LIN 313: Language and Computers

Naive Bayes Practice

P(verify | phish) = 40/100

Consider a corpus of 200 emails annotated as *phishing* or *legitimate*. There are 100 emails in each category. The word *verify* appears in 40 of the phishing emails and 5 of the legitimate emails. The word *account* appears in 60 of the phishing emails and 40 of the legitimate emails. These counts record the number of emails containing each word, regardless of how many times it occurs in an email.

P(phish) = P(legit) = 100/200 = 1/2

Part 1: Using the same corpus

Use this corpus to estimate both the prior probabilities and the word likelihoods. Show your work. You may leave your answers as fractions.

(i) Calculate the probability that an email is phishing given that it contains *verify*.

prior

$$P(\text{phish} \mid \text{verify}) = \frac{P(\text{verify} \mid \text{phish}) \cdot P(\text{phish})}{P(\text{verify})}$$

likelihood; P(verify) → evidence

$$= \frac{40/100 \cdot 1/2}{45/200} = \frac{40}{45} = \frac{8}{9} = 0.889$$

(ii) Calculate the probability that an email is phishing given that it contains *account*.

P(phish | account) =

Solve yourself!

P(legit | verify) = 1 − 0.889 = .111

Final Answer = 3/5 or 0.6 or 60%.

Part 1: Using the same corpus (continued)

(iii) Calculate the probability that an email is phishing given that it contains both *verify* and *account*. Use the Naive Bayes assumption and state what this assumption means. Show your work.

$$P(\text{phish} \mid \text{verify} \,\&\, \text{account})$$

$$= \frac{P(\text{verify} \,\&\, \text{account} \mid \text{phish}) \cdot P(\text{phish})}{P(\text{verify} \,\&\, \text{account})}$$

$$= \frac{P(\text{verify} \mid \text{phish}) \times P(\text{account} \mid \text{phish}) \times P(\text{phish})}{P(\text{verify}) \cdot P(\text{account})}$$

↓ Naive Bayes assumption

= Calculate yourself!

Answer = $\frac{12}{13}$ or 0.923 or 92.3%.

Solve yourself! only Priors change.

Part 2: Using different corpora

You now apply the classifier to a different inbox. A representative sample of 2,000 emails from this inbox contains 100 phishing emails and 1,900 legitimate emails. Use this sample to estimate the prior probabilities, and use the original 200-email corpus from Part 1 to estimate the word likelihoods. Assume that the word probabilities within each category also apply to the new inbox.

Repeat the calculations from Part 1 using these estimates. Show your work.

(i) Calculate the probability that an email is phishing given that it contains *verify*.

Ans = $\frac{8}{27}$

(ii) Calculate the probability that an email is phishing given that it contains *account*.

Ans = $\frac{3}{41}$

Part 2: Using different corpora (continued)

(iii) Calculate the probability that an email is phishing given that it contains both *verify* and *account*. Use the Naive Bayes assumption and the estimates from Part 2. Show your work.

Ans = $\frac{12}{31}$

If the classifier labels an email as phishing when its estimated probability exceeds $\frac{1}{2}$, does its prediction for an email containing both *verify* and *account* change between Parts 1 and 2? Briefly explain why or why not.

Yes! Because the prior has changed and is no longer 50% for phish vs. legit.