Търсене и извличане на информация. Приложение на дълбоко машинно обучение

Стоян Михов

Лекция 7: Клъстеризация във векторно пространство. Вероятно приблизително коректно обучение (PAC-обучение).

План на лекцията

1. **Формалности за курса (5 мин)**
2. Клъстеризация (10 мин)
3. k-means (15 мин)
4. Варианти и подобрения на K-means (15 мин)
5. PAC-обучение (20 мин)
6. Пример за PAC-обучение (25 мин)

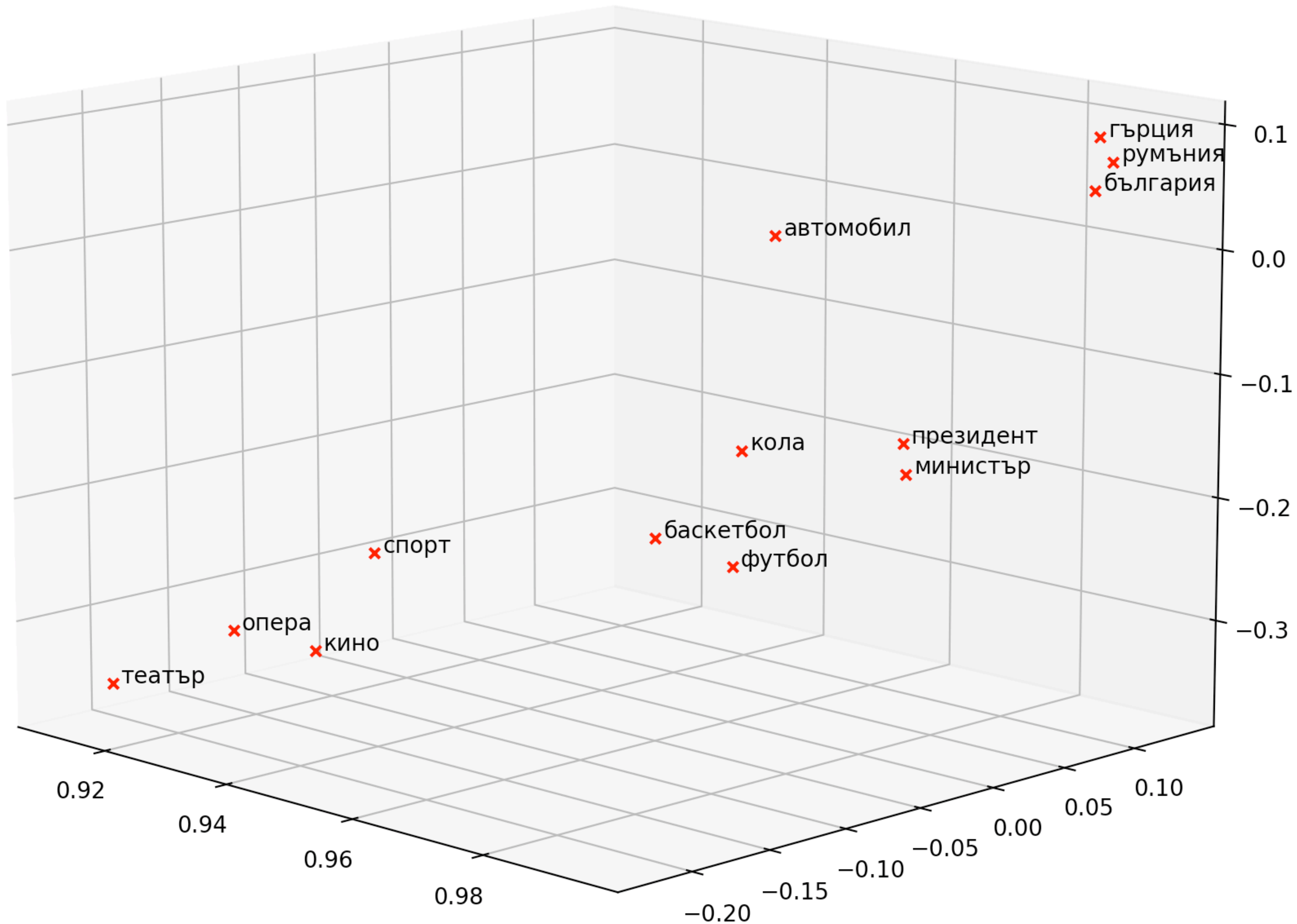
Формалности

- Засега ще провеждаме занятията онлайн всяка сряда от 8:15 до 12:00 часа.
- Засега ще използваме платформата Google meet: meet.google.com/hue-frfx-axb
- Днес ще използваме едновременно слайдове и бяла дъска. Моля следете съответния екран.
- В Moodle на 19.11.2021 г. ще бъде публикувано домашно задание, което следва да бъде предадено до края на деня на 28.11.2021 г.
- Седмата лекция се базира на глава 16 от първия учебник и секция 10.4 от втория учебник.

План на лекцията

1. Формалности за курса (5 мин)
2. **Клъстеризация (10 мин)**
3. k-means (15 мин)
4. Варианти и подобрения на K-means (15 мин)
5. PAC-обучение (20 мин)
6. Пример за PAC-обучение (25 мин)

Семантично пространствени релации



Клъстеризация

- Задачата на клъстеризацията е да намерим “естествено” групиране на обектите в “клъстери”

