Tutorial 7

Tutorial on Naive Bayes

Ben Halstead, Luke Chang

The University of Auckland

May 2021

Topics

1. Multinomial Naive Bayes Classifier

2. Text Preprocessing

3. Extensions

Example 1 - Multinomial Naive Bayes Classifier

Example

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	?

Example 1 - Formulate the Problem

- Attribute: *Outlook* (O), *Temperature* (T), *Humidity* (H), *Wind* (W).
- Output: *Rain* (R) - Binary classification problem
- How can we formulate this task?

Example 1 - Formulate the Problem

- Attribute: *Outlook* (O), *Temperature* (T), *Humidity* (H), *Wind* (W).
- Output: *Rain* (R) - Binary classification problem
- How can we formulate this task?
  - The probability of a given output $r \in \{\text{True}, \text{False}\}$ is:

$$P(R = r | O, T, H, W)$$

Example 1 - Formulate the Problem

- Attribute: *Outlook* (O), *Temperature* (T), *Humidity* (H), *Wind* (W).
- Output: *Rain* (R) - Binary classification problem
- How can we formulate this task?
  - The probability of a given output $r \in \{\text{True}, \text{False}\}$ is:

$$P(R = r | O, T, H, W)$$

  - We want to predict the label with the highest probability.

$$R = \underset{r \in \{\text{T,F}\}}{\operatorname{argmax}} P(R = r | O, T, H, W)$$

Example 1 - Formulate the Problem

- Bayes Theorem:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- Terminology – Prior: $P(Y)$, Likelihood: $P(X|Y)$, Posterior: $P(Y|X)$, Marginal Probability: $P(X)$.
- How to rewrite this expression using Bayes Theorem?

Example 1 - Formulate the Problem

- Bayes Theorem:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- Terminology – Prior: $P(Y)$, Likelihood: $P(X|Y)$, Posterior: $P(Y|X)$, Marginal Probability: $P(X)$.
- How to rewrite this expression using Bayes Theorem?

$$R = \operatorname*{argmax}_{r\in\{\mathsf{T},\mathsf{F}\}} P(R = r|O,T,H,W)$$

$$R = \operatorname*{argmax}_{r\in\{\mathsf{T},\mathsf{F}\}} \frac{P(O,T,H,W|R = r)P(R = r)}{P(O,T,H,W)}$$

- How can we simplify this problem?

Example 1 - Formulate the Problem

- Bayes Theorem:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- Terminology – Prior: $P(Y)$, Likelihood: $P(X|Y)$, Posterior: $P(Y|X)$, Marginal Probability: $P(X)$.
- How to rewrite this expression using Bayes Theorem?

$$R = \underset{r\in\{\text{T,F}\}}{\operatorname{argmax}} P(R=r|O,T,H,W)$$

$$R = \underset{r\in\{\text{T,F}\}}{\operatorname{argmax}} \frac{P(O,T,H,W|R=r)P(R=r)}{P(O,T,H,W)}$$

- How can we simplify this problem?
  $P(Y|X)$ is proportional to: $P(Y|X) \propto P(X|Y)P(Y)$

$$R = \underset{r\in\{\text{T,F}\}}{\operatorname{argmax}} P(O,T,H,W|R=r)P(R=r)$$

- The marginal Probability $P(O,T,H,W)$ is omitted. If we want to know the probability, we can normalise all possible outcomes.

Example 1 - Calculate the Prior $P(R)$

- 10 observations
- 5 days was raining.
- 5 days wasn't.

Example 1 - Calculate the Prior $P(R)$

- 10 observations
- 5 days was raining.
- 5 days wasn't.
- $P(R = \text{True}) = \frac{5}{10} = 0.5$
- $P(R = \text{False}) = \frac{5}{10} = 0.5$

Example 1 - Calculate the Prior $P(R)$

- 10 observations
- 5 days was raining.
- 5 days wasn't.
- $P(R = \text{True}) = \frac{5}{10} = 0.5$
- $P(R = \text{False}) = \frac{5}{10} = 0.5$

Example

Gambler's fallacy: Toss a coin, I see 20 head showed up in a row. The next time it must land on tail, since the probability of a coin landing 21 times is too low, 0.5^{21}.

