Търсене и извличане на информация. Приложение на дълбоко машинно обучение

Стоян Михов

Лекция 3: Мултиномен документен модел и Бейсов класификатор. Избор на характеристики. Линейни класификатори.

План на лекцията

1. **Формалности за курса (5 мин)**
2. Мултиномно разпределение (5 мин)
3. Максимизиране на правдоподобието при Мултиномно разпределение (15 мин)
4. Наивен Бейсов класификатор с мултиномен модел (20 мин)
5. Избор на характеристики (20 мин)
6. Линейни класификатори (20 мин)

Формалности

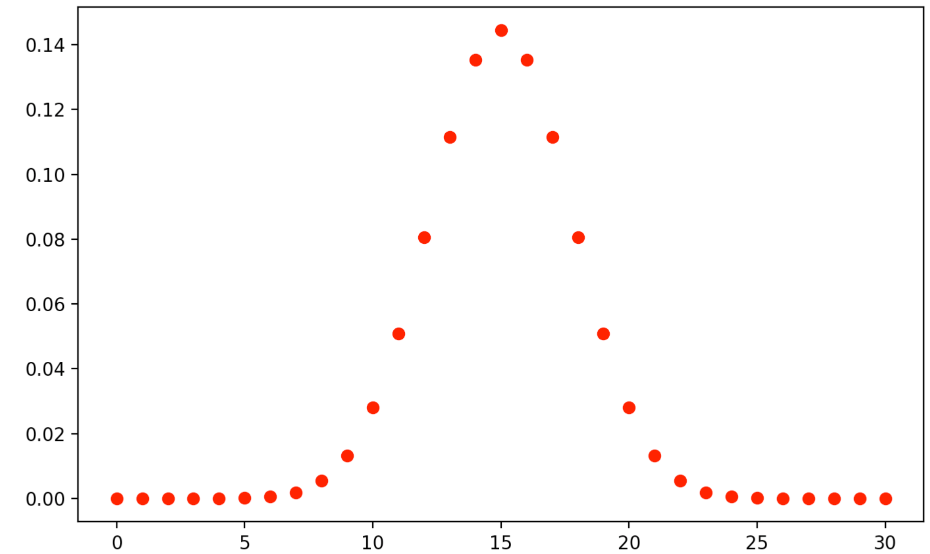
- Упражненията към курса ще се провеждат в компютърна зала 306 на ФМИ от 10:15 часа.
- Настоящата (трета) лекция също се базира на глави 13 и 14 от първия учебник.
- Поставени задачи:
  - Който реши поставена по време на лекции задача ще се радвам да ми изпрати по мейл отговора.

План на лекцията

1. Формалности за курса (5 мин)
2. **Мултиномно разпределение (5 мин)**
3. Максимизиране на правдоподобието при Мултиномно разпределение (15 мин)
4. Наивен Бейсов класификатор с мултиномен модел (20 мин)
5. Избор на характеристики (20 мин)
6. Линейни класификатори (20 мин)

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**

- Обобщение: **Мултиномно разпределение** $M(n, l, p_1, p_2, \ldots, p_l)$

  - Хвърляме n зара с l страни.

  - Случайни величини: $X_1, X_2, \ldots, X_l - X_i$ връща брой хвърляния на “i”.

$$\Pr[X_1 = k_1, X_2 = k_2, \ldots, X_l = k_l] = \frac{n!}{k_1!k_2!\ldots k_l!} p_1^{k_1} p_2^{k_2} \ldots p_l^{k_l} \text{ когато } \sum_{i=1}^{l} k_i = n.$$

План на лекцията

1. Формалности за курса (5 мин)
2. Мултиномно разпределение (5 мин)
3. **Максимизиране на правдоподобието при Мултиномно разпределение (15 мин)**
4. Наивен Бейсов класификатор с мултиномен модел (20 мин)
5. Избор на характеристики (20 мин)
6. Линейни класификатори (20 мин)

