UNIT 5

Machine Learning

Contents

- Concepts of Learning
- Learning by analogy, Inductive Learning, Explanation based Learning
- Supervised Learning (Classification/Regression) – Nearest Neighbor, Naïve Bayes, Logistic Regression, Support Vector Machine, Neural Network
- Unsupervised Learning
  - Clustering K-means
  - Dimensionality Reduction (Principal Component Analysis, Linear Discriminant Analysis)
- Reinforcement Learning
- Fuzzy Learning
- Boltzmann Machines
- Deep Learning

Concept of Learning

Learning

- Acquiring new knowledge – Knowledge acquisition
- Modifying Old Knowledge, behavior, and skills- skills refinement
- May involve synthesizing different types of information
- Learning is also based on feedbacks

Concept of Learning

Learning

- Acquiring new knowledge – Knowledge acquisition
- Modifying Old Knowledge, behavior, and skills- skills refinement
- May involve synthesizing different types of information
- Learning is also based on feedbacks

Concept of Learning

Types of Learning

- Rote Learning
  - Rote learning is a technique which focuses on memorization
  - It avoids understanding the inner complexities and inferences of the subject that is being learned and instead focuses on memorizing the material
  - The major practice involved in rote learning is repetition
- Learning by examples
  - Agent learns by seeing examples and classify the similar object to same class
- Explanation based Learning
  - Describes objects in brief
- Learning by taking advice
- Learning by analogy

Concept of Learning

Types of Learning

- Inductive Learning
  - Determine general pattern, rules and facts
  - Mainly example of experienced based learning
- Deductive Learning
  - Determine specific patterns, rules and facts
  - New rules are generated from old ones

Concept of Learning

Machine Learning

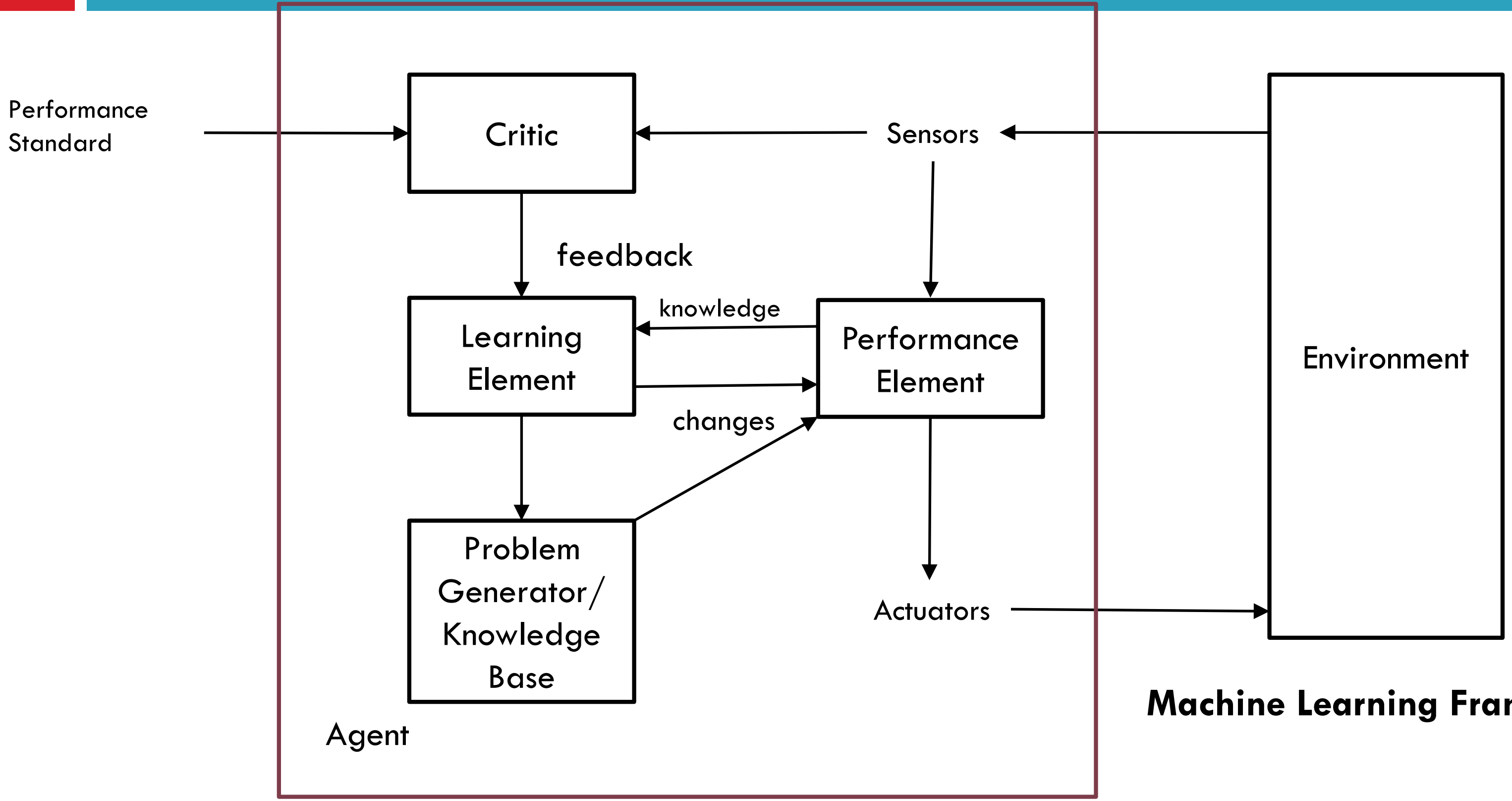
- Branch of AI that use algorithms to allow computer to evolve behaviours based on data collected from database or gathered through sensors
- Focuses on prediction, based on known properties learned from the training data
- The performance is usually evaluated with respect to the ability to reproduce known knowledge