Example 1 - Calculate the Prior $P(R)$

- 10 observations
- 5 days was raining.
- 5 days wasn't.
- $P(R = \text{True}) = \frac{5}{10} = 0.5$
- $P(R = \text{False}) = \frac{5}{10} = 0.5$

Example

Gambler's fallacy: Toss a coin, I see 20 head showed up in a row. The next time it must land on tail, since the probability of a coin landing 21 times is too low, 0.5^{21}.

Problem:

- Frequentist Statistics: Each tail is independent.
- $P(Y) \neq P(Y|X)$
- Bayes Theorem: If the same coin is used, the more evidence you collect, the more likely the coin is loaded.

Example 1 - Calculate the Likelihood

- The first part of the formula, $P(O,T,H,W|R=r)$, is the *likelihood*.
- It describes how *likely* an event occurs, given the outcome.
- How do we calculate the *likelihood*?

Example 1 - Calculate the Likelihood

- The first part of the formula, $P(O,T,H,W|R=r)$, is the *likelihood*.
- It describes how *likely* an event occurs, given the outcome.
- How do we calculate the *likelihood*?
- We could try to calculate $P(O,T,H,W|R=r)$ directly. What is the problem with this approach?
    - We need to compute all possible combinations.
    - We have: $3\times 3\times 2\times 2 = 36$ different combinations of O,T,H,W *per label*.
    - We only have 10 observations – not enough to calculate probabilities for all combinations.
- Naive Bayes makes the assumption that all attributes are independent: allowing us to calculate the likelihood as:

$$P(O,T,H,W|R=r) = P(O|R=r)p(T|R=r)p(H|R=r)p(W|R=r)$$

Example 1 - Calculate the Likelihood

Day	**Outlook** (O)	**Temperature** (T)	**Humidity** (H)	**Wind** (W)	**Rain** (R)
1	Sunny	Hot	High	Weak	True
3	Overcast	Hot	High	Weak	True
4	Rain	Mild	High	Weak	True
5	Rain	Cool	Normal	Weak	True
8	Overcast	Mild	High	Strong	True

$$\sum P(X = x_i | R = r) = 1$$

- $P(O = \text{Sunny} | R = \text{T}) = \frac{1}{5}$
- $P(O = \text{Overcast} | R = \text{T}) = \frac{2}{5}$
- $P(O = \text{Rain} | R = \text{T}) = \frac{2}{5}$

- $P(T = \text{Hot} | R = \text{T}) = \frac{2}{5}$
- $P(T = \text{Mild} | R = \text{T}) = \frac{2}{5}$
- $P(T = \text{Cool} | R = \text{T}) = \frac{1}{5}$

- $P(H = \text{High} | R = \text{T}) = \frac{4}{5}$
- $P(H = \text{Normal} | R = \text{T}) = \frac{1}{5}$

- $P(T = \text{Strong} | R = \text{T}) = \frac{1}{5}$
- $P(T = \text{Weak} | R = \text{T}) = \frac{4}{5}$

Example 1 - Calculate the Likelihood

Day	**Outlook** (O)	**Temperature** (T)	**Humidity** (H)	**Wind** (W)	**Rain** (R)
2	Sunny	Hot	High	Strong	False
6	Rain	Cool	Normal	Strong	False
7	Overcast	Cool	Normal	Strong	False
9	Sunny	Cool	Normal	Weak	False
10	Rain	Mild	Normal	Weak	False

- $P(O = \text{Sunny}|R = \text{F}) = \frac{2}{5}$
- $P(O = \text{Overcast}|R = \text{F}) = \frac{1}{5}$
- $P(O = \text{Rain}|R = \text{F}) = \frac{2}{5}$

- $P(T = \text{Hot}|R = \text{F}) = \frac{1}{5}$
- $P(T = \text{Mild}|R = \text{F}) = \frac{1}{5}$
- $P(T = \text{Cool}|R = \text{F}) = \frac{3}{5}$

- $P(H = \text{High}|R = \text{F}) = \frac{1}{5}$
- $P(H = \text{Normal}|R = \text{F}) = \frac{4}{5}$

- $P(T = \text{Strong}|R = \text{F}) = \frac{3}{5}$
- $P(T = \text{Weak}|R = \text{F}) = \frac{2}{5}$

Example 1 - Calculate the Posterior

Day	**Outlook** (O)	**Temperature** (T)	**Humidity** (H)	**Wind** (W)	**Rain** (R)
11	Sunny	Mild	Normal	Strong	?