Максимизиране на правдоподобието при мултиномно разпределение

- Предполагаме биномна функция на разпределение $M(n, l, p_1, p_2, \ldots, p_l)$ на m н.е.р съвместни случайни величини $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m)}$, където $\mathbf{X}^{(i)} = (X_1^{(i)}, X_2^{(i)}, \ldots, X_l^{(i)})$ с наблюдения $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}$, където $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_l^{(i)})$. Т.е. $\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_l = x_l] = \dfrac{n!}{x_1!x_2!\ldots x_l!} p_1^{x_1} p_2^{x_2} \ldots p_l^{x_l}$

- Нека измежду $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ има $f_{k_1,k_2,\ldots,k_l}$ на брой стойности $(k_1, k_2, \ldots, k_l)$. Търсим: $\hat{\mathbf{p}} = \arg\max_{\mathbf{p}} L(\mathbf{p}) = \arg\max_{\mathbf{p}} \log L(\mathbf{p}) =$

$$= \sum_{k_1+k_2+\ldots+k_l=n} f_{k_1,k_2,\ldots,k_l} \left(\log \frac{n!}{k_1!k_2!\ldots k_l!} + k_1 \log p_1 + k_2 \log p_2 + \ldots + k_l \log p_l \right)$$

- Множител на Лагранж: $g(p_1, p_2, \ldots, p_l) = p_1 + p_2 + \ldots + p_l - 1$

$$\frac{\partial \log L(p_1, p_2, \ldots, p_l) - \lambda g(p_1, p_2, \ldots, p_l)}{\partial p_i} = 0 \Longrightarrow p_i = \frac{\sum_{k_1+k_2+\ldots+k_l=n} k_i f_{k_1,k_2,\ldots,k_l}}{\lambda}$$

- $$\frac{\partial \log L(p_1, p_2, \ldots, p_l) - \lambda g(p_1, p_2, \ldots, p_l)}{\partial \lambda} = 0 \Longrightarrow p_1 + p_2 + \ldots p_l = 1$$

- $$\hat{p}_i = \frac{1}{n} \sum_{j=1}^{m} \frac{x_i^{(j)}}{m}$$

План на лекцията

1. Формалности за курса (5 мин)
2. Мултиномно разпределение (5 мин)
3. Максимизиране на правдоподобието при Мултиномно разпределение (15 мин)
4. **Наивен Бейсов класификатор с мултиномен модел (20 мин)**
5. Избор на характеристики (20 мин)
6. Линейни класификатори (20 мин)

Мултиномен документен модел

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

- На всеки документ съпоставяме M-мерен вектор от естествени числа – $d = (f_1, f_2, \ldots, f_M)$, където f_i е броят на срещанията на терма t_i в документа. Т.е. $\mathbb{X} = \mathbb{N}^M$.

- Това представяне на документите се нарича **Bag of Words**

- Предполагаме, че при условие, че дължината на документа е $n = f_1 + \ldots + f_M$, имаме мултиномно разпределение. Т.е.:

- $\Pr[d] = \Pr[(f_1, f_2, \ldots, f_M)] = K_d \prod_{i=1}^{M} \Pr[t_i]^{f_i}$, където $K_d = \dfrac{n!}{f_1! \ldots f_M!}$

- При произволна (променлива) дължина на документ, коефицента K_d се умножава по вероятността $\Pr[|d| = n]$ – дадения документ d да има дължина n.

