Bayesian theory

Note taker: Daniel Restrepo-Montoya

In classification, Bayes' rule is used to calculate the probabilities of the classes. The main aim is related about how we can make rational decisions to minimize expected risk.

Bayes' theorem provides a way to calculate the probability of a hypothesis based on its prior probability, the probabilities of observing various data given the hypothesis, and the observed data itself.

Probability and inference

Data comes from a process that is not completely known. This lack of knowledge is indicating by modelling the process as a random process.

Bernoulli process: performing multiple experiments.

> *In probability and statistics, a Bernoulli process is a discrete-time stochastic process consisting of a sequence of independent random variables taking values over two symbols. Prosaically, a Bernoulli process is coin flipping, possibly with an unfair coin. A variable in such a sequence may be called a Bernoulli variable. (**wikipedia)***

There is a example, normally speaking if you know you got a fair coin the probability will be .5, but, if you suspect that you got a charge coin you can estimate the probability doing a Bernoulli process.

Classification, probabilistic model

Input/Output

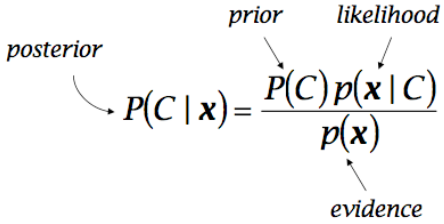
Example related about credit scoring and how to establish 2 classes, basically, High and Low risk costumer. A prediction in a form presented, is homologue to the second one (coin and credit decision).

The focus is basically analyzing past transactions, the bank is planning to identify good and bad customers from their bank accounts. They have the customer yearly income and savings which are fundamental to build the classification model.

Bayes' rule

Explain de posterior, prior (historical data), likelihood (conditional probability, IF-THEN rule), and evidence concepts. Join probability, it has to be exhaustive and excluded.

Bayes' Rule



$$P(C=0)+P(C=1)=1$$
$$p(\boldsymbol{x})=p(\boldsymbol{x} \mid C=1)P(C=1)+p(\boldsymbol{x} \mid C=0)P(C=0)$$
$$p(C=0 \mid \boldsymbol{x})+P(C=1 \mid \boldsymbol{x})=1$$

- *Prior: P(C=1) is call prior probability that C takes the value 1, but it depends in the situation. Is the conditional probability based on the knowledge*
- *Likelihood: Conditional probability that an event belonging to C associated observation value X.*
- *Evidence: is the marginal probability that an observation X is seen, regardless of whether it is a positive or negative example.*
- *Posterior: Combining the prior and what the data tells us using Baye's rule, it is calculated the posterior probability.*

$$\text{Posterior} = \frac{\text{Prior X likelihood}}{\text{Evidence}}$$

The bayes' classifier chooses the class with the highest posterior probability.

Making a decision based on probability
Bayes' rule K>2 Classes

Marginalization concept: Marginalization is the correct Bayesian way of dealing with nuisance variables (recall that for the prediction task, w is nuisance). Again, marginalization is a basic procedure of probability and not Bayesian per se. (Seeger, 2006)

Choose classes, after you calculate de probability of multiple classes, so you will get the highest probability.

Losses and risks
When you do a classification you can include some risk in your decision. Each decision has a cost. Minimizing the risk is part of the key, taking a decision depending in the decision thinking about given or denying a credit.

An action define α_i as the decision to assign the input to class C_1 and λ as the loss incurred for taking action α_i when the input actually belongs to C_k.

The choose action is the one with a minimum risk.
λ: depends on two variables, the first is the action

	λ (s)	C_0 Low risk	C_0 High risk
give	α_0	0	1
deny	α_1	1	0

Losses and risks:
Loss
There are some cases where a decision is not equally good or costly, in some cases it is fundamental take into account potential situations related to the condition.

Choose α_i if $R(\alpha_i|x)=\min R(\alpha_k|x)$

$$1-P(C_i|x)=\min_K 1-P(c_i|x)$$

$$P(c_i|x)=\max_K P(c_i|x)$$

You can get three kinds of actions

$$\lambda \text{ is the cost of rejecting } < \begin{cases} 0 \text{ accepting} \\ 1 \text{ otherwise} \end{cases}$$

Then:
You have to compare the risk of all actions and also must get the action that gives you the minimum risk. On the other hand, you look for the minimum risk and you can decide. At the end you must compare the choose related with the minimum risk and the minimum rejecting and decide:

Reject **Otherwise**

In some cases, wrong decisions (misclassifications) may have a very high cost, and it is required a complex system.

