# INFS 4203 / 7203 Data Mining
# Tutorial 1: Classification

Doris He
d.he@uq.edu.au

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Q : Assume a test dataset contains 1% catfish and 99% other types of fish. For a classifier that always makes prediction that a fish is not a catfish, what' s the accuracy, precision and recall?
- A:

| | | Predicted class | |
|---|---|---|---|
| | | **Yes** | **No** |
| **Actual Class** | **Yes** | 0 (**TP**) | 1% (**FN**) |
| | **No** | 0 (**FP**) | 99% (**TN**) |

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Q : Assume a test dataset contains 1% catfish and 99% other types of fish. For a classifier that always makes prediction that a fish is not a catfish, what' s the accuracy, precision and recall?
- A:

| | | Predicted class | |
|---|---|---|---|
| | | **Yes** | **No** |
| **Actual Class** | **Yes** | 0 (**TP**) | 1% (**FN**) |
| | **No** | 0 (**FP**) | 99% (**TN**) |

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} = 99\%$$

$$Precision = \frac{TP}{TP + FP} = 0$$

$$Recall = \frac{TP}{TP + FN} = 0$$

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Q : Calculate the precision and recall for the following two binary classifiers M1 and M2, and discuss which one is better.

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |

M1

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |

M2

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Q : Calculate the precision and recall for the following two binary classifiers M1 and M2, and discuss which one is better.

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |

M1

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |

M2

| | | Predicted class | |
|---|---|---|---|
| **M1** | | **Red** | **Not Red** |
| **Actual Class** | **Red** | **(TP)** | **(FN)** |
| | **Not Red** | **(FP)** | **(TN)** |

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Q : Calculate the precision and recall for the following two binary classifiers M1 and M2, and discuss which one is better.

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |

M1

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |

M2

| | | Predicted class | |
|---|---|---|---|
| **M1** | | **Red** | **Not Red** |
| **Actual Class** | **Red** | **4 (TP)** | **(FN)** |
| | **Not Red** | **(FP)** | **(TN)** |

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Q : Calculate the precision and recall for the following two binary classifiers M1 and M2, and discuss which one is better.

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |

M1

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |

M2

| | | Predicted class | |
|---|---|---|---|
| **M1** | | **Red** | **Not Red** |
| **Actual Class** | **Red** | **4 (TP)** | **2 (FN)** |
| | **Not Red** | **(FP)** | **(TN)** |

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Q : Calculate the precision and recall for the following two binary classifiers M1 and M2, and discuss which one is better.

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |

M1

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |

M2

| | | Predicted class | |
|---|---|---|---|
| **M1** | | **Red** | **Not Red** |
| **Actual Class** | **Red** | **4** (TP) | **2** (FN) |
| | **Not Red** | **1** (FP) | (TN) |

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Q : Calculate the precision and recall for the following two binary classifiers M1 and M2, and discuss which one is better.

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |

M1

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |

M2

| | | Predicted class | |
|---|---|---|---|
| **M1** | | **Red** | **Not Red** |
| **Actual Class** | **Red** | **4** (TP) | **2** (FN) |
| | **Not Red** | **1** (FP) | **3** (TN) |

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Q : Calculate the precision and recall for the following two binary classifiers M1 and M2, and discuss which one is better.

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |

M1

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |

M2

| | | Predicted class | |
|---|---|---|---|
| **M1** | | **Red** | **Not Red** |
| **Actual Class** | **Red** | **4** (TP) | **2** (FN) |
| | **Not Red** | **1** (FP) | **3** (TN) |

$$Precision = \frac{TP}{TP + FP} = \frac{4}{5}$$

$$Recall = \frac{TP}{TP + FN} = \frac{4}{6}$$

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Q : Calculate the precision and recall for the following two binary classifiers M1 and M2, and discuss which one is better.

