Домашнее задание 10

Андрей Лебедев, группа 424

Задача 1

Условие

1. **Первая рубрика**:
   - Экспертная классификация: 75 документов.
   - Система классифицировала 100 документов, из них 50 правильно.
2. **Вторая рубрика**:
   - Экспертная классификация: 50 документов.
   - Система классифицировала 40 документов, из них 30 правильно.

Решение

Макро-усредненные характеристики

Точность

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Для первой рубрики:

$$P_1 = \frac{50}{100} = 0.5$$

Для второй рубрики:

$$P_2 = \frac{30}{40} = 0.75$$

Макро-точность:

$$P_{\text{macro}} = \frac{P_1 + P_2}{2} = \frac{0.5 + 0.75}{2} = 0.625$$

Полнота

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Для первой рубрики:

$$R_1 = \frac{50}{75} \approx 0.667$$

Для второй рубрики:

$$R_2 = \frac{30}{50} = 0.6$$

Макро-полнота:

$$R_{\text{macro}} = \frac{R_1 + R_2}{2} = \frac{0.667 + 0.6}{2} \approx 0.6335$$

F-мера

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

Для первой рубрики:

$$F_1 = \frac{2 \cdot 0.5 \cdot 0.667}{0.5 + 0.667} \approx 0.571$$

Для второй рубрики:

$$F_2 = \frac{2 \cdot 0.75 \cdot 0.6}{0.75 + 0.6} \approx 0.667$$

Макро F-мера:

$$F_{\text{macro}} = \frac{F_1 + F_2}{2} = \frac{0.571 + 0.667}{2} \approx 0.619$$

Микро-усредненные характеристики

- Суммарное количество TP: $50 + 30 = 80$
- Суммарное количество FP: $(100 - 50) + (40 - 30) = 60$
- Суммарное количество FN: $(75 - 50) + (50 - 30) = 45$

Микро-точность

$$P_{\text{micro}} = \frac{\text{TP}_{\text{sum}}}{\text{TP}_{\text{sum}} + \text{FP}_{\text{sum}}} = \frac{80}{80 + 60} = \frac{80}{140} \approx 0.571$$

Микро-полнота

$$R_{\text{micro}} = \frac{\text{TP}_{\text{sum}}}{\text{TP}_{\text{sum}} + \text{FN}_{\text{sum}}} = \frac{80}{80 + 45} = \frac{80}{125} = 0.64$$

Микро F-мера

$$F_{\text{micro}} = \frac{2 \cdot P_{\text{micro}} \cdot R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}} = \frac{2 \cdot 0.571 \cdot 0.64}{0.571 + 0.64} \approx 0.603$$

Ответ

- **Макро-точность**: 0.625
- **Макро-полнота**: 0.6335
- **Макро F-мера**: 0.619
- **Микро-точность**: 0.571
- **Микро-полнота**: 0.64
- **Микро F-мера**: 0.603

Задача 2

Дано

- Класс Россия:
  - $D1$ = (Москва, Питер, Питер)
  - $D2$ = (Питер, Российский, Питер)
  - $D3$ = (Питер, Казань)
- Класс США:
  - $D4$ = (Москва, Питер, Айдахо)
- Документ для классификации:

$$D5 = (\text{Питер, Питер, Питер, Москва, Айдахо})$$

Решение

Составим частотный словарь для каждого класса.

Для России:

$$\text{tf}_{\text{Москва}} = 1, \quad \text{tf}_{\text{Питер}} = 5, \quad \text{tf}_{\text{Российский}} = 1, \quad \text{tf}_{\text{Казань}} = 1, \quad \text{tf}_{\text{Айдахо}} = 0$$

Для США:

$$\text{tf}_{\text{Москва}} = 1, \quad \text{tf}_{\text{Питер}} = 1, \quad \text{tf}_{\text{Айдахо}} = 1, \quad \text{tf}_{\text{Российский}} = 0, \quad \text{tf}_{\text{Казань}} = 0$$

Априорные вероятности: Для каждого класса вычислим априорные вероятности:

$$P(\text{Россия}) = \frac{3}{4}, \quad P(\text{США}) = \frac{1}{4}$$

Логарифмы:

$$\log P(\text{Россия}) = \log \frac{3}{4} \approx -0.287, \quad \log P(\text{США}) = \log \frac{1}{4} \approx -1.386$$

Multinomial Naive Bayes

Вероятности для России:

$$P(\text{Питер}|\text{Россия}) = \frac{\text{tf}_{\text{Питер}} + 1}{\sum \text{tf} + |V|} = \frac{5+1}{8+5} \approx 0.46$$

$$P(\text{Москва}|\text{Россия}) = \frac{\text{tf}_{\text{Москва}} + 1}{\sum \text{tf} + |V|} = \frac{1+1}{8+5} \approx 0.15$$

$$P(\text{Айдахо}|\text{Россия}) = \frac{\text{tf}_{\text{Айдахо}} + 1}{\sum \text{tf} + |V|} = \frac{0+1}{8+5} \approx 0.08$$

Вероятности для США:

$$P(\text{Питер}|\text{США}) = \frac{\text{tf}_{\text{Питер}} + 1}{\sum \text{tf} + |V|} = \frac{1+1}{3+5} = 0.25$$

$$P(\text{Москва}|\text{США}) = \frac{\text{tf}_{\text{Москва}} + 1}{\sum \text{tf} + |V|} = \frac{1+1}{3+5} = 0.25$$

$$P(\text{Айдахо}|\text{США}) = \frac{\text{tf}_{\text{Айдахо}} + 1}{\sum \text{tf} + |V|} = \frac{1+1}{3+5} = 0.25$$

Итоговые вероятности:

$$\log P(D5|\text{Россия}) = 3 \cdot \log(0.46) + \log(0.15) + \log(0.08) + \log P(\text{Россия}) \approx -3.06$$

$$\log P(D5|\text{США}) = 3 \cdot \log(0.25) + \log(0.25) + \log(0.25) + \log P(\text{США}) \approx -3.61$$

Bernoulli Naive Bayes

Для России:

$$P(\text{Питер}|\text{Россия}) = \frac{3+1}{3+2} = 0.8,$$

$$P(\text{Москва}|\text{Россия}) = \frac{1+1}{3+2} = 0.4,$$

$$P(\text{Айдахо}|\text{Россия}) = \frac{0+1}{3+2} = 0.2$$

Для США:

$$P(\text{Питер}|\text{США}) = \frac{1+1}{1+2} = 0.667,$$

$$P(\text{Москва}|\text{США}) = \frac{1+1}{1+2} = 0.667,$$

$$P(\text{Айдахо}|\text{США}) = \frac{1+1}{1+2} = 0.667$$

Итоговые вероятности:

$$\log P(D5|\text{Россия}) = \log(0.8) + \log(0.4) + \log(0.2) + \log(1-0.4) + \log(1-0.4) + \log P(\text{Россия}) \approx -1.76$$

$$\log P(D5|\text{США}) = \log(0.667) + \log(0.667) + \log(0.667) + \log(1-0.333) + \log(1-0.333) + \log P(\text{США}) \approx$$
$$\approx -1.47$$

Заключение

- Для Multinomial Naive Bayes: документ $D5$ относится к классу «Россия» ($\log P(D5|\text{Россия}) > \log P(D5|\text{США})$).
- Для Bernoulli Naive Bayes: документ $D5$ относится к классу «США» ($\log P(D5|\text{США}) > \log P(D5|\text{Россия})$).

Таким образом, результат зависит от выбранной модели.