- Търсим най-вероятния клас c при условие, че имаме документ d. Т.е. търсим $c_{MAP} = \arg\max_{c \in \mathbb{C}} \Pr[c \mid d]$
- $$\Pr[c \mid d] = \frac{\Pr[d \mid c]\Pr[c]}{\Pr[d]}$$
- $$\Pr[d \mid c] = \Pr[(f_1, f_2, \ldots, f_M) \mid c] = K_d \prod_{i=1}^{M} \Pr[t_i \mid c]^{f_i}$$
- $$c_{MAP} = \arg\max_{c \in \mathbb{C}} \Pr[c] \prod_{i=1}^{M} \Pr[t_i \mid c]^{f_i}$$
- $$c_{MAP} = \arg\max_{c \in \mathbb{C}} \log \Pr[c] + \sum_{i=1}^{M} f_i \log \Pr[t_i \mid c] =$$
  $$= \arg\max_{c \in \mathbb{C}} \log \Pr[c] + \sum_{k=1}^{n} \log \Pr[t_{d_k} \mid c]$$

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
- $T_{c,t}$ - брой на всички срещания на терма t в документи в $\mathbb{D}$ от клас c.
- $\Pr[c] \approx \frac{N_c}{N}$
- $\Pr[t_i \,|\, c] \approx \frac{T_{c,t_i}}{\sum_{t' \in V} T_{c,t'}} \approx \frac{T_{c,t_i} + 1}{\sum_{t' \in V} T_{c,t'} + |V|}$

Алгоритми за наивен Бейсов класификатор чрез мултиномен документен модел

```
TrainMultinomialNB(C, D)
V <- EXTRACTVOCABULARY(D)
N <- COUNTDOCS(D)
for each c in C do
   Nc <- COUNTDOCSINCLASS(D, c)
   prior[c] <- Nc/N
   textc <- CONCATENATETEXTOFALLDOCSINCLASS(D, c)
   for each t in V do
      Tc[t] <- COUNTTOKENSOFTERM(textc, t)
   for each t in V do
      condprob[t][c] <- (Tc[t]+1)/sum(Tc[t’]+1 for t' in V)
return V, prior, condprob

ApplyMultinomialNB(C, V, prior, condprob, d)
W <- EXTRACTTOKENSFROMDOC(V, d)
for each c in C do
   score[c] <- log(prior[c])
   for each t in W do
      score[c] += log(condprob[t][c])
return argmax(c in C, score[c])
```

План на лекцията

1. Формалности за курса (5 мин)
2. Мултиномно разпределение (5 мин)
3. Максимизиране на правдоподобието при Мултиномно разпределение (15 мин)
4. Наивен Бейсов класификатор с мултиномен модел (20 мин)
5. **Избор на характеристики (20 мин)**
6. Линейни класификатори (20 мин)

Избор на характеристики (Feature selection) Защо?

- Редуцира се времето за трениране и прилагане на класификатора.
- Намалява се размерът на модела.
- Подобряват се качествата на модела:
  - елиминира се шум
  - намалява се опасността от пренапасване (overfitting)
  - може да подобри ефективността (F-оценката)
- Важна и нетривиална задача (Feature engeneering)

Методи за избор на характеристики

- Най-прост метод:
  - избор на термовете по честота на срещания,
  - въпреки простотата дава сравнително добри резултати.
- По-сложни (и по-ефективни) методи:
  - Мярка за взаимна информация (MI) — MI измерва доколко присъствието или отсъствието на даден терм допринася за взимането на правилното решение за класификация.
  - χ^2 тест за независимост — тества доколко две събития, в случая срещане на даден терм в документ и документа да е от даден клас, са независими.

Мярка за взаимна информация (MI)

Мярката за взаимна информация количествено определя количеството информация, получено за една случайна променлива чрез наблюдение на другата случайна променлива.

