KOREA UNIVERSITY

1905

COSE361: Artificial Intelligence

Homework 3

Spring 2020 - Dr. Hyunwoo J. Kim

Answer the questions in the spaces provided. If you run out of room for an answer, continue on the back of the page.

Name: 홍성원

Student ID: 2016320187

Instructor's name: 김현우

1. Answer the following questions.

(a) Explain the Bias-Variance Trade-off.

Bias is error that shows how much the model differ from the true model. Variance is error that shows how much the model from different training sets differ from each other. If we have simple model, we can reduce variance but we get high bias. If we have complex model, we can reduce bias but we get high variance. So this trade-off is indispensable.

(b) Draw a confusion matrix and explain the following concepts.

T: TRUE, F: FALSE
P: POSITIVE, N: NEGATIVE

		predictive values: positive	predictive values: negative
ACTUAL VALUE	pos	TP	FN
	neg	FP	TN

(c) Accuracy

Accuracy is a proportion of correct classification. $\therefore Acc = \frac{TP+TN}{TP+TN+FP+FN}$

(d) Recall

Recall is a proportion of TP from Actual positive values. $\therefore Recall = \frac{TP}{TP+FN}$

(e) Precision

Precision is a proportion of TP from predictive positive values. $\therefore Precision = \frac{TP}{TP+FP}$

(f) False Positive

FP is classification that is predicted positive but actual value is negative.

(g) False Negative

FN is classificaion that is predicted negative but actual value is positive.

(h) Explain why both recall and precision need to be considered for evaluation.
(hint. trivial predictions to get the best recall or precision)

next page

(i) Explain 'Occam's Razor' and 'The Curse of Dimensionality'.

next page

(j) GIve the examples of three random variables that are conditionally independent.

$X \perp\!\!\!\perp Y|Z$

X: The ground is wet today

Y: Raining tomorrow

Z: Raining Today.

1. (h)

If we model only to get best precision, we can get the right value well. But among the right values, there still can be a lot of values that is not selected by the model because of low recall.

Then if we model only to get best recall, we can get a lot of right values from total right values. But the model might select wrong value too to increase recall. So there exists recall-precision trade-off. It is important to consider both recall and precision to get good model.

1. (i)

Occam's razor is a principle. This principle claims that you should not have unnecessary assumption a lot and if you can explain something with less logic, you must not have more logic. This occam's razor prefers simpler theory.

Dimensionality means the amount of information. For example, if we only have weights and heights data, dimensionality is two. If we get more information such as age, blood pressure, dimensionality increases and we can manage more information. However, as dimensionality increase, the model needs much more data to learn and it is very hard to learn for the model. This is called 'the curse of dimensionality'.

(k) (Naïve Classifier) The vocabularies of our spam filter are the following.
V="secret", "offer", "low", "price", "valued", "customer", "today", "dollar", "million", "sports", "is", "for", "play", "healthy", "pizza".

We have the following spam emails for training.
"million dollar offer",
"secret offer today",
"secret is secret"

We have the following normal (non-spam) emails for training.
"low price for valued customer"
"play secret sports today"
"sports is healthy"
"low price pizza"

Using the Naïve Bayes Model, calculate the probabilities.
Write intermediate calculation steps as well.

1 $P(\text{secret}|\text{spam}) = ?$
2 $P(\text{spam}) = ?$
3 $P(\text{sports}|\text{non-spam}) = ?$
4 $P(\text{dollar}|\text{spam}) = ?$
5 $P(\text{spam}|\text{"sports is healthy"}) = ?$
6 $P(\text{non-spam}|\text{"sports is healthy"}) = ?$
7 $P(\text{"sports is healthy"}|\text{spam}) = ?$
8 $P(\text{"sports is healthy"}|\text{non-spam}) = ?$

Let spam be S,
non-spam be N.

① $P(\text{secret}|S) = \frac{2}{3}$ ② $P(S) = \frac{3}{7}$ ③ $P(\text{sports}|N) = \frac{2}{4} = \frac{1}{2}$

④ $P(\text{dollar}|S) = \frac{1}{3}$

⑦ $P(\text{"sports is healthy"}|S) = \underbrace{P(\text{sports}|S)}_{0} \cdot \underbrace{P(\text{is}|S)}_{\frac{1}{3}} \cdot \underbrace{P(\text{healthy}|S)}_{0} = 0$

⑧ $P(\text{"sports is healthy"}|N) = P(\text{sports}|N) \cdot P(\text{is}|S) \cdot P(\text{healthy}|S) = \frac{1}{2} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{32}$

⑤ $P(S|\text{"sports is healthy"}) = \frac{P(\text{"sports is healthy"}|S) \cdot P(S)}{P(\text{"sports is healthy"})} = \frac{P(\text{"sports is healthy"}|S) \cdot P(S)}{P(\text{"sports is healthy"}|S) \cdot P(S) + P(\text{"sports is healthy"}|N) \cdot P(N)}$

$= 0$

⑥ $P(N|\text{"sports is healthy"}) = \frac{P(\text{"sports is healthy"}|N) \cdot P(N)}{P(\text{"sports is healthy"})} = \frac{P(\text{"sports is healthy"}|N) \cdot P(N)}{P(\text{"sports is healthy"}|N) \cdot P(N) + P(\text{"sports is healthy"}|S) \cdot P(S)}$

$= \frac{\frac{1}{32} \cdot \frac{4}{7}}{\frac{1}{32} \cdot \frac{4}{7} + 0} = 1$