Now we have:

$$R = \underset{r \in \{\mathsf{T},\mathsf{F}\}}{\operatorname{argmax}} P(O|R=r)P(T|R=r)P(H|R=r)P(W|R=r)P(R=r)$$

$$\begin{aligned} P(R=\mathsf{T}|O,T,H,W) &\propto P(O=\mathsf{S}|R=\mathsf{T})P(T=\mathsf{M}|R=\mathsf{T})P(H=\mathsf{N}|R=\mathsf{T})P(W=\mathsf{S}|R=\mathsf{T})P(R=\mathsf{T}) \\ &\propto \frac{1}{5}\cdot\frac{2}{5}\cdot\frac{1}{5}\cdot\frac{1}{5} = 0.0032 \end{aligned}$$

$$\begin{aligned} P(R=\mathsf{F}|O,T,H,W) &\propto P(O=\mathsf{S}|R=\mathsf{F})P(T=\mathsf{M}|R=\mathsf{F})P(H=\mathsf{N}|R=\mathsf{F})P(W=\mathsf{S}|R=\mathsf{F})P(R=\mathsf{F}) \\ &\propto \frac{2}{5}\cdot\frac{1}{5}\cdot\frac{4}{5}\cdot\frac{3}{5} = 0.0384 \end{aligned}$$

Example 1 - How Likely?

Day	Outlook (O)	Temperature (T)	Humidity (H)	Wind (W)	Rain (R)
11	Sunny	Mild	Normal	Strong	?

$$P(R = \mathsf{T}|O,T,H,W) = \frac{0.0032}{P(O,T,H,W)}$$

$$P(R = \mathsf{F}|O,T,H,W) = \frac{0.0384}{P(O,T,H,W)}$$

$$P(R = \mathsf{T}|O,T,H,W) = \frac{0.0032}{0.0032 + 0.0384} = 0.08$$

$$P(R = \mathsf{F}|O,T,H,W) = \frac{0.0384}{0.0032 + 0.0384} = 0.92$$

Given the information, there is 92% to not rain.

Text Representation - Bag of Words (BoW)

- *Bag of Words* represents a document as an unordered collection of words and their frequencies.
- Since there is no positional information, probabilities can be learned with less data.
- We can represent the set of words in the document, or we can represent the count of words in the document.

BoW Example

- What is the BoW representation of the sentence:
  in 1992 we started assembling an ordered library of cosmid clones from chromosome xiv of the yeast

BoW Example

- What is the BoW representation of the sentence:
  in 1992 we started assembling an ordered library of cosmid clones from chromosome xiv of the yeast
- { ”in”: 1, ”1992”: 1, ”we”: 1, ”started”: 1, ”assembling”: 1, ”an”: 1, ”ordered”: 1, ”library”: 1, ”of”: 2, ”cosmid”: 1, ”clones”: 1, ”from”: 1, ”chromosome”: 1, ”xiv”: 1, ”the”: 1, ”yeast” }

BoW Example

- What is the BoW representation of the sentence:
  John likes to watch movies. “The Watch” is his favorite movie.
- { ”John”: 1, ”likes”: 1, ”to”: 1, ”watch”: 1, ”movies.”: 1, ””“The”: 1, ”Watch””: 1, ”is”: 1, ”his”: 1, ”favorite”: 1, ”movie.”: 1 }
- Is there any problem?

BoW Example

- What is the BoW representation of the sentence:
  John likes to watch movies. “The Watch” is his favorite movie.
- { ”John”: 1, ”likes”: 1, ”to”: 1, ”watch”: 1, ”movies.”: 1, ”“The”: 1, ”Watch””: 1, ”is”: 1, ”his”: 1, ”favorite”: 1, ”movie.”: 1 }
- Is there any problem?
  - We have to consider English grammar.
  - Punctuation
  - Plural
  - Verbs in third-person singular
  - Capital letters

BoW Example

- What is the BoW representation of the sentence:
  John likes to watch movies. “The Watch” is his favorite movie.
- { ”John”: 1, ”likes”: 1, ”to”: 1, ”watch”: 1, ”movies.”: 1, ””The”: 1, ”Watch””: 1, ”is”: 1, ”his”: 1, ”favorite”: 1, ”movie.”: 1 }
- Is there any problem?
  - We have to consider English grammar.
  - Punctuation
  - Plural
  - Verbs in third-person singular
  - Capital letters
- { ”john”: 1, ”like”: 1, ”to”: 1, ”watch”: 2, ”movie”: 2, ””the”: 1, ”is”: 1, ”his”: 1, ”favorite”: 1 }
- Can we do better?