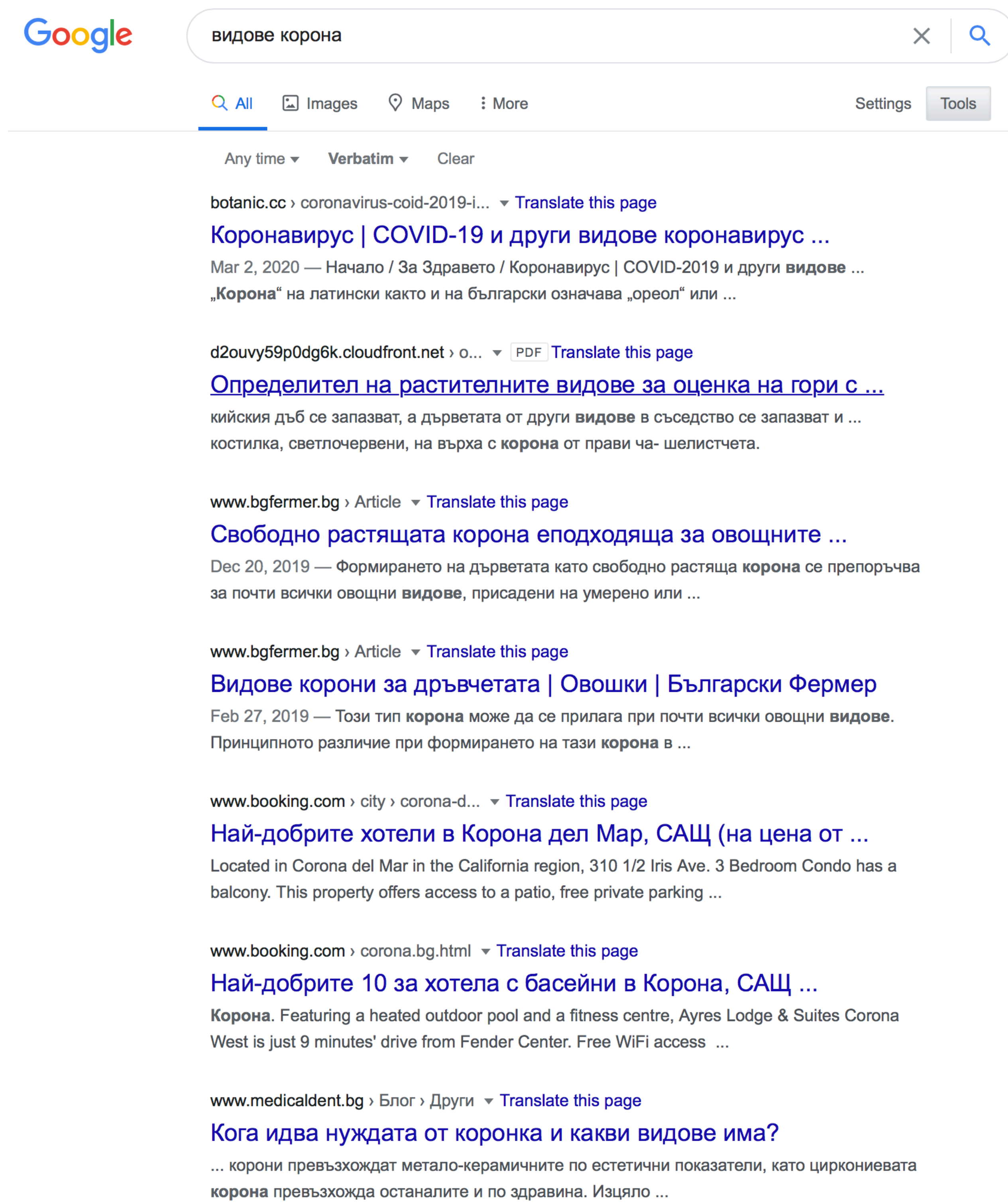
- Целта е:

  - В рамките на един клъстер обектите да са близки

  - Обектите от различни клъстери да са далеч един от друг

Приложение на клъстеризацията в търсенето на информация

- При търсене в Интернет често се връщат много хиляди резултати, като потребителя може да разгледа едва няколко от тях.
- Поради многозначността на езика резултатите могат да бъдат от различни области.
- Чрез клъстеризиране на резултатите от заявката на първата страница се връщат по няколко резултата от всеки от клъстерите.



Класификация ⇔ клъстеризация

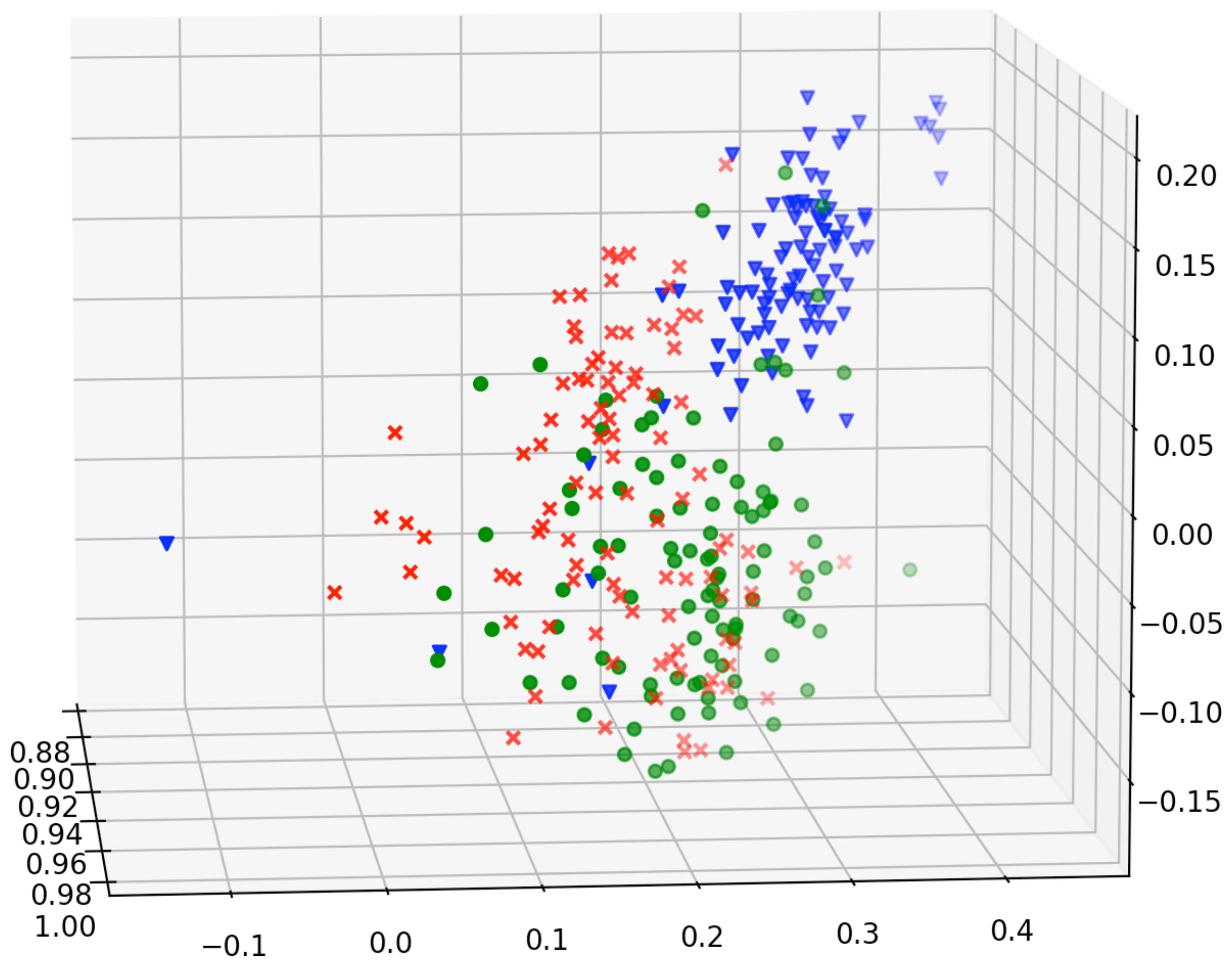
- При класификацията имаме предварително зададени класове и база от класифицирани документи ⇔ при клъстеризацията нямаме нито зададени класове, нито техния брой, нито класифицирани документи.

