Machine Learning

Assignment 8.2

Submitted By: Ranji Raj

December 17, 2020

a) Naive Bayes and relation to MAP hypothesis

The task of finding the **most probable** hypothesis h, from some space H, given the observed training data D, can be modeled by Bayes theorem. The Bayes theorem is given as:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

where,
$P(h|D)$ is the posterior probability,
$P(D|h)$ is the likelihood,
$P(h)$ is the class prior,
$P(D)$ is the evidence.

In many ML scenarios, the learner considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis $h \in H$ given the observed data D (or at least one of the the maximally probable if there are several).

Any such maximally probable hypothesis is called a **Maximum-A-Posteriori (MAP)** hypothesis. To determine the MAP hypotheses, Bayes theorem can be used to calculate the posterior probability of each candidate hypothesis. Therefore,

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} \quad P(h|D)$$

$$\therefore h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \quad \frac{P(D|h)P(h)}{P(D)}$$

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \quad P(D|h)P(h) \tag{1}$$

Dropping the term $P(\mathrm{D})$ since it is a constant independent of h.

b) "naive" in Naive Bayes and its relevance

- **naive** assumption: The features in the dataset are mutually independent to each other.
- **Relevance**: There is no need to rely on exact duplicates in the training set to make classification.
- If two features are actually dependent, say, "flight delay" and "appearance of fog", then assuming they are independent means you get to *double-count evidence*.

c) Estimation of probabilities in Naive Bayes for classification

For each target value v_j,

$$\hat{P}(v_j) \leftarrow estimateP(v_j)$$

For each attribute value a_i of each attribute A,

$$\hat{P}(a_i|v_j) \leftarrow estimateP(a_i|v_j)$$

Then a new instance x is classified as,

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} \quad \hat{P}(v_j) \prod_{a_i \in X} \hat{P}(a_i|v_j)$$

d) Dealing with missing values and zero frequency problem

Missing values

In general, for dealing with missing values when training Naive Bayes classifier either,

- Omit the records with any missing values,
- Omit only the missing attributes.

Zero Frequency problem

Problem: If an individual class label is missing, then the frequency-based probability estimate will be zero. And we obtain a zero when all the probabilities are multiplied.
Solution: **Laplace correction/smoothing**

$$P(A = c|y) = \frac{m_c^y + 1}{n^y + v}$$

m_c^y: number of training instances with value c and class y,
n^y: number of training instances of class y,
v: number of unique values of A.