Stop words

- Does every word in a BoW representation help select a label?
- Common words like ‘a’ or ‘the’ often do not affect our selection of the label at all.
- These common words with little meaning are known as *stop words*.
- It is common to simply remove all stop words from the training data, so we can focus on important words which can actually separate classes.

Stop words

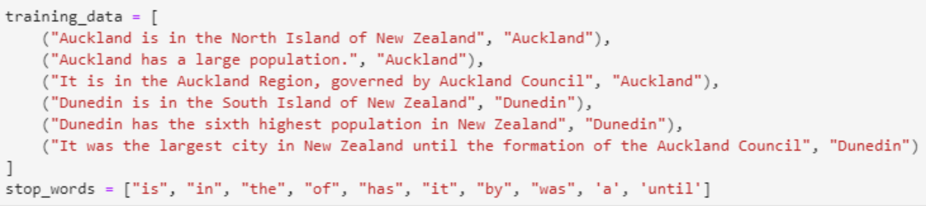
- Does every word in a BoW representation help select a label?
- Common words like ‘a’ or ‘the’ often do not affect our selection of the label at all.
- These common words with little meaning are known as *stop words*.
- It is common to simply remove all stop words from the training data, so we can focus on important words which can actually separate classes.

Given: **John likes to watch movies. “The Watch” is his favorite movie.**
{ ”john”: 1, ”like”: 1, ”watch”: 2, ”movie”: 2, ”his”: 1, ”favorite”: 1 }

BoW Likelihood

- To implement Naive Bayes, we need to learn likelihood values from the training data. How do we do this with a BoW representation?
- Our features are a collection of words w_i each appearing f_i times.
- For each word in this collection, the likelihood under class v_j, $p(w_i|v_j)$, is the probability of selecting w_i if you picked a random word out of all words in the training set labeled with class v_j.
- $p(w_i|v_j) = \frac{n_{ij}}{n_j}$, the number of times word w_i appears in the training set in sentences labeled with class v_j, divided by the number of all words in sentences labeled with class v_j.

BoW Example



```
training_data = [
    ("Auckland is in the North Island of New Zealand", "Auckland"),
    ("Auckland has a large population.", "Auckland"),
    ("It is in the Auckland Region, governed by Auckland Council", "Auckland"),
    ("Dunedin is in the South Island of New Zealand", "Dunedin"),
    ("Dunedin has the sixth highest population in New Zealand", "Dunedin"),
    ("It was the largest city in New Zealand until the formation of the Auckland Council", "Dunedin")
]
stop_words = ["is", "in", "the", "of", "has", "it", "by", "was", 'a', 'until']
```

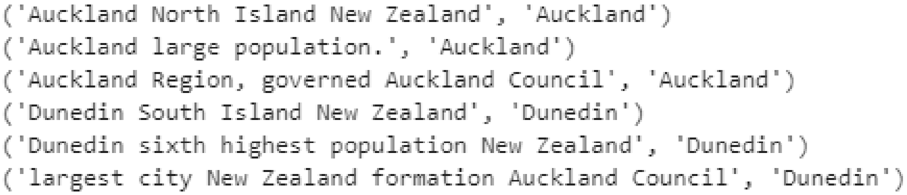
Figure: Six training observations in the format (sentence, label).

BoW Example

What does the training data look like after the stop words have been removed?

BoW Example

What does the training data look like after the stop words have been removed?



```
('Auckland North Island New Zealand', 'Auckland')
('Auckland large population.', 'Auckland')
('Auckland Region, governed Auckland Council', 'Auckland')
('Dunedin South Island New Zealand', 'Dunedin')
('Dunedin sixth highest population New Zealand', 'Dunedin')
('largest city New Zealand formation Auckland Council', 'Dunedin')
```

Figure: Training data with stop words removed.

