A2

Freitag, 8. Mai 2020 08:50

Zhuo Zeng 3489547
Jiaqi Qin 3493229
Jiaqi Wu 3506048

1.1. $P(\text{Apple}) = \frac{1}{2}\times\frac{8}{8+4} + \frac{1}{2}\times\frac{10}{10+2} = \frac{9}{12} = \frac{3}{4}$

$P(\text{box}=1 \mid \text{Apple}) = \frac{P(\text{box}=1, \text{Apple})}{P(\text{Apple})} = \frac{8/24}{3/4} = \frac{4}{9}$

1.2. $E :=$ one is green, one is yellow.

$A :=$ from 1994 bag drew a yellow

$P(E) = P(C_1=g, C_2=y \cup C_1=y, C_2=g) = 0.1\times0.14 + 0.2\times0.8 = 0.54$

$P(A|E) = \frac{P(A,E)}{P(E)} = \frac{P(E|A)\cdot P(A)}{P(E)} = \frac{0.2\times0.2}{0.54} = \frac{20}{27} \approx 74.1\%$

3. 1) text will be presented by the vectors, in which every position is the frequency of appearance from each word $t = [x_1, x_2, x_3 \cdots]$

a) similarity function, the Angle between two vectors.

$sim(t_1, t_2) = \cos(t_1, t_2) = \frac{t_1 \cdot t_2}{|t_1|\cdot|t_2|}$

b) choose K nearest text vectors from training set. according to the similarity function.

for example:

text = ['go', 'to', 'only', 'free', 'entry', 'to' ···]

⇓ vectorized

text_pred = [1, 2, 1, 0, 1, 0, 0, 4 - - -]

text_train[i] = [0, 1, 1, 1, 0, 0, 0, 0 - - -]

$result[i] = sim(text\_pred, text\_train[i]) = \frac{1\times0 + 2\times1 + 1\times1 + 0\times1 \cdots}{\sqrt{1^2+2^2+1^2+0^2+\cdots}\sqrt{0^2+1^2+1^2+\cdots}}$

find k smallest numbers in result
determine the label of text_pred by counting the most common label in these k training set.

disadvantages: extremly slow, especially for texts with many words

advantages: easy to understand.
simple algorithm

4. 1) Because KNN needs to compute the distances along every dimension without an explicit knowledge of classes, the more dimensions it has, the more calculations it needs, and the more sparse the training data will be. which reduces the ability of prediction of KNN, if adding more data, the cost of calculation will be too large or even unable to compute.

2) the best way to circumvent it is using dimensionality reduction (e.g. find the dependency between different dimensions or generate a new feature combining several old features.)