Нека U е случайна величина приемаща стойност 1, ако даден терм t се среща в документ, а C е случайна величина приемаща стойност 1, ако документът е от даден клас c. Тогава **мярката за взаимна информация** се дефинира като:

$$I(U;C) = \sum_{e_t \in \{0,1\}} \sum_{e_C \in \{0,1\}} \Pr[U = e_t, C = e_C] \ \log_2 \frac{\Pr[U = e_t, C = e_C]}{\Pr[U = e_t] \Pr[C = e_C]}$$

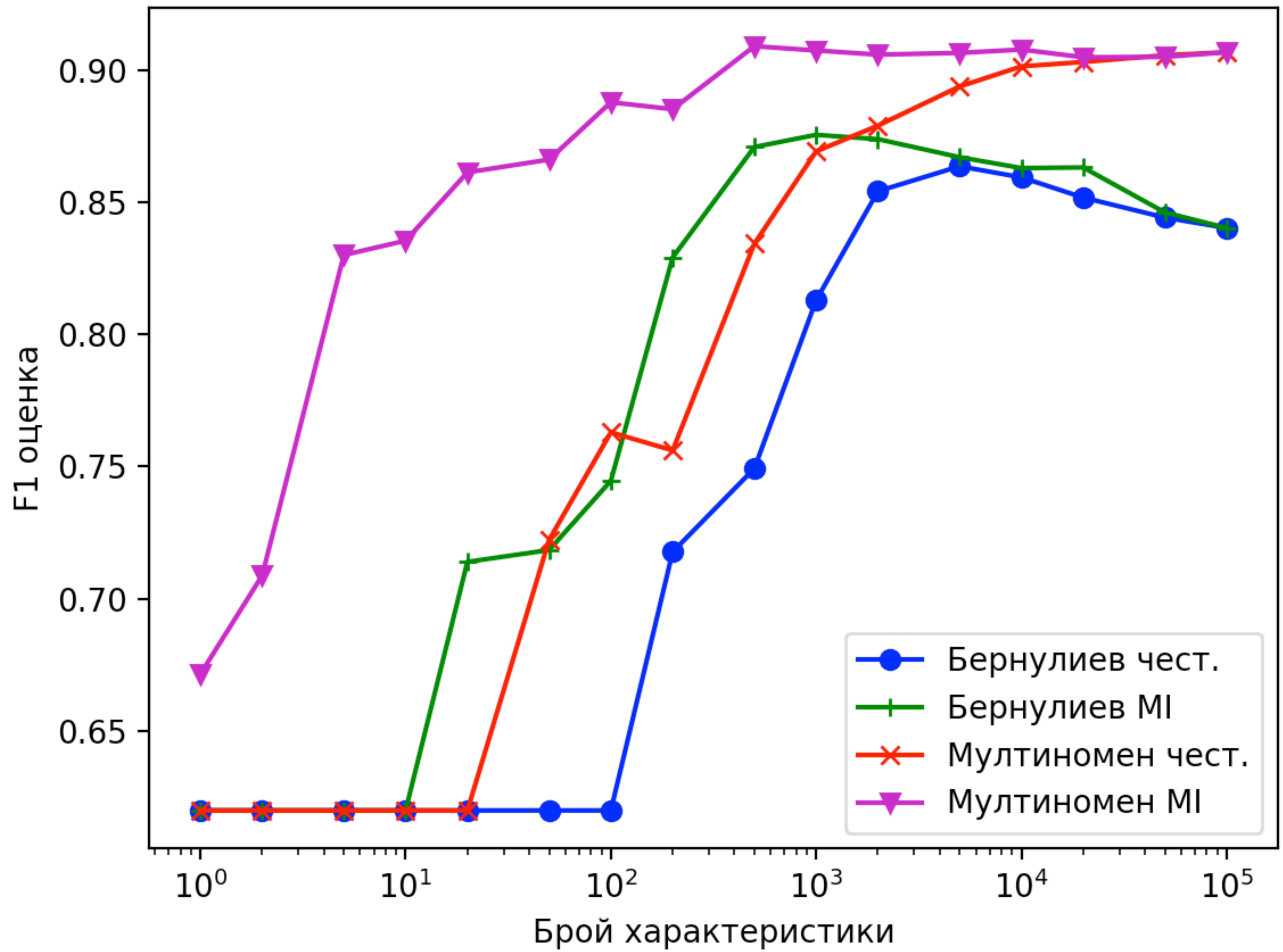
Оценяване на мярката за взаимна информация чрез максимизиране на правдоподобието