BoW Example

What is the vocabulary in this example (the overall set of words)?

{ region, governed, highest, auckland, sixth, population, city, dunedin, formation, large, population, largest, south, council, island, north, zealand, new }

New Zealand and **new population** both add the counter for *new* by 1, but do they mean the same?

BoW Example

What are the priors for Auckland (A) and Dunedin (D)? ($p(A)$ and $p(D)$)

BoW Example

What are the priors for Auckland (A) and Dunedin (D)? ($p(A)$ and $p(D)$)

- 6 training sentences.
- 3 labeled A, so $p(A) = \frac{3}{6} = 0.5$
- 3 labeled D, so $p(D) = \frac{3}{6} = 0.5$

BoW Example

What is the likelihood of seeing the word "Auckland" in a sentence labeled Auckland (A)? Dunedin (D)? ($p(\text{"}Auckland\text{"}|A)$ and $p(\text{"}Auckland\text{"}|D)$)

BoW Example

What is the likelihood of seeing the word "Auckland" in a sentence labeled Auckland (A)? Dunedin (D)? ($p(\text{"}Auckland\text{"}|A)$ and $p(\text{"}Auckland\text{"}|D)$)

- The training data has 13 words in total labeled with "A".
- 4 words labeled "A" are the word "Auckland".
- $p(\text{"}Auckland\text{"}|A) = \frac{4}{13}$

BoW Example

What is the likelihood of seeing the word "Auckland" in a sentence labeled Auckland (A)? Dunedin (D)? ($p(\text{"}Auckland\text{"}|A)$ and $p(\text{"}Auckland\text{"}|D)$)

- The training data has 13 words in total labeled with "A".
- 4 words labeled "A" are the word "Auckland".
- $p(\text{"}Auckland\text{"}|A) = \frac{4}{13}$
- The training data has 18 words in total labeled with "D".
- 1 word labeled "D" is the word "Auckland".
- $p(\text{"}Auckland\text{"}|D) = \frac{1}{18}$

BoW Problem

What is the likelihood of seeing the word ”North” in a sentence labeled Auckland (A)? Dunedin (D)? ($p(\text{''}North\text{''}|A)$ and $p(\text{''}North\text{''}|D)$)

BoW Problem

What is the likelihood of seeing the word "North" in a sentence labeled Auckland (A)? Dunedin (D)? ($p(\text{"}North\text{"}|A)$ and $p(\text{"}North\text{"}|D)$)

- The training data has 13 words in total labeled with "A".
- 1 words labeled "A" are the word "North".
- $p(\text{"}North\text{"}|A) = \frac{1}{13}$
- The training data has 18 words in total labeled with "D".
- 0 words labeled "D" are the word "North".
- $p(\text{"}North\text{"}|D) = \frac{0}{18} = 0$

What happens when we try to calculate probabilities for a sentence with the word "North" in it?

0 Probabilities

- Since we multiply probabilities, and 0 encountered means the final probability will also be 0.
- The sentence "North Dunedin" would have get 0 probability for both Auckland and Dunedin!
- We can fix this with *smoothing*, adding a small constant to the count of each word, label pair. This means we get no 0 likelihoods.

BoW Example

Using smoothing with constant 1, what is the likelihood of seeing the word "Auckland" in a sentence labeled Auckland (A)? Dunedin (D)? ($p(\text{"}Auckland\text{"}|A)$ and $p(\text{"}Auckland\text{"}|D)$)

BoW Example

Using smoothing with constant 1, what is the likelihood of seeing the word "Auckland" in a sentence labeled Auckland (A)? Dunedin (D)? ($p(\text{"}Auckland\text{"}|A)$ and $p(\text{"}Auckland\text{"}|D)$)

- The training data has 13 words in total labeled with "A".
- Smoothing adds 1 to the total count per word in the vocab, so adds 18. Our total count is $13 + 18 = 31$.
- 4 words labeled "A" are the word "Auckland", + 1 from smoothing = 5.
- $p(\text{"}Auckland\text{"}|A) = \frac{5}{31}$

BoW Example

Using smoothing with constant 1, what is the likelihood of seeing the word "Auckland" in a sentence labeled Auckland (A)? Dunedin (D)? ($p(\text{"}Auckland\text{"}|A)$ and $p(\text{"}Auckland\text{"}|D)$)

