

99

Radial Basis Function (RBF)

pra-sâmi

100

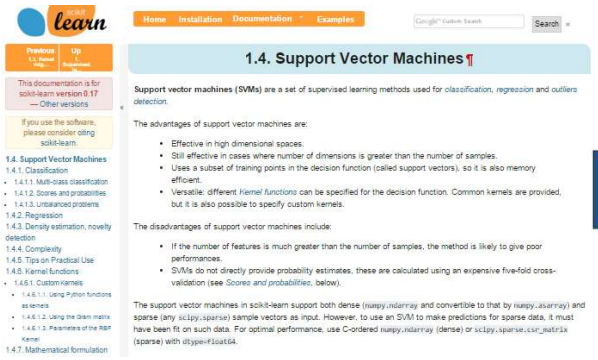
The Radial Kernels

- Another way to deal with previously shown data is Radial Kernel or Radial Basis Function (RBF)
- It finds Support Vector Classifiers in Infinite dimensions
- Intuitively, Radial Kernels behave like a weighted Nearest Neighbor
  - Closest observation has lot of influence on how we classify new observation
- Calculating influence = $e^{-\gamma\,(a-b)^2}$ $e^{-\gamma\,(a-b)^2}$
- a and b refer to two different observations
- γ scales the influence and is determined by cross validation

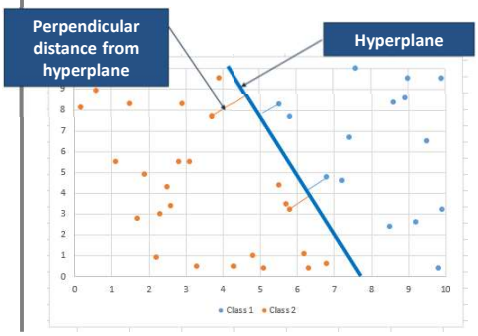
pra-sâmi

101

Support Vector Machines



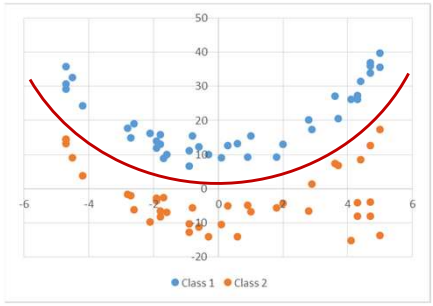
- Assumes that these classes are linearly separable
- Support Vector Machines (SVM) finds a hyperplane to be as far as possible from the closest members of both classes.



pra-sâmi

102

Support Vector Machines

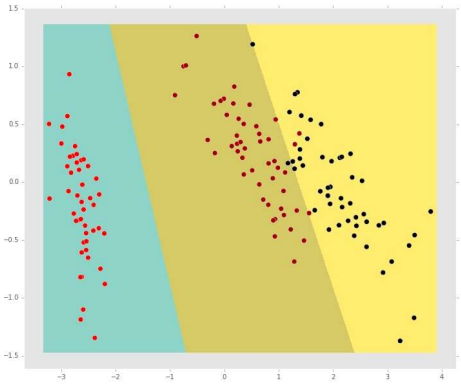


- What if two classes are not linearly separable?
- We can still find a n^{th} order plane which can separate the two classes
  - Use of Kernel

pra-sâmi

103 Support Vector Machines

```
#Lets make predictions using Support Vector Machines
from sklearn.svm import SVC
clf = SVC(kernel='linear')
```



- Basic steps remain the same
- Import svm
- Instantiate a classifier
- Fit the classifier on training data
- Make predictions on test data
- For ease of comparison, the decision surface is also plotted.

pra-sâmi

104 SVM Parameters

- SVC: (sklearn.svm.SVC (C = 1.0, kernel='rbf', degree=3, gamma = 'auto')
- Kernel:
  - Linear
  - RBF
  - Polynomial
- C = Penalty parameter
  - A hyperplane separating all classes can lead to poorly fit models especially if there are unusual cases.
  - C allows for some examples to be "ignored" or placed on the wrong side of the hyperplane.
  - Large C ⇒ low bias and high variance.
  - Small C ⇒ higher bias and lower variance.
- Gamma:
  - Small gamma ⇒ sharp curves in the higher dimensions ⇒ low bias and high variance
  - Large gamma ⇒ a smooth curves ⇒ higher bias and low variance
- Degree:
  - Degree of the polynomial kernel function ('poly'). Ignored by all other kernels.

pra-sâmi

105

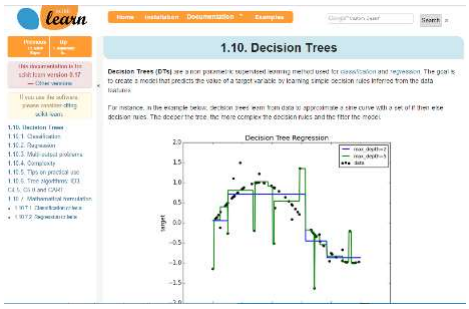
Advantages and Disadvantages

Advantages	Disadvantages
❑ Effective in high dimensional spaces	❑ If the number of features is much greater than the number of samples, the method is likely to give poor performances.
❑ Still effective in cases where number of dimensions is greater than the number of samples	❑ SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation
❑ Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient	
❑ Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.	

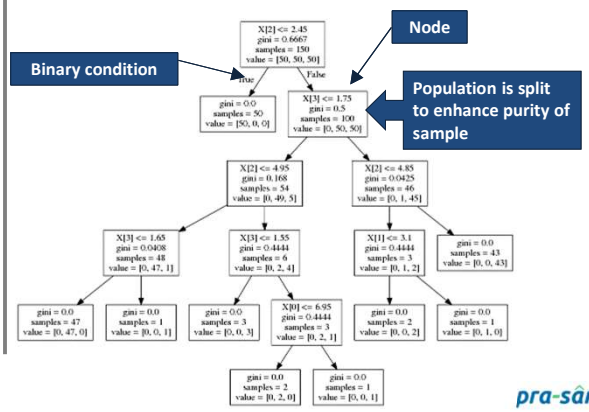
pra-sâmi

106

Decision Trees



- ❑ Decision tree is most intuitive and most popular.
- ❑ Decision Tree is like “20 questions” game… based on yes/no answers, one can find the answer.



pra-sâmi

107 Entropy

- Entropy is measure of impurity of a collection.
  - All the samples are of same type $\Rightarrow$ low impurity
    $\Rightarrow$ entropy = 0
  - Evenly spread sample $\Rightarrow$ high impurity
    $\Rightarrow$ entropy = 1

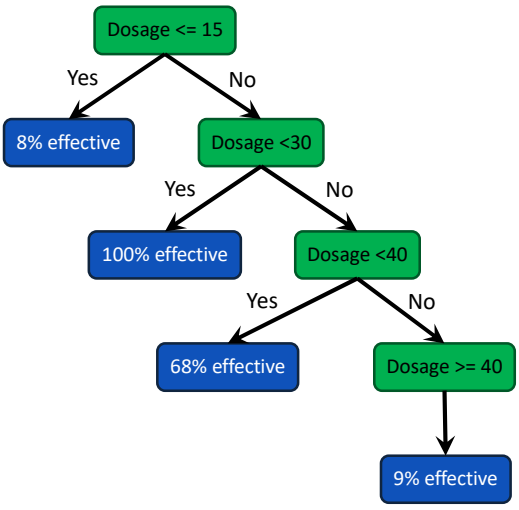
- Circles = 6
- Stars = 6

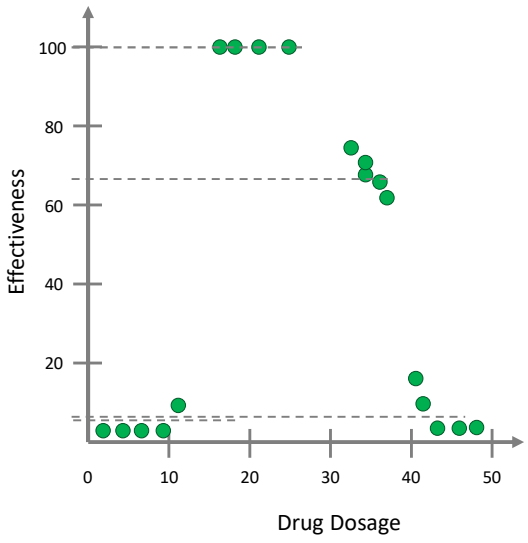
- Entropy = $-\sum p_i \cdot \log_2 p_i$
  - Example: $= -0.5.\log_2 0.5\text{-}0.5.\log_2 0.5 = 0.5 + 0.5' = 1.0$
  - If we split them based on color
    $\Rightarrow$ 6 circles and 3 stars