Нека с $N_{e_t e_c}$ да означим броя на документите, за които $U = e_t$ и $C = e_C$. Например N_{10} е броят на документите, в които се среща термът t и не е от клас c. Тогава: $\Pr[U = e_t, C = e_C] \approx \frac{N_{e_t e_c}}{N}$,

$$\Pr[U = e_t] \approx \frac{N_{e_t 0} + N_{e_t 1}}{N} = \frac{N_{e_t \bullet}}{N}, \Pr[C = e_C] \approx \frac{N_{\bullet e_c}}{N}$$

$$I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1\bullet}N_{\bullet 1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0\bullet}N_{\bullet 1}}$$
$$+\frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1\bullet}N_{\bullet 0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0\bullet}N_{\bullet 0}}$$

Резултат от избора на характеристики



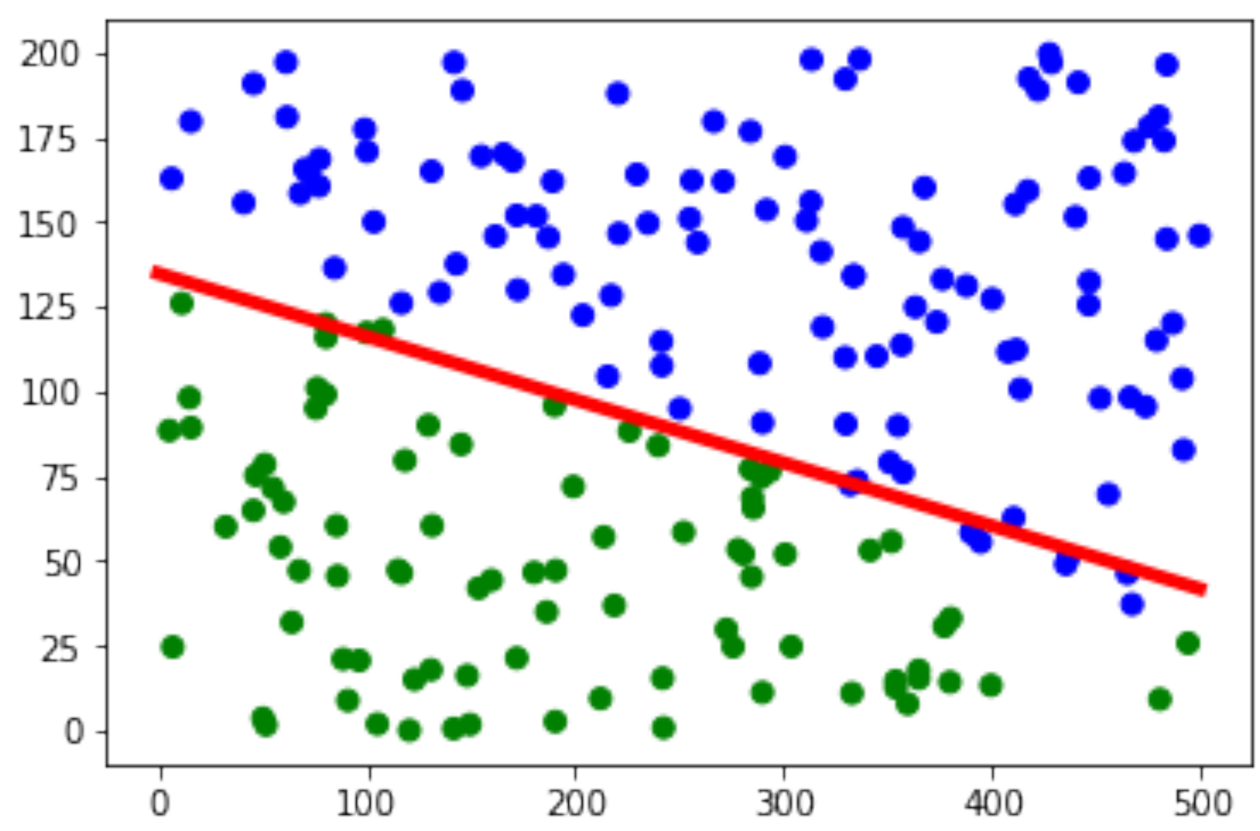
План на лекцията

1. Формалности за курса (5 мин)
2. Мултиномно разпределение (5 мин)
3. Максимизиране на правдоподобието при Мултиномно разпределение (15 мин)
4. Наивен Бейсов класификатор с мултиномен модел (20 мин)
5. Избор на характеристики (20 мин)
6. **Линейни класификатори (20 мин)**

Линеен класификатор

- Предполагаме, че документното пространство $\mathbb{X}$ е подмножество на $\mathbb{R}^n$ и класифицираме в два класа — обикновено $\mathbb{C} = \{-1, 1\}$.

- Класификатор $\gamma : \mathbb{X} \to \mathbb{C}$ наричаме линеен, ако съществуват вектор $\mathbf{w} \in \mathbb{R}^n$ и число $b \in \mathbb{R}$, така че за всяко $d \in \mathbb{X}$ е изпълнено: $\gamma(d) = \mathrm{sign}(\mathbf{w} \,.\, d + b)$