- При класификацията се стремим да намерим функция (класификатор), която да определя класа на даден документ по подобие на класифицираните документи ⇔ при клъстеризацията се търси разбиване на базата на имплицитните закономерности в базата от документи.

- Задачата за класификация е пример за обучение “с учител” (supervised learning) ⇔ задачата за клъстеризация е пример за обучение “без учител” (unsupervised learning)

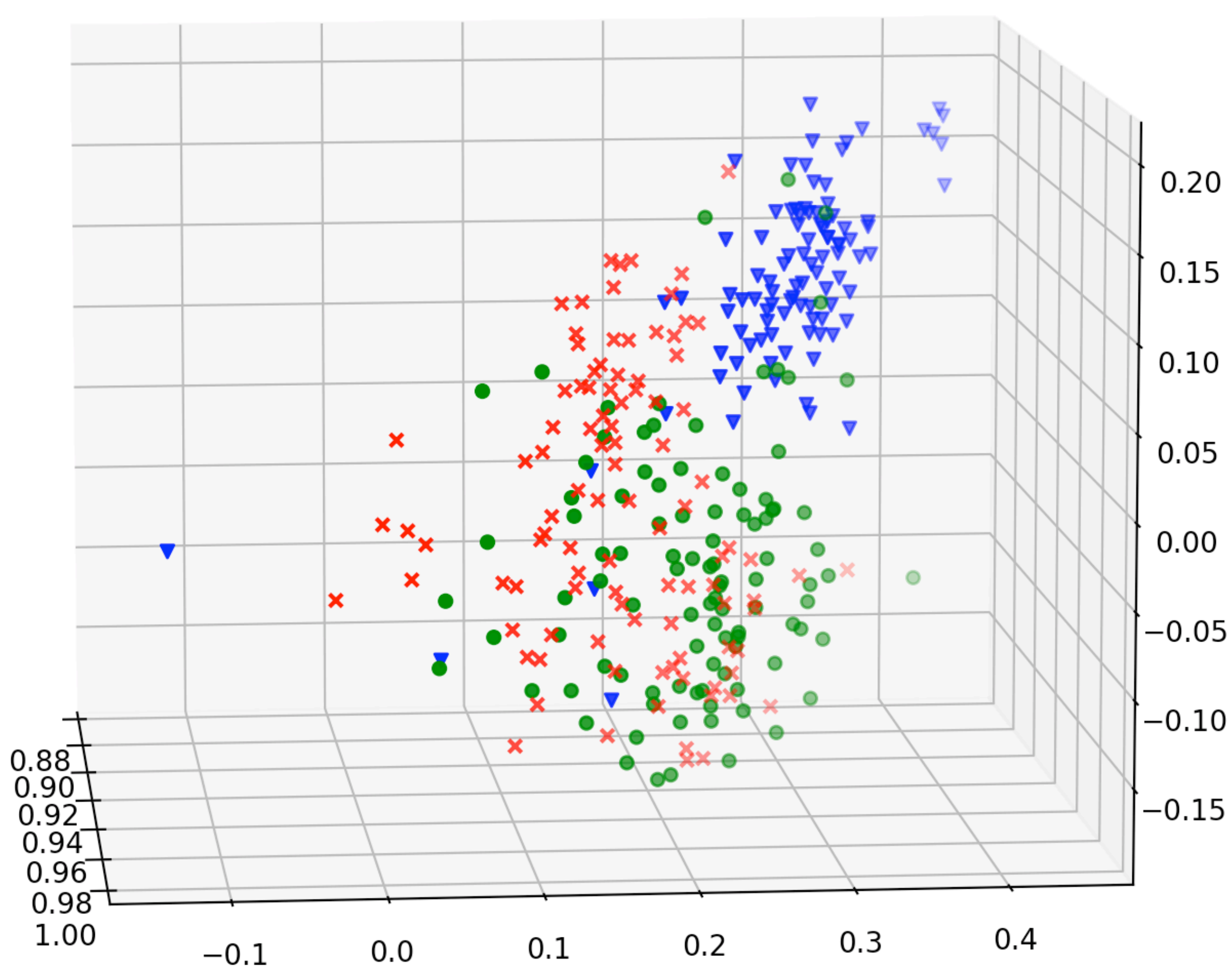
Класификация ⇔ клъстеризация

Класове от документи

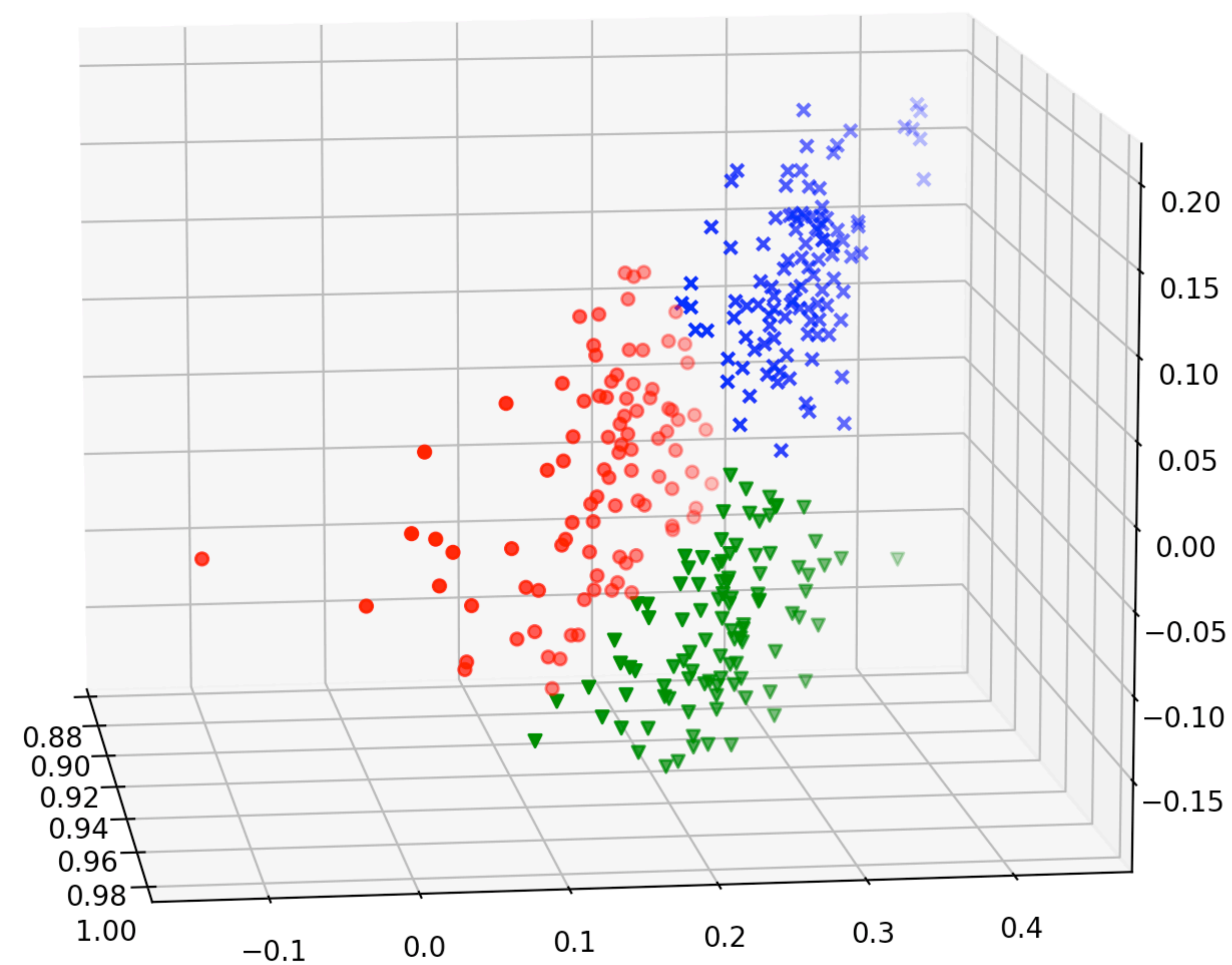


Класификация ⇔ клъстеризация

Класове от документи



Клъстери от документи



Други приложения на клъстеризацията

- Групиране на резултатите от търсене
- Групиране на поток от документи — новини, мейлове, съобщения, …
- Ускоряване на търсене по подобие
- Търсене чрез разбиване-събиране (gather-scatter)

План на лекцията

1. Формалности за курса (5 мин)
2. Клъстеризация (10 мин)
3. **k-means (15 мин)**
4. Варианти и подобрения на K-means (15 мин)
5. PAC-обучение (20 мин)
6. Пример за PAC-обучение (25 мин)

K-means

- Дадено е множество от вектори $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_S\} \subset \mathbb{R}^N$, и число $K \in \mathbb{N}$

- Търсим разбиване на $\mathbf{X}$ в K клъстера $W_1, W_2, \ldots, W_K \subset \mathbf{X}$, $W_1 \cup W_2 \cup \ldots \cup W_K = \mathbf{X}$, така че:
  $\mathrm{RSS}(W_1, W_2, \ldots, W_k) = \sum_{k=1}^{K} \sum_{\mathbf{x} \in W_k} \|\mathbf{x} - \mu_k\|^2$ е минимално,
