Търсене и извличане на информация. Приложение на дълбоко машинно обучение

Стоян Михов

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

План на лекцията

1. **Формалности за курса (2 мин)**
2. Наивен Бейсов класификатор за Бернулиев документен модел (20 мин)
3. Мултиномно разпределение (5 мин)
4. Максимизиране на правдоподобието при Мултиномно разпределение (15 мин)
5. Наивен Бейсов класификатор с мултиномен модел (20 мин)
6. Избор на характеристики (10 мин)
7. Линейни класификатори (15 мин)

Формалности

- Засега не може да ни намерят по-голяма зала…
- 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 (2 мин)
2. **Наивен Бейсов класификатор за Бернулиев документен модел (20 мин)**
3. Мултиномно разпределение (5 мин)
4. Максимизиране на правдоподобието при Мултиномно разпределение (15 мин)
5. Наивен Бейсов класификатор с мултиномен модел (20 мин)
6. Избор на характеристики (10 мин)
7. Линейни класификатори (15 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, $U_i : \mathbb{X} \to \{0,1\}$ са взаимно независими 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

- $\Pr[c \mid d] = \dfrac{\Pr[d \mid c] \Pr[c]}{\Pr[d]}$

- $c_{MAP} = \arg\max_{c \in \mathbb{C}} \Pr[c] \Pr[d \mid c] = \arg\max_{c \in \mathbb{C}} \log \Pr[c] + \log \Pr[d \mid c]$

- $\Pr[d \mid c] = \Pr[(e_1, e_2, \ldots, e_M) \mid c] = \prod_{i=1}^{M} \Pr[U_i = e_i \mid c]$

- $c_{MAP} = \arg\max_{c \in \mathbb{C}} \left(\log \Pr[c] + \sum_{i=1}^{M} \log \Pr[U_i = e_i \mid c] \right)$

Оценяване на параметрите използвайки принципа за максималното правдоподобие

- 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

. $\Pr[c] \approx \frac{N_c}{N}$

. $\Pr[U_i = 1 \mid c] \approx \frac{N_{c,t_i}}{N_c} \approx \frac{N_{c,t_i} + 1}{N_c + 2}$

. $\Pr[U_i = 0 \mid c] = 1 - \Pr[U_i = 1 \mid c]$

Алгоритми за наивен Бейсов класификатор чрез Бернулиев документен модел

```
TrainBernoulliNB(C, D)
1   V <- EXTRACT_VOCABULARY(D)
2   N <- COUNT_DOCS(D)
3   for each c in C do
4       Nc <- COUNT_DOCS_IN_CLASS(D, c)
5       prior[c] <- Nc/N
6       for each t in V do
7           Nct <- COUNT_DOCS_IN_CLASS_CONTAINING_TERM(D, c, t)
8           condprob[t][c] <- (Nct + 1)/(Nc + 2)
9   return V, prior, condprob

ApplyBernoulliNB(C, V, prior, condprob, d)
1   Vd <- EXTRACT_TERMS_FROM_DOC(V, d)
2   for each c in C do
3       score[c] <- log prior[c]
4       for each t in V do
5           if t in Vd then
6               score[c] += log(condprob[t][c])
7           else
8               score[c] += log(1-condprob[t][c])
9   return argmax(c in C, score[c])
```

