Търсене и извличане на информация. Приложение на дълбоко машинно обучение

Стоян Михов

Лекция 7: Клъстеризация във векторно пространство. Логистична регресия. Спускане по градиент.

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. Логистична регресия (15 мин)
6. Обучение чрез спускане по градиента (15 мин)
7. Логистична регресия при много класове (15 мин)

Формалности

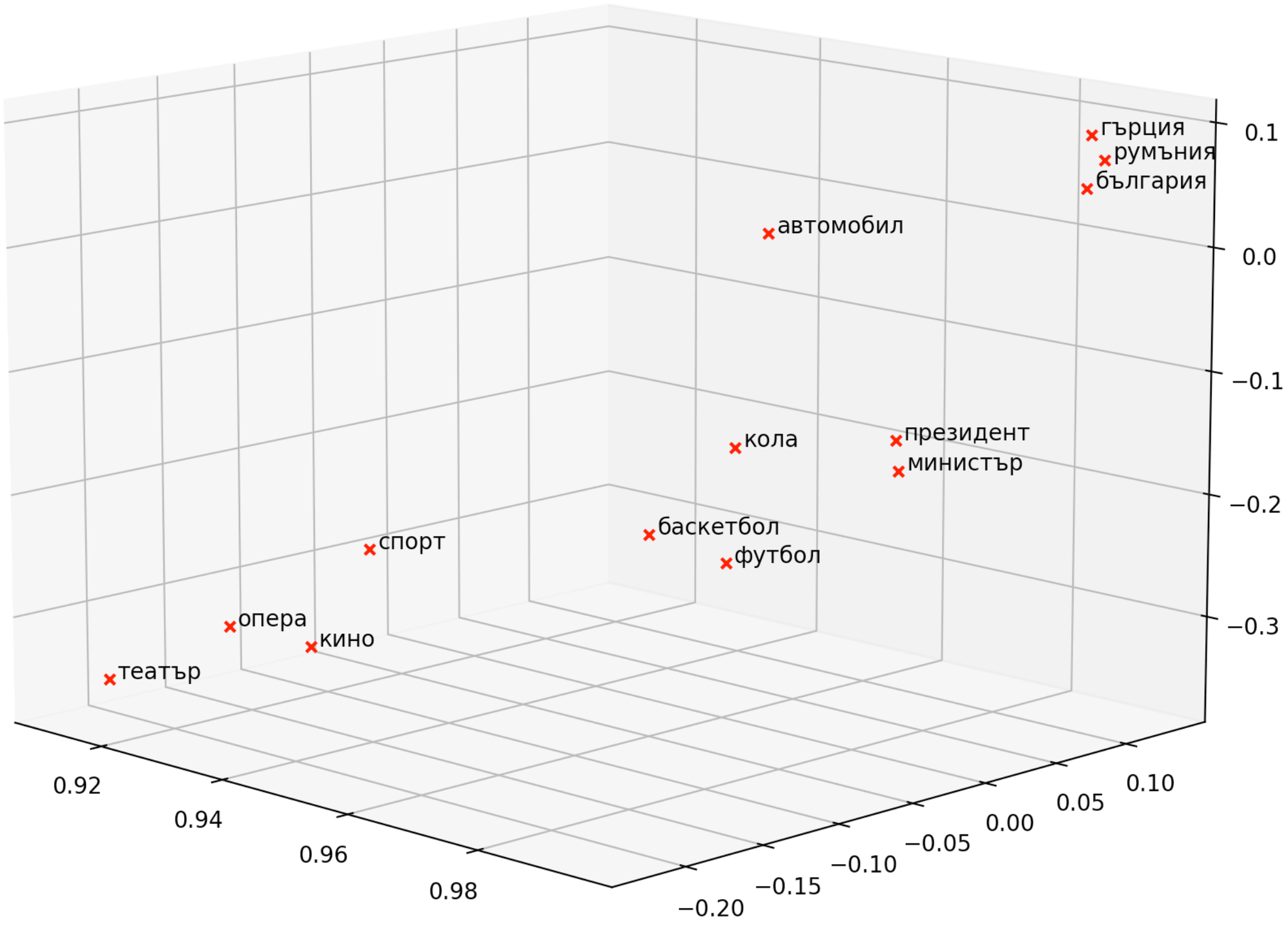
- В Moodle е публикувано Домашно задание 1, което следва да бъде предадено до края на деня на 30.11.2021 г.