Concept of Learning

Machine Learning

- Different cases of machine learning
  - Supervised learning: inputs and corresponding outputs are bind together
  - Unsupervised learning: only inputs are available for the learner
  - Reinforcement learning: learning based on reward or punishment

Concept of Learning



Machine Learning Framework

Concept of Learning

Machine Learning Framework

- Environment
  - It refers to the nature and quality of the information given to the learning element
  - The nature of information depends on its level
    - High level: information is abstract and deals with broad class of problems
    - Low level: information is detailed and deals with single problem
  - Quality of information involves noise free, ordered and reliable information

Concept of Learning

Machine Learning Framework

- Learning Element
  - Acquire new knowledge through learning element
  - Learning may be of any type discussed above
- Problem Generator/ Knowledge Base
  - Stores the information about the problems and solutions are suggested
  - Knowledge Base should be
    - Expressive: Knowledge should be represented in easy and understandable way
    - Modifiable: Must be easy to update or add new data in the knowledge base
    - Extendibility: the knowledge base should have feature to change its structure that should be well defined

Concept of Learning

Machine Learning Framework

- Performance Element
  - This part analyses how complex the learning is and how learning is being performed
  - Complexity depends on the type of task to be performed
  - It must send feedback to the learning system as well so that evaluation of overall performance could be done
  - The learning element should have access to all internal actions of the performance elements
- Sensors and Actuators
  - Sensors collect information from environment
  - Actuators implement the suggested action

Learning by Analogy

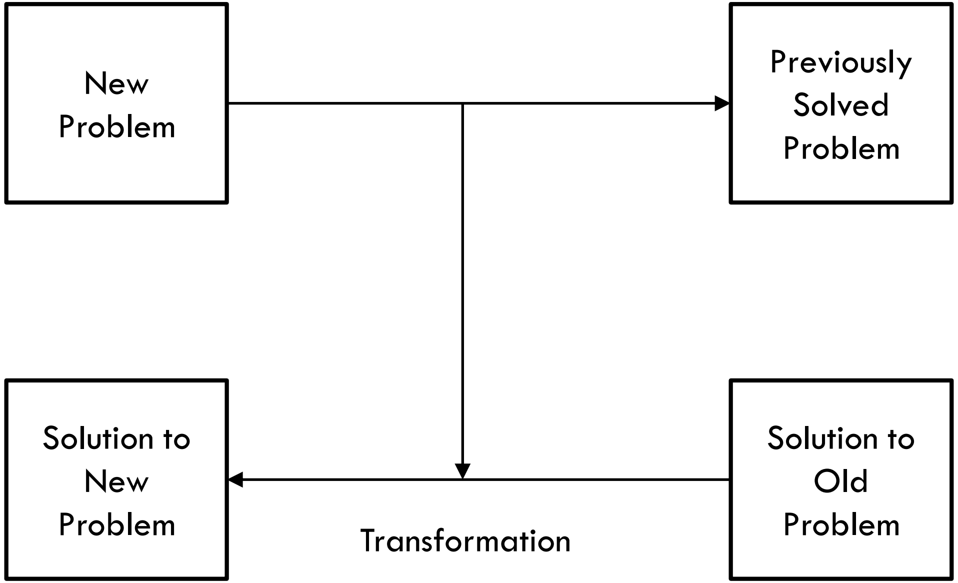
- Analogy is a powerful inference tool
- Our language and reasoning are laden with analogy
- Example:
  - Last month share market was a roller coaster.
  - Bill is like a fire engine
- So, AI must be able to grasp analogy for learning easily
- It is used in different learning strategies like learning by advice taking, learning in problem solving, etc.