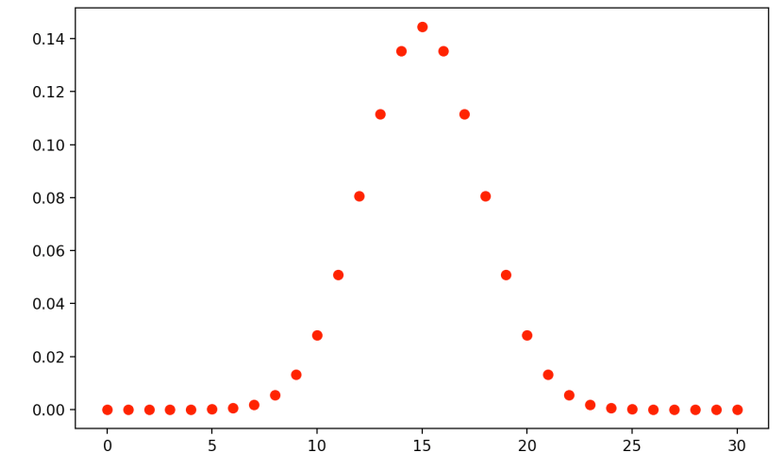
План на лекцията

1. Формалности за курса (2 мин)
2. Наивен Бейсов класификатор за Бернулиев документен модел (20 мин)
3. **Мултиномно разпределение (5 мин)**
4. Максимизиране на правдоподобието при Мултиномно разпределение (15 мин)
5. Наивен Бейсов класификатор с мултиномен модел (20 мин)
6. Избор на характеристики (10 мин)
7. Линейни класификатори (15 мин)

Пример за разпределение на дискретна случайна величина

- Основно пространство: всички възможни резултати при хвърляне на n монети.
- Случайна величина X: брой хвърляния на ези.
- **Биномно разпределение**: $B(n, p)$

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



- Обобщение: **Мултиномно разпределение** $M(n, l, p_1, p_2, \ldots, p_l)$
  - Хвърляме n еднакви зара с по l страни.
  - Случайни величини: $X_1, X_2, \ldots, X_l - X_i$ връща брой хвърляния на "i".

$$\Pr[X_1 = k_1, X_2 = k_2, \ldots, X_l = k_l] = \frac{n!}{k_1! k_2! \ldots k_l!} p_1^{k_1} p_2^{k_2} \ldots p_l^{k_l}$$ когато

$$\sum_{i=1}^{l} k_i = n, \sum_{i=1}^{l} p_i = 1.$$

План на лекцията

1. Формалности за курса (2 мин)
2. Наивен Бейсов класификатор за Бернулиев документен модел (20 мин)
3. Мултиномно разпределение (5 мин)
4. **Максимизиране на правдоподобието при Мултиномно разпределение (15 мин)**
5. Наивен Бейсов класификатор с мултиномен модел (20 мин)
6. Избор на характеристики (10 мин)
7. Линейни класификатори (15 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.

  $$\Pr[X_1^{(i)} = x_1^{(i)}, X_2^{(i)} = x_2^{(i)}, \ldots, X_l^{(i)} = x_l^{(i)}] = \frac{n!}{x_1^{(i)}! x_2^{(i)}! \ldots x_l^{(i)}!} p_1^{x_1^{(i)}} p_2^{x_2^{(i)}} \ldots p_l^{x_l^{(i)}}$$

- Търсим:

  $$\hat{\mathbf{p}} = \arg\max_{\mathbf{p}} L(\mathbf{p}) = \arg\max_{\mathbf{p}} \log L(\mathbf{p}) =$$

  $$= \arg\max_{\mathbf{p}} \sum_{i-1}^{m} \log \Pr[X_1^{(i)} = x_1^{(i)}, X_2^{(i)} = x_2^{(i)}, \ldots, X_l^{(i)} = x_l^{(i)}] =$$

  $$= \arg\max_{\mathbf{p}} \sum_{i=1}^{n} \left(\log \frac{n!}{x_1^{(i)}! x_2^{(i)}! \ldots x_l^{(i)}!} + x_1^{(i)} \log p_1 + x_2^{(i)} \log p_2 + \ldots + x_l^{(i)} \log p_l \right)$$

- **Проблем:** Ако $p_i \to \infty$, то $L(\mathbf{p}) \to \infty$.