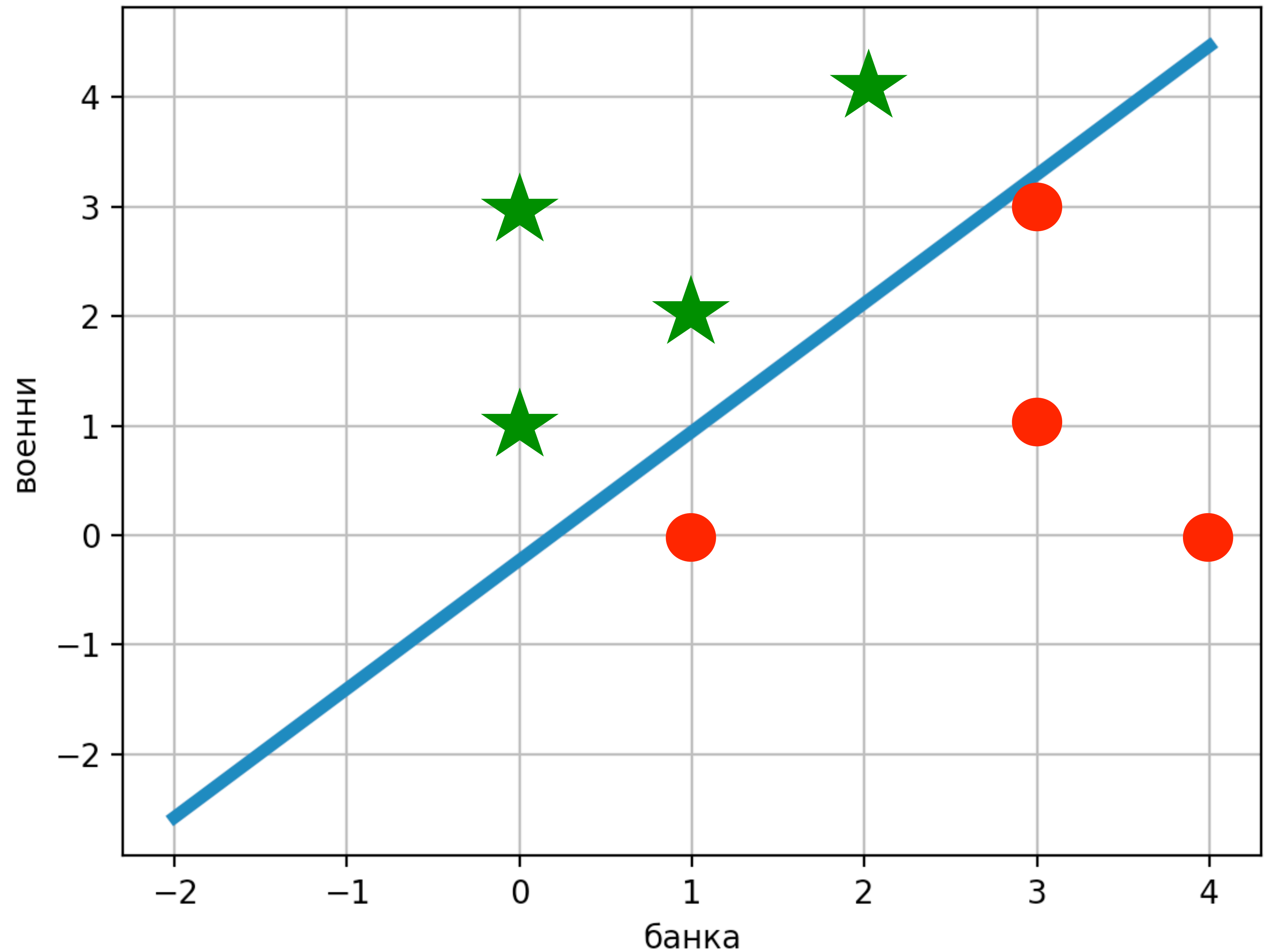


Представяне на мултиномен наивен Бейсов класификатор като линеен класификатор

- $c_{MAP} = \arg\max_{c \in \mathbb{C}} \Pr[c] \prod_{i=1}^{M} \Pr[t_i \mid c]^{f_i}$
- $\log \frac{\Pr[c \mid d]}{\Pr[\bar{c} \mid d]} = \log \frac{\Pr[c]}{\Pr[\bar{c}]} + \sum_{i=1}^{M} f_i \log \frac{Pr[t_i \mid c]}{Pr[t_i \mid \bar{c}]}$
- $d = (f_1, f_2, \ldots, f_M)$
- $\mathbf{w} = \left(\log \frac{Pr[t_1 \mid c]}{Pr[t_1 \mid \bar{c}]}, \log \frac{Pr[t_2 \mid c]}{Pr[t_2 \mid \bar{c}]}, \ldots, \log \frac{Pr[t_M \mid c]}{Pr[t_M \mid \bar{c}]} \right)$
- $b = \log \frac{\Pr[c]}{\Pr[\bar{c}]}$
- **Задача:** Бернулиевият наивен Бейсов класификатор линеен ли е? Защо?

Пример за представяне на мултиномен наивен Бейсов класификатор като линеен класификатор