    pCircles = 6/9 = 0.667
    pStars = 3/9 = 0.333
  - Entropy = $-\frac{6}{9}.\log_2\frac{6}{9} + \frac{3}{9}.\log_2\frac{3}{9} = 0.918$
  - Information Gain = $entropy\ of\ parent - [weighted\ average] entropy\ of\ children$
  - Information Gain = $1 - \frac{9}{12} * 0.918 - \frac{3}{12} * 0.00 = 0.3115$

pra-sâmi

108 Regression Decision Tree



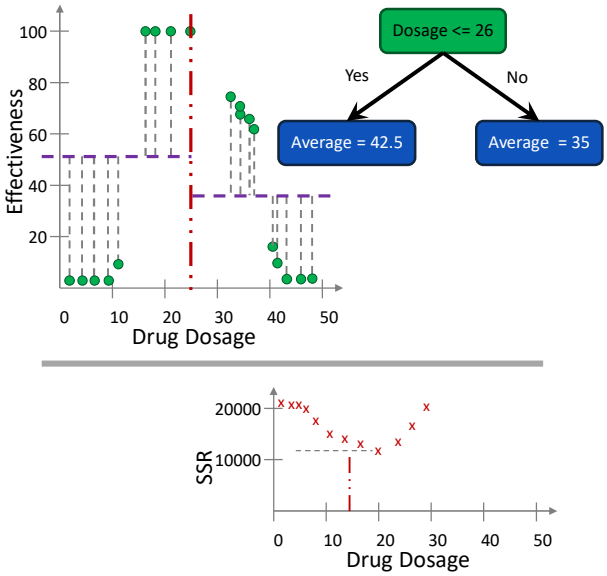


pra-sâmi

109

Identify Root Node

- Iterate from left to right
- Take Dosage <= value at the data point
- Calculate average effectiveness for two partition
- Using average effectiveness calculate 'Sum of Squared Residuals' for all point.
- Plot values for all iterations, pick the lowest value
  - Dosage = 15
  - Hence Dosage 15 becomes root

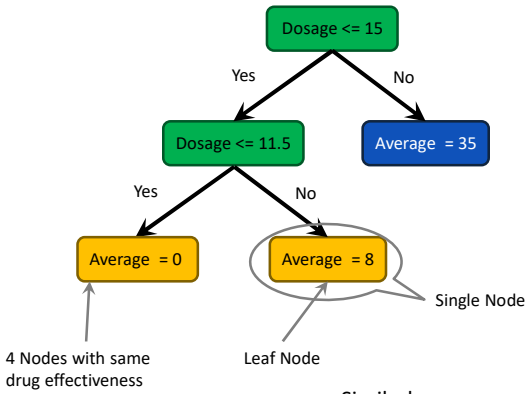


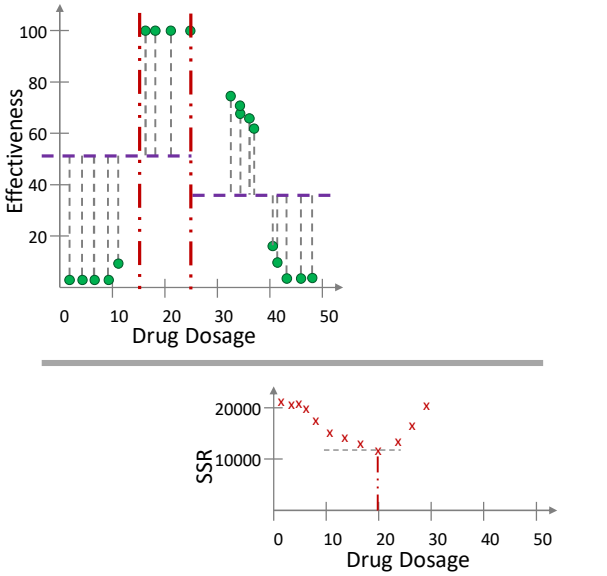
pra-sâmi

110

Expand Left Side of the Tree

- Again Iterate from left to right and calculate Min SSR but only for left side
- This time min SSE is at dosage 11.5





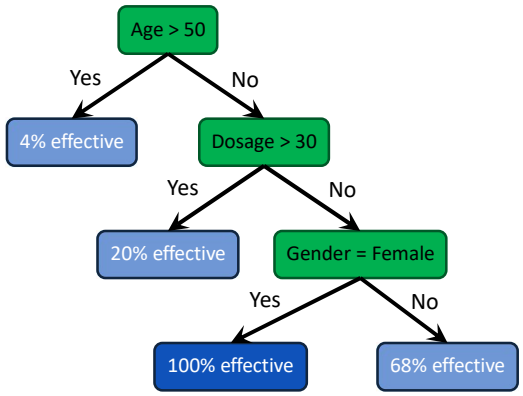
Similarly, we can split the Right side next.

pra-sâmi

111

Regression with Multiple Features

Dosage	Age	Gender	...	Drug Effectiveness (%)
10	25	Female	...	98
20	73	Male	...	4
35	56	Female	...	99
5	13	Male	...	44
...	...	...	...	...

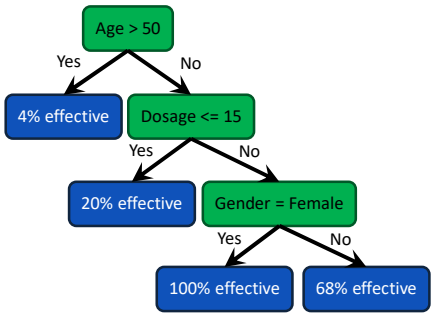


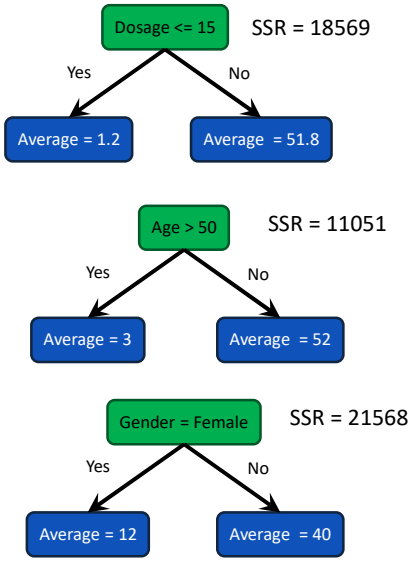
pra-sâmi

112

Regression Steps

- Take each feature and find split giving MinSSR
- Arrange in the order of SSR, Repeat for further splits



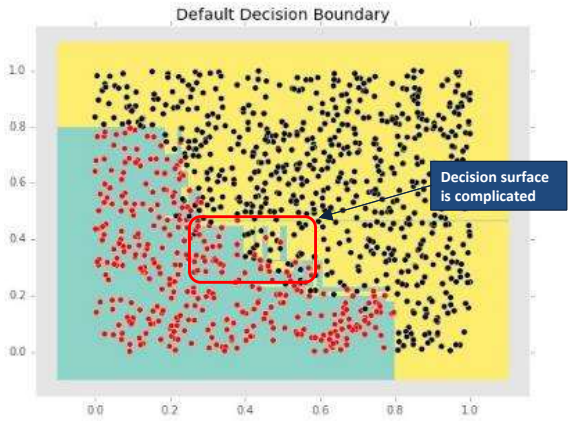


pra-sâmi

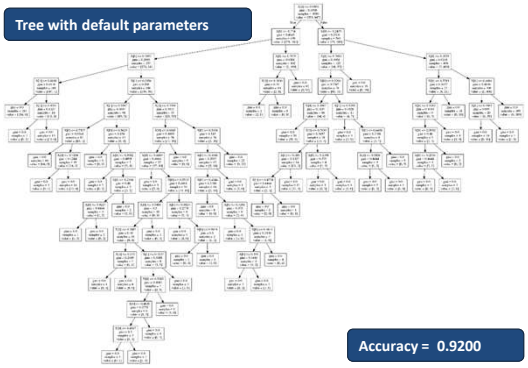
113

Decision Trees

```
#Lets make predictions using Desicion Trees
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier()
```



- Lets analyze iris dataset using decision tree and plot the resultant decision tree.
- Its debatable if area marked belongs to blue datapoints.