Discriminant functions
The aim is establish a model able to discriminate classes. Basically, classification can also be seen as implementing a set of discriminant functions. There ar some ways to partition the space to choose and option related with the data include in the model. (Figure 1)

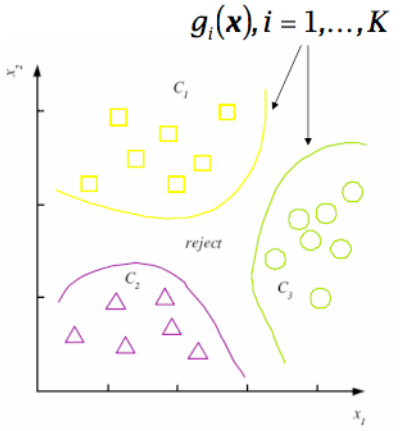


Figure 1: example of decision regions and decision boundaries. (Alpaidyn, 2004)
In the way to discriminate a G or set of G's, it will be given you the discriminated functions.

When there are two classes, we can define a single discriminant:

$$g(x) = g_1(x) - g_2(x)$$

and choose $\begin{cases} C_1 \text{ if } g(x)>0 \\ C_2 \text{ otherwise} \end{cases}$

The two-class learning problem where the positive examples can be taken as C_1 and the negative examples as C_2.

Classification system	
Dichotomizer	K = 2 Classes
Polychotomizer	K ≥ 2 Classes

Utility Theory
It is possible to generalize the problem of the utility theory thinking about the approach related to the expected risk and chose the action that minimizes expected risk. The utility theory is concerned with making rational decisions when we are uncertain about the state.

In the context of classification, decisions correspond to choosing one of the classes, and maximizing the expected utility is equivalent to minimizing expected risk.

Note that maximizing expected utility is just one possibility; one may define other types of rational behaviour, for example, minimizing worst possible loss.

Value of information

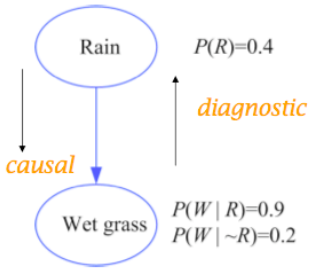
It is relevant to evaluate the quality of the information. So, it is relevant to decide about good and bad information because this is part and is one of the most important related characteristics of the model.

Bayesian networks

This method is also called belief networks or a probabilistic network is the model and is one of the most used methods at the moment.

- Graphical model.
- Representing interaction between variables visually.
- Composed of nodes and arcs between the nodes.
- Each node corresponds to the random variable, X, an has a value corresponding to a probability.
- If there is a direct arc X to Y, means that X has a direct influence on Y.
- Direct acyclic graph (DAG), there are no cycles.
- The nodes and the arcs define the structure of the network.

Causes and Baye's Rule

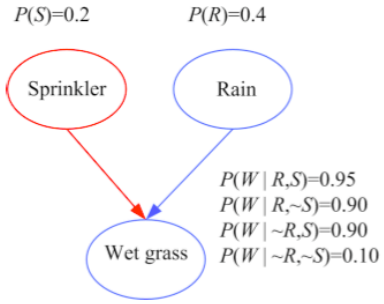


Diagnostic inference: Knowing that the grass is wet, what is the probability that rain is the cause?

$$P(R \mid W) = \frac{P(W \mid R)P(R)}{P(W)}$$
$$= \frac{P(W \mid R)P(R)}{P(W \mid R)P(R) + P(W \mid \sim R)P(\sim R)}$$
$$= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75$$

Bayes' rules allows us to invert the dependencies and have a diagnosis.

Casual vs diagnostic inference



Causal inference: If the sprinkler is on, what is the probability that the grass is wet?

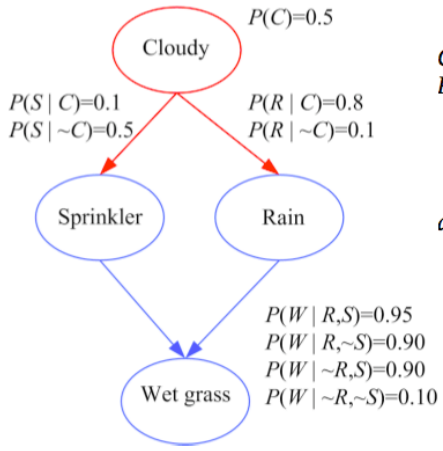
$$P(W|S) = P(W|R,S)\,P(R|S) + P(W|\sim R,S)\,P(\sim R|S)$$
$$= P(W|R,S)\,P(R) + P(W|\sim R,S)\,P(\sim R)$$
$$= 0.95\ 0.4 + 0.9\ 0.6 = 0.92$$

Diagnostic inference: If the grass is wet, what is the probability that the sprinkler is on? $P(S|W) = 0.35 > 0.2\ P(S)$

$P(S|R,W) = 0.21$ *Explaining away: Knowing that it has rained decreases the probability that the sprinkler is on.*

R and S are independent, then, it is possible to calculate the probability that the sprinkler is on, given the grass is wet. Note also that R and S are independent, however we may think that they are actually dependent in the presence of another variable.