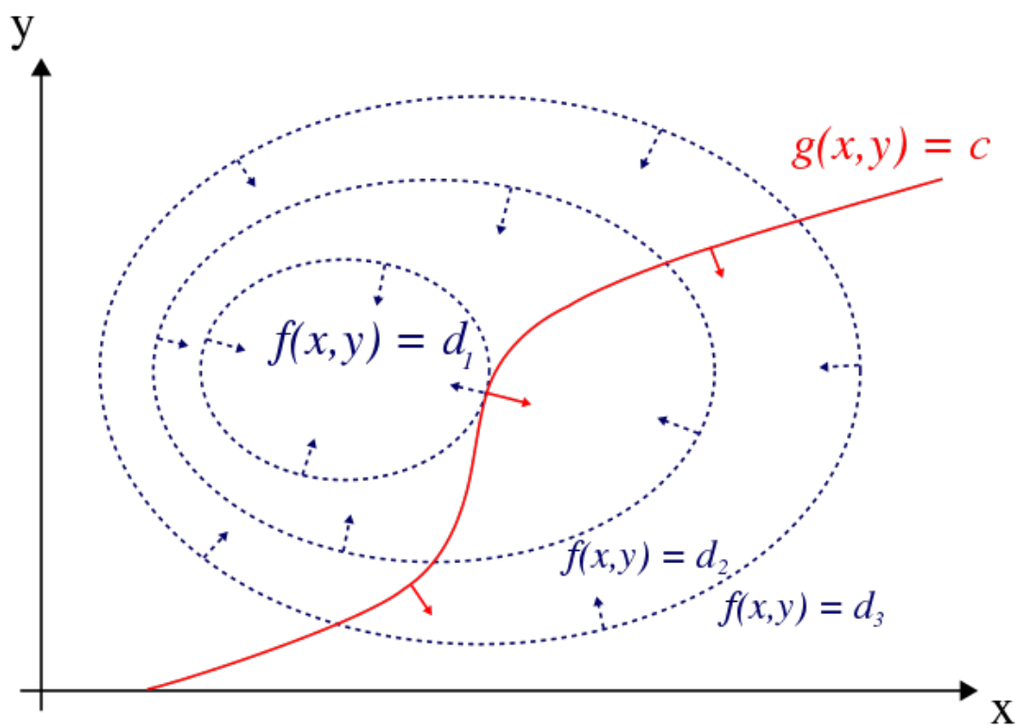
- **Решение:** Трябва да намерим $\hat{\mathbf{p}} = \arg\max_{\mathbf{p}} \log L(\mathbf{p})$ при ограничение $\sum_{i=1}^{l} p_i = 1$.

Намиране на локални екстремуми при ограничения — множители на Лагранж

- Търсим: $\left| \begin{array}{l} \hat{\mathbf{x}} = \arg\max_{\mathbf{x}} f(\mathbf{x}) \\ g(\mathbf{x}) = c \end{array} \right.$

- Необходимо условие: $\left| \begin{array}{l} \nabla_{\mathbf{x}} f(\mathbf{x}) - \lambda \nabla_{\mathbf{x}} g(\mathbf{x}) = 0 \\ g(\mathbf{x}) = c \end{array} \right.$

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}) \\ \frac{\partial}{\partial x_2} f(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_n} f(\mathbf{x}) \end{bmatrix}$$



- Еквивалентно условие: $\nabla_{\mathbf{x},\lambda} \mathscr{L}(\mathbf{x}, \lambda) = 0$, където $\mathscr{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda(g(\mathbf{x}) - c)$.

- По-подробно — например в учебника “Диференциално смятане” на проф. Тагамлицки: https://store.fmi.uni-sofia.bg/fmi/or/_private/Calculus1.pdf

Максимизиране на правдоподобието при мултиномно разпределение

- Дадени са m н.е.р l-мерни 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 $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_l^{(i)})$ и $\Pr[X_1^{(i)} = x_1^{(i)}, X_2^{(i)} = x_2^{(i)}, \ldots, X_l^{(i)} = x_l^{(i)}] = \frac{n!}{x_1^{(i)}!x_2^{(i)}!\ldots x_l^{(i)}!} p_1^{x_1^{(i)}} p_2^{x_2^{(i)}} \ldots p_l^{x_l^{(i)}}$
- 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:
$\hat{\mathbf{p}} = \arg\max_{\mathbf{p}} L(\mathbf{p}) = \arg\max_{\mathbf{p}} \log L(\mathbf{p}) =$

$$= \arg\max_{\mathbf{p}} \sum_{k_1+k_2+\ldots+k_l=n} f_{k_1,k_2,\ldots,k_l} \left(\log \frac{n!}{k_1!k_2!\ldots k_l!} + k_1 \log p_1 + k_2 \log p_2 + \ldots + k_l \log p_l \right)$$, при ограничение: $\sum_{i=1}^{l} p_i = 1,$