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |

M1

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |

M2

| | | Predicted class | |
|---|---|---|---|
| **M1** | | **Red** | **Not Red** |
| **Actual Class** | **Red** | **4** (TP) | **2** (FN) |
| | **Not Red** | **1** (FP) | **3** (TN) |

$$Precision = \frac{TP}{TP + FP} = \frac{4}{5}$$

$$Recall = \frac{TP}{TP + FN} = \frac{4}{6}$$

| | | Predicted class | |
|---|---|---|---|
| **M2** | | **Red** | **Not Red** |
| **Actual Class** | **Red** | **3** (TP) | **3** (FN) |
| | **Not Red** | **0** (FP) | **4** (TN) |

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Q : Calculate the precision and recall for the following two binary classifiers M1 and M2, and discuss which one is better.

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |

M1

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |

M2

| M1 | | Predicted class | |
|---|---|---|---|
| | | Red | Not Red |
| Actual Class | Red | 4 (TP) | 2 (FN) |
| | Not Red | 1 (FP) | 3 (TN) |

$$Precision = \frac{TP}{TP + FP} = \frac{4}{5}$$

$$Recall = \frac{TP}{TP + FN} = \frac{4}{6}$$

| M2 | | Predicted class | |
|---|---|---|---|
| | | Red | Not Red |
| Actual Class | Red | 3 (TP) | 3 (FN) |
| | Not Red | 0 (FP) | 4 (TN) |

$$Precision = \frac{TP}{TP + FP} = \frac{3}{3}$$

$$Recall = \frac{TP}{TP + FN} = \frac{3}{6}$$

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- A : Calculate the precision and recall

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |

M1

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |

M2

M1 :
- Precision: 80%
- Recall: 67%

M2 :
- Precision: 100%
- Recall: 50%

# T1-Q1

## Accuracy, Precision, Recall, F-measure

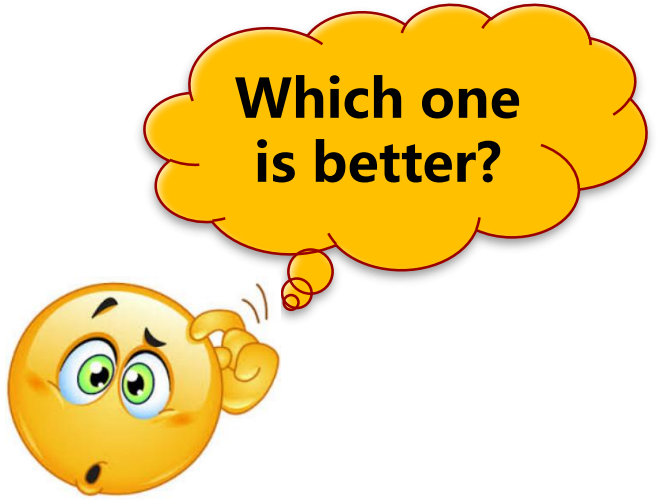
- A : Calculate the precision and recall

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |

M1

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |

M2




M1 :
- Precision: 80%
- Recall: 67%

M2 :
- Precision: 100%
- Recall: 50%

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Precision
  - the percentage of relevant items in the retrieved items
- Recall
  - the percentage of retrieved item in the relevant items

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Precision
  - the percentage of relevant items in the retrieved items
- Recall
  - the percentage of retrieved item in the relevant items

sometimes high precision but low recall

sometimes high recall but low precision

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Precision
  - the percentage of relevant items in the retrieved items
- Recall
  - the percentage of retrieved item in the relevant items
- F-measure
  - harmonic mean of precision and recall

sometimes high precision but low recall

sometimes high recall but low precision

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Precision
  - the percentage of relevant items in the retrieved items
- Recall
  - the percentage of retrieved item in the relevant items
- F-measure
  - harmonic mean of precision and recall

sometimes high precision but low recall

sometimes high recall but low precision

a single measure that trades off precision versus recall

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- Precision
  - the percentage of relevant items in the retrieved items
- Recall
  - the percentage of retrieved item in the relevant items
- F-measure
  - harmonic mean of precision and recall

sometimes high precision but low recall

sometimes high recall but low precision

a single measure that trades off precision versus recall

$$F-measure = \frac{2}{\frac{1}{precision}+\frac{1}{recall}} = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- A : Calculate the precision, recall and F-measure and discuss

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |

M1

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |

M2

M1 :

- F-measure = $\frac{2\times80\%\times67\%}{80\%+67\%} = 73\%$

M2 :