  където $\mu_k = \frac{1}{|W_k|} \sum_{\mathbf{x} \in W_k} \mathbf{x}$ за $k = 1,2,\ldots,K$.

- μ_k е центроида (центъра на тежестта) на W_k

Алгоритъм K-means

1. Започваме с първоначални центроиди $\mu_1, \mu_2, \ldots, \mu_K$.

2. За всеки от центроидите μ_k намираме клъстера W_k
$$W_k = \{\mathbf{x} \in \mathbf{X} \mid \arg\min_i \|\mathbf{x} - \mu_i\|^2 = k\}.$$

3. Намираме новите стойности на центроидите:
$$\mu_k = \frac{1}{|W_k|} \sum_{\mathbf{x} \in W_k} \mathbf{x}$$

4. Докато не се изпълни условие за край повтаряме стъпките 2-4

Алгоритъм K-means

```
K-means ( {x[1],…,x[S] } , K )
  (μ[1],…,μ[K] ) <- SelectSeeds({x[1],…,x[S]}, K )
  while stopping criterion has not been met do
     for k <- 1 to K do
        ω[k] <- {}
     for i <- 1 to S do
        k <- argmin_j |μ[j] – x[i]|
        ω[k] <- ω[k] ∪ {x[i]}
     for k <- 1 to K do
        μ[k] <- 1/|ω[k]| ∑_{x∈ω[k]} x
   return{μ[1],…,μ[K]}
```

Коректност

Твърдение: Нека са дадени вектори $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^N$ и $W_1, W_2, \ldots, W_K \subset \mathbf{X}$ са дефинирани като $W_k = \{\mathbf{x} \in \mathbf{X} \mid \arg\min_i \|\mathbf{x} - \mu_i\|^2 = k\}$, за $k = 1,2,\ldots,K$. Нека $W'_1, W'_2, \ldots, W'_K \subset \mathbf{X}$ е произволно разбиване на $\mathbf{X}$. Тогава: $\mathrm{RSS}(W_1, W_2, \ldots, W_k) \leq \mathrm{RSS}(W'_1, W'_2, \ldots, W'_k)$.