Learning by Analogy

- Two methods of Analogy Problem solving are:
  - Transformational Analogy : focuses on final solution
  - Derivational Analogy: focuses on process of problem solving

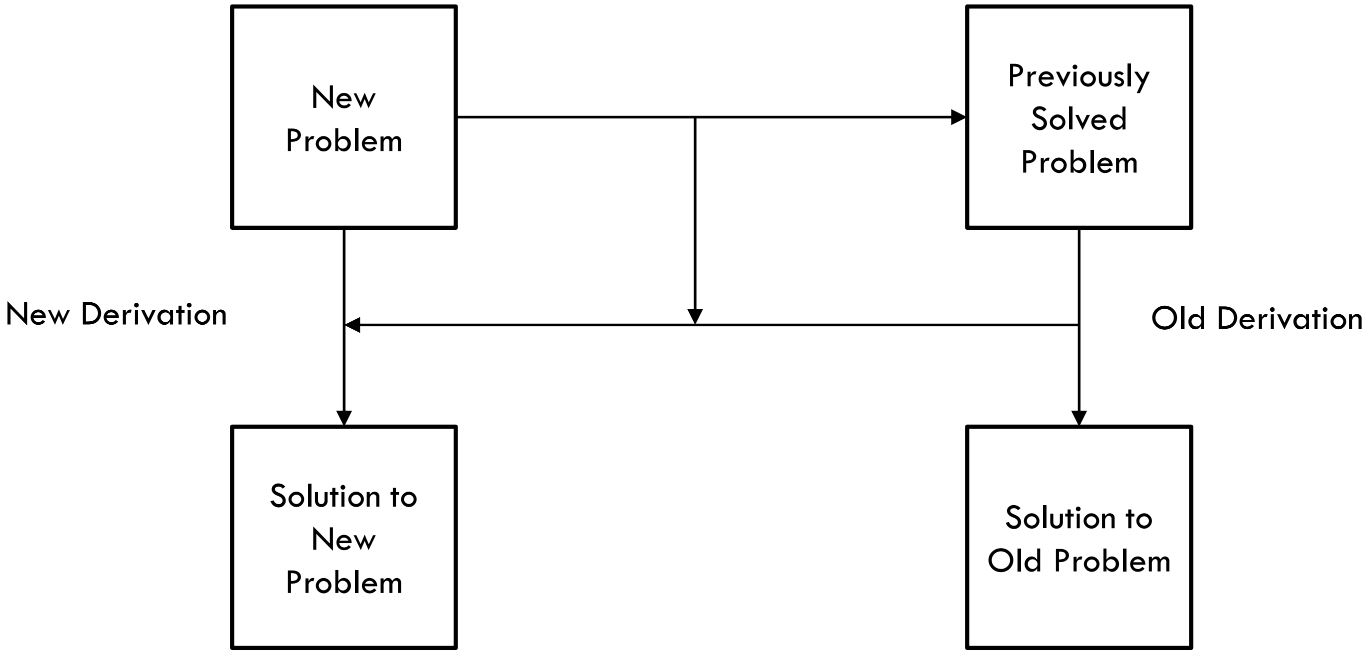
Learning by Analogy

- Transformational Analogy



Learning by Analogy

Derivational Analogy



Inductive Learning (Learning By Examples

- Based on classification of problems prior to solving
- Based on induction of result
- Based on generic facts, rules and patterns
- Also called concept learning
- Approaches:
  - Winston's Learning Program
  - Version Space
  - Decision Tree

Inductive Learning (Learning By Examples

Winston's Learning Program

- An early structural concept learning program
- Operated in simple blocks world domain
- Goal was to construct representations of the definitions of concepts in the block domain
- A near miss in Winston's learning program is an object that is not an instance of the concept in question but that is very similar to instances

Inductive Learning (Learning By Examples

Winston's Learning Program: Approach

- Begins with a structural description of one known instance of the concept
- Examine descriptions of other known instances of the concept
- Examine descriptions of near misses of the concept and restrict the definition to exclude this
- For example: refer to page number 459-461 of Artificial Intelligence, Rich and Knight

Explanation Based Learning

- *Explanation-based learning (EBL)* uses a domain theory to construct an explanation of the training example, usually a proof that the example logically follows from the theory
- Using this proof the system filters the noise, selects only the relevant to the proof aspects of the domain theory, and organizes the training data into a systematic structure
- This makes the system more efficient in later attempts to deal with the same or similar examples.

Explanation Based Learning

Basic Concept of EBL

- *Target concept*. The task of the learning system is to find an effective definition of this concept. Depending on the specific application the target concept could be a classification, theorem to be proven, a plan for achieving goal, or heuristic to make a problem solver more efficient.
- *Training example*. This is an instance of the target concept
- *Domain theory*. Usually this is a set of rules and facts representing domain knowledge. They are used to explain how the training example is an instance of the target concept.
- *Operationality criteria*. Some means to specify the form of the concept definition.

