Artificial Intelligence and Machine Learning

Probability Based Learning

Big Idea

- We can use estimates of likelihoods to determine the most likely prediction that should be made.
- More importantly, we revise these predictions based on data we collect and whenever extra evidence becomes available.

Classification Problem Example

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	TRUE	TRUE	FALSE	FALSE
2	FALSE	TRUE	FALSE	FALSE
3	TRUE	FALSE	TRUE	FALSE
4	TRUE	FALSE	TRUE	FALSE
5	FALSE	TRUE	FALSE	TRUE
6	TRUE	FALSE	TRUE	FALSE
7	TRUE	FALSE	TRUE	FALSE
8	TRUE	FALSE	TRUE	TRUE
9	FALSE	TRUE	FALSE	FALSE
10	TRUE	FALSE	TRUE	TRUE

HEADACHE	FEVER	VOMITING	MENINGITIS
TRUE	FALSE	TRUE	?

Probability

- Basic element: random variable/feature
- Domain values for a feature must be
  - Exhaustive
  - Mutually exclusive
- Elementary propositions are constructed by the assignment of a value to a random feature.
- Prior or unconditional probability associated with a proposition is the degree of belief accorded to it in the absence of any other information.

Probability (II)

- A probability function, P(x=v), returns the probability of a feature, x, taking a specific value, v.
- A joint probability refers to the probability of an assignment of specific values to multiple different features.
- A conditional probability refers to the probability of one feature taking a specific value given that we already know the value of a different feature.
- A probability distribution is a data structure that describes the probability of each possible value a feature can take. The sum of a probability distribution must equal to 1.
- A joint probability distribution is a probability distribution over more than one feature assignment and is written as a multi-dimensional matrix in which each cell lists the probability of a particular combination of feature values being assigned. The sum of all the cells in a joint probability distribution must equal to 1.

Fundamentals

- Bayes' Theorem

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

- Example:
  A patient has tested positive for a serious disease. The test is 99% accurate. However, the disease is extremely rare, striking only 1 in 10,000 people. What is the actual probability that the patient has the disease?

Application of Bayes' Theorem

$$P(disease|test+) = \frac{P(test+|disease)P(disease)}{P(test+)}$$

$$P(test+) = P(test+|disease)P(disease)$$

$$+P(test+|\neg disease)P(\neg disease)$$

= 0.99 * 0.0001 + 0.01 * 0.9999

= 0.0101

$$P(disease|test+) = \frac{0.99 * 0.0001}{0.0101} = 0.0098$$

Generalized Bayes’ Theorem

$$P(t = l|q[1], \dots, q[m]) = \frac{P(q[1], \dots, q[m]|t = l)P(t = l)}{P(q[1], \dots, q[m])}$$

- Chain Rule (apply to a conditional probability):

$$P(q[1], \dots, q[m]|t = l) = P(q[1]|t = l) \times P(q[2]|q[1], t = l) \times \cdots \times P(q[m]|q[m-1], \dots, q[1], t = l)$$

Bayesian MAP (maximum posteriori) Prediction Model

$$H_{MAP}(q) = \operatorname*{argmax}_{l \in levels(t)} P(t = l \mid q[1],\ldots,q[m])$$

$$= \operatorname*{argmax}_{l \in levels(t)} \frac{P(q[1],\ldots,q[m] \mid t = l) \times P(t = l)}{P(q[1],\ldots,q[m])}$$

$$= \operatorname*{argmax}_{l \in levels(t)} P(q[1],\ldots,q[m] \mid t = l) \times P(t = l)$$

Advantages of Bayesian Model

- Probabilistic learning: Calculate explicit probabilities for hypothesis, among the most practical approaches to certain types of learning problems
- Incremental: Each training example can incrementally increase/decrease the probability that a hypothesis is correct. Prior knowledge can be combined with observed data.
- Probabilistic prediction: Predict multiple hypotheses, weighted by their probabilities
- Standard: Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured

Disadvantages

- Practical difficulty: require initial knowledge of many probabilities, significant computational cost
- Curse of Dimensionality: As the number of descriptive features grows, the number of potential conditioning events grows. Consequently, an exponential increase is required in the size of the dataset as each new descriptive feature is added to ensure that for any conditional probability, there are enough instances in the training dataset matching the conditions so that the resulting probability is reasonable.
- If dataset is not large enough, model is over-fitting to the training data.