Доказателство:

Нека $k(\mathbf{x}) = \arg\min_i \|\mathbf{x} - \mu_i\|^2$ и $l(\mathbf{x}) = l \leftrightarrow \mathbf{x} \in W'_l$. Тогава:

$$\mathrm{RSS}(W'_1, W'_2, \ldots, W'_k) = \sum_{k=1}^{K} \sum_{\mathbf{x} \in W'_k} \|\mathbf{x} - \mu_k\|^2 = \sum_{\mathbf{x} \in \mathbf{X}} \|\mathbf{x} - \mu_{l(\mathbf{x})}\|^2 \geq \sum_{\mathbf{x} \in \mathbf{X}} \|\mathbf{x} - \mu_{k(\mathbf{x})}\|^2$$

$$= \sum_{k=1}^{K} \sum_{\mathbf{x} \in W_k} \|\mathbf{x} - \mu_k\|^2 = \mathrm{RSS}(W_1, W_2, \ldots, W_k)$$

Твърдение: Нека е дадено множество $W \subset \mathbb{R}^N$. Тогава:

$$\arg\min_{\mathbf{y} \in \mathbb{R}^N} \sum_{\mathbf{x} \in W} \|\mathbf{x} - \mathbf{y}\|^2 = \frac{1}{|W|} \sum_{\mathbf{x} \in W} \mathbf{x}$$

Доказателство:

$$\frac{\partial}{\partial \mathbf{y}} \sum_{\mathbf{x} \in W} \|\mathbf{x} - \mathbf{y}\|^2 = 2|W|\mathbf{y} - 2\sum_{\mathbf{x} \in W} \mathbf{x} = 0$$

Следствие: На всяка стъпка от алгоритъма RSS намалява.

План на лекцията

1. Формалности за курса (5 мин)
2. Клъстеризация (10 мин)
3. k-means (15 мин)
4. **Варианти и подобрения на K-means (15 мин)**
5. PAC-обучение (20 мин)
6. Пример за PAC-обучение (25 мин)

Условия за край

- Функцията RSS не е изпъкнала. Намирането на глобален екстремум е NP пълен проблем.
- Евристично решение:
    - Когато RSS спре да се подобрява
    - Когато подобрението на RSS е под определен праг
    - След извършване на предварително фиксиран брой итерации

Начални центроиди

Оказва се, че резултатът от клъстеризирането с k-means силно зависи от началните центроиди.

Варианти:

1. Избираме първите K вектора от $\mathbf{X}$ и изпълняваме алгоритъма k-means — най-просто, но наивно.

2. Избираме с равномерно случайно разпределение K вектора от $\mathbf{X}$ и изпълняваме алгоритъма k-means.

3. Повтаряме няколко пъти точка 2 и избираме резултата с най-добър RSS.

4. Избираме началните центроиди, така че да ги раздалечим — виж k-means++

K-means++

1. Избираме първия центроид μ_1 с равномерно случайно разпределение от $\mathbf{X}$.

2. Нека сме избрали центроиди $\mu_1, \mu_2, \ldots, \mu_l$. Нека $D(\mathbf{x}) = \min_{i=1}^{l} \|\mathbf{x} - \mu_i\|$. Дефинираме случайно разпределение върху $\mathbf{X}$ като за всеки вектор $\mathbf{x} \in \mathbf{X}$ дефинираме $\Pr_l[\mathbf{x}] = \dfrac{D(\mathbf{x})^2}{\sum_{\mathbf{x}' \in \mathbf{X}} D(\mathbf{x}')^2}$. Избираме центроида μ_{l+1} от $\mathbf{X}$ със случайно разпределение $\Pr_l[\mathbf{x}]$.

3. Повтаряме стъпки 2-3, докато изберем K центроида.

4. С избраните центроиди изпълняваме алгоритъма k-means

Доказва се, че очакваното за RSS при k-means++ е по малко от $O(\log K)$ по глобалния минимум за RSS:
Arthur, D.; Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding". Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035.

План на лекцията

1. Формалности за курса (5 мин)
2. Клъстеризация (10 мин)
3. k-means (15 мин)
4. Варианти и подобрения на K-means (15 мин)
5. **PAC-обучение (20 мин)**
6. Пример за PAC-обучение (25 мин)

Какво ще разбираме под “машинно обучение”

- Съществуват различни подходи (алгоритми), за машинно обучение: вероятностни, градиентни, алгебрични, комбинаторни, …
- Машинното обучение може да се разглежда като “с учител” и “без учител”
- За формалното изследване на различните алгоритми е целесъобразно въвеждането на формална рамка, която да ни позволява да сравняваме и оценяваме свойствата на различните алгоритми за машинно обучение.
- Ще разгледаме един от разпространените подходи за формализиране на понятието машинно обучение — рамката “Вероятно приблизително коректно” обучение (Probably Approximately Correct PAC-обучение)

Формализиране на задачата за машинно обучение

- Нека с $\mathscr{X}$ означим множеството от всички възможни **наблюдения** (примери) и ще наричаме входно пространство или домейн.
- Нека с $\mathscr{Y}$ означим множеството от възможни **класове**. Ще разглеждаме само крайни множества от класове.
- Функция $c : \mathscr{X} \rightarrow \mathscr{Y}$ ще наричаме **класификатор (концепт)**. Когато $\mathscr{Y} = \{0,1\}$ ни е даден бинарен класификатор, който може да разглеждаме като подмножеството $\mathscr{X}$, за което c, връща стойност 1.
- Под **клас от класификатори** ще разбираме конкретно множество от класификатори, които ще искаме да научим и ще означаваме с $\mathscr{C}$. Например, ако $\mathscr{X}$ е двумерното пространство и разглеждаме бинарни класификатори, то класът от класификатори би могъл да бъде множеството от всички триъгълници в равнината.
- Предполагаме, че ни е дадено фиксирано, но неизвестно **вероятностно разпределение** $\mathscr{D}$ върху $\mathscr{X}$, и че всички наблюдения от $\mathscr{X}$, които правим са независими и еднакво разпределени с разпределение $\mathscr{D}$.

Задачата на машинното обучение

- Обучаемият получава извадка $S = (x_1, x_2, \ldots, x_m)$ от независими и идентично разпределени с разпределение $\mathscr{D}$ наблюдения, заедно със съответни етикети $(c(x_1), c(x_2), \ldots, c(x_m))$ относно конкретен класификатор $c \in \mathscr{C}$, който следва да се научи.

- Задачата на обучаемият е въз основа на наблюденията S класификатор да избере h_S, с минимална грешка при обобщение спрямо целевия класификатор $c \in \mathscr{C}$.