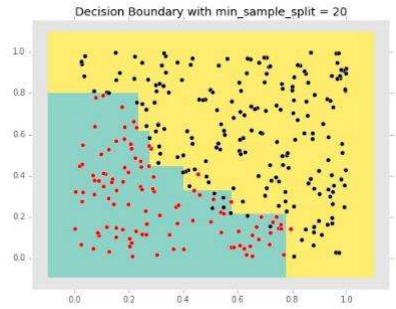


pra-sâmi

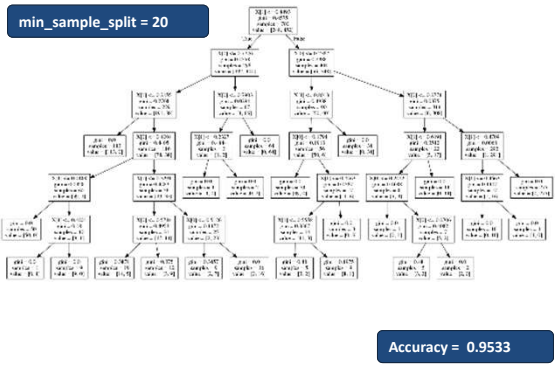
114

Decision Trees

```
#Lets tune the parameters.
#Test what if we restrict the min_sample_split to 20!
clf = DecisionTreeClassifier(min_samples_split = 20)
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
```



- By restricting min_sample_split, we could simplify or decision boundary and achieve higher accuracy.



pra-sâmi

115

Decision Trees

- Are simple to understand and interpret. People are able to understand decision tree models after a brief explanation.
- Have value even with little hard data. Important insights can be generated based on experts describing a situation (its alternatives, probabilities, and costs) and their preferences for outcomes.
- Allow the addition of new possible scenarios
- Help determine worst, best and expected values for different scenarios
- Use a white box model. If a given result is provided by a model.
- Can be combined with other decision techniques

pra-sâmi

116

Advantages and Disadvantages

Advantages	Disadvantages
❑ Easy to understand. ❖ Human inspection (physical) is possible.	❑ Extremely sensitive to change in data : can result in a drastically different tree.
❑ Can give valuable results even with little data	❑ Prone to overfitting : Can be negated by validation methods and pruning.
❑ Scalable : addition of new possible scenarios	❑ Difficulty in dealing with multicollinearity: when two variables both explain the same thing, a decision tree will greedily choose the best one, other is ignored.
❑ Help determine worst, best and expected values for different scenarios	❑ Lack of a principled probabilistic framework.
❑ Can be combined with other decision techniques	❑ Poor resolution if data has complex relationships among the features
	❑ Practically Limited to Classification ❖ Poor Resolution With Continuous Expectation Variables

pra-sâmi

117

Wisdom of Crowds

- Ensemble methods
  - Combine multiple classifiers to make “better ” classifier
  - Predictions are averaged out – reducing error
  - Can use weighted combinations
  - Can use even different classifiers

- Two types of ensemble:
  - Bagging
    - Weak learners are trained in parallel
    - Used on weak learners that exhibit high variance and low bias
    - Avoids overfitting
    - Being leveraged for loan approval processes and statistical genomics
  - Boosting
    - They learn sequentially➔ a series of models are constructed and with each new model iteration, the weights of the misclassified data in the previous model are increased.
    - Redistribution of weights helps to focus on the parameters which will lead to performance improvement
    - Leveraged when low variance and high bias is observed
    - Prone to overfitting
    - while boosting has been used more within image recognition apps and search engines

118

Random Forest

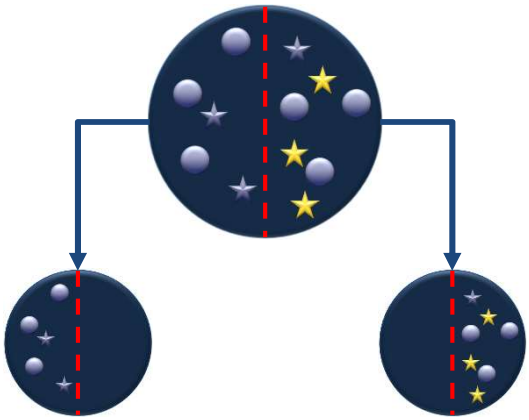
- Random Forests are ensembles of decision trees

- Random Record Selection
  - Each tree is built from a separate random sample (bootstrap sample)
  - Create hundreds of trees each - gain diversity from examining different examples
  - Over fit the data extensively.

- Random Feature Selection
  - A random subset of variables
  - Number of variables to consider is $\sqrt{}$(features) - its configurable though
  - Forces the trees to find alternative ways to predict the target

119

Split of sample



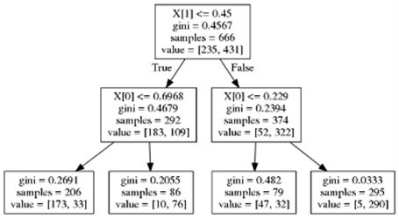
Decision trees involve greedy, recursive partitioning.

pra-sâmi

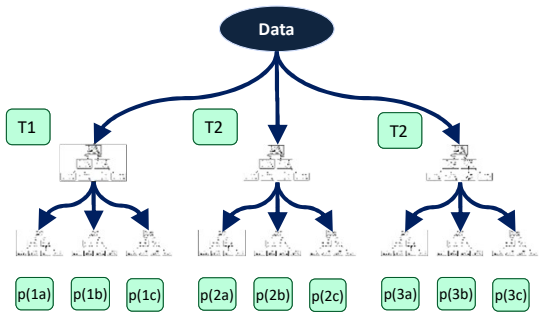
120

Random Forest

Decision Tree



Random Forest



- Split at point of maximum info gain from random sample
- Prediction = average probabilities

pra-sâmi

121 Advantages and Disadvantages

Advantages

- Most accurate learning algorithms available.
  - For many data sets, it produces a highly accurate classifier.
- Efficient even on large datasets.
- Can handle thousands of features.
- Estimates feature importance as well.
- Unbiased estimate of the generalization error.
- Effective method for estimating missing data
  - Maintains accuracy even if significant chunk of data is missing.
- Balances out the error in class population unbalanced data sets.

Disadvantages

- Tend to over-fit for some datasets
  - especially noisy classification/regression tasks.
- If dataset has significantly higher instances of missing some feature in categorical targets, random forests are biased against these features.
  - Feature importance may not be reliable

122 Summary

- Fast to train, even faster in prediction.
- Not special need for cross validation, it is part of the algorithm
- Support parallelism.
- Resistance to over tuning
- Cluster identification can be used to generate tree-based clusters through sample proximity
- Can be easily be distributed to different hardware clusters

123

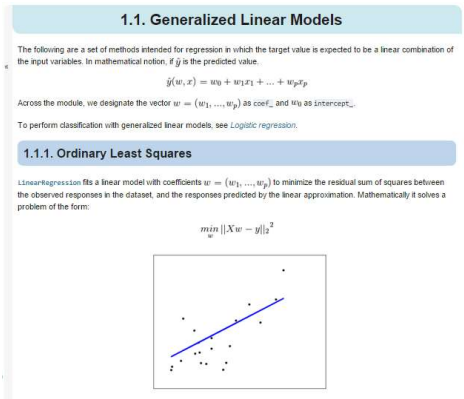
Gradient Boosting

- Gradient Descent
  - In theory : it an algorithm to minimize functions
  - It is achieved by moving in negative direction of the slope
  - Take linear equation for example:
    - y = m * x +c or y = f(x)
    - For any value y_i, error = y_i-m * x_i –c
    - Standard approach is to build an error function: for each line calculated sum of squared error
    - Best fitted line will have least error
  - In general, a function may have more than one minima
    - There are approaches available to mitigate this risk
    - Its particularly useful when number of features are in couple of thousands
- Gradient Boosting
  - Gradient boosting is Gradient Descent + Boosting

pra-sâmi

125

Linear Regression



1.1. Generalized Linear Models

The following are a set of methods intended for regression in which the target value is expected to be a linear combination of the input variables. In mathematical notion, if $\hat{y}$ is the predicted value.

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p$$

Across the module, we designate the vector $w = (w_1, \ldots, w_p)$ as `coef_` and w_0 as `intercept_`.

To perform classification with generalized linear models, see *Logistic regression*.