Bayesian networks: Causes



Causal inference:

$$P(W|C) = P(W|R,S)\,P(R,S|C) + P(W|\sim R,S)\,P(\sim R,S|C) + P(W|R,\sim S)\,P(R,\sim S|C) + P(W|\sim R,\sim S)\,P(\sim R,\sim S|C)$$

and use the fact that

$$P(R,S|C) = P(R|C)\,P(S|C)$$

Diagnostic: $P(C|W) = ?$

For the model given, C, R and S are independent, this is part of the advantage of Bayesian networks, which explicitly encode independencies and allow breaking down inference into calculation over small groups of variables.

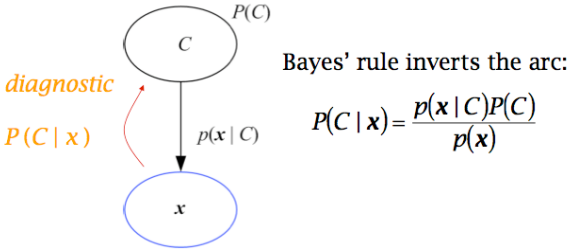
The graphical representation is visual and helps understanding. The network represents conditional independence statements and allows us to break down the problem of representing the joint distribution of many variables into local structures; this eases both analysis and computation.

Bayesian networks, inference.
Belief propagation
It is an efficient algorithm that is used for inference when the network is a tree.

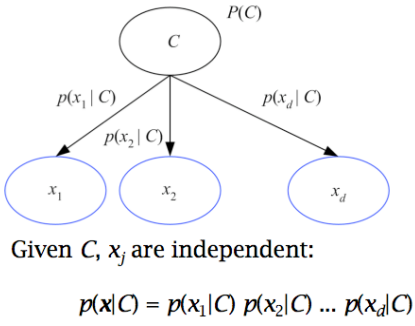
Junction Tree
An algorithm, which converts a given directed acyclic graph to a tree by clustering variables, so, that belief propagation can be done.

One of the best advantage of using Bayesian networks is that we do not need to designate explicitly certain variables as input and certain others as output. The value of any set of variables can be established through evidence and the probabilities of any othes set of variables can be inferred, and the differences between unsupervised and supervised learning becomes blurry.

Bayesian networks, classification



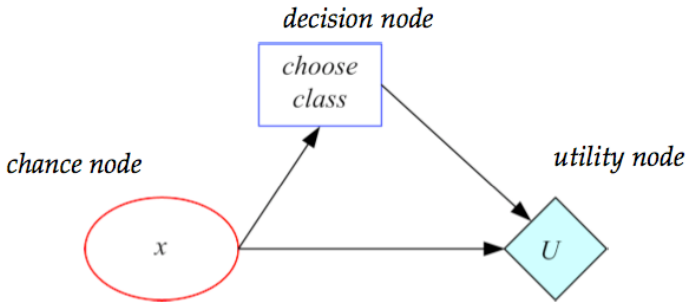
A



B

A: This is a classical Bayesian network for classification. B: Naïve Bayes' classifier is a Bayesian network for classification assuming independent inputs.

Influence diagrams



Influence diagrams are graphical models that allow the generalization of Bayesian networks to include decisions and utilities. And influence diagram contains chance nodes representing random variables that we use in Bayesian networks. A decision node represents a choice of actions. A utility node is where the utility is calculated. Decisions may be based on chance nodes and may affect other chance nodes and the utility node.

Association Rules

An association rule is an implication of the form X → Y.
There are two measures:

Confidence: confidence of association rule X → Y.

$$\text{Confidence } (X \rightarrow Y): \quad P(Y \mid X) = \frac{P(X,Y)}{P(X)} = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$

Conditional probability, P(Y|X), which is what we normally calculate.

Support: support of the association rule X → Y.

$$\text{Support } (X \rightarrow Y): \quad P(X,Y) = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers}\}}$$

Support shows the statistical significance of the rule whereas confidence shows the strength of the rule.

Reference

ALPAYDIN, E. Introduction to Machine Learning. The MIT Press, October 2004, ISBN 0-262-01211-1.

Seeger, M. Bayesian Modelling for Data Analysis and Learning from Data. Max-Planck Institute for Biological Cybernetics. March 18, 2006.These notes provide clarifying remarks and definitions complementing the course Bayesian Modelling for Data Analysis and Learning from Data, to be held at IK 2006. (http://www.kyb.tuebingen.mpg.de/bs/people/seeger/papers/handout.pdf)