- F-measure = $\frac{2\times100\%\times50\%}{100\%+50\%} = 67\%$

# T1-Q1

## Accuracy, Precision, Recall, F-measure

- A : Calculate the precision, recall and F-measure and discuss

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |

M1

| | |
|---|---|
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |
| ● | Not Red |

M2

M1 :

- Precision: 80%
- Recall: 67%
- F-measure: 73%

M2 :

- Precision: 100%
- Recall: 50%
- F-measure: 67%

M1 :

- F-measure = $\frac{2\times80\%\times67\%}{80\%+67\%} = 73\%$

M2 :

- F-measure = $\frac{2\times100\%\times50\%}{100\%+50\%} = 67\%$

# T1-Q2

## Naïve Bayesian classifier

| RID | AGE | INCOME | STUDENT | RATING | CLASS |
|---|---|---|---|---|---|
| 1 | Youth | High | No | Fair | No |
| 2 | Youth | High | No | Excellent | No |
| 3 | Middle-aged | High | No | Fair | Yes |
| 4 | Senior | Medium | No | Fair | Yes |
| 5 | Senior | Low | Yes | Fair | Yes |
| 6 | Senior | Low | Yes | Excellent | No |
| 7 | Middle-aged | Low | Yes | Excellent | Yes |
| 8 | Youth | Medium | No | Fair | No |
| 9 | Youth | Low | Yes | Fair | Yes |
| 10 | Senior | Medium | Yes | Fair | Yes |
| 11 | Youth | Medium | Yes | Excellent | Yes |
| 12 | Middle-aged | Medium | No | Excellent | Yes |
| 13 | Middle-aged | High | Yes | Fair | Yes |
| 14 | Senior | Medium | No | Excellent | No |

# T1-Q2

## Naïve Bayesian classifier

| RID | AGE | INCOME | STUDENT | RATING | CLASS |
|---|---|---|---|---|---|
| 1 | Youth | High | No | Fair | No |
| 2 | Youth | High | No | Excellent | No |
| 3 | Middle-aged | High | No | Fair | Yes |
| 4 | Senior | Medium | No | Fair | Yes |
| 5 | Senior | Low | Yes | Fair | Yes |
| 6 | Senior | Low | Yes | Excellent | No |
| 7 | Middle-aged | Low | Yes | Excellent | Yes |
| 8 | Youth | Medium | No | Fair | No |
| 9 | Youth | Low | Yes | Fair | Yes |
| 10 | Senior | Medium | Yes | Fair | Yes |
| 11 | Youth | Medium | Yes | Excellent | Yes |
| 12 | Middle-aged | Medium | No | Excellent | Yes |
| 13 | Middle-aged | High | Yes | Fair | Yes |
| 14 | Senior | Medium | No | Excellent | No |
| X | Youth | Medium | Yes | Fair | ? |

# T1-Q2

## Naïve Bayesian classifier

- Input :$X$: $(x_1 = youth,\ x_2 = medium,\ x_3 = yes,\ x_4 = fair)$
- Output: $C$ $(yes/no)$

| RID | AGE | INCOME | STUDENT | RATING | CLASS |
|---|---|---|---|---|---|
| 1 | Youth | High | No | Fair | No |
| 2 | Youth | High | No | Excellent | No |
| 3 | Middle-aged | High | No | Fair | Yes |
| 4 | Senior | Medium | No | Fair | Yes |
| 5 | Senior | Low | Yes | Fair | Yes |
| 6 | Senior | Low | Yes | Excellent | No |
| 7 | Middle-aged | Low | Yes | Excellent | Yes |
| 8 | Youth | Medium | No | Fair | No |
| 9 | Youth | Low | Yes | Fair | Yes |
| 10 | Senior | Medium | Yes | Fair | Yes |
| 11 | Youth | Medium | Yes | Excellent | Yes |
| 12 | Middle-aged | Medium | No | Excellent | Yes |
| 13 | Middle-aged | High | Yes | Fair | Yes |
| 14 | Senior | Medium | No | Excellent | No |