Independent Events

- If knowledge of one event has no effect on the probability of another event, and vice versa, then the two events are independent of each other.
- If two events X and Y are independent then:
  $P(X|Y) = P(X)$
  $P(X, Y) = P(X) \times P(Y)$
- Full independence between events is quite rare.
- Two, or more, events are independent when a third event has happened, then these events are conditionally independent.

Naïve Bayes' Classifier

- Naïve Bayes' Classifier assumes that the attributes are conditionally independent:

$$P(q[1], \ldots, q[m] | t = l) = \prod_{i=1}^{m} P(q[i] | t = l)$$

- Naïve Bayes' Classifier:

$$H_{MAP}(q) = \underset{l \in levels(t)}{argmax} \left(\prod_{i=1}^{m} P(q[i] | t = l)\right) \times P(t = l)$$

Play-tennis example: estimating $P(x_i|C)$

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

outlook	
P(sunny|p) = 2/9	P(sunny|n) = 3/5
P(overcast|p) = 4/9	P(overcast|n) = 0
P(rain|p) = 3/9	P(rain|n) = 2/5
temperature	
P(hot|p) = 2/9	P(hot|n) = 2/5
P(mild|p) = 4/9	P(mild|n) = 2/5
P(cool|p) = 3/9	P(cool|n) = 1/5
humidity	
P(high|p) = 3/9	P(high|n) = 4/5
P(normal|p) = 6/9	P(normal|n) = 2/5
windy	
P(true|p) = 3/9	P(true|n) = 3/5
P(false|p) = 6/9	P(false|n) = 2/5

P(p) = 9/14

P(n) = 5/14

Play-tennis example: classifying X

- An unseen sample
  X = <rain, hot, high, false>
- P(X|p)·P(p) =
  P(rain|p)·P(hot|p)·P(high|p)·P(false|p)·P(p) =
  3/9·2/9·3/9·6/9·9/14 = 0.010582
- P(X|n)·P(n) =
  P(rain|n)·P(hot|n)·P(high|n)·P(false|n)·P(n) =
  2/5·2/5·4/5·2/5·5/14 = 0.018286
- Sample X is classified in class n (don't play)

The independence hypothesis…

- … makes computation possible
- … yields optimal classifiers when satisfied
- … but is seldom satisfied in practice, as attributes (variables) are often correlated.
- Surprisingly, given the naivety and strength of the assumption it depends upon, a Naïve Bayes' model often performs reasonably well.

Handling Insufficient Data

- Problem:
  - Probability of some case is 0 because of lacking of sample
- Solution – using smoothing:
  - Smoothing takes some of the probability from the events with lots of the probability share and gives it to the other probabilities in the set.
  - There are several different ways to smooth probabilities.
  - Laplacian smoothing is one of them:

$$P(f = v|t) = \frac{count(f = v|t) + k}{count(f|t) + (k \times |domain(f)|)}$$

Handling Continuous Features -- PDF

- A probability density function (PDF) represents the probability distribution of a continuous feature using a mathematical function.
- A PDF defines a density curve and the shape of the curve is determined by:
  - The statistical distribution that is used to define the PDF
  - The values of the statistical distribution parameters

Some Standard Probability Distributions

- Normal distribution

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Exponential distribution

$$E(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & for\ x > 0 \\ 0 & otherwise \end{cases}$$