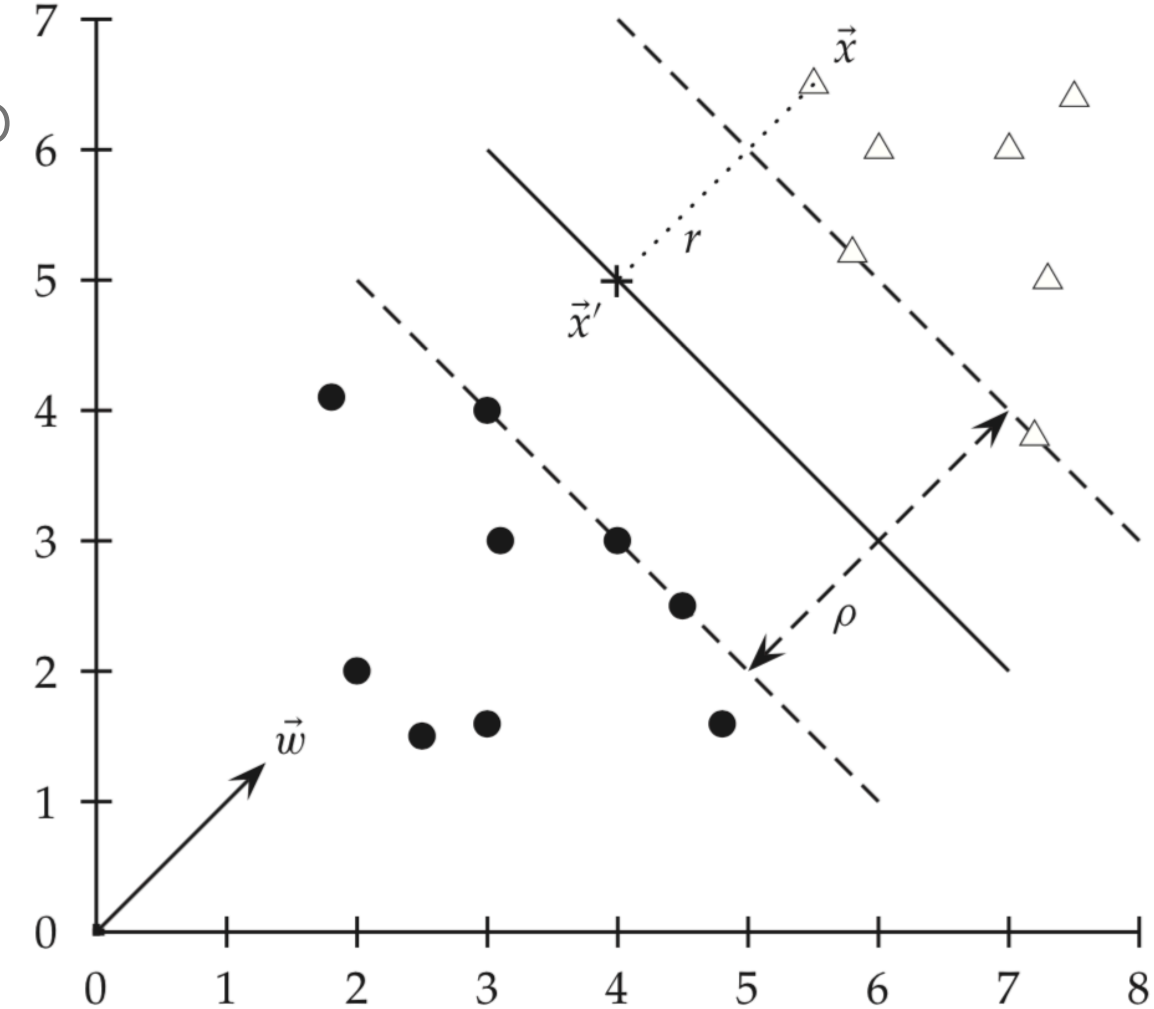
Класификатор за разграничаване между икономически и военни новини.



$$\gamma(d) = \mathrm{sign}(4.55 \times \#\text{банка} - 3.88 \times \#\text{военни} - 0.89)$$

Линеен класификатор SVM (Support Vector Machine)

Търсим разделителна хиперравнина зададена с уравнение $\overrightarrow{w} \,.\, \overrightarrow{x} + b = 0$, така че минималното разстояние ρ от хиперравнината до точка от $\mathbb{D}$ да е максимално



Решаваме квадратична оптимизационна задача:

- $\min \frac{1}{2} \overrightarrow{w} \,.\, \overrightarrow{w}$
- $y_i(\overrightarrow{w} \,.\, \overrightarrow{x_i} + b) \geq 1$, for all $(\overrightarrow{x_i}, y_i) \in \mathbb{D}$

Заключение

- Влагането на документите в многомерно числово документно пространство:
    - цели да заменим семантичното подобие с геометрична близост
    - позволявя прилагането на геометрични и алгебрични методи - например линейни класификатори
- Селекцията на характеристики може съществено да подобри качествата на даден класификатор