Supervised learning (Classification/Regression) - Nearest Neighbor, Naïve Bayes, Logistic Regression, Support Vector Machine, Neural Networks

Supervised Learning

Classification

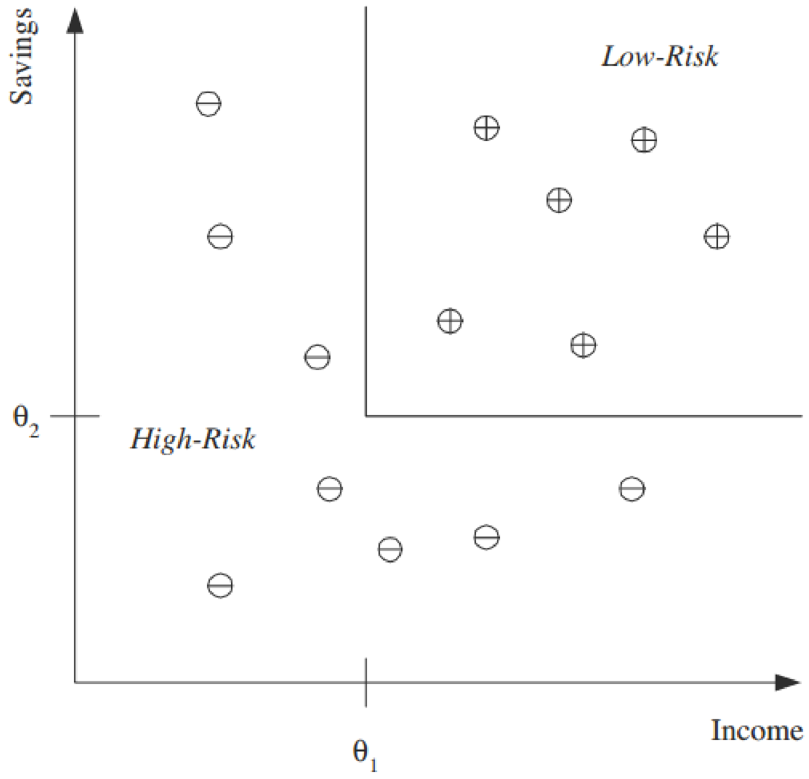
- The Basic cognitive process of arranging into classes or categories of the same types. [Dictionary Definition]
- The fundamental task of classification is prediction
- Ex: IF income > θ_1 AND saving > θ_2 THEN Low-risk ELSE high-Risk

for suitable value of θ_1 and θ_2

Supervised Learning

Classification

- Given new application with a certain income and saving we can decide whether it is high risk or low risk
- Instead of making 0/1(low-risk/high-risk) type decision we can calculate probability, P(X/Y), where X are customer attributes and Y is 0 or 1



Reasoning about Uncertainty Probability

- One of the most common characteristics of the human information available is its imperfection due to partial observability, non deterministic or combination of both
- An agent may not know what state it is in or will be after certain sequence of actions
- Agent can cope with these defects and make rational judgments and rational decisions to handle such uncertainty and draw valid conclusions

Reasoning about Uncertainty Probability

What is uncertainty?

- The lack of the exact knowledge that would enable us to reach a perfectly reliable conclusion
- Classical Logic permits only exact reasoning i.e. perfect knowledge always exists

IF A is true **THEN A is not false**	**and**	**IF B is true** **THEN B is not false**

- In Real world such clear cut knowledge could not be provided to systems

Reasoning about Uncertainty Probability

Sources of Uncertain Knowledge

- **Weak Implication:** Domain experts and knowledge engineer have rather painful or hopeless task of establishing concrete correlation between IF(Condition) and THEN(action) part of rules. Vague Data.
- **Imprecise Language** : NLP is ambiguous and imprecise. We define facts in terms of **often, sometimes, frequently, hardly ever**. Such can affect IF-THEN implication
- **Unknown Data**: incomplete and missing data should be processes to an approx. reasoning with this values
- **Combining the views of different experts**: Large system uses data from many experts

Reasoning about Uncertainty Probability

- The basic Concept of probability plays significant role in our life like we try to determine the probability of rain, prospect of promotion, likely hood of winning in Black Jack
- The probability of an event is the proportion of cases in which the event occurs (Good, 1959)
- Probability, mathematically, is indexed between 0 and 1
- Most events have probability index strictly between 0 and 1, which means that each event has at lease two possible outcomes: favorable outcome or success and unfavorable outcomes or failure

$$\mathrm{P}(\text{success}) = \frac{\mathit{The\ number\ of\ successes}}{\mathit{The\ number\ of\ possible\ outcomes}}$$

$$\mathrm{P}(failure) = \frac{\mathit{The\ number\ of\ failure}}{\mathit{The\ number\ of\ possible\ outcomes}}$$

Reasoning about Uncertainty Probability

- If *s* is the number of success and *f* is the number of failure then:

$$\mathrm{P}(success) = \frac{s}{s+f}$$

$$\mathrm{P}(failure) = \frac{f}{s+f}$$

and

$$p + q = 1$$

Reasoning about Uncertainty Probability

- Let us consider classical examples with a coin and a dice. If we throw a coin, the probability of getting a head will be equal to the probability of getting a tail. In a single throw, $s = f = 1$, and therefore the probability of getting a head (or a tail) is 0.5.
- Consider now a dice and determine the probability of getting a 6 from a single throw. If we assume a 6 as the only success, then $s = 1$ and $f = 5$, since there is just one way of getting a 6, and there are five ways of not getting a 6 in a single throw. Therefore, the probability of getting a 6 is