- Метод на Лагранж: $g(p_1, p_2, \ldots, p_l) = p_1 + p_2 + \ldots + p_l$, $\mathscr{L}(\mathbf{p}, \lambda) = \log L(\mathbf{p}) - \lambda(g(\mathbf{p}) - 1)$.
- Търсим $\nabla_{\mathbf{p},\lambda} \mathscr{L}(\mathbf{p}, \lambda) = 0$.

$$\frac{\partial \log L(p_1, p_2, \ldots, p_l) - \lambda(g(p_1, p_2, \ldots, p_l) - 1)}{\partial p_i} = 0 \Longrightarrow p_i = \frac{\sum_{k_1+k_2+\ldots+k_l=n} k_i f_{k_1,k_2,\ldots,k_l}}{\lambda}$$

- $$\frac{\partial \log L(p_1, p_2, \ldots, p_l) - \lambda(g(p_1, p_2, \ldots, p_l) - 1)}{\partial \lambda} = 0 \Longrightarrow p_1 + p_2 + \ldots p_l = 1$$
- $\hat{p}_i = \frac{\sum_{j=1}^{m} x_i^{(j)}}{nm}$

План на лекцията

1. Формалности за курса (2 мин)
2. Наивен Бейсов класификатор за Бернулиев документен модел (20 мин)
3. Мултиномно разпределение (5 мин)
4. Максимизиране на правдоподобието при Мултиномно разпределение (15 мин)
5. **Наивен Бейсов класификатор с мултиномен модел (20 мин)**
6. Избор на характеристики (10 мин)
7. Линейни класификатори (15 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 $K_d = \frac{n!}{f_1! \ldots f_M!}$
- 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 $c_{MAP} = \arg\max_{c \in \mathbb{C}} \Pr[c \mid d]$
- $\Pr[c \mid d] = \dfrac{\Pr[d \mid c] \Pr[c]}{\Pr[d]}$
- $\Pr[d \mid c] = \Pr[(f_1, f_2, \ldots, f_M) \mid c] = K_d \prod_{i=1}^{M} \Pr[t_i \mid c]^{f_i}$
- $c_{MAP} = \arg\max_{c \in \mathbb{C}} \Pr[c] \prod_{i=1}^{M} \Pr[t_i \mid c]^{f_i}$
- $c_{MAP} = \arg\max_{c \in \mathbb{C}} \log \Pr[c] + \sum_{i=1}^{M} f_i \log \Pr[t_i \mid c] =$
- $= \arg\max_{c \in \mathbb{C}} \log \Pr[c] + \sum_{k=1}^{n} \log \Pr[t_{d_k} \mid c]$

Оценяване на параметрите използвайки принципа за максималното правдоподобие

- 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.

. $\Pr[c] \approx \frac{N_c}{N}$

. $\Pr[t_i \mid c] \approx \frac{T_{c,t_i}}{\sum_{t' \in V} T_{c,t'}} \approx \frac{T_{c,t_i} + 1}{\sum_{t' \in V} T_{c,t'} + \mid V \mid}$

