Търсене и извличане на информация. Приложение на дълбоко машинно обучение

Стоян Михов

Лекция 2: Класифициране на документи. Вероятностни модели. Принцип на максимално правдоподобие. Наивен Бейсов класификатор за Бернулиев документен модел.

План на лекцията

1. Формалности за курса (5 мин)
2. Задача класифициране на документи (15 мин)
3. Оценяване на информационна система — прецизност, обхват и F-оценка (10 мин)
4. Вероятностно пространство, събития, величини (30 мин)
5. Принцип на максимално правдоподобие (10 мин)
6. Наивен Бейсов класификатор за Бернулиев документен модел (20 мин)

Формалности

- Слайдове и материали от упражненията се качват в Moodle
- Позволява се използването на преносими компютри по време на упражненията
- Не всички теми са развити по същия начин в трите учебника, които препоръчах — следете лекциите
  - първата лекция е компилация на глави 1, 2 и 3 на първия учебник
  - втората лекция са базира на глава 13 и 14 от първия учебник
- Първото домашно ще бъде дадено след 2 седмици

План на лекцията

1. Формалности за курса (5 мин)
2. **Задача класифициране на документи (15 мин)**
3. Оценяване на информационна система — прецизност, обхват и F-оценка (10 мин)
4. Вероятностно пространство, събития, величини (30 мин)
5. Принцип на максимално правдоподобие (10 мин)
6. Наивен Бейсов класификатор за Бернулиев документен модел (20 мин)

Информационни нужди, при които се налага класифициране на информация

- Филтриране на поток от информация:
  - Google allert;
  - Спам филтър на електронна поща;
  - Разпознаване на фалшиви новини, ревюта, постове и т.н.
  - Класифициране на пристигащи писма (лични, служебни, …);
  - Блокиране на страници с нежелано съдържание (защита на малолетни).
- Структуриране на колекция от документи по съдържание:
  - Библиотечни каталози;
  - Разбиване по теми и категории (онлайн новинарски сайтове);
  - Класифициране емоционален заряд (sentiment) — положителен, отрицателен неутрален (при рецензии, оценки, коментари, …).

Подходи за класифициране на документи

- Ръчно от хора (например библиотекарите в библиотеките)
  - сравнително прецизно и надежно,
  - но бавно, скъпо и трудно за скалиране.
- Чрез създаване на правила — например разширени булеви изрази върху ключови думи (позиционни, с шаблони и близости)
  - прецизността и обхватът зависят силно от експертността на създателите на правилата,
  - трудни и скъпи за разработване и поддържане.
- Чрез машинно обучение — от наличен корпус от класифицирани документи автоматично се “обучава” класификатор
  - при определени условия може да достигне добра прецизност и обхват,
  - поддържането е лесно и сравнително евтино.

Постановка на задачата

- Дадени са:
  - фиксирано множество от класове: $\mathbb{C} = \{c_1, c_2, \ldots, c_j\}$,
  - пространство от представяния на документите $\mathbb{X}$ (обикновено това е многомерно векторно пространство).
- Търсим класификатор: $\gamma : \mathbb{X} \to \mathbb{C}$
  - на документа представен с $d \in \mathbb{X}$ класификаторът γ съпоставя класа $\gamma(d) \in \mathbb{C}$.

Формална постановка на машинното обучение на класификатор

- Дадени са документно пространство $\mathbb{X}$ и множество от класове $\mathbb{C}$.
- Корпус за обучение (трениране) $\mathbb{D} \subset \mathbb{X} \times \mathbb{C}$ е крайна последователност от двойки документ/клас $(d, c) \in \mathbb{X} \times \mathbb{C}$.
- Метод (алгоритъм) за обучение на класификатор ще наричаме функция Γ, която по даден корпус за обучение $\mathbb{D}$ връща класифицираща функция $\gamma = \Gamma(\mathbb{D})$, т.е. $\Gamma(\mathbb{D}) : \mathbb{X} \rightarrow \mathbb{C}$.

План на лекцията

1. Формалности за курса (5 мин)
2. Задача класифициране на документи (15 мин)
3. **Оценяване на информационна система — прецизност, обхват и F-оценка (10 мин)**
4. Вероятностно пространство, събития, величини (30 мин)
5. Принцип на максимално правдоподобие (10 мин)
6. Наивен Бейсов класификатор за Бернулиев документен модел (20 мин)

Оценяване на информационна система

Проблем: Как да оценим ефективността на дадена информационна система?

За да оценим една информационна система са ни необходими:

- Корпус от документи,
- Множество от информационни нужди — заявки или класове,
- Оценка за релевантност на документите спрямо всяка една от информационните нужди.

