Ut:4 -9M

- Suppose you are given a dataset containing information about whether emails are spam or not spam, along with two features:
  i. the presence of the word "offer" (1: present, 0: absent)
  ii. the presence of the word "free" (1: present, 0: absent)

You are tasked with classifying a new email with the following features values: "offer" = 1 & "free" = 1.

Given the training dataset:

| Email | Offee | Free | Spam |
|---|---|---|---|
| 1 | 1 | 0 | No |
| 2 | 0 | 1 | yes |
| 3 | 1 | 1 | yes |
| 4 | 0 | 1 | No |
| 5 | 1 | 1 | yes |

Calculate probability that the new email is spam using Naive Bayes.

→

i. Given:

Total emails: 5

Spam = yes : 3

Spam = No : 2

ii. Prior Probabilities:

$P(\text{Spam}) = 3/5$

$P(\text{Not Spam}) = 2/5$

iii. Calculate likelihoods/conditional probabilities:

a. for "spam = yes" (3 mails):

- "offer" = 1, occurs in 2 mails (3 & 5)
- "free" = 1, occurs in 3 mails (2, 3, & 5)

$\therefore P(\text{offer} = 1 / \text{spam} = \text{yes}) = 2/3$

$P(\text{free} = 1 / \text{spam} = \text{yes}) = 3/3 = 1$

b. for spam = No (2 mails):

- offer = 1, occurs in 1 out of 2 mails.
- free = 1, occurs in 1 out of 2 mails.

$\therefore P(\text{offer} = 1 / \text{spam} = \text{No}) = 1/2$
$P(\text{free} = 1 / \text{spam} = \text{No}) = 1/2$

iv. Apply Naive Bayes:

a. $P(\text{spam} = \text{yes} / \text{offer} = 1, \text{free} = 1) \propto P(\text{spam}) \cdot P(\text{offer} = 1 / \text{spam} = \text{yes}) \cdot P(\text{free} = 1 / \text{spam} = \text{yes})$

$= \frac{3}{5} \cdot \frac{2}{3} \cdot 1 = \frac{2}{5}$

b. $P(\text{Spam} = \text{No} / \text{offer}=1, \text{free}=1) \propto P(\text{Not spam}) \cdot P(\text{offer}=1 / \text{spam}=\text{No}) \cdot P(\text{free}=1 / \text{spam}=\text{No})$

$$= \frac{2}{5} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{10}$$

v. Normalize:

- Normalizing constant $= \frac{2}{5} + \frac{1}{10} = \frac{1}{2} = 0.5$

- Now, divide each by constant to get final probability:

$$P(\text{Spam}) = \frac{2/5}{0.5} = \frac{4}{5} = 0.8$$

$$P(\text{Not spam}) = \frac{1/10}{0.5} = \frac{0.1}{0.5} = 0.2$$

Ans:

The probability that the new email with offer=1 & free=1, is spam is 0.8 (80%).

Ut:5 -9M

– Suppose you have the following dataset containing the coordinates of points in a 2-dimensional space.

| Point | X-coordinate | Y-coordinate |
|---|---|---|
| A | 2 | 3 |
| B | 4 | 7 |
| C | 3 | 5 |
| D | 6 | 9 |
| E | 8 | 6 |
| F | 7 | 8 |

perform Kmeans clustering. Assume initial centroid: (2,3) & (8,6).

→

i. Initial centroids:

$C_1$ : (2, 3)

$C_2$ : (8, 6)

Iteration 1: (Assign to nearest cluster)

| Point | Coordinates ($x_1$, $y_1$) | Distance to $C_1$ (2, 3) | $C_2$ (8, 6) | Cluster |
|---|---|---|---|---|
| A | 2, 3 | 0.00 | 6.70 | $C_1$ |
| B | 4, 7 | 4.47 | 4.12 | $C_2$ |
| C | 3, 5 | 2.23 | 5.09 | $C_1$ |
| D | 6, 9 | 7.21 | 3.60 | $C_2$ |
| E | 8, 6 | 6.70 | 0.00 | $C_2$ |
| F | 7, 8 | 7.07 | 2.23 | $C_2$ |

Find distance using euclidean distance: $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

ex: point A (2,3) & $C_1$ (2,3)

$$\text{distance} = \sqrt{(2-2)^2+(3-3)^2} = 0.00$$

point A (2,3) & $C_2$ (8,6)

$$d = \sqrt{(8-2)^2+(6-3)^2} = 6.70$$

Recompute centroids:

Cluster 1 ($C_1$): Points A(2, 3), C(3, 5)

$\therefore$ New centroid : $\left(\frac{2+3}{2}, \frac{3+5}{2}\right) = (2.5, 4)$