- **Грешка при обобщение** $R(h)$ или истинска грешка между класификатор $h \in \mathscr{H}$ и целевия класификатор $c \in \mathscr{C}$ дефинираме: $R(h) = \Pr_{x \sim \mathscr{D}}[\{x \mid h(x) \neq c(x)\}]$.

Дефиниция на PAC-обучение

Казваме, че класът от класификатори $\mathscr{C}$ е **PAC-обучаем** ако:

- съществува алгоритъм $\mathscr{A}$, връщащ на извадка от наблюдения S класификатор $h_S = \mathscr{A}(S)$, и
- полиномиална функция $poly : \mathbb{R}^2 \to \mathbb{R}$,

така че за всяко $\varepsilon > 0$ и всяко $\delta > 0$ ако $m > poly(1/\varepsilon, 1/\delta)$ и S е извадка с поне m наблюдения, то:

$$\Pr_{S\sim\mathscr{D}^m}[\{S \mid R(h_S) \leq \varepsilon\}] \geq 1 - \delta.$$

Т.е. Клас от класификатори $\mathscr{C}$ е PAC-обучаем, ако хипотезата, върната от алгоритъма след извадка от наблюдения, чийто брой е полиномиален спрямо $1/\varepsilon$ и $1/\delta$, е приблизително правилна (грешка най-много ε) с голяма вероятност (поне $1 - \delta$) , което оправдава терминологията “Вероятно приблизително коректно”.

План на лекцията

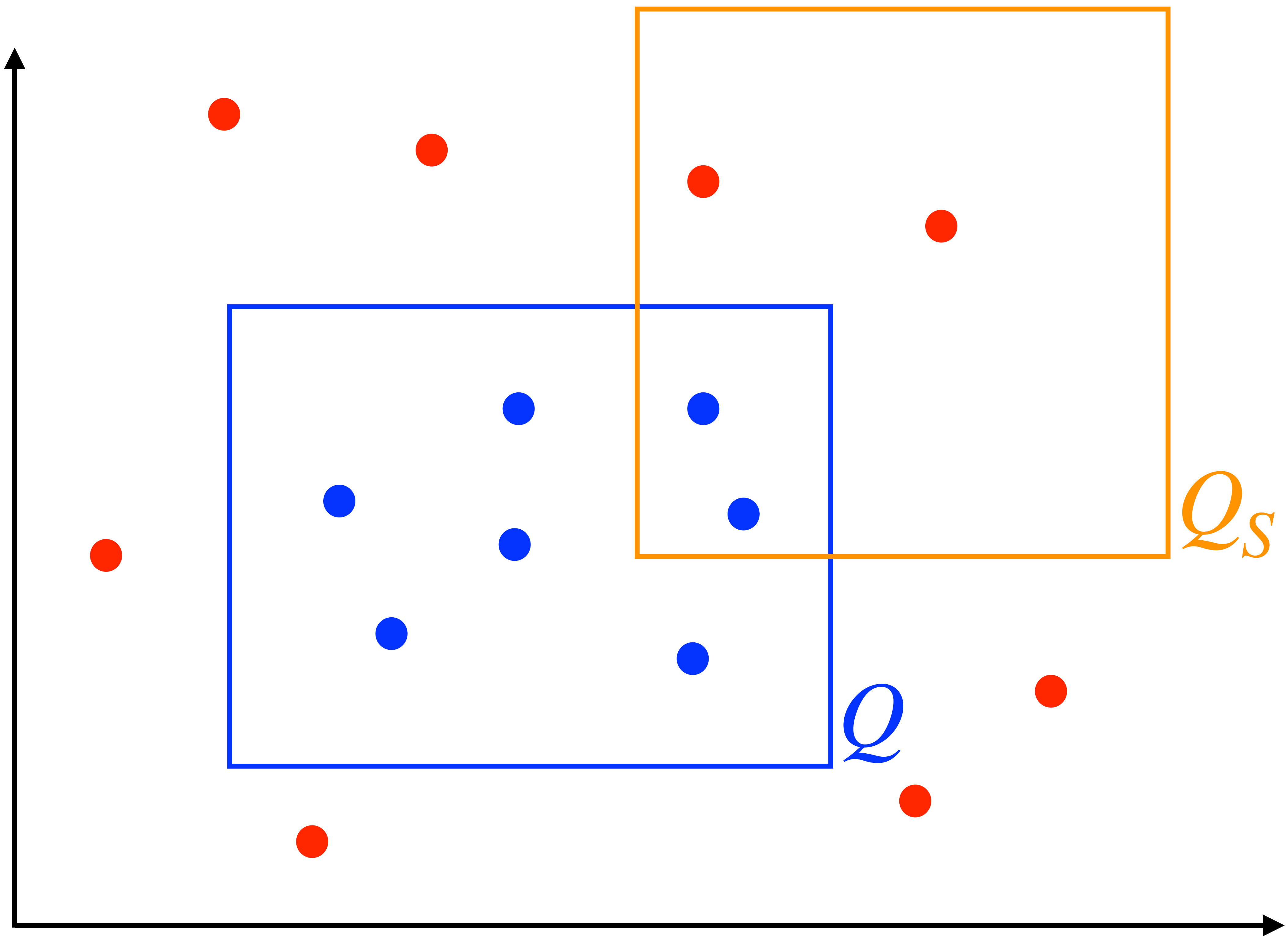
1. Формалности за курса (5 мин)
2. Клъстеризация (10 мин)
3. k-means (15 мин)
4. Варианти и подобрения на K-means (15 мин)
5. PAC-обучение (20 мин)
6. **Пример за PAC-обучение (25 мин)**

Пример за PAC-обучаем класификатор

- Нека $\mathscr{X} = \mathbb{R}^2$ – пространството от всички точки в равнината.

- Нека $\mathscr{C}$ множеството на всички правоъгълници, чиито страни са успоредни на координатните оси. Т.е. всеки класификатор $c \in \mathscr{C}$ е множеството от влизащи в правоъгълник със страни успоредни на осите.

- Задачата на обучаемия е по извадка от наблюдения с техните етикети да избере правоъгълник, който е максимално близък до целевия.



