School of Computing and Information Systems
The University of Melbourne
COMP90049 Introduction to Machine Learning (Semester 1, 2023)
Week 7

1. Consider a Naive Bayes model trained using the following familiar weather dataset:

| *ID* | *Outl* | *Temp* | *Humi* | *Wind* | PLAY |
|---|---|---|---|---|---|
| A | s | h | n | F | N |
| B | s | h | h | T | N |
| C | o | h | h | F | Y |
| D | r | m | h | F | Y |
| E | r | c | n | F | Y |
| F | r | c | n | T | N |

Suppose that you made additional observations of days and their features. But you don't have the label for the PLAY in these days:

| *ID* | *Outl* | *Temp* | *Humi* | *Wind* | PLAY |
|---|---|---|---|---|---|
| G | o | m | n | T | ? |
| H | s | m | h | F | ? |

How could you incorporate this information into your Naïve Bayes model without manually annotating the labels? If necessary, recompute your model parameters.

For incorporating the new unlabeled instances, we can use Self Training method to increase our training data. The Self-training steps are as follow:

(a) Train the learner on the currently labelled instances.
(b) Use the learner to predict the labels of the unlabeled instances.
(c) Where the learner is very confident, add newly labelled instances to the training set.
(d) Repeat until all instances are labelled, or no new instances can be labelled confidently.

For step (a) let's train our model using the labelled instances. In NB we need the probability of each label (the prior probabilities):

$P(Play = Y) = \frac{1}{2}$ $P(Play = N) = \frac{1}{2}$

And all the conditional probabilities between the labels of class (PLAY) and other attribute values.

$P(Outl = s \mid N) = \frac{2}{3}$ $P(Outl = o \mid N) = 0$ $P(Outl = r \mid N) = \frac{1}{3}$
$P(Outl = s \mid Y) = 0$ $P(Outl = o \mid Y) = \frac{1}{3}$ $P(Outl = r \mid Y) = \frac{2}{3}$
$P(Temp = h \mid N) = \frac{2}{3}$ $P(Temp = m \mid N) = 0$ $P(Temp = c \mid N) = \frac{1}{3}$
$P(Temp = h \mid Y) = \frac{1}{3}$ $P(Temp = m \mid Y) = \frac{1}{3}$ $P(Temp = c \mid Y) = \frac{1}{3}$
$P(Humi = n \mid N) = \frac{2}{3}$ $P(Humi = h \mid N) = \frac{1}{3}$
$P(Humi = n \mid Y) = \frac{1}{3}$ $P(Humi = h \mid Y) = \frac{2}{3}$
$P(Wind = T \mid N) = \frac{2}{3}$ $P(Wind = F \mid N) = \frac{1}{3}$
$P(Wind = T \mid Y) = 0$ $P(Wind = F \mid Y) = 1$

For step (b) using Laplace smoothing method we can find the label for instances G and H and check the confidence of NB model for classifying them.

For G, this will look like:

$$N: \quad P(N) \times P(Outl = o \mid N)\, P(Temp = m \mid N) P(Humi = n \mid N)\, P(Wind = T \mid N)$$
$$= \frac{1}{2} \times \frac{0+1}{3+3} \times \frac{0+1}{3+3} \times \frac{2+1}{3+2} \times \frac{2+1}{3+2} = \frac{1}{2} \times \frac{1}{6} \times \frac{1}{6} \times \frac{3}{5} \times \frac{3}{5} = \mathbf{0.005}$$

$$Y: \quad P(Y) \times P(Outl = o \mid Y)\, P(Temp = m \mid Y) P(Humi = n \mid Y)\, P(Wind = T \mid Y)$$
$$= \frac{1}{2} \times \frac{1+1}{3+3} \times \frac{1+1}{3+3} \times \frac{1+1}{3+2} \times \frac{0+1}{3+2} = \frac{1}{2} \times \frac{2}{6} \times \frac{2}{6} \times \frac{2}{5} \times \frac{1}{5} \cong 0.004$$

Using these results G will be classified as N with probability of 0.005.