$$P = \frac{1}{1+5} = 0.1666$$

Likewise, the probability of not getting 6 is

$$q = \frac{5}{1+5} = 0.8333$$

Reasoning about Uncertainty Probability

- Above instances are for independent events i.e. mutually exclusive events which can not happen simultaneously
- In the dice experiment, the two events of obtaining a 6 and, for example, a 1 are mutually exclusive because we cannot obtain a 6 and a 1 simultaneously in a single throw. However, events that are not independent may affect the likelihood of one or the other occurring. Consider, for instance, the probability of getting a 6 in a single throw, knowing this time that a 1 has not come up. There are still five ways of not getting a 6, but one of them can be eliminated as we know that a 1 has not been obtained. Thus,

$$p = \frac{1}{1 + (5 - 1)}$$

Reasoning about Uncertainty Probability

- Let A and B be two not mutually exclusive events, but occur conditionally on the occurrence of other.
- The probability of event A will occur if event B occurs is called conditional Probability

$$\mathrm{p(A|B)} = \frac{\textit{the number of times A and B can occur}}{\textit{the number of times B can occur}}$$

The probability of both A and B will occur is called joint probability $(\boldsymbol{A} \cap \boldsymbol{B})$

$\mathrm{p(A|B)} = \frac{p(A \cap B)}{p(B)}$, the probability of A occurring given B has occurred

$\mathrm{p(B|A)} = \frac{p(B \cap A)}{p(A)}$, the probability of B occurring given A has occurred

Reasoning about Uncertainty Probability

Joint probability is commutative, thus

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

Therefore,

$$p(A \cap B) = p(B|A) * p(A)$$

Now the final equation becomes:

$$p(A|B) = \frac{p(B|A)*p(A)}{p(B)}$$ --(a)

Where:

p(A|B) is the conditional probability that event A occurs given event B has occurred

p(B|A)is the conditional probability that event B occurs given event A has occurred

p(A)is the probability of event A occurring   p(B)is the probability of event B occurring

The above equation (a) is known as **Bayesian Rule**

Reasoning about Uncertainty Probability

- For *n* number of mutually exclusive event B we have

$$p(A \cap B_1) = p(A|B_1) \times p(B_1)$$
$$p(A \cap B_2) = p(A|B_2) \times p(B_2)$$
$$\vdots$$
$$p(A \cap B_n) = p(A|B_n) \times p(B_n)$$

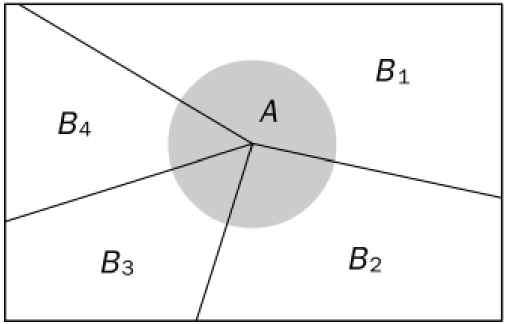
or when combined:

$$\sum_{i=1}^{n} p(A \cap B_i) = \sum_{i=1}^{n} p(A|B_i) \times p(B_i)$$

Reasoning about Uncertainty Probability

- Summed over an exhaustive list of events for Bi, we get :

$$\sum_{i=1}^{n} p(A \cap B_i) = p(A)$$



- $$p(A) = \sum_{i=1}^{n} p(A|B_i) \times p(B_i)$$

Reasoning about Uncertainty Probability

- If the occurrence of A depends on only two mutually exclusive events, i.e. B and NOT B, then above equation becomes

$$p(A) = p(A|B) \times p(B) + p(A|\neg B) \times p(\neg B)$$

- Similarly,

$$p(B) = p(B|A) \times p(A) + p(B|\neg A) \times p(\neg A)$$

- Substituting above equations in Bayesian Equation, We get:

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B|A) \times p(A) + p(B|\neg A) \times p(\neg A)}$$

Bayesian Networks

Why Bayesian Network???

- To represent the probabilistic relationship between two different classes
- To avoid dependencies between values of attributes by joint conditional probability distribution
- In Naïve Bayes classifier, attributes are conditionally independent

Bayesian Networks

- Bayesian Network are also known as **Bayes Network, Belief Networks** and **Probabilistic Networks**

- A BN is defined by two parts, **Directed Acyclic Graph (DAG)** and **Conditional Probability Tables (CPT)**