Cluster 2 ($C_2$): Points B(4, 7), D(6, 9), E(8, 6), F(7, 8)

$\therefore$ New centroid: $\left(\frac{4+6+8+7}{4}, \frac{7+9+6+8}{4}\right)$

$= (6.25, 7.5)$

Iteration 2:

| Point | coordinates | Distance to $C_1(2.5, 4)$ | $C_2(6.25, 7.5)$ | New cluster |
|---|---|---|---|---|
| A | 2, 3 | 1.11 | 6.18 | 1 |
| B | 4, 7 | 3.35 | 2.30 | 2 |
| C | 3, 5 | 1.11 | 4.10 | 1 |
| D | 6, 9 | 6.10 | 1.52 | 2 |
| E | 8, 6 | 5.85 | 2.30 | 2 |
| F | 7, 8 | 6.02 | 0.90 | 2 |

Centroids remain the same as in iteration 1

→ convergence reached.

$\therefore$ Final Clusters & centroids.

Cluster 1 centroid ($C_1$): (2.5, 4)

Cluster 2 centroid ($C_2$): (6.25, 7.5)

with assigments as:

Cluster 1: A, C

Cluster 2: B, D, E, F

- 9M

Ut:4

– Calculate the support & confidence value for all the possible item sets.

| Transaction ID | Items Bought |
|---|---|
| 1 | Onion, Potato, Cold drink |
| 2 | Onion, Burger, Colddrink |
| 3 | Eggs, Onion, Colddrink |
| 4 | Potato, Milk, Eggs |
| 5 | Potato, Burger, Colddrink, Milk, Eggs |

→

i. List all items & count 1-itemset support count

| item | Count (C) | Support = C/Total transactions |
|---|---|---|
| Onion | 3 | 3/5 = 60% |
| Potato | 3 | 3/5 = 0.6 = 60% |
| Colddrink | 4 | 4/5 = 0.8 = 80% |
| Burger | 2 | 2/5 = 0.4 = 40% |
| Eggs | 3 | 3/5 = 0.6 = 60% |
| Milk | 2 | 2/5 = 0.4 = 40% |

ii. 2-itemset support Count

| itemset | Count | Support |
|---|---|---|
| {Onion, potato} | 1 | 1/5 = 20% |
| {Onion, Colddrink} | 3 | 60% |
| {Onion, Burger} | 1 | 20% |
| {Onion, Eggs} | 1 | 20% |
| {Potato, Colddrink} | 2 | 40% |
| {Potato, Burger} | 1 | 20% |
| {Potato, Eggs} | 2 | 40% |
| {Potato, Milk} | 2 | 40% |
| ~~{Potato, Eggs}~~ | | |
| {Colddrink, Burger} | 2 | 40% |
| {colddrink, Eggs} | 2 | 40% |
| {colddrink, Milk} | 1 | 20% |
| {Burger, Eggs} | 1 | 20% |
| {Burger, Milk} | 1 | 20% |
| {Milk, Eggs} | 2 | 40% |

iii. Confidence Calculations:

Formula:

Confidence(A→B) = Support(AUB) / Support(A)

| Rule | Support(AUB) | Support(A) | Confidence |
|---|---|---|---|
| Onion → Colddrink | 3 | 3 | 100% |
| Colddrink → Onion | 3 | 4 | 75% |
| Potato → colddrink | 2 | 3 | 66.67% |
| Colddrink → Potato | 2 | 4 | 50% |
| Potato → Eggs | 2 | 3 | 66.67% |
| Eggs → Potato | 2 | 3 | 66.67% |
| Potato → Milk | 2 | 3 | 66.67% |
| Milk → Potato | 2 | 2 | 100% |
| Colddrink → Burger | 2 | 4 | 50% |
| Burger → colddrink | 2 | 2 | 100% |
| Colddrink → Eggs | 2 | 4 | 50% |
| Eggs → colddrink | 2 | 3 | 66.67% |
| Milk → Eggs | 2 | 2 | 100% |
| Eggs → Milk | 2 | 3 | 66.67% |
| Onion → Potato | 1 | 3 | 33.34% |

Ut:5 -9M

Suppose that the given data the taste is to cluster points (with (x,y) representing location) into three cluster, where the points are:

A1(2,10), A2(2,5), A3(8,4),
B1(5,8), B2(7,5), B3(6,4),
C1(1,2), C2(4,9).

The distance function is Euclidean distance, suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. use the Kmeans algorithm to show only the 3 cluster centers after the first round of execution with steps.

