Machine Learning

Classification using Bayesian Belief Networks

Satishkumar L. Varma

Department of Information Technology
SVKM's Dwarkadas J. Sanghvi College of Engineering, Vile Parle, Mumbai.
ORCID | Scopus | Google Scholar | Google Site | Website

Outline

- Classification
  - Bayesian Belief Networks
  - Hidden Markov Models
  - Support Vector Machine
    - Maximum Margin Linear Separators
    - Quadratic Programming solution to finding maximum margin separators
    - Kernels for learning non-linear functions
  - Classification using k Nearest Neighbour Algorithm

Bayesian Belief Networks

- Naive Bayes Classifier OR Naive Bayesian Classifier
  - It is useful for text classification and spam filtering
- Bayesian Network OR Bayesian Belief Networks OR Belief Networks or Bayes Nets
  - It is useful in medical diagnosis, bioinformatics, and NLP

Naive Bayesian Classifier

- **Naive Bayesian Classifier**
- It is based on Bayes' theorem with independence assumptions between predictors.
- It requires less data to get a good result in many cases.
- A Naive Bayesian model is easy to build,
  - with no complicated iterative parameter estimation which
  - makes it particularly useful for very large datasets.
- Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and
- It is widely used because it often outperforms more sophisticated classification methods.
- Use it if you're only interested in solving a prediction task: use Naive Bayes.
- Algorithm:
  - Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c).
  - Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.
  - This assumption is called class conditional independence.

Naive Bayesian Classifier

- P(c|x) is posterior prob. of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the prob. of predictor given class.
- P(x) is the prior probability of predictor.

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Naive Bayesian Classifier

- Example:
- The posterior prob. can be calculated by first, constructing a frequency table for each attribute against target.
- Then, transforming the frequency tables to likelihood tables and
- Finally use the Naive Bayesian equn to calculate the posterior probability for each class.
- The class with the highest posterior probability is the outcome of prediction.

$$P(x \mid c) = P(Sunny \mid Yes) = 3/9 = 0.33$$

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$$P(x) = P(Sunny) = 5/14 = 0.36$$

$$P(c) = P(Yes) = 9/14 = 0.64$$

Posterior Probability: $P(c \mid x) = P(Yes \mid Sunny) = 0.33 \times 0.64 \div 0.36 = 0.60$

Naive Bayesian Classifier

- The zero-frequency problem
  - Add 1 to the count for every attribute value-class combination (Laplace estimator) when an attribute value (Outlook=Overcast) doesn't occur with every class value (Play Golf=no).
- Numerical Predictors
  - Numerical variables need to be transformed to their categorical counterparts (binning) before constructing their frequency tables.
  - Other option is using the distribution of the numerical variable to have a good guess of the frequency.
  - For example, one common practice is to assume normal distributions for numerical variables.
  - The PDF for the normal distribution is defined by two parameters (mean and standard deviation).

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Mean

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5}$$

Standard deviation

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

Normal distribution

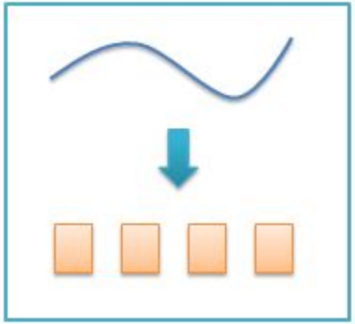
		Humidity	Mean	StDev
Play Golf	yes	86 96 80 65 70 80 70 90 75	79.1	10.2
	no	85 90 70 95 91	86.2	9.7

$$P(\text{humidity} = 74 \mid \text{play} = \text{yes}) = \frac{1}{\sqrt{2\pi}(10.2)} e^{-\frac{(74-79.1)^2}{2(10.2)^2}} = 0.0344$$

$$P(\text{humidity} = 74 \mid \text{play} = \text{no}) = \frac{1}{\sqrt{2\pi}(9.7)} e^{-\frac{(74-86.2)^2}{2(9.7)^2}} = 0.0187$$

Naive Bayesian Classifier

- **Binning**
- Binning or discretization is the process of transforming numerical variables into categorical counterparts.
- An example is to bin values for Age into categories such as 20-39, 40-59, and 60-79.
- Numerical variables are usually discretized in the modeling methods based on frequency tables
  - e.g., decision trees
- Moreover, binning may improve accuracy of the predictive models by reducing the noise or non-linearity.
- Finally, binning allows easy identification of outliers, invalid and missing values of numerical variables.
- There are two types of binning, unsupervised and supervised.