Оценяване на система без ранкиране на документите

- $\text{Прецизност} = \dfrac{\#(\text{извлечени релевантни документи})}{\#(\text{всички извлечени документи})}$

- $\text{Обхват} = \dfrac{\#(\text{извлечени релевантни документи})}{\#(\text{всички релевантни документи})}$

	Релевантни	Нерелевантни
Извлечени	верни позитивни (tp)	грешни позитивни (fp)
Неизвлечени	грешни негативни (fn)	верни негативни (tn)

$$P = \frac{tp}{tp + fp}$$

- $R = \dfrac{tp}{tp + fn}$

F-оценка (F-score)

- Представлява претеглено средно хармонично между прецизността и обхвата.
- $$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R},$$
- където $\beta^2 = \dfrac{1-\alpha}{\alpha}$
- Най-често се използва $\beta = 1$ т.е. $\alpha = \dfrac{1}{2}$. В такъв случай се използва означението $F_{\beta=1}$ или $F1$ или само F.

План на лекцията

1. Формалности за курса (5 мин)
2. Задача класифициране на документи (15 мин)
3. Оценяване на информационна система — прецизност, обхват и F-оценка (10 мин)
4. **Вероятностно пространство, събития, величини (30 мин)**
5. Принцип на максимално правдоподобие (10 мин)
6. Наивен Бейсов класификатор за Бернулиев документен модел (20 мин)

Формална дефиниция

- **Вероятностно пространство** Ω — множество от всички възможни елементарни събития.

- σ **- алгебра** $\mathscr{F}$ — множество от подмножества на Ω, съдържащо Ω и затворено относно допълнение и изброими обединения. Елементите на $\mathscr{F}$ наричаме събития (не елементарни) и бележим с главни букви A, B, C

- **Вероятностно разпределение** $\Pr$ — функция от $\mathscr{F}$ в $[0,1]$, така че:

  - $\Pr[\Omega] = 1$

  - За всяка крайна последователност от непресичащи се (дизюнктни) събития $A_1, A_2, \ldots, A_n$ е изпълнено: $\Pr[A_1 \cup \ldots \cup A_n] = \sum_{i=1}^{n} \Pr[A_i]$

Пример

- Честен зар:
  - $\Omega = \{1,2,3,4,5,6\}$
  - $\mathscr{F} = 2^{\Omega}$
  - $\Pr[A] = \dfrac{|A|}{6}$
- Ако събитието A дефинираме като числото получено при хвърляне на зара да по-малко от 5, то A е с вероятност $\Pr[A] = \dfrac{2}{3}$.

- Вероятностно пространство Ω наричаме **дискретно,** ако Ω е крайно или изброимо. В такъв случай ще предполагаме, че синглетоните на Ω са събития, от което следва, че $\mathscr{F} = 2^{\Omega}$.

- Засега разглеждаме само дискретни вероятностни пространства.

- **Случайна величина (променлива)** наричаме функция $X : \Omega \to \mathbb{R}$, така че за всеки реален интервал $I \subset \mathbb{R}$ първообразът му е събитие т.е. $X^{-1}(I) \in \mathscr{F}$ (винаги е изпълнено при дискретни вероятностни пространства).

- **Функция на разпределение на дискретна случайна величина** наричаме функцията: $x \mapsto \Pr[X = x]$.

- **Съвместна функция на разпределение на дискретните случайни величини** X, Y наричаме функцията: $(x, y) \mapsto \Pr[X = x, Y = y]$.

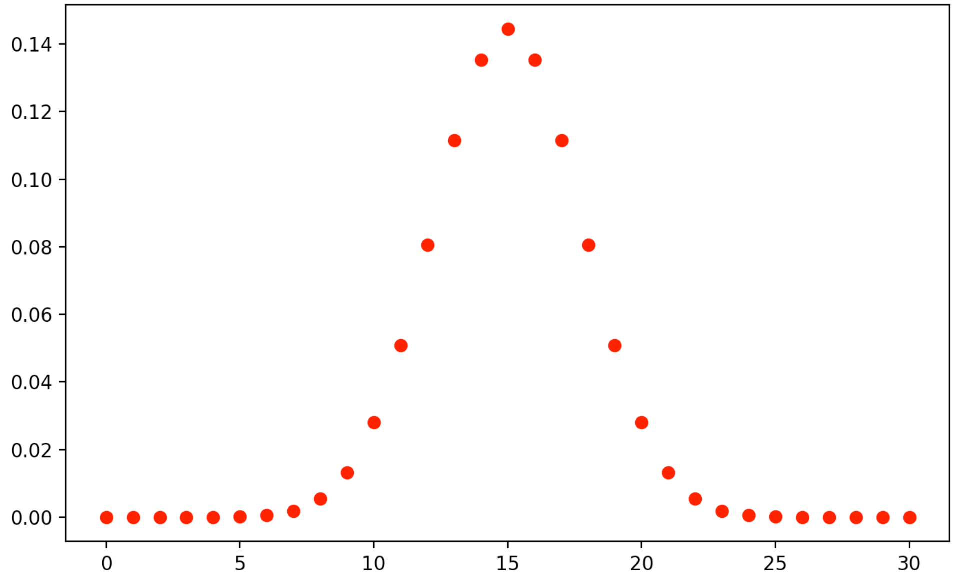
Пример за разпределение на дискретна случайна величина