Алгоритми за наивен Бейсов класификатор чрез мултиномен документен модел

```
TrainMultinomialNB(C, D)
1   V <- EXTRACTVOCABULARY(D)
2   N <- COUNTDOCS(D)
3   for each c in C do
4       Nc <- COUNTDOCSINCLASS(D, c)
5       prior[c] <- Nc/N
6       textc <- CONCATENATETEXTOFALLDOCSINCLASS(D, c)
7       for each t in V do
8           Tc[t] <- COUNTTOKENSOFTERM(textc, t)
9       for each t in V do
10          condprob[t][c] <- (Tc[t]+1)/sum(Tc[t’]+1 for t' in V)
11  return V, prior, condprob

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

План на лекцията

1. Формалности за курса (2 мин)
2. Наивен Бейсов класификатор за Бернулиев документен модел (20 мин)
3. Мултиномно разпределение (5 мин)
4. Максимизиране на правдоподобието при Мултиномно разпределение (15 мин)
5. Наивен Бейсов класификатор с мултиномен модел (20 мин)
6. **Избор на характеристики (10 мин)**
7. Линейни класификатори (15 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:

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

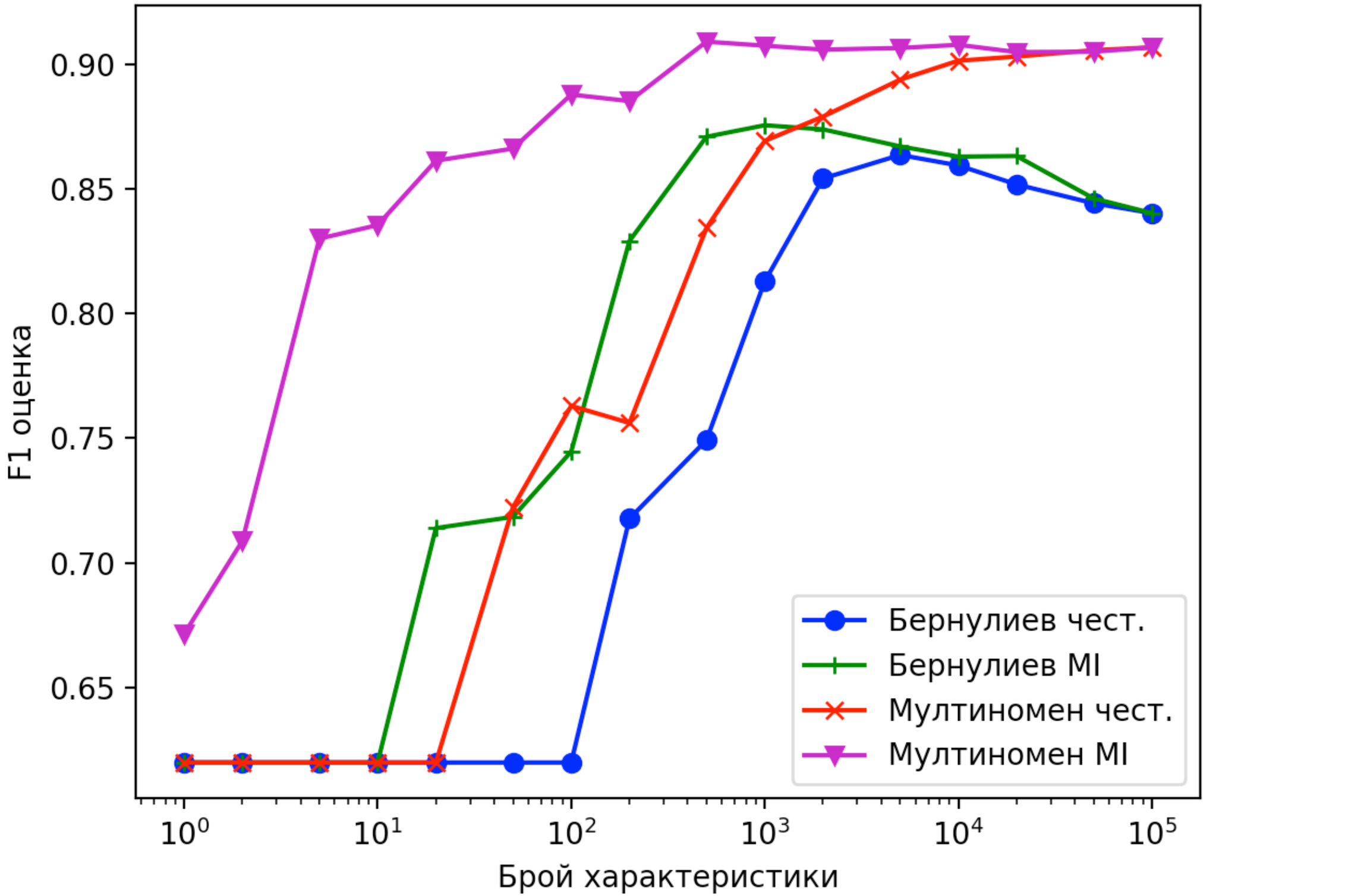
Оценяване на мярката за взаимна информация чрез максимизиране на правдоподобието

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: $\Pr[U = e_t, C = e_C] \approx \frac{N_{e_t e_c}}{N}$,

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

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