1.1.1. Ordinary Least Squares

`LinearRegression` fits a linear model with coefficients $w = (w_1, \ldots, w_p)$ to minimize the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation. Mathematically it solves a problem of the form:

$$\min_w ||Xw - y||_2^2$$

- Regressions predict expected value of the new record.
- Results are continuous
- Hypothesis: can be represented as
  - $h_\theta = \theta_1 * X + \theta_0$
- Cost Function:
  - $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta^i(x) - \hat{y})^2$
  - Optimize the cost function

pra-sâmi

126

Reflect…

- What does the "Bayes" in Naive Bayes refer to in terms of probability theory?
  - A. Bayesian networks
  - B. Bayes' theorem
  - C. Bayesian regression
  - D. Bayesian optimization
- Which type of Naive Bayes classifier is suitable for text classification tasks?
  - A. Gaussian Naive Bayes
  - B. Multinomial Naive Bayes
  - C. Bernoulli Naive Bayes
  - D. Complement Naive Bayes
- When should Laplace smoothing (additive smoothing) be used in Naive Bayes?
  - A. When the feature space is too large
  - B. When the training data is small and sparse
  - C. When the feature distribution is Gaussian
  - D. Laplace smoothing is not used in Naive Bayes
- Which step is NOT a part of the Naive Bayes training process?
  - A. Estimating class priors
  - B. Calculating the likelihood of features
  - C. Applying Laplace smoothing
  - D. Making predictions on new data

127

Reflect…

- In Naive Bayes, what is the purpose of the prior probabilities?
  - A. To estimate feature independence
  - B. To compute the posterior probabilities
  - C. To reduce the dimensionality of the feature space
  - D. To calculate the likelihood of features
- Which performance metric is commonly used to evaluate the performance of Naive Bayes classifiers?
  - A. R-squared
  - B. Mean squared error
  - C. Accuracy
  - D. F1-score

128 Reflect…

- What is Naive Bayes primarily used for?
  - A. Image classification
  - B. Natural language processing
  - C. Regression analysis
  - D. Time series forecasting
- In Naive Bayes, what assumption is made about the independence of features?
  - A. Features are always dependent on each other
  - B. Features are conditionally independent given the class
  - C. Features are independent of the class
  - D. Features are irrelevant in classification
- Which probability distribution is commonly used for modeling the likelihood in Gaussian Naive Bayes?
  - A. Normal distribution
  - B. Poisson distribution
  - C. Exponential distribution
  - D. Binomial distribution
- In a binary classification problem, how many class labels does Naive Bayes predict?
  - A. 1
  - B. 2
  - C. 3
  - D. It depends on the number of features.

pra-sâmi

129 Reflect…

- What does the "Bayes" in Naive Bayes refer to in terms of probability theory?
  - A. Bayesian networks
  - B. Bayes' theorem
  - C. Bayesian regression
  - D. Bayesian optimization
- Which type of Naive Bayes classifier is suitable for text classification tasks?
  - A. Gaussian Naive Bayes
  - B. Multinomial Naive Bayes
  - C. Bernoulli Naive Bayes
  - D. Complement Naive Bayes
- When should Laplace smoothing (additive smoothing) be used in Naive Bayes?
  - A. When the feature space is too large
  - B. When the training data is small and sparse
  - C. When the feature distribution is Gaussian
  - D. Laplace smoothing is not used in Naive Bayes
- Which step is NOT a part of the Naive Bayes training process?
  - A. Estimating class priors
  - B. Calculating the likelihood of features
  - C. Applying Laplace smoothing
  - D. Making predictions on new data

pra-sâmi

130

Reflect…

- In Naive Bayes, what is the purpose of the prior probabilities?
  - A. To estimate feature independence
  - B. To compute the posterior probabilities
  - C. To reduce the dimensionality of the feature space
  - D. To calculate the likelihood of features
- Which performance metric is commonly used to evaluate the performance of Naive Bayes classifiers?
  - A. R-squared
  - B. Mean squared error
  - C. Accuracy
  - D. F1-score

131

Reflect…

- What is the primary objective of linear regression analysis?
  - A. To classify data into different categories
  - B. To predict a continuous outcome variable based on one or more predictor variables
  - C. To identify outliers in the dataset
  - D. To calculate the mean of the data
- In simple linear regression, how many predictor variables are used to predict the outcome variable?
  - A. None
  - B. One
  - C. Two or more
  - D. It depends on the dataset
- What is the main goal of linear regression when fitting a line to data points?
  - A. Minimize the sum of squared residuals
  - B. Maximize the R-squared value
  - C. Minimize the number of predictor variables
  - D. Maximize the p-value
- What does the coefficient of determination (R-squared) measure in linear regression?
  - A. The strength of the relationship between predictor variables
  - B. The significance of the predictor variables
  - C. The proportion of the variation in the dependent variable explained by the independent variables
  - D. The p-value of the regression model

132 Reflect…

- Which of the following assumptions is not typically required for linear regression to be valid?
  A. Linearity
  B. Homoscedasticity
  C. Independence of errors
  D. Multicollinearity

- When performing multiple linear regression with several predictor variables, how is the relationship between the predictor variables and the outcome variable typically represented?
  A. In a scatterplot
  B. In a correlation matrix
  C. In a regression equation
  D. In a probability distribution

- What is the primary purpose of logistic regression?
  A. To predict a continuous outcome variable
  B. To predict a categorical outcome variable
  C. To identify outliers in the dataset
  D. To perform dimensionality reduction

- In logistic regression, what type of outcome variable is typically predicted?
  A. Continuous
  B. Binary (two categories)
  C. Multinomial (more than two categories)
  D. Ordinal (ordered categories)

pra-sâmi

133 Reflect…

- Which function is used to model the relationship between predictor variables and the log-odds of the outcome variable in logistic regression?
  A. Linear function
  B. Sigmoid function
  C. Exponential function
  D. Polynomial function

- What is the output of logistic regression called, representing the probability of an observation belonging to a particular category?
  A. Odds ratio
  B. Log-odds
  C. Risk ratio
  D. Logistic probability

- In logistic regression, what is the likelihood function used to estimate?
  A. Mean squared error
  B. Maximum likelihood of the observed data
  C. Residuals
  D. R-squared value

- Which of the following evaluation metrics is commonly used to assess the performance of a logistic regression model?
  A. Mean Absolute Error (MAE)
  B. Root Mean Squared Error (RMSE)
  C. R-squared (R^2)
  D. Area Under the Receiver Operating Characteristic (ROC-AUC)

pra-sâmi

134 Strengths and Weaknesses of Neighborhood-Based Methods

Advantages	Disadvantages
❑ Easy to implement and debug and justify	❑ Offline phase can sometimes be impractical in large-scale settings ❖ User-based method requires at least O(m^2) time and space
❑ The recommendations are relatively stable with the addition of new items and users	❑ Limited coverage because of sparsity
❑ It is also possible to create incremental approximations of these methods	

pra-sâmi

135 Reflect…

- ❑ What is the primary objective of a Support Vector Machine (SVM)?
  - A. Classification
  - B. Regression
  - C. Clustering
  - D. Dimensionality reduction
- ❑ In an SVM, what are support vectors?
  - A. Data points closest to the decision boundary
  - B. Data points farthest from the decision boundary
  - C. Data points with the highest feature values
  - D. Data points with missing values
- ❑ Which kernel function is commonly used in SVM for non-linear classification?
  - A. Linear kernel
  - B. Polynomial kernel
  - C. Radial basis function (RBF) kernel
  - D. Sigmoid kernel
- ❑ The margin in an SVM represents:
  - A. The distance between support vectors
  - B. The distance between data points in different classes
  - C. The width of the decision boundary
  - D. The number of support vectors

pra-sâmi

136

Reflect…

- What is the primary purpose of the regularization parameter (C) in SVM?
  A. To control the trade-off between maximizing the margin and minimizing classification errors
  B. To determine the kernel function to be used
  C. To set the number of support vectors
  D. To adjust the learning rate in training

- In the context of SVM, what does "hard margin" refer to?
  A. When the SVM allows some misclassification errors
  B. When the SVM enforces strict separation between classes
  C. When the SVM uses a non-linear kernel function
  D. When the SVM has a small margin

- What is the key difference between a linear SVM and a logistic regression classifier?
  A. SVM uses a different optimization technique
  B. SVM can only perform binary classification
  C. SVM finds the maximum-margin decision boundary
  D. Logistic regression is a supervised learning algorithm

- In a multi-class classification problem, how can SVM be used?
  A. Train a separate binary SVM for each class
  B. SVM cannot be used for multi-class problems
  C. Use a softmax function with a single SVM
  D. Use clustering algorithms instead

pra-sâmi

137

Reflect…

- Which of the following is a disadvantage of Support Vector Machines?
  A. They work well with small datasets only
  B. They are highly sensitive to outliers
  C. They always find the global minimum during training
  D. They are not suitable for non-linear problems

- Which of the following is not a common application of Support Vector Machines?
  A. Image classification
  B. Text classification
  C. Anomaly detection
  D. Principal Component Analysis (PCA)

pra-sâmi

138

pra-sâmi

ADDITIONAL MATERIAL

pra-sâmi

Use of Machine Learning
Probability
Agent in uncertain environment

140

Use of Machine Learning

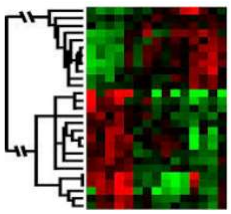
pra-sâmi

141

When Do We Use Machine Learning?