- 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. Логистична регресия (15 мин)
6. Обучение чрез спускане по градиента (15 мин)
7. Логистична регресия при много класове (15 мин)

Семантично пространствени релации



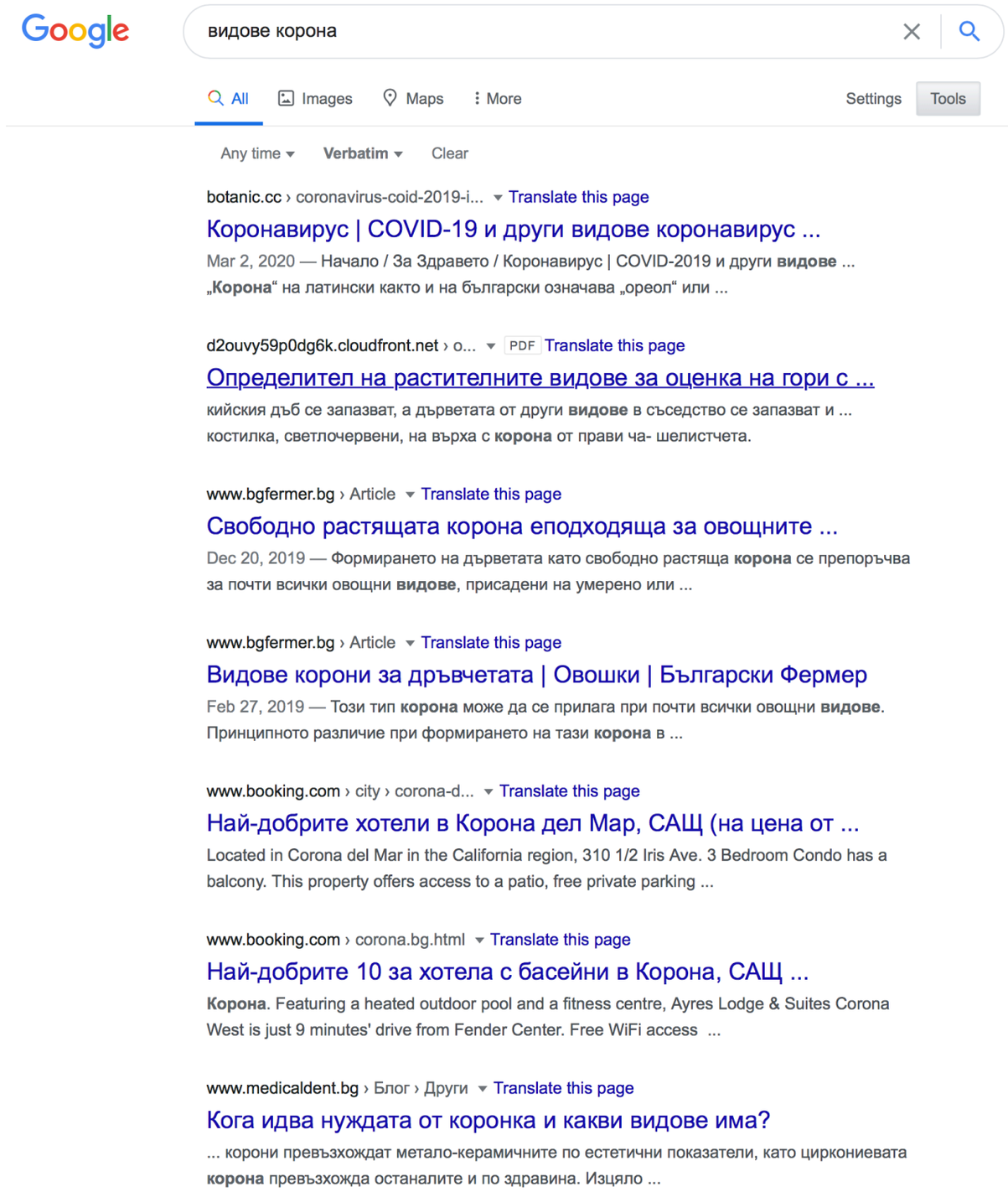
Клъстеризация