# T1-Q2

## Naïve Bayesian classifier

- Input :$X$: $(x_1 = youth,\ x_2 = medium,\ x_3 = yes,\ x_4 = fair)$
- Output: $C$ $(yes/no)$
- Maximize $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$

| RID | AGE | INCOME | STUDENT | RATING | CLASS |
|---|---|---|---|---|---|
| 1 | Youth | High | No | Fair | No |
| 2 | Youth | High | No | Excellent | No |
| 3 | Middle-aged | High | No | Fair | Yes |
| 4 | Senior | Medium | No | Fair | Yes |
| 5 | Senior | Low | Yes | Fair | Yes |
| 6 | Senior | Low | Yes | Excellent | No |
| 7 | Middle-aged | Low | Yes | Excellent | Yes |
| 8 | Youth | Medium | No | Fair | No |
| 9 | Youth | Low | Yes | Fair | Yes |
| 10 | Senior | Medium | Yes | Fair | Yes |
| 11 | Youth | Medium | Yes | Excellent | Yes |
| 12 | Middle-aged | Medium | No | Excellent | Yes |
| 13 | Middle-aged | High | Yes | Fair | Yes |
| 14 | Senior | Medium | No | Excellent | No |

# T1-Q2

## Naïve Bayesian classifier

- Input :$X$: $(x_1 = youth,\ x_2 = medium,\ x_3 = yes,\ x_4 = fair)$
- Output: $C$ $(yes/no)$
- Maximize $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$
- $P(C = yes) = \frac{9}{14} = 0.643$
- $P(C = no) = \frac{5}{14} = 0.357$

| RID | AGE | INCOME | STUDENT | RATING | CLASS |
|---|---|---|---|---|---|
| 1 | Youth | High | No | Fair | No |
| 2 | Youth | High | No | Excellent | No |
| 3 | Middle-aged | High | No | Fair | Yes |
| 4 | Senior | Medium | No | Fair | Yes |
| 5 | Senior | Low | Yes | Fair | Yes |
| 6 | Senior | Low | Yes | Excellent | No |
| 7 | Middle-aged | Low | Yes | Excellent | Yes |
| 8 | Youth | Medium | No | Fair | No |
| 9 | Youth | Low | Yes | Fair | Yes |
| 10 | Senior | Medium | Yes | Fair | Yes |
| 11 | Youth | Medium | Yes | Excellent | Yes |
| 12 | Middle-aged | Medium | No | Excellent | Yes |
| 13 | Middle-aged | High | Yes | Fair | Yes |
| 14 | Senior | Medium | No | Excellent | No |

# T1-Q2

## Naïve Bayesian classifier

- Input :$X$: $(x_1 = youth,\ x_2 = medium,\ x_3 = yes,\ x_4 = fair)$
- Output: $C$ $(yes/no)$
- Maximize $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$
- $P(C = yes) = \frac{9}{14} = 0.643$
- $P(C = no) = \frac{5}{14} = 0.357$

| RID | AGE | INCOME | STUDENT | RATING | CLASS |
|---|---|---|---|---|---|
| 1 | Youth | High | No | Fair | No |
| 2 | Youth | High | No | Excellent | No |
| 3 | Middle-aged | High | No | Fair | Yes |
| 4 | Senior | Medium | No | Fair | Yes |
| 5 | Senior | Low | Yes | Fair | Yes |
| 6 | Senior | Low | Yes | Excellent | No |
| 7 | Middle-aged | Low | Yes | Excellent | Yes |
| 8 | Youth | Medium | No | Fair | No |
| 9 | Youth | Low | Yes | Fair | Yes |
| 10 | Senior | Medium | Yes | Fair | Yes |
| 11 | Youth | Medium | Yes | Excellent | Yes |
| 12 | Middle-aged | Medium | No | Excellent | Yes |
| 13 | Middle-aged | High | Yes | Fair | Yes |
| 14 | Senior | Medium | No | Excellent | No |