- Human expertise does not exist (navigating on Mars)
- Humans can't explain their expertise (speech recognition)
- Models must be customized (personalized medicine)
- Models are based on huge amounts of data (genomics)



- Learning isn't always useful:
  - There is no need to "learn" to calculate payroll

pra-sâmi

142

Classic Example

- It is very hard to say what makes a '2'

Slide credit: Geoffrey Hinton

143

More Examples

- Recognizing patterns:
  - Facial identities or facial expressions
  - Handwritten or spoken words
  - Medical images
- Generating patterns:
  - Generating images or motion sequences
- Recognizing anomalies:
  - Unusual credit card transactions
  - Unusual patterns of sensor readings in a nuclear power plant
- Prediction:
  - Future stock prices or currency exchange rates

144

Sample Applications

- Web search
- Computational biology
- Finance
- E-commerce
- Space exploration
- Robotics
- Information extraction
- Social networks
- Debugging software

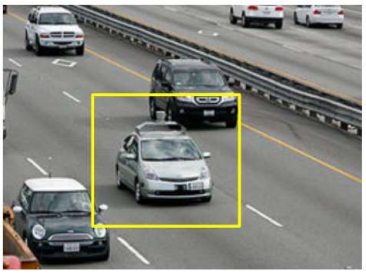
Your Favorite Area

pra-sâmi

145

Autonomous vehicles

- Nevada made it legal for autonomous cars to drive on roads in June 2011
- • As of 2013, four states (Nevada, Florida, California, and Michigan) have legalized autonomous cars

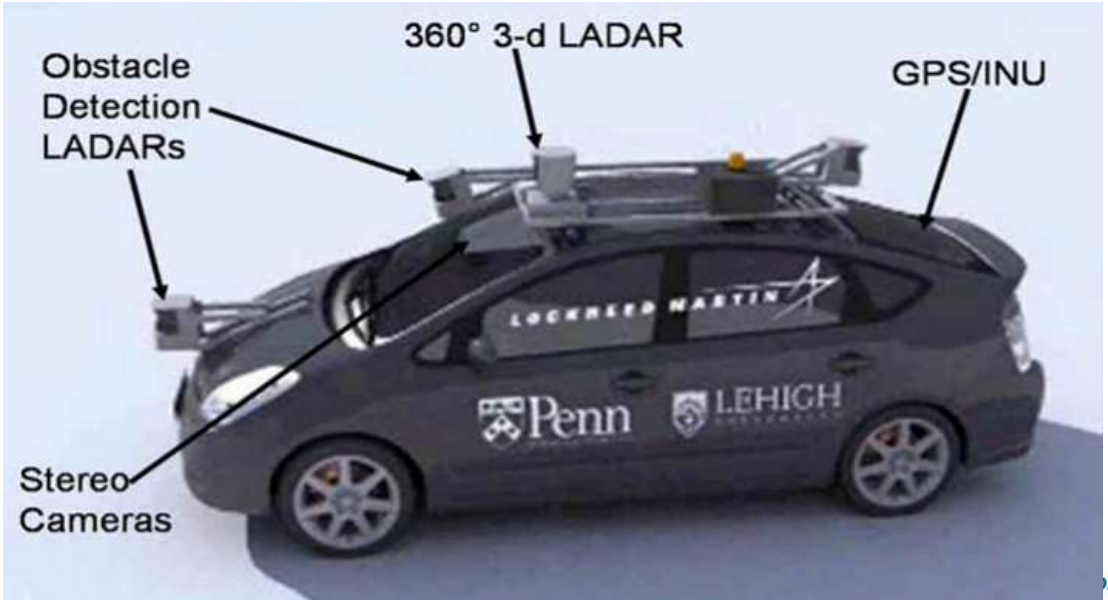


Penn's Autonomous Car (Ben Franklin Racing Team)

pra-sâmi

146

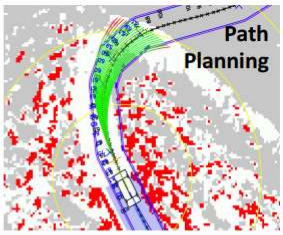
Autonomous Car Sensors

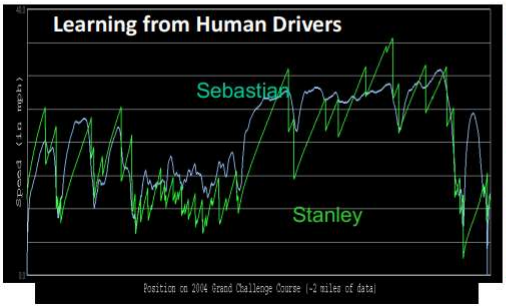


pra-sâmi

147

Autonomous Car Technology

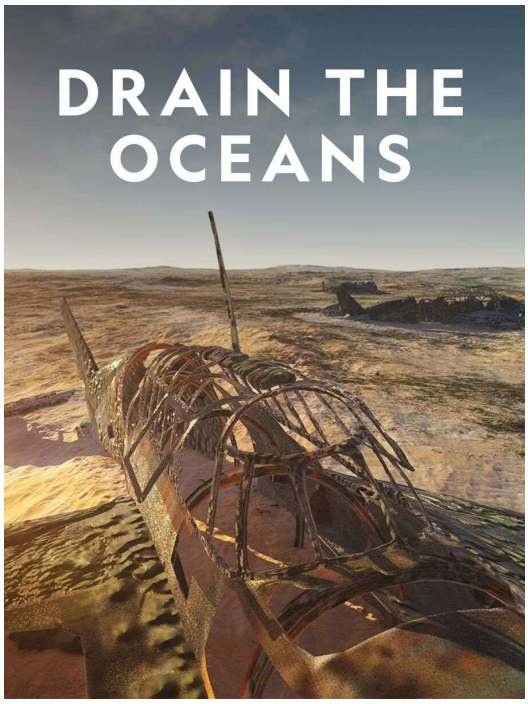




Images taken from Sebastian Thrun's web site

pra-sâmi

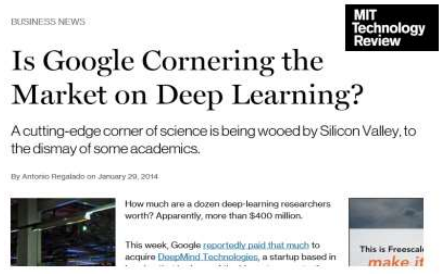
148



pra-sâmi

149

Deep Learning in the Headlines


BUSINESS NEWS

MIT Technology Review

Is Google Cornering the Market on Deep Learning?

A cutting-edge corner of science is being wooed by Silicon Valley, to the dismay of some academics.

By Antonio Regalado on January 29, 2014

How much are a dozen deep-learning researchers worth? Apparently, more than $400 million.

This week, Google reportedly paid that much to acquire DeepMind Technologies, a startup based in

BloombergBusinessweek
Technology

Acquisitions

The Race to Buy the Human Brains Behind Deep Learning Machines

By Ashlee Vance | January 27, 2014

intelligence projects. "DeepMind is bona fide in terms of its research capabilities and depth," says Peter Lee, who heads Microsoft Research.