Резултат от избора на характеристики



План на лекцията

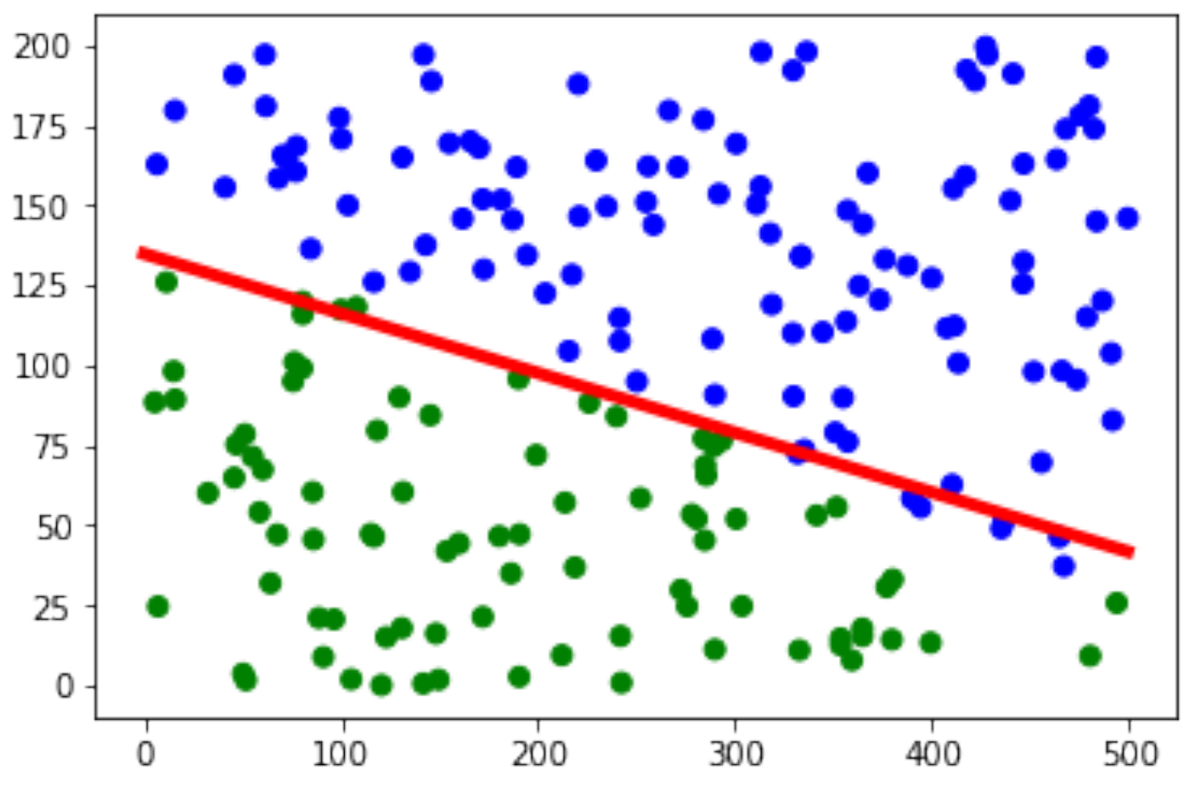
1. Формалности за курса (2 мин)
2. Наивен Бейсов класификатор за Бернулиев документен модел (20 мин)
3. Мултиномно разпределение (5 мин)
4. Максимизиране на правдоподобието при Мултиномно разпределение (15 мин)
5. Наивен Бейсов класификатор с мултиномен модел (20 мин)
6. Избор на характеристики (10 мин)
7. **Линейни класификатори (15 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 $\gamma : \mathbb{X} \rightarrow \mathbb{C}$ 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:

$$\gamma(d) = \text{sign}(\mathbf{w} \,.\, d + b)$$