- 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

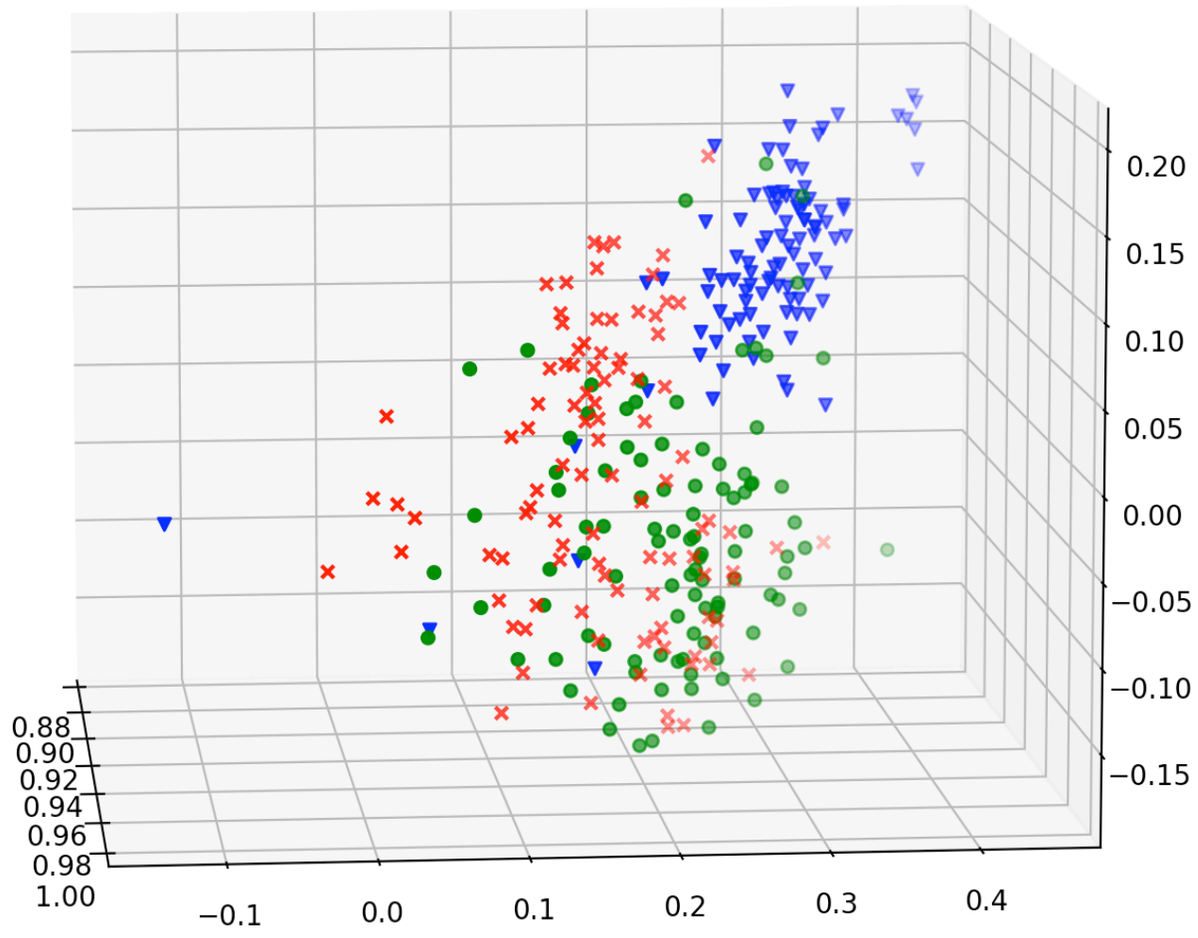
- При класификацията имаме предварително зададени класове и база от класифицирани документи ⇔ при клъстеризацията нямаме нито зададени класове, нито техния брой, нито класифицирани документи.