According to Lee, Microsoft, Facebook (FB), and Google find themselves in a battle for deep learning talent. Microsoft has gone from four full-time deep learning experts to 70 in the past three years. "We would have more if the talent was there to


WIRED GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIG

INNOVATION INSIGHTS | community content | featured

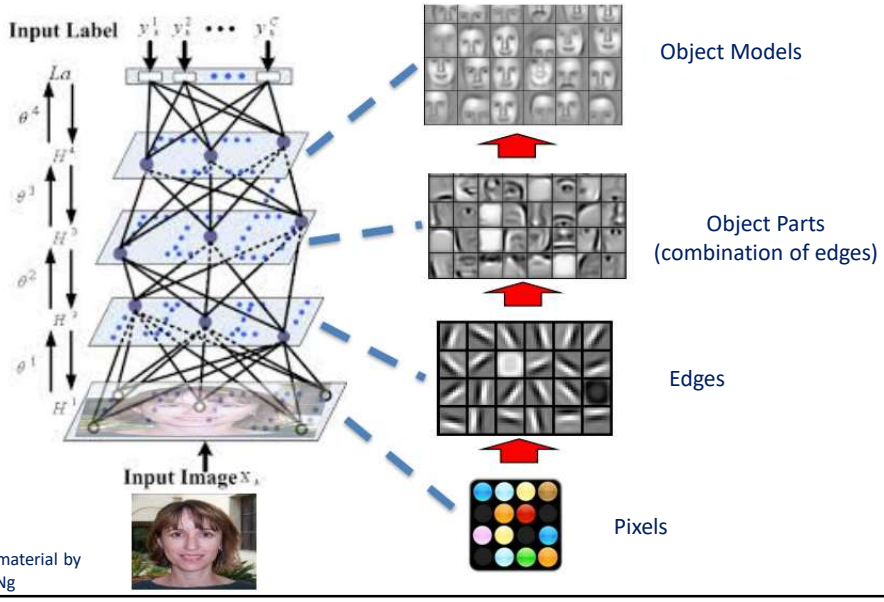
Deep Learning's Role in the Age of Robots

BY JULIAN GREEN, JETPAC 05.02.14 2:56 PM

pra-sâmi

150

Deep Belief Net on Face Images

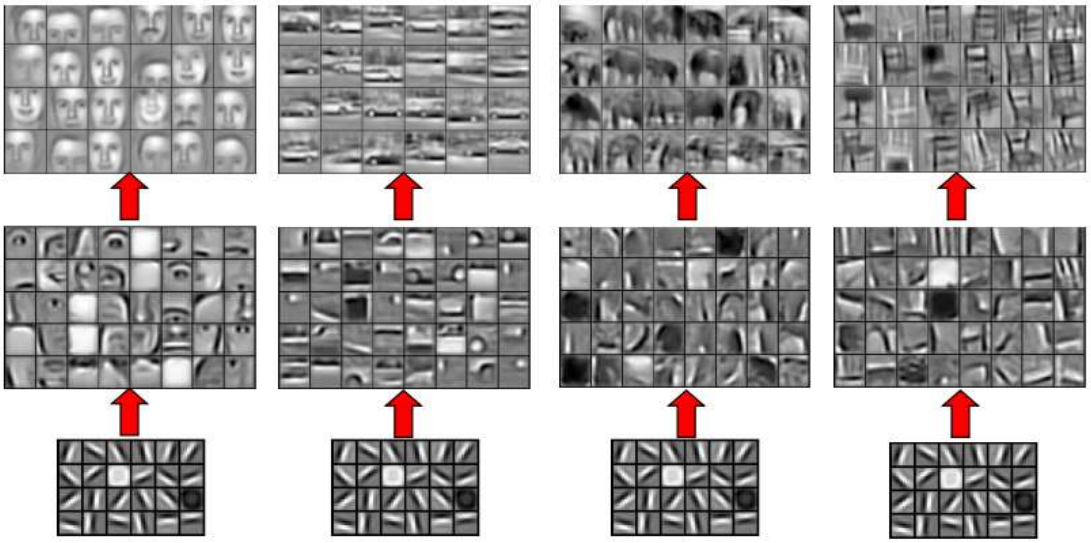


Based on Course material by Andrew Ng

pra-sâmi

151

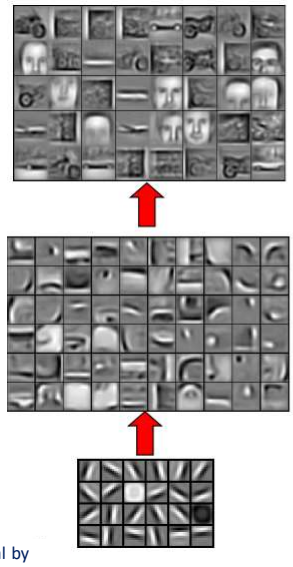
Learning of Object Parts



pra-sâmi

152

Training on Multiple Objects



Trained on 4 classes (cars, faces, motorbikes, airplanes).

Second layer: Shared-features and object-specific features.

Third layer: More specific features.

Based on Course material by Andrew Ng

18

pra-sâmi

153

Scene Labeling via Deep Learning

[Farabet et al. ICML 2012, PAMI 2013]

pra-sâmi

154

Impact of Deep Learning in Speech Technology

Slide credit: Li Deng, MS Research

pra-sâmi

155

Agent in Uncertain Environment

pra-sâmi

156

Uncertainty Everywhere

- The agent forms a plan, A90,
  - Involves leaving home 90 minutes before the flight departs
  - Driving at a reasonable speed
- Are you certain "*Plan A90 will get us to the airport in time*."?
  - Not in absolute sense but with some riders

pra-sâmi

157

Uncertainty Everywhere

- How about other plans, such as A180,
  - Might increase the agent's belief that it will get to the airport on time,
  - But also increase the likelihood of a long wait
- Probability is an agent's measure of belief in some proposition — subjective probability.
- An agent's belief depends on its prior belief and what it observes.

pra-sâmi

158

Uncertain Reasoning

- Simple rule: Toothache ⇒ Cavity 
- Not all patients with toothaches have cavities
  - Some of them have gum disease, an abscess, or one of several other problems
  - Toothache ⇒ Cavity ∨ Gum Problem ∨ Abscess ... 
- Let's try other way round Cavity ⇒ Toothache
  - But this rule is not right either; not all cavities cause pain
- Trying to use logic to cope with a domain like medical diagnosis thus fails
  - Laziness: It is too much work to list the complete set of antecedents or consequents needed to ensure an exception-free rule and too hard to use such rules
  - Theoretical ignorance: Medical science has no complete theory for the domain.
  - Practical ignorance: Even if we know all the rules, we might be uncertain about a particular patient because not all the necessary tests have been or can be run.

159

Agent In Uncertain Environment

- Agents don't have complete knowledge about the world.
- Agents need to make (informed) decisions given their uncertainty.
- It isn't enough to assume what the world is like.
  - Example: wearing a seat belt.
- An agent needs to reason about its uncertainty.
- When an agent makes an action under uncertainty, it is gambling ⇒ probability

160

Probability

pra-sâmi

161

Probability Theory: Variables and Events

- A random variable can be an observation, outcome or event
  - Value of which is uncertain
  - e.g. a coin.
  - Throw as the random variable denoting the outcome when we toss the coin
- A Boolean random variable has two outcomes.
- The set of possible outcomes for a random variable is called its domain.
  - The domain of Throw is {head, tail}
  - Cavity has the domain {true, false}
  - Toothache has the domain {true, false}

pra-sâmi

162

Simple Probability

If there are *n* elementary events associated with a random experiment and *m* of *n* of them are favorable to an event *A*, then the probability of happening or occurrence of *A* is

$$P(A) = \frac{m}{n}$$

pra-sâmi

163

Simple Probability

- Suppose, A and B are any two events and P(A), P(B) denote the probabilities that the events A and B will occur, respectively.
- Mutually Exclusive Events:
  - Two events are mutually exclusive, if the occurrence of one precludes the occurrence of the other.
  - Example: Tossing a coin (two events) or a dice (six events)
- Can you give an example, so that two events are not mutually exclusive?
  - Hint: Tossing two identical coins, Weather (sunny, foggy, warm)

pra-sâmi

164

Simple Probability

- Independent events: Two events are independent if occurrences of one does not alter the occurrence of other.
  - Example: Tossing both coin and ludo cube together (How many events are here?)
- Can you give an example, where an event is dependent on one or more other events(s)?
  - Hint: Receiving a message (A) through a communication channel (B) over a computer (C), rain and dating.

165

Joint Probability

If P(A) and P(B) are the probability of two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A and B are mutually exclusive, then $P(A \cap B) = 0$

If A and B are independent events, then $P(A \cap B) = P(A).P(B)$

Thus, for mutually exclusive events

$$P(A \cup B) = P(A) + P(B)$$

166

Conditional Probability

If events are dependent, then their probability is expressed by conditional probability. The probability that A occurs given that B is denoted by $P(A|B)$.

