COMPSCI 762 2022 S1 Week 7 Solution

Luke Chang

April 7, 2022

Question 1

- Attribute: *Outlook* (O), *Temperature* (T), *Humidity* (H), *Wind* (W);
- Output: *Rain* (R) - Binary classification problem;
- 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: $R = \text{argmax}_{r\in\{\text{T,F}\}} P(R = r|O, T, H, W)$
- Bayes Theorem: $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$
- How to rewrite this expression using Bayes Theorem?

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

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

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

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

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

$$P(R = \text{True}) = \frac{5}{10} = 0.5$$
$$P(R = \text{False}) = \frac{5}{10} = 0.5$$

- Calculate the Likelihood:
  - $P(O = \text{Sunny}|R = \text{T}) = \frac{1}{5}$
  - $P(O = \text{Overcast}|R = \text{T}) = \frac{2}{5}$
  - $P(O = \text{Rain}|R = \text{T}) = \frac{2}{5}$
  - $P(T = \text{Hot}|R = \text{T}) = \frac{2}{5}$
  - $P(T = \text{Mild}|R = \text{T}) = \frac{2}{5}$
  - $P(T = \text{Cool}|R = \text{T}) = \frac{1}{5}$
  - $P(H = \text{High}|R = \text{T}) = \frac{4}{5}$
  - $P(H = \text{Normal}|R = \text{T}) = \frac{1}{5}$
  - $P(T = \text{Strong}|R = \text{T}) = \frac{1}{5}$
  - $P(T = \text{Weak}|R = \text{T}) = \frac{4}{5}$

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

- Let's predict:

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

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

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

$$\begin{aligned} P(R=\mathrm{T}|O,T,H,W) &= \frac{0.0032}{P(O,T,H,W)} \\ P(R=\mathrm{F}|O,T,H,W) &= \frac{0.0384}{P(O,T,H,W)} \\ P(R=\mathrm{T}|O,T,H,W) &= \frac{0.0032}{0.0032+0.0384} = 0.08 \\ P(R=\mathrm{F}|O,T,H,W) &= \frac{0.0384}{0.0032+0.0384} = 0.92 \end{aligned}$$

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

Question 2

- What does the training data look like after transforming to lower case, and removing the stop words and punctuations?

Index	Sentence	Word Count	Label
1	auckland north island new zealand	5	Auckland
2	auckland large population	3	Auckland
3	auckland region governed auckland concil	5	Auckland
4	dunedin south island new zealand	5	Dunedin
5	dunedin sixth highest population new zealand	6	Dunedin
6	largest city new zealand formation auckland council	7	Dunedin

- 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 }

- Priors:

$$p(A) = \frac{3}{6} = 0.5$$

$$p(D) = \frac{3}{6} = 0.5$$

- Likelihood of seeing the word “auckland”:

$$p(\text{“auckland”}|A) = \frac{4}{13}$$

$$p(\text{“auckland”}|D) = \frac{1}{18}$$

- Likelihood of seeing the word “north”:

$$p(\text{“north”}|A) = \frac{1}{13}$$

$$p(\text{“north”}|D) = \frac{0}{18} = 0$$

- 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.

$$p(\text{“auckland”}|A) = \frac{5}{31}$$

- 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.

$$p(\text{“auckland”}|D) = \frac{2}{36}$$

- For $p(\text{“north”}|A)$, and $p(\text{“north”}|D)$, 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.

$$p(\text{“north”}|A) = \frac{2}{31}$$

$$p(\text{“north”}|D) = \frac{1}{36}$$

- Predict “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:

$$p(\text{“auckland”}|v) \times p(\text{“north”}|v) \times p(\text{“north”}|v) \times p(\text{“island”}|v) \times p(v)$$

$$= p(\text{“auckland”}|v)^1 \times p(\text{“north”}|v)^2 \times p(\text{“island”}|v)^1 \times p(v)$$

  - 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.