For H, we will have:

$$N: \quad P(N) \times P(Outl = s \mid N)\, P(Temp = m \mid N) P(Humi = h \mid N)\, P(Wind = F \mid N)$$
$$= \frac{1}{2} \times \frac{2+1}{3+3} \times \frac{0+1}{3+3} \times \frac{1+1}{3+2} \times \frac{1+1}{3+2} = \frac{1}{2} \times \frac{3}{6} \times \frac{1}{6} \times \frac{2}{5} \times \frac{2}{5} \cong 0.007$$

$$Y: \quad P(Y) \times P(Outl = s \mid Y)\, P(Temp = m \mid Y) P(Humi = h \mid Y)\, P(Wind = F \mid Y)$$
$$= \frac{1}{2} \times \frac{0+1}{3+3} \times \frac{1+1}{3+3} \times \frac{2+1}{3+2} \times \frac{3+1}{3+2} = \frac{1}{2} \times \frac{1}{6} \times \frac{2}{6} \times \frac{3}{5} \times \frac{4}{5} \cong \mathbf{0.013}$$

Using these results H will be classified as Y with probability of 0.013.

In step (c), we are adding the instances that our model is confident about their label to our training set. Let's assume the probability of higher than 0.01 as our confidentiality threshold. According to the results our model is not confident about instance G label ($0.005 \not> 0.01$) but rather confident about the instance H label ($0.013 > 0.01$). So, we add H to our model training data and recompute our model parameters and repeat the steps again. Our new dataset would be as follows:

| *ID* | *Outl* | *Temp* | *Humi* | *Wind* | PLAY |
|---|---|---|---|---|---|
| A | s | h | n | F | N |
| B | s | h | h | T | N |
| C | o | h | h | F | Y |
| D | r | m | h | F | Y |
| E | r | c | n | F | Y |
| F | r | c | n | T | N |
| H | s | m | h | F | Y |

This time our prior probabilities will be:

P(Play = Y) = $\frac{4}{7}$ P(Play = N) = $\frac{3}{7}$

Our conditional probabilities will also change to

P(Outl = s | N) = $\frac{2}{3}$ P(Outl = o | N) = 0 P(Outl = r | N) = $\frac{1}{3}$

P(Outl = s | Y) = $\frac{1}{4}$ P(Outl = o | Y) = $\frac{1}{4}$ P(Outl = r | Y) = $\frac{2}{4}$

P(Temp = h | N) = $\frac{2}{3}$ P(Temp = m | N) = 0 P(Temp = c | N) = $\frac{1}{3}$

P(Temp = h | Y) = $\frac{1}{4}$ P(Temp = m | Y) = $\frac{2}{4}$ P(Temp = c | Y) = $\frac{1}{4}$

P(Humi = n | N) = $\frac{2}{3}$ P(Humi = h | N) = $\frac{1}{3}$

P(Humi = n | Y) = $\frac{1}{4}$ P(Humi = h | Y) = $\frac{3}{4}$

P(Wind = T | N) = $\frac{2}{3}$ P(Wind = F | N) = $\frac{1}{3}$

P(Wind = T | Y) = 0 P(Wind = F | Y) = 1

Using these new parameters, we will recalculate the probability of labels for instance G.

$$N: \quad P(N) \times P(Outl = o \mid N)\, P(Temp = m \mid N) P(Humi = n \mid N)\, P(Wind = T \mid N)$$
$$= \frac{3}{7} \times \frac{0+1}{3+3} \times \frac{0+1}{3+3} \times \frac{2+1}{3+2} \times \frac{2+1}{3+2} = \frac{3}{7} \times \frac{1}{6} \times \frac{1}{6} \times \frac{3}{5} \times \frac{3}{5} \cong \mathbf{0.0042}$$

$$Y: \quad P(Y) \times P(Outl = o \mid Y)\, P(Temp = m \mid Y) P(Humi = n \mid Y)\, P(Wind = T \mid Y)$$
$$= \frac{4}{7} \times \frac{1+1}{4+3} \times \frac{2+1}{4+3} \times \frac{1+1}{4+2} \times \frac{0+1}{4+2} = \frac{4}{7} \times \frac{2}{7} \times \frac{3}{7} \times \frac{2}{6} \times \frac{1}{6} \cong 0.0038$$

Instance G will still be classified as N with probability of 0.0042 but it doesn't pass our confidentiality threshold (0.01) and therefor our self-training algorithm stops in this stage.