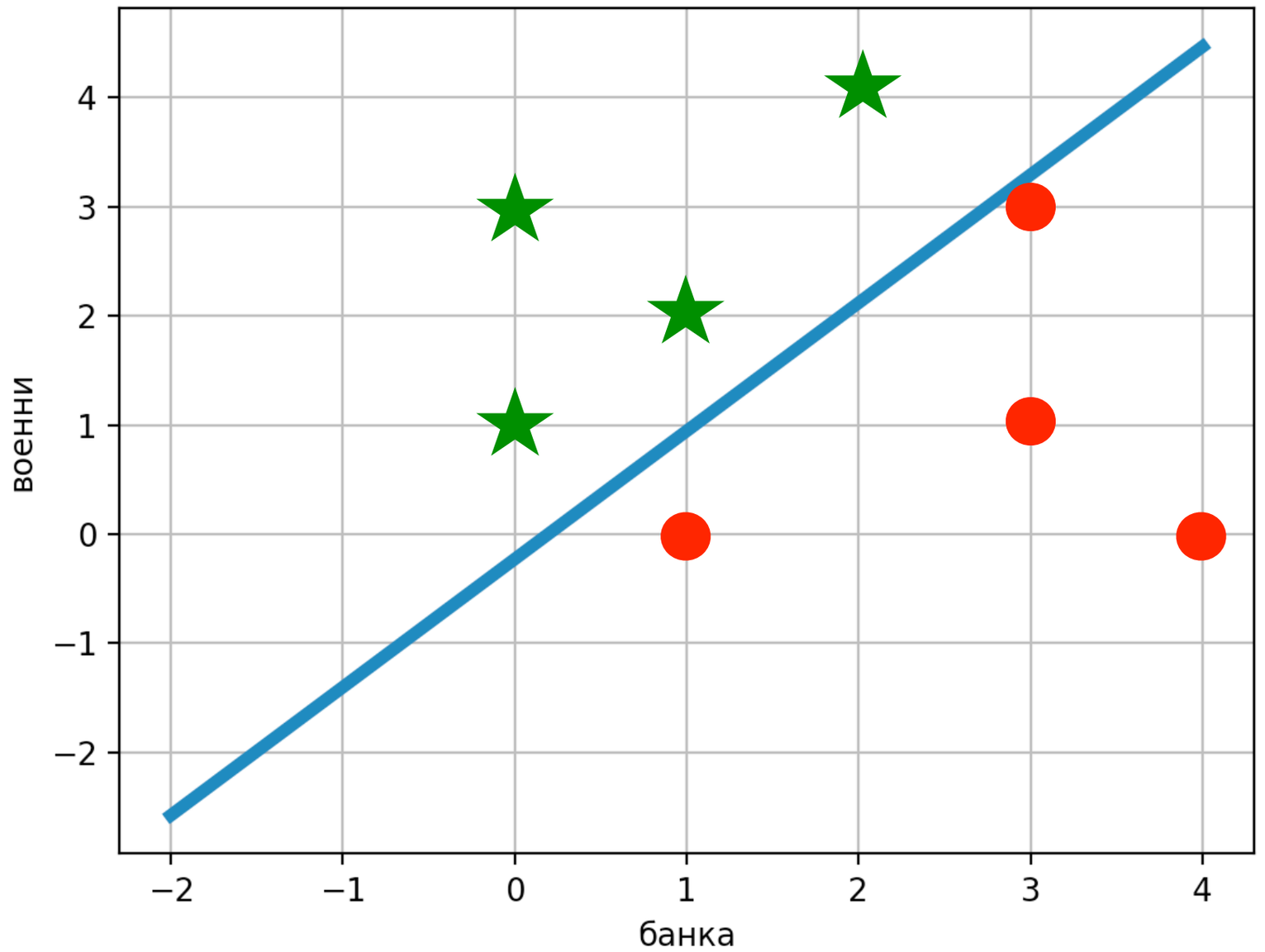


Представяне на мултиномен наивен Бейсов класификатор като линеен класификатор

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

Пример за представяне на мултиномен наивен Бейсов класификатор като линеен класификатор

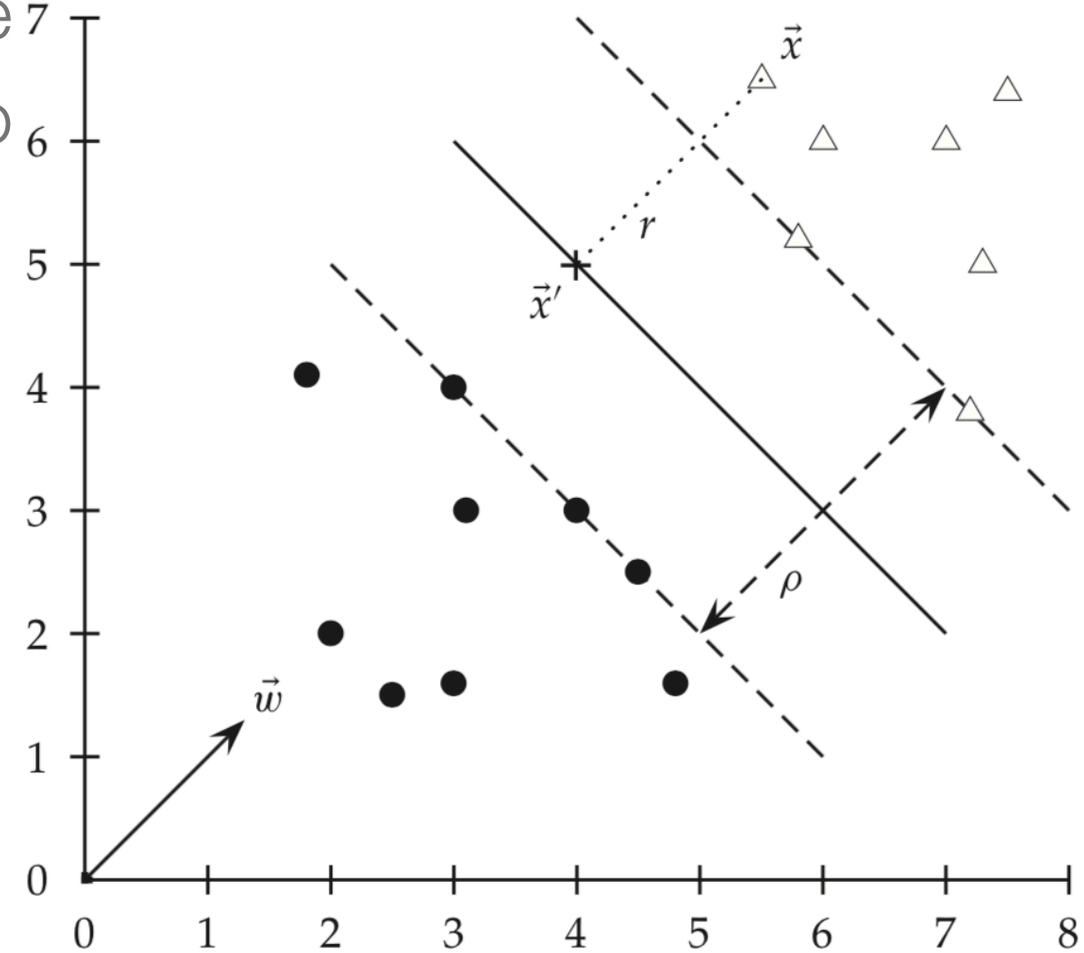
Класификатор за разграничаване между икономически и военни новини.



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

Линеен класификатор SVM (Support Vector Machine)

Търсим разделителна хиперравнина зададена с уравнение $\overrightarrow{w} . \vec{x} + b = 0$, 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 замениме 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