- The training data has 13 words in total labeled with "A".
- Smoothing adds 1 to the total count per word in the vocab, so adds 18. Our total count is 13 + 18 = 31.
- 4 words labeled "A" are the word "Auckland", + 1 from smoothing = 5.
- $p(\text{"}Auckland\text{"}|A) = \frac{5}{31}$
- The training data has 18 words in total labeled with "D". We add 18 from smoothing, so the total count = 36.
- 1 word labeled "D" is the word "Auckland", + 1 from smoothing = 2.
- $p(\text{"}Auckland\text{"}|D) = \frac{2}{36}$

BoW Example

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 $p("North"|D)$)

BoW Example

Using smoothing with constant 1, what is the likelihood of seeing the word "North" in a sentence labeled Auckland (A)? Dunedin (D)? ($p(\text{"}North\text{"}|A)$ and $p(\text{"}North\text{"}|D)$)

- The training data has 13 words in total labeled with "A".
- Smoothing adds 1 to the total count per word in the vocab, so adds 18. Our total count is $13 + 18 = 31$.
- 1 words labeled "A" are the word "North", $+$ 1 from smoothing $=$ 2.
- $p(\text{"}North\text{"}|A) = \frac{2}{31}$

BoW Example

Using smoothing with constant 1, what is the likelihood of seeing the word "North" in a sentence labeled Auckland (A)? Dunedin (D)? ($p(\text{"}North\text{"}|A)$ and $p(\text{"}North\text{"}|D)$)

- The training data has 13 words in total labeled with "A".
- Smoothing adds 1 to the total count per word in the vocab, so adds 18. Our total count is $13 + 18 = 31$.
- 1 words labeled "A" are the word "North", + 1 from smoothing = 2.
- $p(\text{"}North\text{"}|A) = \frac{2}{31}$
- The training data has 18 words in total labeled with "D". We add 18 from smoothing, so the total count = 36.
- 0 words labeled "D" are the word "North", + 1 from smoothing = 1.
- $p(\text{"}North\text{"}|D) = \frac{1}{36}$

Now we can still calculate probabilities, even with words not seen in the training data for a label.

BoW Example

Using smoothing, what label would Naive Bayes predict for the sentence: ”Auckland is in the north of the North Island”?

BoW Example

Using smoothing, what label would Naive Bayes predict for the sentence: ”Auckland is in the north of the North Island”?

- The BoW representation of the sentence is {”Auckland”: 1, ”North”: 2, ”Island”: 1}
- For label v, the posterior probability is given by:
- $p(\text{''}Auckland\text{''}|v) \times p(\text{''}North\text{''}|v) \times p(\text{''}North\text{''}|v) \times p(\text{''}Island\text{''}|v) \times p(v)$
- $= p(\text{''}Auckland\text{''}|v)^1 \times p(\text{''}North\text{''}|v)^2 \times p(\text{''}Island\text{''}|v)^1 \times p(v)$

BoW Example

Using smoothing, what label would Naive Bayes predict for the sentence: "Auckland is in the north of the North Island"?