Suppose, A and B are two events associated with a random experiment. The probability of A under the condition that B has already occurred and $P(B) \neq 0$ is given by

$$P(A|B) = \frac{\text{Number of events in } B \text{ which are favourable to } A}{\text{Number of events in } B}$$

$$= \frac{\text{Number of events favourable to } A \cap B}{\text{Number of events favourable to } B}$$

$$= \frac{P(A \cap B)}{P(B)}$$

pra-sâmi

167

Conditional Probability

$P(A \cap B) = P(A).P(B|A), \quad if\ P(A) \neq 0$

or $P(A \cap B) = P(B).P(A|B), \quad if\ P(B) \neq 0$

For three events A, B and C

$$P(A \cap B \cap C) = P(A).P(B).P(C|A \cap B)$$

For n events A1, A2, …, An and if all events are mutually independent to each other

$$P(A_1 \cap A_2 \cap \ \ldots\ldots\ldots\ldots \cap A_n) = P(A_1).P(A_2)\ldots\ldots\ldots\ldots P(A_n)$$

Note:

$P(A|B) = 0$ if events are mutually exclusive

$P(A|B) = P(A)$ if A and B are independent

$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$ otherwise,

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

pra-sâmi

168

Total Probability

Let $E_1, E_2, \ldots \ldots E_n$ be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with E_1 *or* E_2 *or* $\ldots \ldots E_n$, then

$$P(A) = P(E_1).P(A|E_1) + P(E_2).P(A|E_2) + \cdots \ldots \ldots \ldots + P(E_n).P(A|E_n)$$

pra-sâmi

169

Probabilities

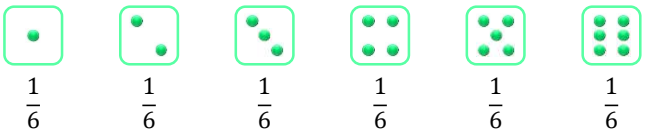
- Let's say we have a set S
  - Heads – Tails of a coin; 6 possible location of a dice
- X ∈ S is a sample from the sample S
- Probability distribution P(X|S)
  - X is a random variable
    - Discrete
    - Continuous
  - Such that $0 \leq P(X) \leq 1$ and $\sum P(X) = 1$
- An event A of any subset of S
  - A = 'flipped a Heads'

pra-sâmi

170

Dice World

- Single dice
  - Six states
  - P(1) = P(2) = P(3) … = P(6) = $\frac{1}{6}$
- Therefore $\sum_{\omega \in W} P(\omega) = 1$

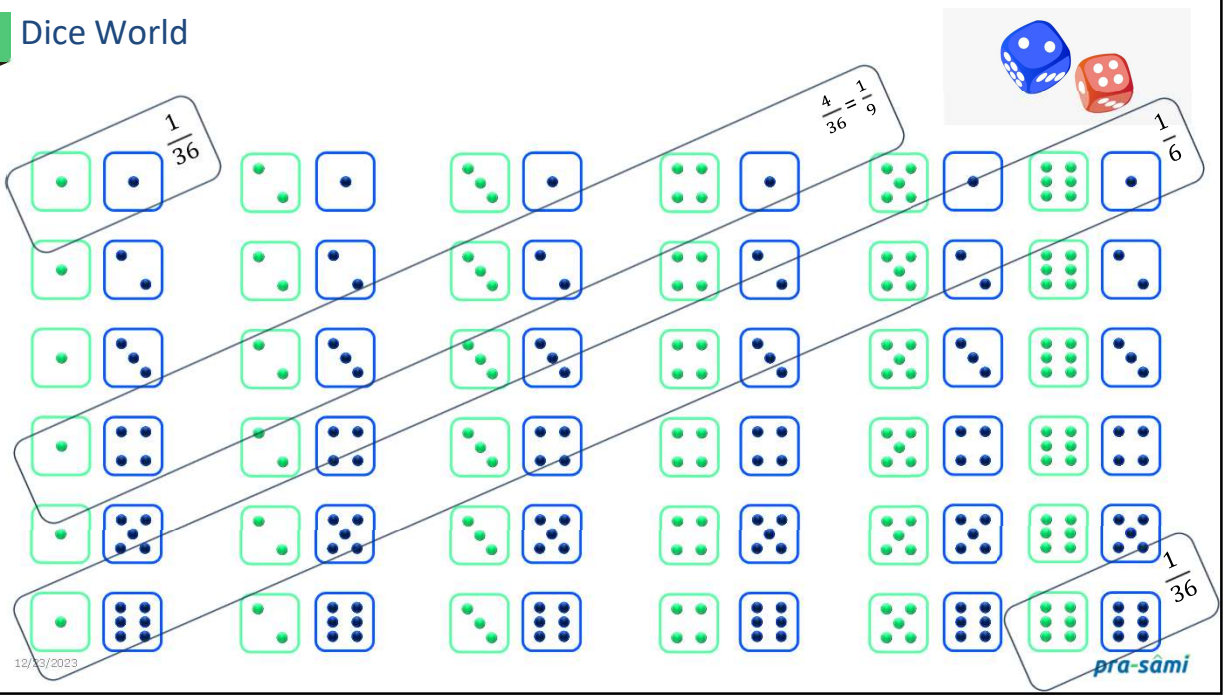


- Fairly straight forward when we have single dice… What if we have two dice…

pra-sâmi

171

Dice World

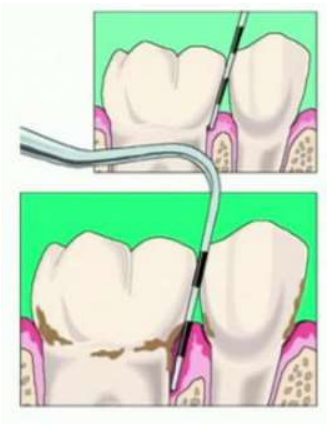


pra-sâmi

172

Dentist's Dilemma

- Mr. Mario is at the dentist office
- He feels that he has tooth ache and he might have cavity.
- The dentist may probe his teeth and then put it to test.
- And we say the probe catches if the probe has found some microorganisms.



pra-sâmi

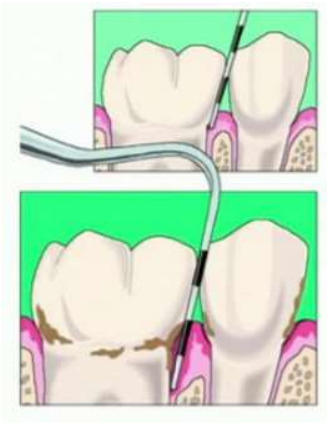
173

Inference by Enumeration

- Start with inference by enumeration

	Toothache		¬ Toothache	
	Catch	¬ Catch	Catch	¬ Catch
Cavity	0.108	0.012	0.072	0.008
¬ Cavity	0.016	0.064	0.144	0.576

- For any proposition Ø sum the atomic events where it is true:
  - $P(\emptyset) = \sum_{\omega:\omega \models \emptyset} p(\omega)$
- P (toothache) = 0.108 + 0.012 + 0.016 + 0.064
  = 0.20



pra-sâmi

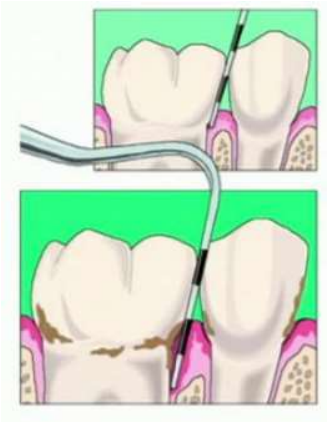
174

Inference by Enumeration

- What is probability of toothache or cavity

	Toothache		¬ Toothache	
	Catch	¬ Catch	Catch	¬ Catch
Cavity	0.108	0.012	0.072	0.008
¬ Cavity	0.016	0.064	0.144	0.576

- P (toothache ∨ cavity) = 0.02 + 0.072 + 0.008
  = 0.28



pra-sâmi

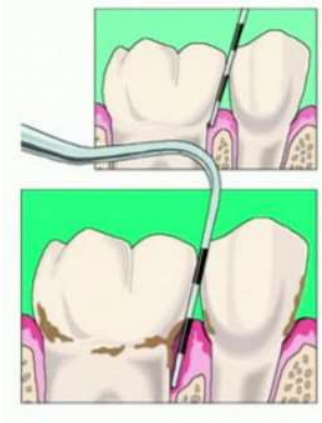
175

Inference by Enumeration

- What is probability of toothache and no cavity

	Toothache		¬ Toothache	
	Catch	¬ Catch	Catch	¬ Catch
Cavity	0.108	0.012	0.072	0.008
¬ Cavity	0.016	0.064	0.144	0.576

- P (¬ Cavity | toothache) = $\frac{P(\neg \text{ Cavity} \wedge Tootheache)}{P(Toothache)}$
- P (¬ Cavity | toothache) = $\frac{0.016+ .064}{0.018+0.012+0.016+ .064}$ = 0.4
- Time complexity = O(d^n)
- Space Complexity = O(d^n)



pra-sâmi

176

Probability Theory: Atomic events

- We can create new events out of combinations of the outcomes of random variables
- An atomic event is a complete specification of the values of the random variables of interest
  - e.g. if dentist world consists of only two Boolean random variables {Toothache, Cavity}, then the world has a four possible atomic events