2. What is the main assumption of self-training? What is the main assumption of Active Learning?

The main assumption of self-training is that similar instances are likely to have the same label. That's why we find the most similar instances to our labelled data and if the similarity is certain enough (better than our defined threshold) we will give them the same label and add them to the cluster.

As for the Active learning main assumption, we assume that instances near class boundaries are the most informative for learning. That's why we find the instances that we are most uncertain about (using different methods such as QBC or Uncertainty Sampling) and send them to the human annotator (Oracle). The assumption here is that having these instances correct label would be most beneficial for forming the correct clusters.

3. (a) Describe the rationale and key principles behind the Query-by-Committee algorithm. (b) Use QBC to determine the instance that our active learner would select first in the following scenario.

| classifier | Instance 1 | | | Instance 2 | | | Instance 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_3$ | $y_1$ | $y_2$ | $y_3$ | $y_1$ | $y_2$ | $y_3$ |
| $C_1$ | 0.2 | 0.7 | 0.1 | 0.2 | 0.7 | 0.1 | 0.6 | 0.1 | 0.3 |
| $C_2$ | 0.1 | 0.3 | 0.6 | 0.2 | 0.6 | 0.2 | 0.21 | 0.21 | 0.58 |
| $C_3$ | 0.8 | 0.1 | 0.1 | 0.05 | 0.9 | 0.05 | 0.75 | 0.01 | 0.24 |
| $C_4$ | 0.3 | 0.5 | 0.2 | 0.1 | 0.8 | 0.1 | 0.1 | 0.28 | 0.62 |

(a) The goal of active learning is to achieve high accuracy with as few queries from the oracle as possible, by selecting the most informative or uncertain examples to query. One of the strategies for query sampling is query-by-committee (QBC), where a suite of classifiers is trained over a fixed training set, and the instance that results in the highest disagreement amongst the classifiers, is selected for querying. The idea is that the models will have different strengths and weaknesses, and by combining their predictions, the algorithm can leverage their collective intelligence. The rationale behind Query-by-Committee is that diversity in the committee is crucial for achieving better accuracy. QBC uses the equation below, which captures vote entropy, to determine the instance that our active learner would select first.

$$x_{VE}^{*} = \underset{x}{\operatorname{argmax}}\left(-\sum_{y_i} \frac{V(y_i)}{C} log_2 \frac{V(y_i)}{C}\right)$$

In this equation $y_i$, $V(y_i), and\ C$ are respectively the possible labels, the number of "votes" that a label receives from the classifiers, and the total number of classifiers.

(b) In this scenario, we have four classifiers ($C_1$, $C_2$, $C_3$, and $C_4$) that predict labels for three instances (instance 1 to 3), where each instance can be labeled as $y_1$, $y_2$, or $y_3$. Each classifier generates a probability for each label, and the label with the highest probability is chosen as the label predicted by that classifier for each instance. To facilitate further analysis, we can transform the given table into a

more convenient format as shown below. For each instance, we then calculate the total number of votes received by each label class:

| classifier | Instance 1 Votes | | | Instance 2 | | | Instance 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_3$ | $y_1$ | $y_2$ | $y_3$ | $y_1$ | $y_2$ | $y_3$ |
| $C_1$ | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| $C_2$ | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| $C_3$ | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| $C_4$ | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| | V(1)=1 | V(2)=2 | V(3)=1 | V(1)=0 | V(2)=4 | V(3)=0 | V(1)=2 | V(2)=0 | V(3)=2 |

We have 4 classifiers in total, and after placing the vote values in the vote entropy, we get the following for each instance:

Instance 1: $H = -\left(\frac{1}{4}log_2\frac{1}{4} + \frac{2}{4}log_2\frac{2}{4} + \frac{1}{4}log_2\frac{1}{4}\right) = 1.5$

Instance 2: $H = -(1log_2 1) = 0$

Instance 3: $H = -\left(\frac{2}{4}log_2\frac{2}{4} + \frac{2}{4}log_2\frac{2}{4}\right) = 1$

Using the QBC the instance that we select is instance 1, for which the classifiers have the highest disagreement. This sample is most difficult to classify and may lie on the boundary between the three classes, therefore by querying this instance, we might learn more about the data space.