- Poisson distribution

$$P(x = k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

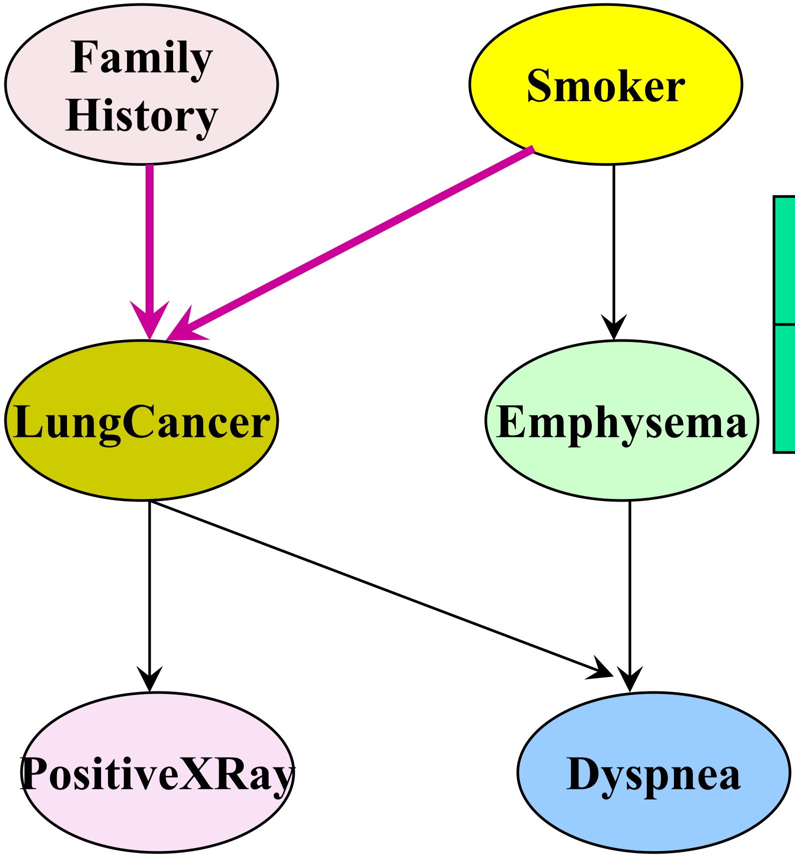
Handling Continuous Features -- Binning

- Use equal-width or equal-frequency techniques to bin continuous features into categorical features.

Bayesian Belief Networks

- Bayesian networks use a graph-based representation to encode the structural relationships between subsets of features in a domain.
- Consequently, a Bayesian network representation is generally more compact than a full joint distribution, yet is not forced to assert global conditional independence between all descriptive features.
- A Bayesian Belief Network is a directed acyclical graph that is composed of three basic elements:
  - Nodes (variables)
  - Edges (causal links)
  - Conditional probability tables (CPT)

Bayesian Belief Network Example



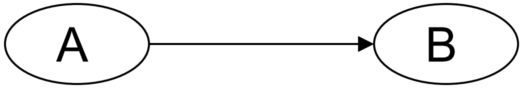
	(FH, S)	(FH, ~S)	(~FH, S)	(~FH, ~S)
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

The conditional probability table for the variable LungCancer

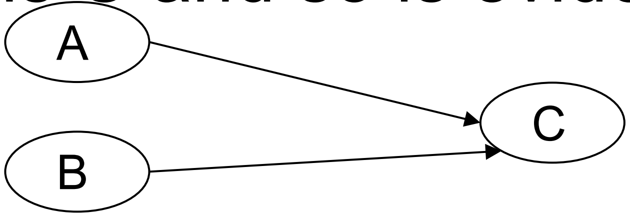
Bayesian Belief Networks

The Meaning of Edges in BBN

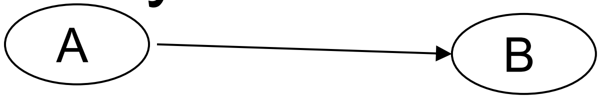
- Causal: A causes B, increased probability of A makes B more likely.