- Вероятностно пространство: всички възможни резултати при хвърляне на n монети.

- Случайна величина X: брой хвърляния на ези.

- **Биномно разпределение**: $B(n, p)$

$$\Pr[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$



- Частен случай: при $n = 1$ получаваме **Бернулиевото разпределение**

Условни вероятности и независимост

- **Условната вероятност** на събитието A при условие събитието B е дефинирана като
$\Pr[A \,|\, B] = \dfrac{\Pr[A \cap B]}{\Pr[B]}$, когато $\Pr[B] \neq 0$

- Събитията A и B наричаме **независими,** ако
$\Pr[A \cap B] = \Pr[A]\Pr[B]$

- Ако събитията A и B са независими, то
$\Pr[A \,|\, B] = \Pr[A]$

- Последователност от случайни величини наричаме **независими и еднакво разпределени,** ако са взаимно независими и имат една и съща функция на разпределение

Основни свойства

- $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$ **(Правило за сумиране)**
- $\Pr[\bigcup_{i=1}^{n} A_i] \leq \sum_{i=1}^{n} \Pr[A_i]$ **(Горна граница на обединение)**
- $\Pr[A \mid B] = \dfrac{\Pr[B \mid A] \Pr[A]}{\Pr[B]}$ **(Формула на Бейс)**
- $\Pr[\bigcap_{i=1}^{n} A_i] = \Pr[A_1] \Pr[A_2 \mid A_1] \Pr[A_3 \mid A_1 \cap A_2] \ldots \Pr[A_n \mid \bigcap_{i=1}^{n-1} A_i]$ **(Верижно правило)**
- $\Omega = A_1 \cup A_2 \cup \ldots \cup A_n$, където $A_i \cap A_j = \varnothing$ за $i \neq j \Rightarrow$

  $\Pr[B] = \sum_{i=1}^{n} \Pr[B \mid A_i] \Pr[A_i]$ **(Теорема за пълна вероятност)**

План на лекцията

1. Формалности за курса (5 мин)
2. Задача класифициране на документи (15 мин)
3. Оценяване на информационна система — прецизност, обхват и F-оценка (10 мин)
4. Вероятностно пространство, събития, величини (30 мин)
5. **Принцип на максимално правдоподобие (10 мин)**
6. Наивен Бейсов класификатор за Бернулиев документен модел (20 мин)

Принцип на максималното правдоподобие

- Дадена е последователност от m независими и еднакво разпределени случайни величини $X_1, X_2, \ldots, X_m$ с функция на разпределение $\Pr[X = x \mid \theta]$, която зависи от параметър θ.

- Наблюдавали (измерили) сме съответни стойности $x_1, x_2, \ldots, x_m$ за последователността от случайните величини $X_1, X_2, \ldots, X_m$.

- Правдоподобието да сме направили съответното наблюдение е:

  $$L(\theta) = \Pr[X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m \mid \theta] = \prod_{i=1}^{m} \Pr[X_i = x_i \mid \theta]$$

- Максимално правдоподобие получаваме, като намерим за каква стойност на θ правдободобието $L(\theta)$ е максимално. Т.е.

