# COMPSCI 762 2022 S1 Week 7 Questions – Bayesian Learning

Luke Chang

April 3, 2022

## Question 1

You come to Fiji for a holiday. 10 days later, you realise the weather forecast here isn't very accurate. Based on the information you gettered so far and today's weather report, you want to know "Will it rain this afternoon?"

| **Day** | **Outlook** ($O$) | **Temperature** ($T$) | **Humidity** ($H$) | **Wind** ($W$) | **Rain** ($R$) |
|---|---|---|---|---|---|
| 1 | Sunny | Hot | High | Weak | True |
| 2 | Sunny | Hot | High | Strong | False |
| 3 | Overcast | Hot | High | Weak | True |
| 4 | Rain | Mild | High | Weak | True |
| 5 | Rain | Cool | Normal | Weak | True |
| 6 | Rain | Cool | Normal | Strong | False |
| 7 | Overcast | Cool | Normal | Strong | False |
| 8 | Overcast | Mild | High | Strong | True |
| 9 | Sunny | Cool | Normal | Weak | False |
| 10 | Rain | Mild | Normal | Weak | False |
| 11 | Sunny | Mild | Normal | Strong | ? |

Answer the question using a Multinomial Naive Bayes Classifier. You should compute it by hand, and explain how you formulate the task and the steps for getting the solution.

## Question 2

### 2.1 Example of Bag of Words

*Bag of Words* (BoW) represents ad document as an unordered collection of words and their frequencies. Since there is no positional information, probabilities can be learned with less data.

Let's consider the example as the following:

**John likes to watch movies. "The Watch" is his favorite movie.**

If we split the sentence above by space, we have:

```
{ "John": 1, "likes": 1, "to": 1, "watch": 1, "movies.": 1, "‘‘The": 1,
"Watch’’": 1, "is": 1, "his": 1, "favorite": 1, "movie.": 1 }
```

Can you identify the problem? We haven't consider the English grammar. For example:

- Punctuation, e.g. *"movies."* should become *"movies"*.

- Plural, e.g., *"movies"* and *"movie"* are the same word.
- Verbs in third-person singular, e.g., *"likes"* and *"like"* are the same word.
- Capital letters, e.g., *"Watch"* and *"watch"* are the same word.
- Stopping words, e.g., *"a"* and *"the"* provide no additional information for predicting the label.

Once we've taken care of the issues above, we have:

```
{ "john":  1, "like":  1, "watch":  2, "movie":  2, "his":  1, "favorite":  1 }
```

### 2.2 Naive Bayes on BoW Representation

Given 6 training observations in the format (sentence, label), answer the following questions by hand:

| Index | Sentence | Label |
|---|---|---|
| 1 | Auckland is in the North Island of New Zealand. | Auckland |
| 2 | Auckland has a large population. | Auckland |
| 3 | It is in the Auckland Region, governed by Auckland Council. | Auckland |
| 4 | Dunedin is in the South Island of New Zealand. | Dunedin |
| 5 | Dunedin has the sixth highest population in New Zealand. | Dunedin |
| 6 | It was the largest city in New Zealand until the formation of the Auckland Council. | Dunedin |

1. What does the training data look like after transforming to lower case, and removing the stop words and punctuations?
2. What is the vocabulary in this example (the overall set of words)?
3. What are the priors for Auckland (A) and Dunedin (D)? ($p(A)$ and $p(D)$)
4. What is the likelihood of seeing the word "Auckland" in a sentence labeled Auckland (A)? Dunedin (D)? ($p("Auckland"|A)$ and $p("Auckland"|D)$)
5. What is the likelihood of seeing the word "North" in a sentence labeled Auckland (A)? Dunedin (D)? ($p("North"|A)$ and $p("North"|D)$)
6. Since we multiply probabilities, and 0 encountered means the final probability will also be 0. Therefore, **Laplace smoothing** is required when applying Naive Bayes on a BoW representation. Using smoothing with constant 1, what is the likelihood of seeing the word "North" in a sentence labeled Auckland (A)? Dunedin (D)? ($p("North"|A)$ and $p("North"|D)$)
7. What label would Naive Bayes predict for the sentence: **"Auckland is in the north of the North Island"**?

**Note:** You should solve both questions without computer aid. However, you may code in Python to check your solutions, but it is not part of the question.