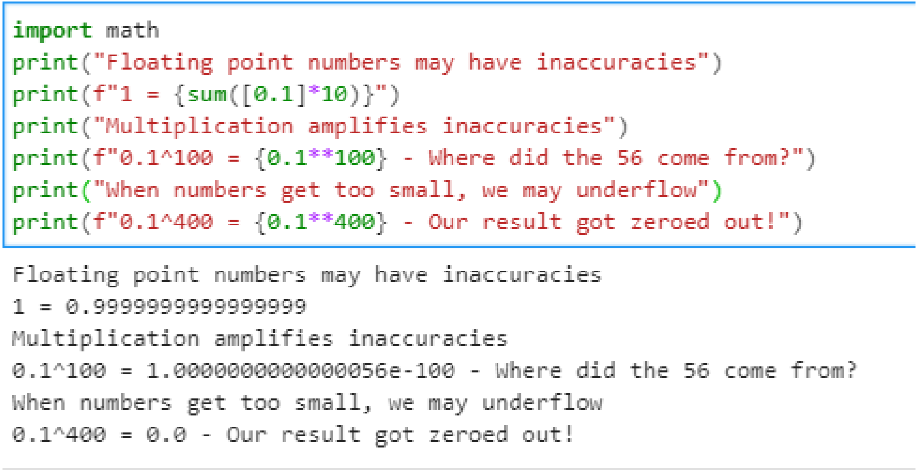
- The BoW representation of the sentence is {"Auckland": 1, "North": 2, "Island": 1}
- For label v, the posterior probability is given by:
- $p(\text{"}Auckland\text{"}|v) \times p(\text{"}North\text{"}|v) \times p(\text{"}North\text{"}|v) \times p(\text{"}Island\text{"}|v) \times p(v)$
- $= p(\text{"}Auckland\text{"}|v)^1 \times p(\text{"}North\text{"}|v)^2 \times p(\text{"}Island\text{"}|v)^1 \times p(v)$
- For $v = Auckland$: $\frac{5}{31}\frac{2}{31}^{2}\frac{2}{31} \times 0.5 = 0.00002$
- For $v = Dunedin$: $\frac{1}{36}\frac{1}{36}^{2}\frac{2}{36} \times 0.5 = 0.0000006$
- The maximum posterior probability is for "v=Auckland", so we label the observations as "Auckland".

Naive Bayes problem - Numeric Instability

A common problem encountered when implementing a Naive Bayes classifier is numeric instability.

- Computers store the likelihoods we calculate as floating point numbers, which may have small inaccuracies.
- Because we are multiplying many small floating point numbers together, we can encounter **underflow**.
- This is when the number we calculate is too small to be properly represented.
- Underflow can lead to our calculating returning a 0 probability, even though it should be a small number.
- This may be more of a problem as the input text gets longer, or the total number of words gets larger, as both push the posterior probability to smaller results which are more likely to underflow.

Underflow



```
import math
print("Floating point numbers may have inaccuracies")
print(f"1 = {sum([0.1]*10)}")
print("Multiplication amplifies inaccuracies")
print(f"0.1^100 = {0.1**100} - Where did the 56 come from?")
print("When numbers get too small, we may underflow")
print(f"0.1^400 = {0.1**400} - Our result got zeroed out!")
```

```
Floating point numbers may have inaccuracies
1 = 0.9999999999999999
Multiplication amplifies inaccuracies
0.1^100 = 1.0000000000000056e-100 - Where did the 56 come from?
When numbers get too small, we may underflow
0.1^400 = 0.0 - Our result got zeroed out!
```

Figure: Example of underflow in python

Log probabilities

- Underflow often occurs in Naive Bayes because we are multiplying many small numbers. How can we avoid it?
- We can covert our multiplication into an addition using log, as well as making our small probabilities larger numbers.
- $log(A \times B) = log(A) + log(B)$
- Logging an expression also maintains relative sizes. If $A > B$ then $log(A) > log(B)$.
- This means that if we take the log of our Naive Bayes expression, it won't change the relative posterior probabilities of the labels. I.E, we will still end up with the same selected label.
- $$\underset{v}{\operatorname{argmax}}\, p(v) \prod_{w_j} p(w_j|v) = \underset{v}{\operatorname{argmax}}\, log(p(v)) + \sum_{w_j} log(p(w_j|v))$$

BoW Example

Using log probabilities, what label would Naive Bayes predict for the sentence: ”Auckland is in the north of the North Island”?

BoW Example

Using log probabilities, what label would Naive Bayes predict for the sentence: ”Auckland is in the north of the North Island”?

- The BoW representation of the sentence is {”Auckland”: 1, ”North”: 2, ”Island”: 1}
- For label v, the posterior probability is given by:
- $log(p(\text{”}Auckland\text{”}|v)) + log(p(\text{”}North\text{”}|v)) + log(p(\text{”}North\text{”}|v)) + log(p(\text{”}Island\text{”}|v)) + log(p(v))$

BoW Example

Using log probabilities, what label would Naive Bayes predict for the sentence: "Auckland is in the north of the North Island"?

