CS60050 Machine Learning - Weekly Report

Nisarg Upadhyaya (19CS30031)

Week 6: 15-17 September

1 Topics covered

- Minimum Description Length Principle
- Bayes Optimal Classifier and Gibbs Algorithm
- Discriminant Functions
- Naive Bayes Classifier
- Bayesian Network
- Losses and Risks, Generalisation to Utility Theory
- Association Rules
- Apriori Algorithm

2 Summary

2.1 Minimum Description Length Principle

We interpret the definition of h_{MAP} as follows

$$h_{MAP} = argmax_{h \in H} P(D|h)P(h) = argmin_{h \in H} - log_2 P(D|h) - log_2 P(h)$$

This shows that short hypotheses are preferred as it optimises description length encoding for h and $D|h$

2.2 Bayes Optimal Classifier and Gibbs Algorithm

If we want to find the most probable classification of a new instance given the training data. We consider all $h \in H$ weighted by their posterior. If the new instance can take any value from the set V we have the Bayes Optimal Classifier as

$$argmax_{v \in V} \sum_{h \in H} P(v|h) \cdot P(h|D)$$

The Bayes Optimal Classifier is very expensive to apply. Gibbs Algorithm randomly chooses a hypothesis $h \in H$ according to the posterior probability distribution. It can be shown that the misclassification error is at most twice that of the Bayes Optimal Classifier when the prior has a uniform distribution.

2.3 Discriminant Functions

Bayesian classifiers can be expressed in the framework of classification based on a set of discriminant functions $g_i(x)$ according to the rule that we assign class C_i if $g_i(x) > g_k(x)$ for all k except i.

2.4 Naive Bayes Classifier

It is based on the assumption that the attribute values are conditionally independent given the target value. If the attributes for the new instance are in the set A the naive Bayes Classifier is

$$argmax_{v \in V} P(v) \prod_{a \in A} P(a|v)$$

For discrete variables $P(a|v)$ is estimated by fraction of times the value occurred in class and for a continuous variable we model a Gaussian distribution for the same. We use Laplacian correction in case any of the conditional probabilities become zero. This classifier is easy to implement and gives good results mostly however it assumes independence between attribute values which is not always true in real life.

2.5 Bayesian Network

It models dependencies using a directed acyclic graph. A node X is a random variable with probability $P(X)$. A directed arc from X to Y means X influences Y with probability $P(Y|X)$.

2.5.1 Conditional Independence

We say that X is conditionally independent of Y given Z if the probability distribution governing X is independent of the value of Y given a value for Z, i.e., if $P(X, Y|Z) = P(X|Z) \cdot P(Y|Z)$. This is used to compute probabilities of all possible combinations of other variables, given a value of a leaf node.

2.5.2 Inference

Given any subset X_i, we can calculate the probability distribution of some other subset of X_i by marginalizing over the joint.

2.6 Losses and risks in Bayesian Decision making and generalisation to utility theory

Let the i-th action be a_i which is assigning class C_i to x. Let l_{ik} be the loss due to a_i if x belongs to C_k. We choose a_i which minimizes the risk $R(a_i|x) = \sum_k l_{ik} \cdot P(C_k|x)$. This can be generalised by considering gain U_{ik} instead of loss. We choose a_i which maximises the expected utility $EU(a_i|x) = \sum_k U_{ik} \cdot P(C_k|x)$.

2.7 Association rules

An association rule is an implication $X \rightarrow Y$. The three useful measures are *Confidence* $= P(Y|X)$, *Support* $= P(X, Y)$ and *Lift* $= \frac{P(Y|X)}{P(Y)}$. Confidence indicates strength of the rule, support shows statistical significance and lift shows the dependency between the variables.

2.8 Apriori algorithm

It is used to get association rules with high support and confidence from a database. It also allows generalising associations among more than two variables. The steps of the algorithm are

1. Finding frequent item sets.
2. Converting them to rules with enough confidence.

A point to note here is that $X \rightarrow Y$ indicates association and not causality.

3 Challenging concepts

Conditional Independence and Bayesian Networks

4 Interesting concepts

None

5 Concepts not understood

None

6 Ideas

We learnt to model dependencies in Bayesian Networks. Giving users good recommendations is an essential part of search engines. However this is restricted to the personal preference of a user. If we can use the social graph of people who know each other and then model that as a Bayesian Network it would be interesting to observe if from the knowledge of social connections it is possible to improve search recommendations.