# T1-Q2

## Naïve Bayesian classifier

- Input :$X$: $(x_1 = youth,\ x_2 = medium,\ x_3 = yes,\ x_4 = fair)$
- Output: $C$ $(yes/no)$
- Maximize $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$
- $P(C = yes) = \frac{9}{14} = 0.643$
- $P(C = no) = \frac{5}{14} = 0.357$

| RID | AGE | INCOME | STUDENT | RATING | CLASS |
|---|---|---|---|---|---|
| 1 | Youth | High | No | Fair | No |
| 2 | Youth | High | No | Excellent | No |
| 3 | Middle-aged | High | No | Fair | Yes |
| 4 | Senior | Medium | No | Fair | Yes |
| 5 | Senior | Low | Yes | Fair | Yes |
| 6 | Senior | Low | Yes | Excellent | No |
| 7 | Middle-aged | Low | Yes | Excellent | Yes |
| 8 | Youth | Medium | No | Fair | No |
| 9 | Youth | Low | Yes | Fair | Yes |
| 10 | Senior | Medium | Yes | Fair | Yes |
| 11 | Youth | Medium | Yes | Excellent | Yes |
| 12 | Middle-aged | Medium | No | Excellent | Yes |
| 13 | Middle-aged | High | Yes | Fair | Yes |
| 14 | Senior | Medium | No | Excellent | No |

# T1-Q2

## Naïve Bayesian classifier

- $P(x_1 = youth|C = yes) = \frac{2}{9} = 0.222$
- $P(x_1 = youth|C = no) = \frac{3}{5} = 0.600$
- $P(x_2 = medium|C = yes) = \frac{4}{9} = 0.444$
- $P(x_2 = medium|C = no) = \frac{2}{5} = 0.400$
- $P(x_3 = yes|C = yes) = \frac{6}{9} = 0.667$
- $P(x_3 = yes|C = no) = \frac{1}{5} = 0.2$
- $P(x_4 = fair|C = yes) = \frac{6}{9} = 0.667$
- $P(x_4 = fair|C = no) = \frac{2}{5} = 0.400$

| RID | AGE | INCOME | STUDENT | RATING | CLASS |
|---|---|---|---|---|---|
| 1 | Youth | High | No | Fair | No |
| 2 | Youth | High | No | Excellent | No |
| 3 | Middle-aged | High | No | Fair | Yes |
| 4 | Senior | Medium | No | Fair | Yes |
| 5 | Senior | Low | Yes | Fair | Yes |
| 6 | Senior | Low | Yes | Excellent | No |
| 7 | Middle-aged | Low | Yes | Excellent | Yes |
| 8 | Youth | Medium | No | Fair | No |
| 9 | Youth | Low | Yes | Fair | Yes |
| 10 | Senior | Medium | Yes | Fair | Yes |
| 11 | Youth | Medium | Yes | Excellent | Yes |
| 12 | Middle-aged | Medium | No | Excellent | Yes |
| 13 | Middle-aged | High | Yes | Fair | Yes |
| 14 | Senior | Medium | No | Excellent | No |

# T1-Q2

## Naïve Bayesian classifier

- $P(x_1 = youth|C = yes) = \frac{2}{9} = 0.222$
- $P(x_1 = youth|C = no) = \frac{3}{5} = 0.600$
- $P(x_2 = medium|C = yes) = \frac{4}{9} = 0.444$
- $P(x_2 = medium|C = no) = \frac{2}{5} = 0.400$
- $P(x_3 = yes|C = yes) = \frac{6}{9} = 0.667$
- $P(x_3 = yes|C = no) = \frac{1}{5} = 0.2$
- $P(x_4 = fair|C = yes) = \frac{6}{9} = 0.667$
- $P(x_4 = fair|C = no) = \frac{2}{5} = 0.400$

$$\boldsymbol{P(X|C = yes)}$$
$$= P(x_1 = youth|C = yes) * P(x_2 = medium|C = yes)$$
$$* P(x_3 = yes|C = yes) * P(x_4 = fair|C = yes)$$
$$= 0.222 * 0.444 * 0.667 * 0.667$$
$$= \mathbf{0.044}$$