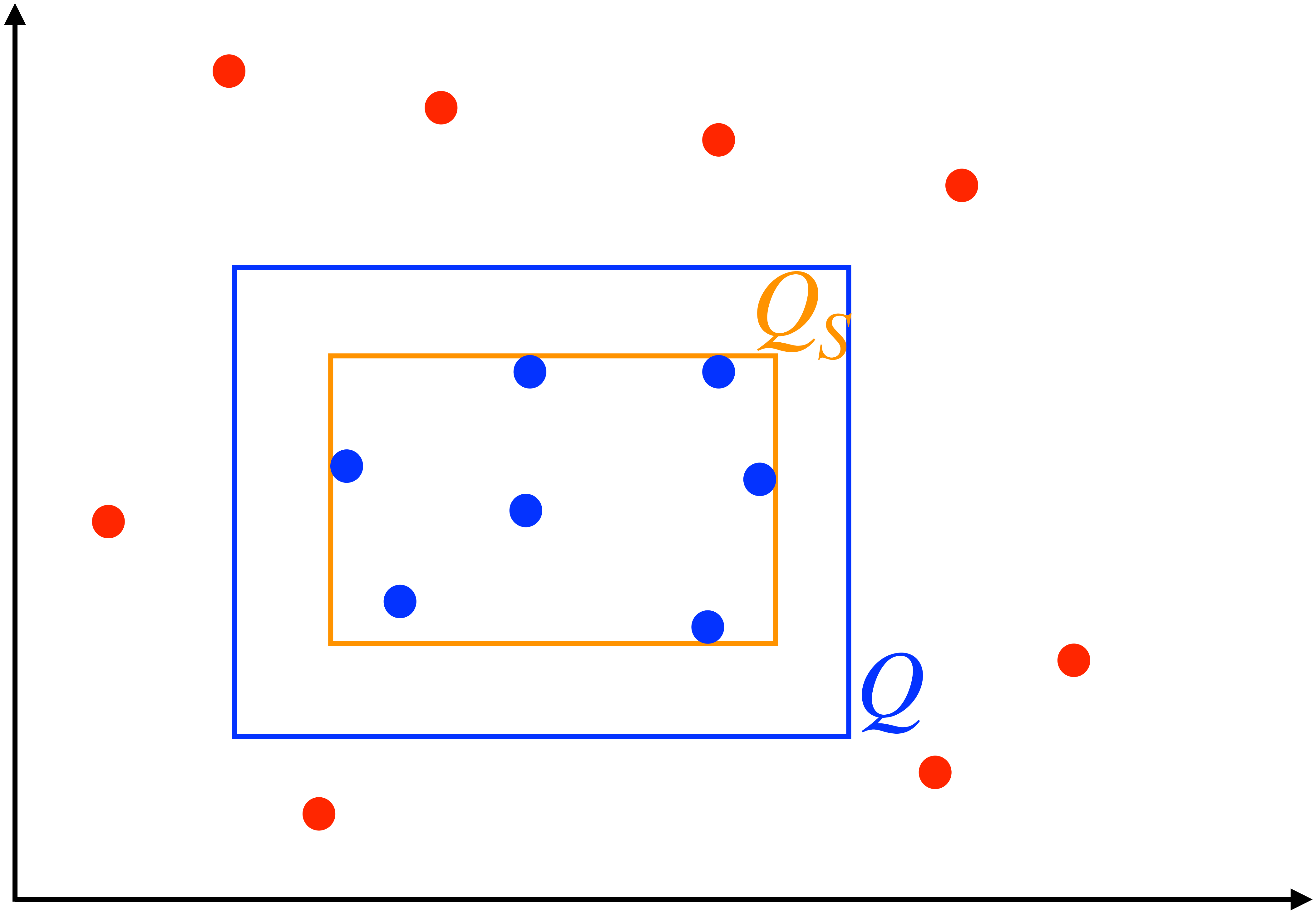
Q_S
Q

Решение

- Нека за дадена наблюдения $S = (x_1, x_2, \ldots, x_m)$ връщаме най-малкия правоъгълник от $\mathscr{C}$, който съдържа всички положителни наблюдения. Т.е. ако
  $P = \{x \in S \mid c(x) = 1\}$ то
  $Q_S = [\min \mathrm{Proj}_1(P), \max \mathrm{Proj}_1(P)] \times [\min \mathrm{Proj}_2(P), \max \mathrm{Proj}_2(P)]$
  т.е. $h_S(x) = \delta_{x \in Q_S}$

- Ясно е, че ако $Q = \{x \in \mathbb{R}^2 \mid c(x) = 1\}$ то $Q \supset Q_S$.

- Следователно: $c(x) \neq h_S(x) \implies x \in Q \backslash Q_S$. Т.е. всички грешки ще са вътре в Q.



Q_S
Q

- За да докажем, че нашия алгоритъм удовлетворява условието за PAC-обучаемост, нека е дадено $\varepsilon > 0$. Нека за търсения класификатор Q, вероятността дадена точка да влезе в Q спрямо разпрелението $\mathscr{D}$ бележим с $\Pr[Q] = \Pr[\{x \mid x \in Q\}]$. Ако $\varepsilon \geq \Pr[Q]$ то вероятността за грешка $R(h_S) = \Pr[\{x \mid h_S(x) \neq c(x)\}] = \Pr[Q \setminus Q_S] \leq \Pr[Q] \leq \varepsilon$. Т.е. винаги (с вероятност 1) грешката ще е по-малка от ε и твърдението е изпълнено.