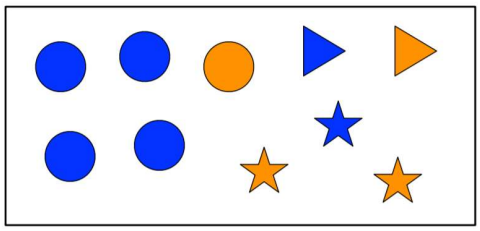
```
Toothache = true ^ Cavity = true
Toothache = true ^ Cavity = false
Toothache = false ^ Cavity = true
Toothache = false ^ Cavity = false
```

- The set of all possible atomic events has two properties:
  - It is mutually exhaustive (nothing else can happen)
  - It is mutually exclusive (only one of the four can happen at one time)
- We can assign probabilities to the outcomes of a random variable.
  - P(Throw = heads) = 0.5
  - P(Mary_Calls = true) = 0.1
  - P(a) = 0.3

177

Semantics

- Possible world

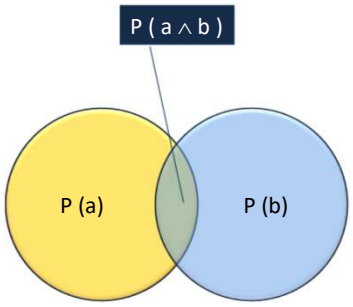


- Suppose the measure of each singleton world is 0.1.
  - What is the probability of circle? 5/10 = 0.5
  - What us the probability of star? 3/10 = 0.3
  - What is the probability of orange? 4/10 = 0.4
  - What is the probability of orange and star? 2/10 = 0.2
  - What is the probability of orange and circle? 1/10 = 0.1

178

Probability Theory: Probabilities

- Some simple rules governing probabilities
- All probabilities are between 0 and 1 inclusive
  - $0 \geq P(a) \geq 1$
- If something is necessarily true it has probability 1
  - P(true) = 1
- The probability of a disjunction being true is
  - $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$
- From these laws all of probability theory can be derived.
- These axioms are sound and complete with respect to the semantics.



pra-sâmi

179

Probability Theory: Conditional Probability

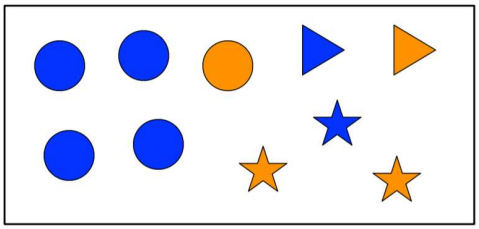
- Probabilistic conditioning specifies how to revise beliefs based on new information.
- An agent builds a probabilistic model taking all background information into account.
  - This gives the prior probability.
- All other information must be conditioned on.
- If evidence e is the all of the information obtained subsequently, the conditional probability P(h | e) of h given e is the posterior probability of h.
  - Evidence e rules out possible worlds incompatible with e.
  - $P(h \mid e) = \frac{P(h \wedge e)}{P(e)}$

pra-sâmi

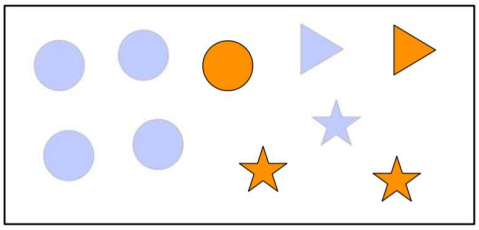
180

Probability Theory: Conditional Probability

- Possible world
- Observed color = orange



- P(Shape=circle | Color=orange) = 0.25
- P(Shape=star | Color=orange) = 0.5



pra-sâmi

181

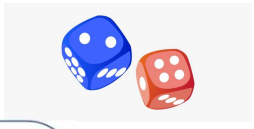
Independence

- A and B are independent iff
- P (A | B) = P (A)
- Or P (B | A) = P (B)
- Or P (A, B) = P (A) . P (B)
- As we can see in case of Two dices

pra-sâmi

182

Conditional Probability



$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)}$$

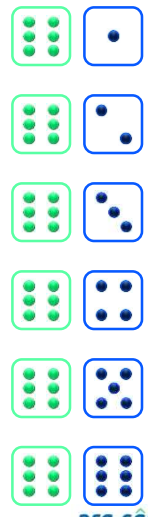
$P(\text{sum} = 12 \wedge$  $) = \frac{1}{36}$

$P($  $) = \frac{1}{6}$

$P(\text{sum} = 12 \mid$  $) = \frac{1}{6}$

Rolling dice is also independent of each other.

Knowing value of one dice does not change the behavior of second dice

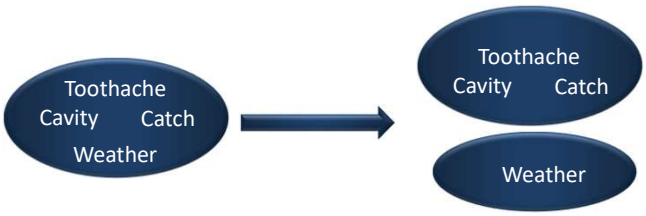


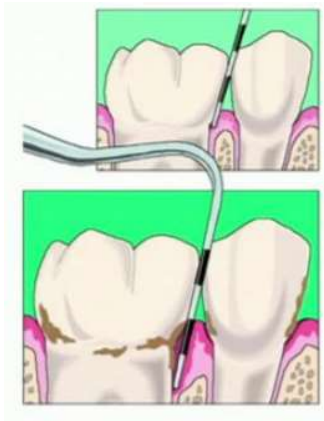
pra-sâmi

183

Independence

- What about toothache





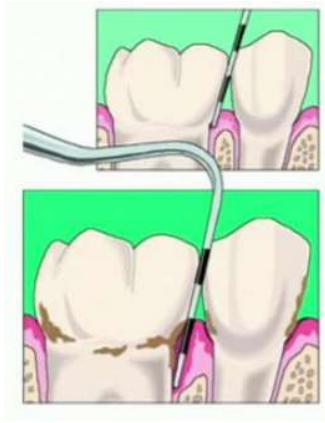
- P(Toothache, Cavity, Catch, Weather) = P(Toothache, Cavity, Catch) * P(Weather)
- With four values of Weather and two values each for Toothache, Cavity, and Catch, there were total 32 states ➔ we can live with 31 as sum is always 1.
- Considering weather to be independent, how many variable : $(2^3 - 1) + (4 - 1) = 10$

pra-sâmi

184

Conditional Independence

- P(Toothache, Cavity, Catch) have 2^3 - 1 = 7 independent states
- If Cavity is true, the probability that probe will catch in it does not depend on Toothache
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)
  - Or
  - P(Catch | Toothache, ¬ Cavity) = P(Catch | ¬ Cavity)
- Catch is conditionally independent of Toothache given Cavity
  - Hence we need on 5 values and not 7



pra-sâmi

185

Probability Theory: Conditional Probability

Flu	Sneeze	Snore	μ
true	true	true	0.064
true	true	false	0.096
true	false	true	0.016
true	false	false	0.024
false	true	true	0.096
false	true	false	0.144
false	false	true	0.224
false	false	false	0.336

- What is:
  - P(flu ∧ sneeze) = 0.064 + 0.096 = 0.16
  - P(flu ∧ ¬sneeze) = 0.016 + 0.024 = 0.04
  - P(flu) = 0.064 + 0.096 + 0.016 + 0.024 = 0.2
  - P(sneeze | flu) = $\frac{P(flu \wedge sneeze)}{P(flu)} = \frac{0.16}{0.2}$ = 0.8
  - P(¬flu ∧ sneeze)
  - P(flu | sneeze)
  - P(sneeze | flu∧snore)
  - P(flu | sneeze∧snore)

pra-sâmi

186

Chain Rule

- Semantics of conditioning gives : $P(h \mid e) = \frac{P(h \wedge e)}{P(e)}$
  - $P(h \wedge e) = P(h \mid e) * p(e)$

- For $P(f_n \wedge f_{n-1} \wedge f_{n-2} \wedge \ldots \wedge f_1)$
  - $= P(f_n \mid f_{n-1} \wedge f_{n-2} \wedge \ldots \wedge f_1) * P(f_{n-1} \wedge f_{n-2} \wedge \ldots \wedge f_1)$
  - $= P(f_n \mid f_{n-1} \wedge f_{n-2} \wedge \ldots \wedge f_1) * P(f_{n-1} \mid f_{n-2} \wedge \ldots \wedge f_1) * P(f_{n-2} \wedge \ldots \wedge f_1)$
  - $= \prod_{i=1}^{n} P(f_i \mid f_1 \wedge f_2 \wedge \cdots \wedge f_{i-1})$

pra-sâmi