→

Initial Clusters: (centroids)

i. cluster 1 ($C_1$) : A1 (2, 10)
ii. Cluster 2 ($C_2$) : B1 (5, 8)
iii. Cluster 3 ($C_3$) : C1 (1, 2)

Iteration 1:

| Point | Coordinates | Distance to $C_1$(2, 10) | Distance to $C_2$(5, 8) | Distance to $C_3$(1, 2) | Cluster |
|---|---|---|---|---|---|
| A1 | 2, 10 | 0 | 3.61 | 8.06 | $C_1$ |
| A2 | 2, 5 | 5.0 | 4.24 | 3.16 | $C_3$ |
| A3 | 8, 4 | 8.49 | 5.0 | 7.28 | $C_2$ |
| B1 | 5, 8 | 3.61 | 0 | 7.21 | $C_2$ |
| B2 | 7, 5 | 7.07 | 3.61 | 6.08 | $C_2$ |
| B3 | 6, 4 | 7.21 | 4.12 | 5.39 | $C_2$ |
| C1 | 1, 2 | 8.06 | 7.21 | 0 | $C_3$ |
| C2 | 4, 9 | 2.24 | 1.41 | 7.07 | $C_2$ |

Compute centroids.

Cluster 1: $C_1$ : $A_1(2,10)$

∴ New centroid : (2, 10)

Cluster 2: $C_2$ : A3(8,4), B1(5,8), B2(7,5), B3(6,4), C2(4,9)

∴ New centroid : $\left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right)$

$= (6.6)$

Cluster 3: $C_3$ : $A_2(2,5)$, C1(1,2)

∴ New centroid = $\left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

∴ After 1st round of execution,
3 cluster centers are:

C1 (2, 10)
C2 (6.6)
C3 (1.5, 3.5)

Page No.:
Date: / /

Ut:5 -8M

- Given the confusion matrix, calculate accuracy, Recall, Error rate with description on heart attack risk.

Predicted classes

| | Classes | Heart Attack Risk-yes | Heart Attack Risk No |
|---|---|---|---|
| Actual classes | Heart Attack Risk-yes | 80 | 220 |
| | Heart Attack Risk-NO | 150 | 9500 |

→

Confusion Matrix

| Actual \ Predicted | Yes (Risk) | No (Risk) | Total |
|---|---|---|---|
| Yes (Risk) | 80 (TP) | 220 (FN) | 300 |
| NO (Risk) | 150 (FP) | 9500 (TN) | 9650 |
| Total | 230 | 9720 | 9950 |

- TP : 80 (correctly predicted "yes" for heart attack risk)
- FN : 220 (Incorrectly predicted "No" when actual was "yes")
- FP : 150 (Incorrectly predicted "yes" when actual was "No")
- TN : 9500 (Correctly predicted "NO" for NO risk).

i. Accuracy: Measures correctness of the model.

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}} = \frac{80 + 9500}{9950} = \frac{9580}{9950}$$

$$\approx 0.9628 = 96.28\%$$

The model is 96.28% correct across both classes.

ii. Precision:

Measures how many predicted "yes" cases were actually correct.

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{80}{80 + 150} = \frac{80}{230}$$

$$\approx 0.3478 = 34.78\%$$

Only 34.78% of predicted "yes" cases truly had heart attack risk.

iii. Recall: how many actual yes cases correctly predicted.

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{80}{80 + 220} = \frac{80}{300}$$

$$\approx 0.2667 = 26.67\%$$

The model captures only 26.67% of actual heart attack cases.

iv. Error Rate: Measures overall incorrect predictions.

$$\text{Error Rate} = 1 - \text{Accuracy} = 1 - 0.9628$$

$$= 0.0372 = 3.72\%$$

3.72% of predictions are wrong.

- Ut:5 -8M

Given the confusion matrix. Calculate Accuracy, Precision, Recall, Error Rate with description on Diabetic Risk.

| | | Predicted classes: Diabetic Risk-Yes | Diabetic Risk-NO |
|---|---|---|---|
| Actual classes | Diabetic Risk (Yes) | 90 | 210 |
| | Diabetic Risk (NO) | 140 | 9560 |

→

Confusion Matrix

| Actual \ Predicted | Yes (Risk) | NO (Risk) | Total |
|---|---|---|---|
| Yes (Risk) | 90 (TP) | 210 (FN) | 300 |
| NO (Risk) | 140 (FP) | 9560 (TN) | 9700 |
| Total | 230 | 9770 | 10,000 |

i. Accuracy : $\frac{TP+TN}{Total} = \frac{90+9560}{10,000} = 0.965 = 96.5\%$

ii. Precision : $\frac{TP}{TP+FP} = \frac{90}{90+140} \approx 0.3913 = 39.13\%$

iii. Recall : $\frac{TP}{TP+FN} = \frac{90}{90+210} = 0.3 = 30\%$

iv. Error Rate : 1 - Accuracy = 1 - 0.965 = 0.035 = 3.5%