- $$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

Максимизиране на правдоподобието при биномно разпределение

- Предполагаме биномна функция на разпределение $B(n, p)$ на m н.е.р случайни величини $X_1, X_2, \ldots, X_m$ с наблюдения $x_1, x_2, \ldots, x_m$. Т.е. $\Pr[X = x \,|\, p] = \binom{n}{x} p^x (1-p)^{n-x}$

- Нека измежду $x_1, x_2, \ldots, x_m$ има f_k на брой стойности k. Търсим:
$\hat{p} = \arg\max_p L(p) = \arg\max_p \log L(p) =$

$$= \sum_{k=1}^{n} f_k \left(\log \binom{n}{k} + k \log p + (n-k) \log(1-p) \right)$$

$$\frac{\partial \log L(p)}{\partial p} = \sum_{k=1}^{n} \left(\frac{k f_k}{p} - \frac{(n-k) f_k}{1-p} \right) = 0$$

$$(1-p) \sum_{k=1}^{n} k f_k = p \sum_{k=1}^{n} (n-k) f_k$$

- $$\hat{p} = \frac{1}{n} \sum_{i=1}^{m} \frac{x_i}{m}$$

План на лекцията

1. Формалности за курса (5 мин)
2. Задача класифициране на документи (15 мин)
3. Оценяване на информационна система — прецизност, обхват и F-оценка (10 мин)
4. Вероятностно пространство, събития, величини (30 мин)
5. Принцип на максимално правдоподобие (10 мин)
6. **Наивен Бейсов класификатор за Бернулиев документен модел (20 мин)**

Бернулиев документен модел

- Нека е даден речник $V = \{t_1, t_2, \ldots, t_M\}$.

- На всеки документ съпоставяме M-мерен вектор от нули и единици $d = (e_1, e_2, \ldots, e_M)$, където $e_i = 1$ ако термът t_i се среща в документа и $e_i = 0$, в противен случай. Т.е. $\mathbb{X} = \{0,1\}^M$.

- Предполагаме, че U_i са <u>взаимно независими</u> случайни величини с бернулиеви разпределения, такива че U_i ни дава i-тата проекция на елементите на $\mathbb{X}$.

- В такъв случай:

$$\Pr[d] = \Pr[(e_1, e_2, \ldots, e_M)] = \Pr[U_1 = e_1, U_2 = e_2, \ldots, U_M = e_M] = \prod_{i=1}^{M} \Pr[U_i = e_i]$$

- Търсим най-вероятния клас c при условие, че имаме документ d. Т.е. търсим $c_{MAP} = \arg\max_{c \in \mathbb{C}} \Pr[c \mid d]$
- MAP = maximum a posteriori
- $$\Pr[c \mid d] = \frac{\Pr[d \mid c] \Pr[c]}{\Pr[d]}$$
- $$\Pr[d \mid c] = \Pr[(e_1, e_2, \ldots, e_M) \mid c] = \prod_{i=1}^{M} \Pr[U_i = e_i \mid c]$$
- $$c_{MAP} = \arg\max_{c \in \mathbb{C}} \Pr[c] \prod_{i=1}^{M} \Pr[U_i = e_i \mid c]$$
- $$c_{MAP} = \arg\max_{c \in \mathbb{C}} \log \Pr[c] + \sum_{i=1}^{M} \log \Pr[U_i = e_i \mid c]$$

Оценяване на параметрите използвайки принципа за <u>максималното правдоподобие</u>

- N - брой документи в $\mathbb{D}$
- N_c - брой документи в $\mathbb{D}$ от клас c
- $N_{c,t}$ - брой документи в $\mathbb{D}$ от клас c, в които се среща терма t
- $\Pr[c] \approx \frac{N_c}{N}$
- $\Pr[U_i = 1 | c] \approx \frac{N_{c,t_i}}{N_c} \approx \frac{N_{c,t_i} + 1}{N_c + 2}$
- $\Pr[U_i = 0 | c] = 1 - \Pr[U_i = 1 | c]$

Алгоритми за наивен Бейсов класификатор чрез Бернулиев документен модел

```
TrainBernoulliNB(C, D)
V <- EXTRACTVOCABULARY(D)
N <- COUNTDOCS(D)
for each c in C do
    Nc <- COUNTDOCSINCLASS(D, c)
    prior[c] <- Nc/N
    for each t in V do
        Nct <- COUNTDOCSINCLASSCONTAININGTERM(D, c, t)
        condprob[t][c] <- (Nct + 1)/(Nc + 2)
return V, prior, condprob

ApplyBernoulliNB(C, V, prior, condprob, d)
Vd <- EXTRACTTERMSFROMDOC(V, d)
for each c in C do
    score[c] <- log prior[c]
    for each t in V do
        if t in Vd then
            score[c] += log(condprob[t][c])
        else
            score[c] += log(1-condprob[t][c])
return argmax(c in C, score[c])
```

Заключение

Наивният Бейсов класификатор:

- не е много наивен,
- може да се включват и други булеви характеристики (не само ключови думи) — например при електронна поща наличие на прикачен файл, дали подателя е в контактната листа и т.н.
- сравнително надежден е и се използва широко например за филтриране на спам,
- изисква малко ресурси и много бързо се обучава и прилага — всичко се свежда до просто броене.