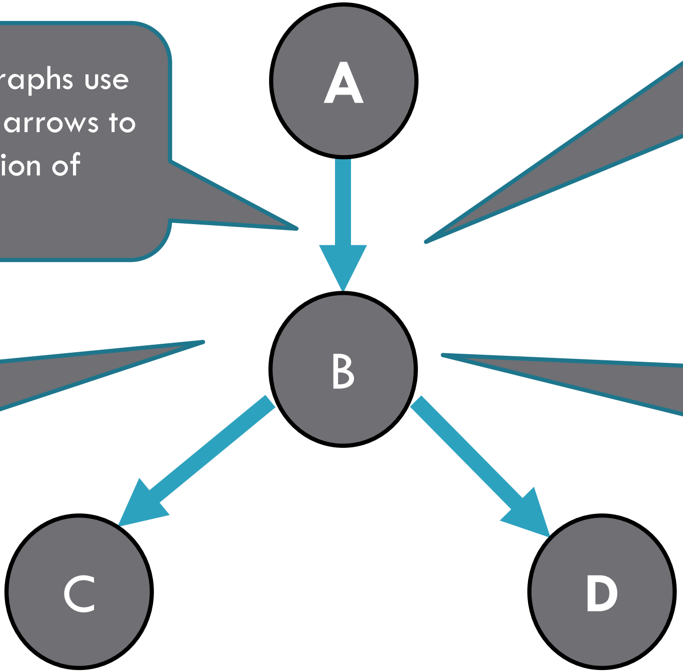
  Nodes → Random Variables

  Arcs→ Indicates Probabilistic dependencies between nodes

Bayesian Networks

Directed acyclic graphs use only unidirectional arrows to show the direction of causation

Follow the general graph principles such as a node A is parent of another node B, if there is an arrow from node A to node B



Each node in graph represents a random variable

Informally, an arrow from node X to node Y means X has a direct influence on Y

Bayesian Networks

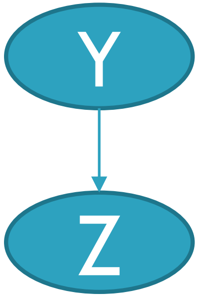
A BN is a directed graph with the following properties:

- **Nodes:** Set of Random Variables which may be discrete or continuous
- **Directed Links (Arcs) :** The real meaning od a link from node X to node Y is that X has a direct influence on Y
- Each node has a Conditional Probability Distribution $\mathbf{P}(X_i|Parents(X_i))$ that quantifies the effects that the parent have on the node
- The graph has no directed cycles

Bayesian Networks

A BN is a directed graph with the following properties (contd…)

- If an arc is drawn from Y to Z, then Y is a parent or immediate predecessor of Z, and Z is a descendant of Y



- Each variable is conditionally independent of its non-descendants in the graph, given its parents

Bayesian Networks

Incremental Network Construction:

1. ***Nodes***: First determine the set of variables that are required to model the domain. Now order them, $\{X_1, X_2, \ldots, X_n\}$. Any order will work, but the resulting network will be more compact if the variables are ordered such that causes precede effects

2. ***Links*** : for $i = 1\ to\ n \quad do$:
    1. Choose, from $X_1, \ldots, X_{n-1}$, a minimal set of parents for X_i such that equation $\mathbf{P}(X_i | X_{i-1}, \ldots, X_1) = \mathbf{P}(X_i | Parents(X_i))$ is satisfied
    2. For each parent insert a link from the parent to X_i
    3. CPTs: Write down the Conditional Probability Table, $\mathbf{P}(X_i | Parents(X_i))$

Bayesian Networks

Conditional Independence:

$$\mathbf{P}(X_1, X_2, \dots, X_n) = \mathbf{P}(X_n | X_{n-1}, \dots, X_1)\, \mathbf{P}(X_{n-1}, \dots, X_1)$$

$$= \mathbf{P}(X_n | X_{n-1}, \dots, X_1)$$

$$\mathbf{P}(X_{n-1}, \dots, X_1) \dots \mathbf{P}(X_2 | X_1) \mathbf{P}(X_1)$$

$$= \sum_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i))$$

A BN represents Conditional Independence

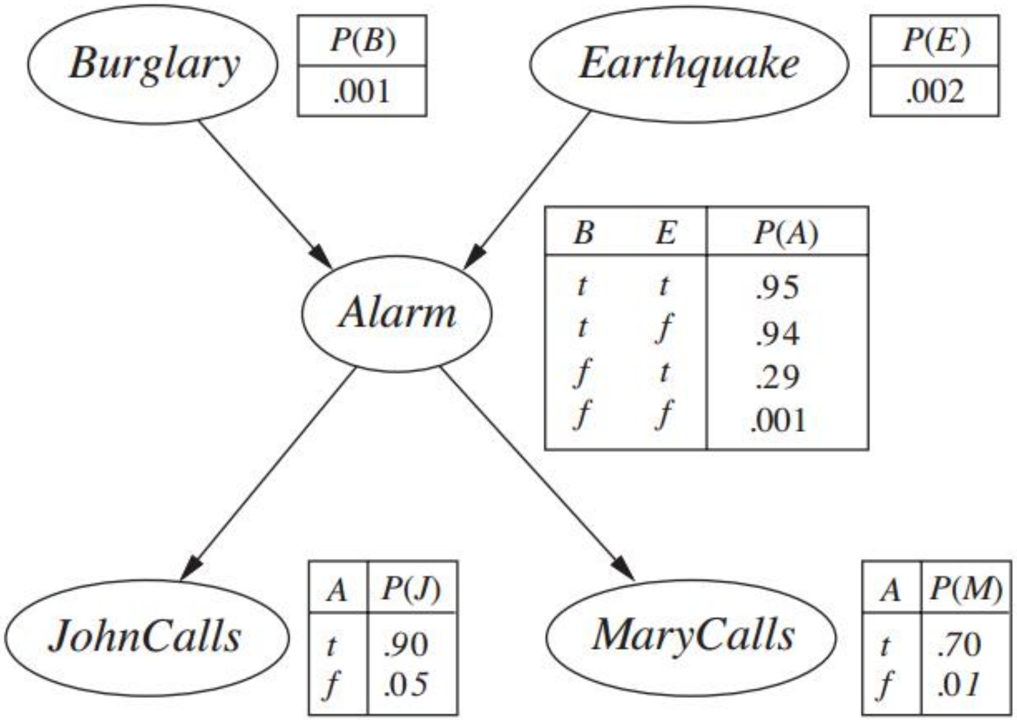
$$\mathbf{P}(X_i | X_{i-1}, \dots, X_1) = \mathbf{P}(X_i | Parents(X_i))$$