Bayesian Belief Networks

- Bayesian Network
- A Naive Bayes classifier is a simple model that describes particular class of Bayesian network
  - where all of the features are class-conditionally independent.
- Because of this, there are certain problems that Naive Bayes cannot solve XOR problem.
- Example: XOR
  - You have a learning problem with binary features x1 and x2 and a target variable y = x1 XOR x2.
  - In a Naive Bayes classifier, x1 and x2 must be treated independently
    - so you would compute things like "The probability that y = 1 given that x1 = 1"
    - hopefully you can see that this isn't helpful,
    - because x1 = 1 doesn't make y = 1 any more or less likely.
    - Since a Bayesian network does not assume independence, it would be able to solve such a problem.
- Bayesian Network Models
  - It model relationships between features in a very general way.
  - If you know what these relationships are, or have enough data to derive them,
  - then it may be appropriate to use a Bayesian Network.

Bayesian Belief Networks

- Bayesian Network OR Bayesian Belief Networks OR Belief Networks or Bayes Nets
- It is a type of probabilistic graphical model.
- It represents the relationship between variables in the form of a directed acyclic graph (DAG)
- It is used for reasoning, learning and decision making.
- It can handle uncertainty.
- It can incorporate prior knowledge into the model.
- It is useful in medical diagnosis, bioinformatics, and NLP

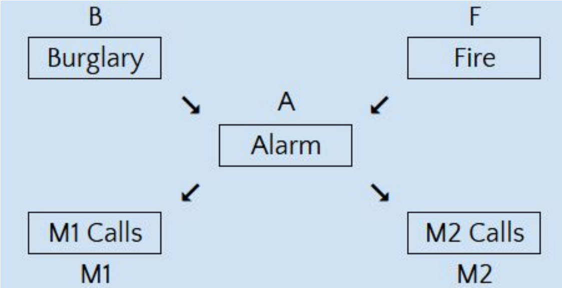
Bayesian Belief Networks

- Construction of a Bayesian network
- Let us assume that the problem can be defined in terms of n random variables.
- Steps for the construction of a Bayesian network:
  - 1. Choose an ordering of variables X1, … ,Xn
  - 2. For i = 1 to n
    - add Xi to the network
    - select parents from X1, … ,Xi−1 such that
    - P (Xi | Parents(Xi)) = P (Xi | X1, … Xi−1)
  - This choice of parents guarantees:
    - P (X1, … ,Xn) = πi =1P (Xi | X1, … , Xi−1) (chain rule)
    - = πi =1P (Xi | Parents(Xi)) (by construction)

$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid Parents(X_i))$$

Bayesian Belief Networks

- Step 1:
- 1.1 Determine what the propositional (random) variables should be.
- 1.2 Determine causal (or another type of influence) relationships and develop the topology of the network.
- Variables are
  - Burglary, Earthquake, Alarm, M1 Calls, M2 Calls
- The Network topology reflecting the “causal” knowledge is as follows:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call
- The resulting Topology of the Bayesian Network is shown here:



Bayesian Belief Networks

- Step 2: To specify a conditional probability table or CPT for each node.
- Burglary B:
  - P (B=T) = 0.001 (‘B’ is TRUE i.e burglary has occurred)
  - P (B=F) = 0.999 (‘B’ is FALSE i.e burglary has not occurred)
- Fire F:
  - P (F=T) = 0.002 (‘F’ is TRUE i.e fire has occurred)
  - P (F=F) = 0.998 (‘F’ is FALSE i.e fire has not occurred)

Alarm A		P(A|B,E)	
B	**F**	**P (A=T)**	**P (A=F)**
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	**0.999**

Alarm A node can be TRUE or FALSE (i.e may have rung or may not have rung).
It has two parent nodes burglary B and fire F which can be TRUE or FALSE
i.e it may have occurred or may not have occurred depending upon diff. conditions.

Person M1	P(M1|A)	
A	**P (M1=T)**	**P (M1=F)**
T	**0.95**	0.05
F	0.05	0.95

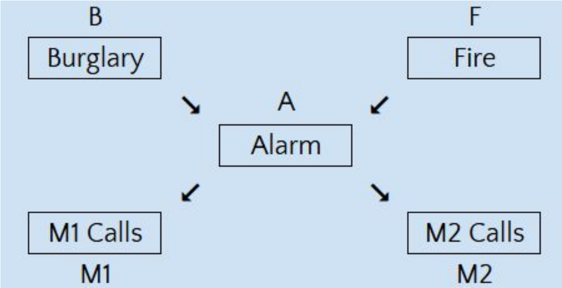
Person M1 node can be TRUE or FALSE (i.e may have called person XYZ or not).
It has a parent node, the alarm A, which can be TRUE or FALSE;
That is it may have rung or may not have rung, upon burglary B or fire F.

Person M2	P(M2|A)	
A	**P (M2=T)**	**P (M2=F)**
T	**0.8**	0.2
F	0.01	0.99

Person M2 node can be TRUE or FALSE (i.e may have called person XYZ or not).
It has a parent node, the alarm A, which can be TRUE or FALSE;
That is it may have rung or may not have rung, upon burglary B or fire F.