- Inter-causal: A and B can each cause C. B explains C and so is evidence against A.



- Evidential: Increased probability of B makes A more likely. B is evidence for A. A depends on B.



Querying the Bayesian Belief Networks

- Each query asks for a joint probability which is computed by applying the chain rule (multiplying corresponding conditional probabilities for each variable involved in the query and its dependents).
- This is because all conditional probabilities or each node given its parent are in CPTs, and each query for conditional probability of a parent given its children can be computed using Bayes theorem.

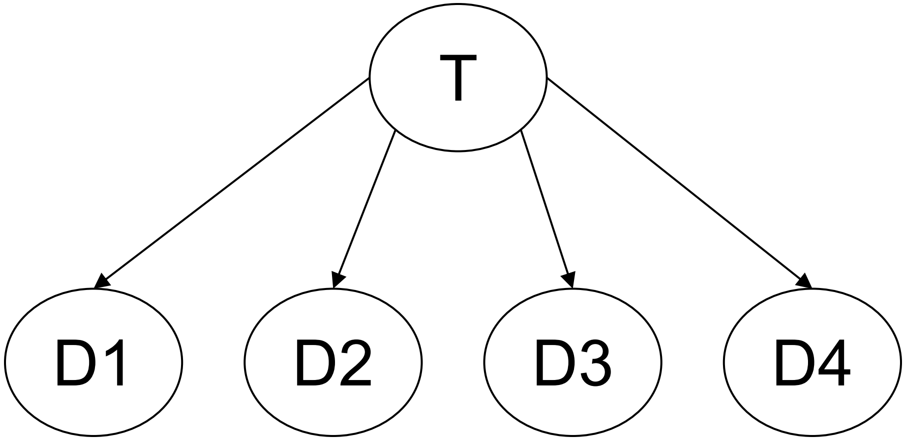
Markov Blanket

- For conditional independence, we need to take into account not only the parents of a node but also the state of its children and their parents.
- The set of nodes in a graph that make a node independent of the rest of the graph are known as Markov blanket of a node.
- The conditional independence of a node x_i in a graph with n nodes is defined as:

$$P(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$
$$= P(x_i|Parents(x_i)) \times \prod_{j \in children(x_i)} P(x_j|Parents(x_j))$$

Naïve Bayes as a Special case of Bayesian Belief Network

- A naïve Bayes classifier is a Bayesian belief network with a specific topological structure.



Missing Parent Value

- Computing a conditional probability for a node becomes more complex if the value of one or more of the parent nodes is unknown.
- They are the hidden variables
- Solution: for each hidden variable consider all possible values of this variable and perform summation by substituting this variable with all possible values in turn
- The power of BBN: When complete knowledge of the state of all the nodes in the network is not available, we can still proceed with the nodes that we do have knowledge of and sum out the unknown nodes.

Learning with Bayesian Belief Networks

- Topology of the network is given (done by human experts)
  - The learning task then involves inducing the CPT from the data
- Automated learning of network topology from the data

Constructing Bayesian Network

- Choose an ordering of variables $x_1, \dots, x_n$
- For i = 1 to n
  - Add x_i to the network
  - Select parents from $x_1, \dots, x_{i-1}$ such that
    $$P(x_i | Parents(x_i)) = P(x_i | x_1, \dots, x_{i-1})$$

This choice of parents guarantees

$$P\left(x_q, \dots, x_n\right) = \prod_{i=1}^{n} P(x_i |\, x_1, \dots, x_{i-1})$$

(by chain rule)

$$= \prod_{i=1}^{n} P(x_i | Parents(x_i))$$

$$(by\ construction)$$

Summary

- Bayesian Belief Networks provide a clean, clear, manageable language and methodology for expressing what we are certain and uncertain about.
- Deciding conditional independence is hard in non-causal directions.
- Causal models and conditional independence seem hardwired for humans ➔ generally easy for domain experts to construct the network topology.
- Easy for machine to learn the CPT from the training dataset given the topology of the network.