Bayesian Networks

Example

- Burglar Alarm at Home
  - Fairly reliable at detecting a Burglary
  - Also Respond at times of Earthquake
- Two neighbors (John and Mary) on hearing Alarm calls you
  - John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then too
  - Mary likes aloud music and sometimes misses the alarm altogether

Bayesian Networks



P(B)
.001

P(E)
.002

B	E	P(A)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	P(J)
t	.90
f	.05

A	P(M)
t	.70
f	.01

Bayesian Networks

- Inference from Effect to cause; given Burglary, what is P(J|B)?

  $P(J \mid B) = ?$

  first calculate probability of Alarm ringing on burglary:

  $P(A \mid B) = P(B)P(\neg E)P(B\cap \neg E) + P(B)P(E)P(B\cap E)$

  $P(A \mid B) = 1*(0.998)*(0.94) + 1*(0.002)*(0.95)$

  $P(A \mid B) = 0.94$

  Now, Let us calculate $P(J \mid B)$

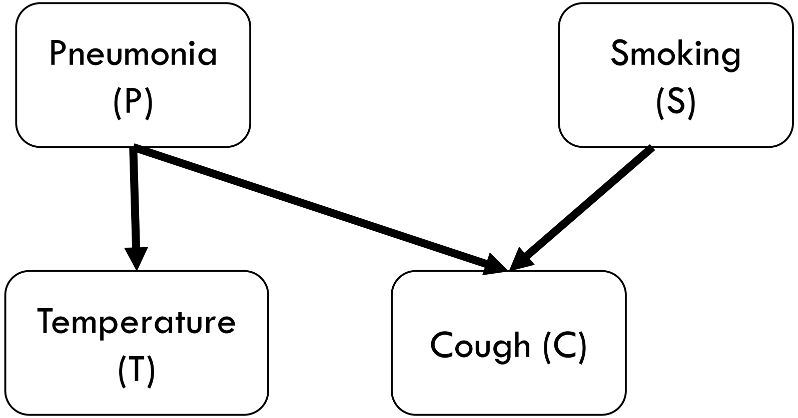
  $P(J \mid B) = P(A \mid B)*P(J) + P(\neg(A \mid B))*P(\neg J)$

  $P(J \mid B) = (0.94) * (0.9) + (0.06) * (0.05) = 0.85$

- Also calculate $P(M|B) = ?$

Bayesian Network

Pneumonia	
True	0.1
False	0.9



smoking	
True	0.2
False	0.8

Find:

a. P(C| S ^ P)

b. P(S | C)

	Temperature	
Pneumonia	Yes	No
Yes	0.9	0.1
No	0.2	0.8

Pneumonia	Smoking	Cough	
		True	False
True	Yes	0.95	0.05
True	No	0.8	0.2
False	Yes	0.6	0.4
False	No	0.05	0.95

Bayesian Networks

Benefits of BN:

- It can readily handle incomplete data sets
- It allows one to learn about causal relationships
- It readily facilitate use of prior knowledge
- It Provide a natural representation for conditional independence
- It is more complex to construct the graph

Supervised Learning

Regression

- The cause and effect relationship
- Example: price of second hand car depends on brand, year, engine, capacity, mileage, etc
- $y = wx + w_0$
- The above equation is known as simple linear regression
- Ex: is there a relationship between advertising budget and sales?

Supervised Learning

Logistic Regression

- Above equation solves the equation directly in the form 0 or 1 i.e. True or false but can not show the degree of dependence
- **Logistic Regression:** models the probability that "Y" belongs to a particular category
- For probability of data "*default*" given *balance* could be written as:

$$P(default = True \mid balance)$$

Supervised Learning

Logistic Regression

- The value of $P(default = true \mid balance)$ ranges between 0 and 1
- For any given value of *balance*, a **prediction** can be made for *default*

Supervised Learning

Logistic Regression seeks to:

- **Model** the probability of an event occurring depending on the values of the independent variables, which can be categorical or numerical
- **Estimates** the probability that an event occurs for the randomly selected observation versus the probability that the event does not occur
- **Predict** the effect of a series of variables on a binary response variable
- **Classify** observations by estimating the probability that an observation is in particular category