| RID | AGE | INCOME | STUDENT | RATING | CLASS |
|---|---|---|---|---|---|
| 1 | Youth | High | No | Fair | No |
| 2 | Youth | High | No | Excellent | No |
| 3 | Middle-aged | High | No | Fair | Yes |
| 4 | Senior | Medium | No | Fair | Yes |
| 5 | Senior | Low | Yes | Fair | Yes |
| 6 | Senior | Low | Yes | Excellent | No |
| 7 | Middle-aged | Low | Yes | Excellent | Yes |
| 8 | Youth | Medium | No | Fair | No |
| 9 | Youth | Low | Yes | Fair | Yes |
| 10 | Senior | Medium | Yes | Fair | Yes |
| 11 | Youth | Medium | Yes | Excellent | Yes |
| 12 | Middle-aged | Medium | No | Excellent | Yes |
| 13 | Middle-aged | High | Yes | Fair | Yes |
| 14 | Senior | Medium | No | Excellent | No |

# T1-Q2

## Naïve Bayesian classifier

- $P(x_1 = youth|C = yes) = \frac{2}{9} = 0.222$
- $P(x_1 = youth|C = no) = \frac{3}{5} = 0.600$
- $P(x_2 = medium|C = yes) = \frac{4}{9} = 0.444$
- $P(x_2 = medium|C = no) = \frac{2}{5} = 0.400$
- $P(x_3 = yes|C = yes) = \frac{6}{9} = 0.667$
- $P(x_3 = yes|C = no) = \frac{1}{5} = 0.200$
- $P(x_4 = fair|C = yes) = \frac{6}{9} = 0.667$
- $P(x_4 = fair|C = no) = \frac{2}{5} = 0.400$

$$\boldsymbol{P(X|C = no)}$$
$$= P(x_1 = youth|C = no) * P(x_2 = medium|C = no)$$
$$* P(x_3 = yes|C = no) * P(x_4 = fair|C = no)$$
$$= 0.600 * 0.400 * 0.200 * 0.400$$
$$= \mathbf{0.019}$$

| RID | AGE | INCOME | STUDENT | RATING | CLASS |
|---|---|---|---|---|---|
| 1 | Youth | High | No | Fair | No |
| 2 | Youth | High | No | Excellent | No |
| 3 | Middle-aged | High | No | Fair | Yes |
| 4 | Senior | Medium | No | Fair | Yes |
| 5 | Senior | Low | Yes | Fair | Yes |
| 6 | Senior | Low | Yes | Excellent | No |
| 7 | Middle-aged | Low | Yes | Excellent | Yes |
| 8 | Youth | Medium | No | Fair | No |
| 9 | Youth | Low | Yes | Fair | Yes |
| 10 | Senior | Medium | Yes | Fair | Yes |
| 11 | Youth | Medium | Yes | Excellent | Yes |
| 12 | Middle-aged | Medium | No | Excellent | Yes |
| 13 | Middle-aged | High | Yes | Fair | Yes |
| 14 | Senior | Medium | No | Excellent | No |

# T1-Q2

## Naïve Bayesian classifier

- To find the class, $C_i$, that maximizes $P(X|C_i)P(C_i)$, we compute
  - $P(X|C = yes)P(C = yes) = 0.044 * 0.643 = 0.028$
  - $P(X|C = no)P(C = no) = 0.019 * 0.357 = 0.007$
- Therefore, the naïve Bayesian classifier predicts yes for tuple X

# Naïve Bayesian Classifier

## Summary

- If we have an n-D attribute vector $X = (x_1, x_2, \dots, x_n)$ and there are $m$ classes $C_1, \dots, C_m$
- Bayesian theorem:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

- A simplified assumption: attributes are conditionally independent, then

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) * P(x_2|C_i) *\dots* P(x_n|C_i)$$

# Naïve Bayesian Classifier

## Summary

- If we have an n-D attribute vector $X = (x_1, x_2, \dots, x_n)$ and there are $m$ classes $C_1, \dots, C_m$
- Bayesian theorem:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

- A simplified assumption: attributes are conditionally independent, then

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) * P(x_2|C_i) * \dots * P(x_n|C_i)$$

- Advantages:
  - Easy to implement
  - Effective and easy to understand
  - Good results obtained in most of the case, even with small dataset
- Disadvantages:
  - Assumption of "independency"

# Thanks for your attention