Bayesian Belief Networks

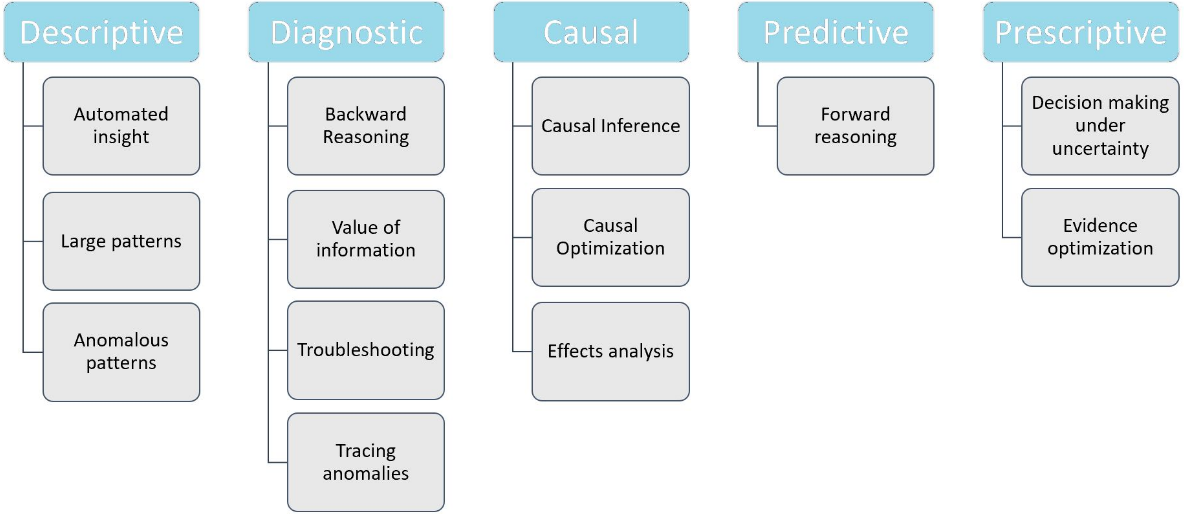
- Example: Calculating Conditional Probability of Events using Bayesian Belief Networks
- Question:
  - Find the probability that M1 is TRUE (M1 has called XYZ), M2 is TRUE (M2 has called XYZ)
  - when the alarm ‘A’ rang, but no burglary ‘B’ and fire ‘F’ has occurred.
  - i.e Find the probability => P (M1, M2, A, ~B, ~F)
  - M1, M2 & A are TRUE events and ~B & ~F are FALSE events
- Solution PDF



$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid Parents(X_i))$$

Bayesian Belief Networks

- Capabilities in terms of the analytics disciplines
  - Descriptive analytics
  - Diagnostic analytics
  - Causal AI.
  - Predictive analytics
  - Prescriptive analytics



Bayesian Belief Networks

- Applications of Bayesian Belief Networks
- Medical Diagnosis:
  - To model relationships between symptoms, diseases, and risk factors.
- Risk Assessment and Decision-Making:
  - To assess risks by modeling dependencies among factors such as
    - market volatility, economic indicators, and credit scores.
- Machine Learning and Data Mining:
  - To discover patterns to predict outcomes, such as fraud detection in banking,
  - by analyzing dependencies between variables.

Bayesian Belief Networks

- **Advantages of Bayesian Belief Networks**
- Handling Uncertainty:
  - To handle uncertainty by updating probabilities dynamically when new evidence becomes available.
- Flexibility and Scalability:
  - To scale to larger networks through incremental updates.
- Incorporating Expert Knowledge:
  - To integrate expert knowledge with data-driven models.

- **Challenges and Limitations**
- Computational Complexity:
  - Computationally expensive as the number of variables and dependencies increases.
- Scaling Issues:
  - Flexible but scaling to very large networks introduces challenges.
- Defining Accurate Priors:
  - Assigning accurate prior probabilities is crucial for the network's reliability.

References

Text books:

1. Ethem Alpaydin, “Introduction to Machine Learning , 4th Edition, The MIT Press, 2020.
2. Peter Harrington, “Machine Learning in Action”, 1st Edition, Dreamtech Press, 2012."
3. Tom Mitchell, “Machine Learning”, 1st Edition, McGraw Hill, 2017.
4. Andreas C, Müller and Sarah Guido, “Introduction to Machine Learning with Python: A Guide for Data Scientists”, 1ed, O'reilly, 2016.
5. Kevin P. Murphy, “Machine Learning: A Probabilistic Perspective”, 1st Edition, MIT Press, 2012."

Reference Books:

6. Aurélien Géron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”, 2nd Edition, Shroff/O'Reilly, 2019.
7. Witten Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal., “Data Mining: Practical machine learning tools and techniques”, 1st Edition, Morgan Kaufmann, 2016.
8. Han, Kamber, “Data Mining Concepts and Techniques”, 3rd Edition, Morgan Kaufmann, 2012.
9. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, “Foundations of Machine Learning”, 1ed, MIT Press, 2012.
10. H. Dunham, “Data Mining: Introductory and Advanced Topics”, 1st Edition, Pearson Education, 2006.

Thank You.