Unsupervised Learning

Principal Components Analysis

- When faced with a **large set of correlated variables,** principal components allow us to **summarize** this set with a smaller number of representative variables **that collectively explain** most of the variability in the original set
- *Principal component analysis* (PCA) refers to *the process by which principal components are computed, and the subsequent use of these components in understanding the data*
- PCA also serves as a tool for data visualization (visualization of the observations or visualization of the variables)
- Visualizing correlation

Unsupervised Learning

Clustering

- Broad set of techniques for finding subgroups or clusters in a dataset
- Both clustering and PCA seek to *simplify the data via a small number of summaries*, but their mechanisms are different:
  - PCA looks to find a low-dimensional representation of the observations that explain a good fraction of the variance;
  - Clustering looks to find homogeneous subgroups among the observations.

Unsupervised Learning

K-Means Clustering

- *K*-means clustering is a simple and elegant approach for partitioning a data set into *K* distinct, non-overlapping clusters
- Let $C_1, \ldots, C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:
  - $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the *K* clusters.
  - $Ck \cap Ck' = \emptyset$ for all $k \neq k'$. In other words, the clusters are noverlapping: no observation belongs to more than one cluster.

Unsupervised Learning

K-Means Clustering

Algorithm for K-means Clustering

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   1. For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   2. Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

Unsupervised Learning

Linear Discriminant Analysis

- Logistic regression involves directly modeling $P(Y = k | X = x)$ using the logistic function for the case of two response classes
- In statistical jargon, we model the conditional distribution of the response Y, given the predictor(s) X
- LDA is less direct approach to estimate these probabilities
- In this approach, *we model the distribution of the predictors X separately in each of the response classes (i.e. given Y), and then use Bayes' theorem to flip these around into estimates for* $P(Y = k | X = x)$

Unsupervised Learning

Linear Discriminant Analysis

Need for LDA when we have logistic regression

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If *n* is small and the distribution of the predictors *X* is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model
- Linear discriminant analysis is popular when we have more than two response classes.

Reinforcement Learning

- Consider, for example, the problem of learning to play chess. A supervised learning agent needs to be told the correct move for each position it encounters, but such feedback is
  seldom available. In the absence of feedback from a teacher, an agent can learn a transition model for its own moves and can perhaps learn to predict the opponent’s moves, but *without some feedback about what is good and what is bad, the agent will have no grounds for deciding which move to make*
- The agent needs to know that something good has happened when it (accidentally) checkmates the opponent, and that something bad has happened when it is checkmated—or vice versa, if the game is suicide chess
- This kind of feedback is called reward or reinforce and such learning approach is called reinforcement learning

Reinforcement Learning

The basic reinforcement learning model consists

1. A set of environment states S
2. A set of actions A
3. Rules for transitioning between states
4. Rules that determine the scalar immediate reward of a transition
5. Rules that describes what agent observes

Fuzzy Learning

- Is a form of many valued logic
- In contrast with traditional logic theory, where binary sets have two valued logic, fuzzy logic variable may have a truth value that ranges in degree of 0 to 1
- Fuzzy logic is a concept of partial truth where true value may range between completely true or completely false

Fuzzy Learning

- Fuzzy logic reflects how two people think it attempts to model our sense of our decision making and our common sense
- In fuzzy theory, fuzzy set A of universe X is defined by function $U_a(x)$ called membership function of set A
- $U_a(x)$:X→[0,1],

  where, $U_a(x)$ =1 if x is totally in A

  =0 if x is not in A

  $0 < U_a(x) < 1$, if x is partly in A.

Fuzzy Learning

- Fuzzy logic reflects how two people think it attempts to model our sense of our decision making and our common sense
- In fuzzy theory, fuzzy set A of universe X is defined by function $U_a(x)$ called membership function of set A
- $U_a(x):X \rightarrow [0,1]$,

  where, $U_a(x)$ =1 if x is totally in A

  =0 if x is not in A

  $0< U_a(x)<1$, if x is partly in A.

Fuzzy Learning

- For any element x of X, membership function $U_a(x)$ equals the degree to which x is an element of set A.
- This degree ranges from 0 to 1 representing degree of membership called membership value of element x is set A
- A fuzzy set is often denoted by its membership function
- Fuzzy Inferences
  - Mamdani inferences: fuzzification of input variables->rule evaluation->aggregation of the rule outputs->defuzzification
  - Sugeno fuzzy inferences

Boltzmann Machine

- It's a network of symmetrically connected, neuron like units that make stochastic decision about whether to be on or off
- Has simple learning algorithm that allows it to discover interesting features that represent complex regularities in training data
- Very slow learning
- For searching, Boltzmann machine has fixed weights on the connection but for learning weights use small update

Boltzmann Machine

- Output y is given by,

$$y = b_i + \sum x_j w_{ij}$$

- Where x_i=1 for on state and 0 for off
- Probability of being on is

$$Prob(x_j = 1) = \frac{1}{1 + e^{-y}}$$

Deep Learning

- On Your Own…

68

Thank You