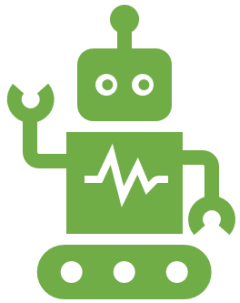
Универзитет „Св. Кирил и Методиј“ во Скопје
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И КОМПЈУТЕРСКО ИНЖЕНЕРСТВО

Наивен Баесов Класификатор



Аудиториски вежби по курсот
Вештачка интелигенција

2023/2024

Типови алгоритми

- Надгледувано учење – учење од записи користејќи класен атрибут
- Ненадгледувано учење – учење од записи каде немаме класен атрибут
- Учење со поттикнување – тренирање на модел да прави специфични одлуки

Класификатори

- Класификација – изгради функција $Y = f(X)$ која зема вектор од карактеристики X (влезови) и ја предвидува лабелата Y (класа или излез)

- Карактеристиките ни се познати, лабелите треба да ги предвидиме

- Пример вектори

Gender	High Temperature	Coughing	Respiratory problems	Corona virus
M	Yes	Yes	Yes	Yes
F	Yes	Yes	No	No

Податочно множество

- Податочно множество е збирот на сите вектори (записи) кои ни се достапни

- За да го евалуираме класификаторот, множеството се дели на два дела:
  - тренинг множество – со ова множество се тренира класификаторот
  - тест множество – со ова множество се евалуира класификаторот

Баесов класификатор

- Пример за надгледувано учење
- Статистички класификатор
- Базиран на Баесова теорема
- Инкрементален – со секој нов запис може да се зголеми/намали веројатноста дека хипотезата е точна

Баесова Теорема

- Условна веројатност

$$P(C \mid A) = \frac{P(A, C)}{P(A)}$$

- Баесова теорема

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Пример на баесова теорема

- Докторот знае дека корона вирус предизвикува покачена телесна температура во 50% од случаите
- Веројатноста пациент да има корона е 1/50,000
- Веројатноста пациент да има покачена температура е 1/1000

- Ако некој пациент има покачена температура, која е веројатноста да има корона вирус?

- $$P(C|T) = \frac{P(T|C) * P(C)}{P(T)} = \frac{0.5 * \frac{1}{50000}}{\frac{1}{1000}} = 0.01$$

Баесов класификатор

- Секој атрибут и класната лабела ги смета како случајни променливи

- Ако имаме даден запис со атрибути $(X_1,X_2,\ldots,X_n)$
  - целта ни е да ја предвидиме класата C
  - поточно, сакаме да ја најдеме C која има максимална вредност за $P(C \mid X_1,X_2,\ldots,X_n)$

- $P(C \mid X_1,X_2,\ldots,Xn) = \frac{P(x_1,x_2,\ldots,xn \mid c) * P(c)}{P(x_1,x_2,\ldots,xn)}$

Наивен Баесов класификатор

- Претпоставува независност помеѓу атрибутите X_i
  - $P(x_1,x_2,...,xn \mid c) = P(x_1 \mid c) * P(x_2 \mid c) ... P(x_n \mid c)$
- Записот припаѓа на онаа класа за која има највисока веројатност
- Многу едноставно работи кога немаме непрекинати вредности

Модулот sklearn

- Инсталација:

```
>>pip install scikit-learn
```

- Содржи голем број на готови имплементации на модели на машинско учење, како и бројни методи за претпроцесирање и евалуација

- Ќе го користиме подмодулот **naive_bayes**

Categorical Naive Bayes

- Имплементација на наивен баесов класификатор кој работи **само над категориски атрибути**
- Метода **fit()** – тренирање на моделот
- Метода **predict()** – прави класификација на запис
- Метода **predict_proba()** – ги враќа веројатностите запис да припаѓа во секоја класа

Gaussian Naive Bayes

- Categorical Naive Bayes работи само над категориски атрибути

- Доколку имаме непрекинати вредности, треба да ги дискретизираме вредностите
  - Подели го рангот на вредности во бинови. Постави една вредност да биде репрезентација на бинот
  - Подели на два дела $(X < v)$ или $(X > v)$

Gaussian Naive Bayes

- Гаусовиот наивен баесов класификатор работи над непрекинати вредности
- Претпоставува дека секој атрибут има нормална распределба
- Методите се исти како и кај Categorical Naive Bayes

Евалуација на класификатор

- Точност – број на записи кои се точно предвидени од вкупниот број на записи
- Прецизност – делот од релевантни записи помеѓу вратените записи
- Одзив - делот од релевантни записи кои се вратени меѓу вкупниот број на релевантни записи