- The BoW representation of the sentence is {"Auckland": 1, "North": 2, "Island": 1}
- For label v, the posterior probability is given by:
- $log(p(\text{"}Auckland\text{"}|v)) + log(p(\text{"}North\text{"}|v)) + log(p(\text{"}North\text{"}|v)) + log(p(\text{"}Island\text{"}|v)) + log(p(v))$
- For $v = Auckland$: $log(\frac{5}{31}) + 2log(\frac{2}{31}) + log(\frac{2}{31}) + log(0.5) = -4.66$
- For $v = Dunedin$: $log(\frac{1}{36}) + 2log(\frac{1}{36}) + log(\frac{2}{36}) + log(0.5) = -6.22$
- The maximum posterior probability is for "v=Auckland", so we label the observations as "Auckland".

N-grams

- One problem we encounter with Naive Bayes is that we lose all interactions between features.
- Often interactions change the meaning of words, e.g., A news article with the words "machine" and "learning" could be to do with learning at school, so classified as 'education' whereas an article with the phrase "machine learning" is much more likely to be classified as 'tech'.
- We can introduce interactions by considering N-grams as features, instead of individual words.
- An N-gram is a sub-sequence of N words found in a text. Commonly we use bigrams (subsets of 2 words) and trigrams (subsets of 3 words).

Bigram example

- What are the Bigrams of the sentence ”Auckland is in the north of the North Island”? (including stop words)

Bigram example

- What are the Bigrams of the sentence "Auckland is in the north of the North Island"? (including stop words)
- {"Auckland is", "is in", "in the", "the north", "north of", "of the", "the North", "North Island}
- We have now captured the phrase "North Island" which is much more informative than "North" and "Island" separately!

N-gram Problem

- What is a problem with N-grams?

N-gram Problem

- What is a problem with N-grams?
- We need *more* training data to learn proper probabilities. We need to see multiple examples of all possible N length subsets for each label.
- It is much more likely to encounter an N-gram we haven't seen before than a single word!

TF-IDF

- Intuition - If a word appears in *every* text, it is useless for telling the difference between classes.
- Alternatively, most highly informative words will not appear in very many texts. E.G, the word "rugby" may be relatively rare, but if we see it, it can tell us a lot about the sentence.
- We would like some way to place less weight on very common words, and more weight on rare words.
- This leads us to the idea of the "document frequency" of a word. This is the proportion of documents (e.g, sentences in the training set) which contained a given word., with range [0, 1]. 0 Means the word appeared in no documents, 1 means it appeared in all documents.
- The log of the *inverse* document frequency, can be thought of as the rarity of a word. $IDF_i = log(\frac{N}{D_i})$, where N is the total number of documents and D_i is the number of documents containing word i. We take the log so that IDF is 0 when the word appears in all documents, and grows slower as the word becomes rarer.

TF-IDF

- In our BoW representation, we already calculate the frequency of each term.
- If we weight this count with the IDF for each word, we can take into account the rarity of each word.

TF-IDF example

What are the IDF values associated with each word in the BoW example on slide 26?

TF-IDF example

What are the IDF values associated with each word in the BoW example on slide 26?

- Auckland: $log(\frac{6}{4}) = 0.18$
- North: $log(\frac{6}{1}) = 0.78$
- Island: $log(\frac{6}{2}) = 0.48$
- New: $log(\frac{6}{4}) = 0.18$
- Zealand: $log(\frac{6}{4}) = 0.18$
- large: $log(\frac{6}{1}) = 0.78$
- population: $log(\frac{6}{2}) = 0.48$
- region: $log(\frac{6}{1}) = 0.78$
- governed: $log(\frac{6}{1}) = 0.78$
- council: $log(\frac{6}{2}) = 0.48$
- Dunedin: $log(\frac{6}{2}) = 0.48$
- south: $log(\frac{6}{1}) = 0.78$
- sixth: $log(\frac{6}{1}) = 0.78$
- highest: $log(\frac{6}{1}) = 0.78$
- ...

TF-IDF example

- In order to use TF-IDF, we just replace any use of a words frequency with its frequency * IDF.
- Using smoothing we calculated $p(w_j|v) = \frac{f_{jv}+1}{\sum_j(f_{jv})+|V|}$, where w_j is word j, f_{jv} is the frequency of word j in class v, and $|V|$ is the number of unique words in the vocabulary.
- Including TF-IDF, we calculate:

$$p(w_j|v) = \frac{f_{jv} \times IDF_j + 1}{\sum_j(f_{jv} \times IDF_j) + |V|}$$