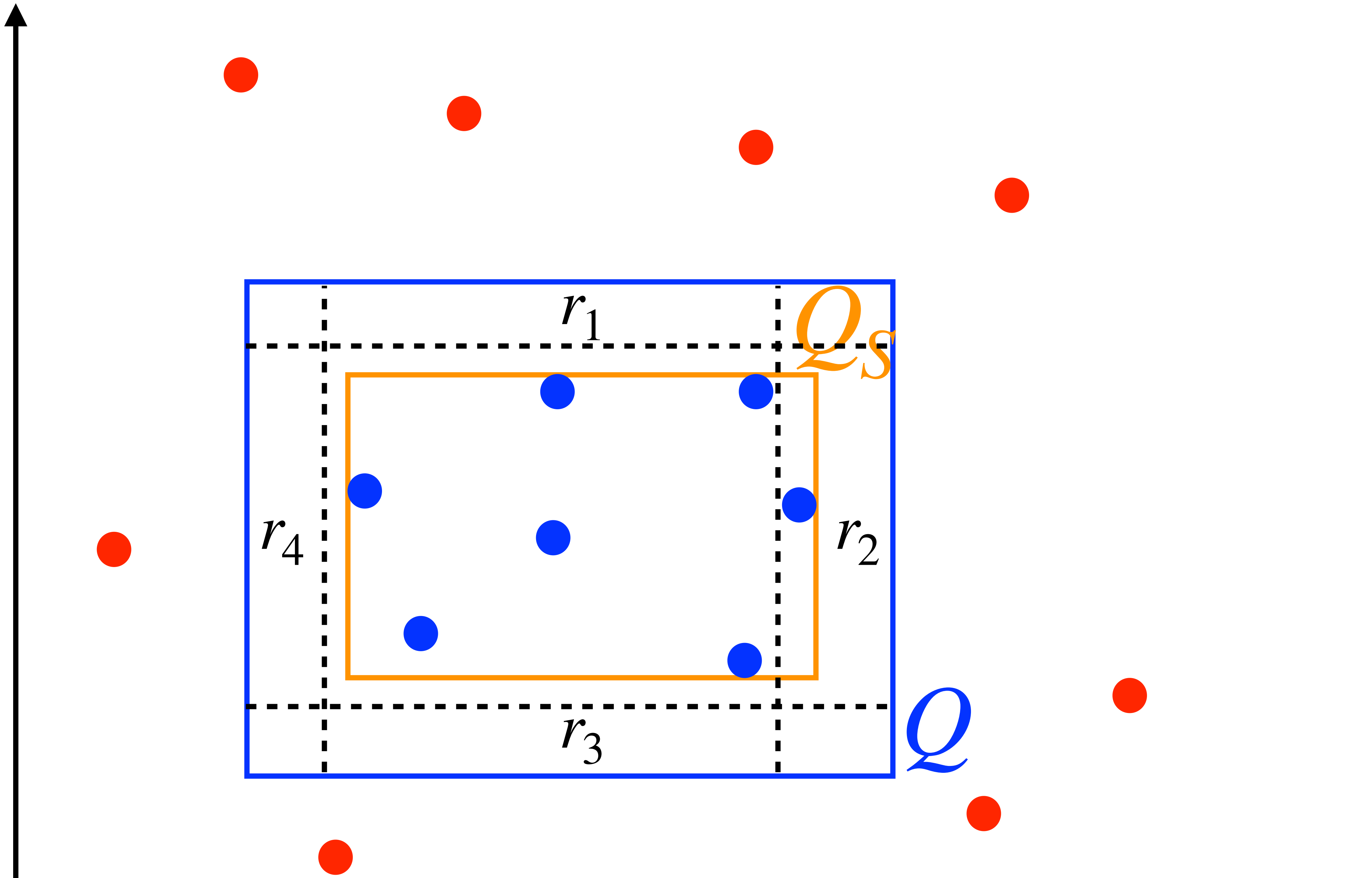
- Нека сега $\varepsilon < \Pr[Q]$. Нека $Q = [l, r] \times [b, t]$. Дефинираме правоъгълниците:

$$r_1 = [l, r] \times [b', t], b' = \sup\{s \,|\, \Pr[[l, r] \times [s, t]] \geq \varepsilon/4\}$$

$$r_2 = [l', r] \times [b, t], l' = \sup\{s \,|\, \Pr[[s, r] \times [b, t]] \geq \varepsilon/4\}$$

$$r_3 = [l, r] \times [b, t'], t' = \inf\{s \,|\, \Pr[[l, r] \times [b, s]] \geq \varepsilon/4\}.$$

$$r_4 = [l, r'] \times [b, t], r' = \inf\{s \,|\, \Pr[[l, s] \times [b, t]] \geq \varepsilon/4\}$$


r_1
Q_S
r_4
r_2
Q
r_3

- Ако допуснем, че Q_S има непразно сечение с всеки от правоъгълниците r_1, r_2, r_3, r_4, то за грешката получаваме:
$$R(h_S) = \Pr[Q \backslash Q_S] \leq \Pr[\bigcup_{i=1}^{4} r_i] \leq \sum_{i=1}^{4} \Pr[r_i] = \varepsilon.$$

- Ако допуснем, че Q_S има празно сечение с поне един от правоъгълниците r_1, r_2, r_3, r_4, то за грешката получаваме:
$$\begin{aligned}\Pr_{S\sim\mathcal{D}^m}[\{S \mid R(h_S) > \varepsilon\}] &\leq \Pr_{S\sim\mathcal{D}^m}[\,\cup_{i=1}^{4}\,\{S \mid Q_S \cap r_i = \emptyset\}] \\ &\leq \sum_{i=1}^{4} \Pr_{S\sim\mathcal{D}^m}[\{S \mid Q_S \cap r_i = \emptyset\}] \\ &\leq 4(1-\varepsilon/4)^m\end{aligned}$$

- Помощно неравенство: $1 - x \leq e^{-x}$. Разглеждаме $f(x) = e^{-x} + x - 1,\ f'(x) = 1 - e^{-x}, f''(x) = e^{-x}$. Следователно функцията е изпъкнала и има единствен минимум при $x = 0$. Но $f(0) = 0$ откъдето следва неравенството.

- $\Pr_{S\sim\mathcal{D}^m}[\{S \mid R(h_S) > \varepsilon\}] \leq 4(1 - \varepsilon/4)^m \leq 4e^{-m\varepsilon/4}$.

- Следователно, за дадено $\delta > 0$, ако изберем $m > \dfrac{4}{\varepsilon}\log\dfrac{4}{\delta}$ получаваме: $\delta > 4e^{-m\varepsilon/4} \geq \Pr_{S\sim\mathcal{D}^m}[\{S \mid R(h_S) > \varepsilon\}]$, откъдето следва $\Pr_{S\sim\mathcal{D}^m}[\{S \mid R(h_S) \leq \varepsilon\}] \geq 1 - \delta$, което трябваше да се покаже. Освен това m е в порядък от $O\left(\dfrac{1}{\varepsilon}\log\dfrac{1}{\delta}\right)$.

Заключение

- Понятието “Вероятно приблизително коректно” PAC-обучение е рамка за математически анализ на алгоритмите за машинното обучение.
- Чрез това и свързаните с него понятия става възможно теоретичното изследване на свойствата и ограниченията на методите за машинно обучение.
- По-задълбочено този подход се разглежда в курса “Теория на машинното обучение и някои нейни приложения в невронните мрежи“
- Няма много практически приложения на този подход, поради което няма да го разглеждаме по-нататък в курса.