- 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 въз основа 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)

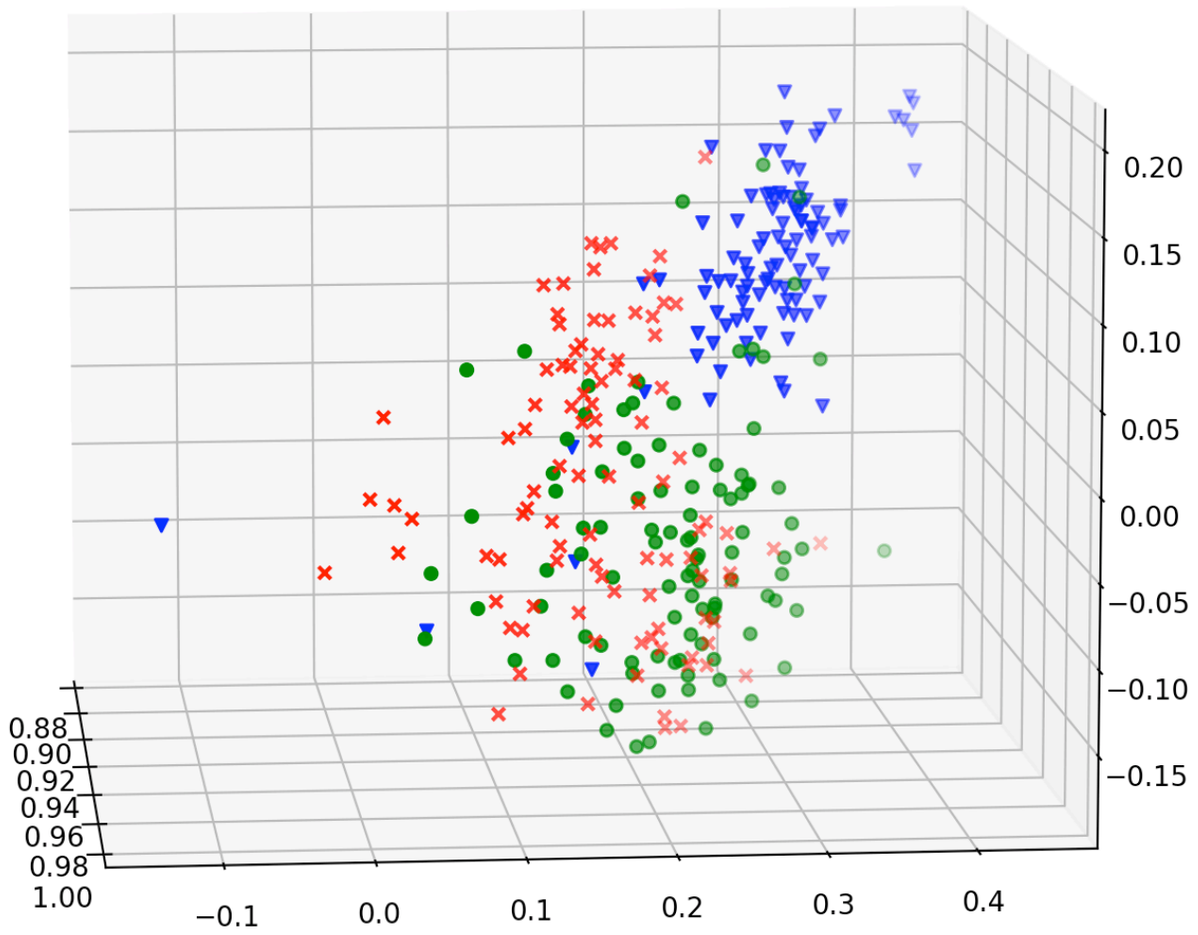
Класификация ⇔ клъстеризация

Класове от документи

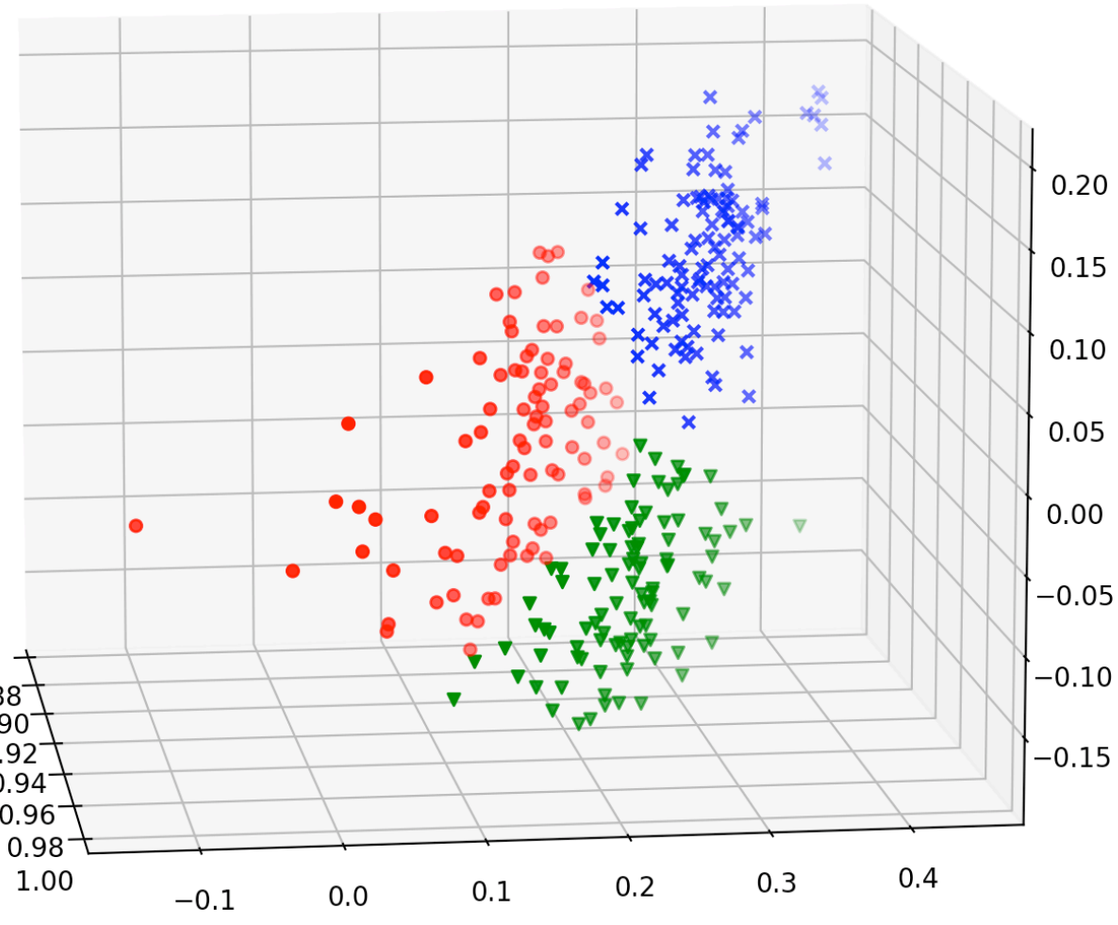


Класификация ⇔ клъстеризация

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. Логистична регресия (15 мин)
6. Обучение чрез спускане по градиента (15 мин)
7. Логистична регресия при много класове (15 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 $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$.

2. За всеки от центроидите μ_k намираме клъстера W_k

$$W_k = \{\mathbf{x} \in \mathbf{X} \mid \arg\min_{i=1}^{K} \|\mathbf{x} - \mu_i\|^2 = k\}.$$

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 )
1    (μ[1],…,μ[K] ) <- SelectSeeds({x[1],…,x[S]}, K )
2    while stopping criterion has not been met do
3        for k <- 1 to K do
4            ω[k] <- {}
5        for i <- 1 to S do
6            k <- argmin_j(μ[j] − x[i])**2
7            ω[k] <- ω[k] U {x[i]}
8        for k <- 1 to K do
9            μ[k] <- 1/|ω[k]| ∑_{x∈ω[k]} x
10   return{μ[1],…,μ[K]}
```

Коректност 1

Твърдение 1: Нека са фиксирани 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 друго разбиване на $\mathbf{X}$. Тогава:

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

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:

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

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

Коректност 2

Твърдение 2: Нека е дадено множество (клъстер) $W \subset \mathbb{R}^N$. 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:

$$\frac{\partial}{\partial \mathbf{y}} \sum_{\mathbf{x}\in W} \|\mathbf{x}-\mathbf{y}\|^2 = 2|W|\mathbf{y} - 2\sum_{\mathbf{x}\in W}\mathbf{x} = 0 \implies \mathbf{y} = \frac{1}{|W|}\sum_{\mathbf{x}\in W}\mathbf{x}$$

$$\frac{\partial^2}{\partial \mathbf{y}^2} \sum_{\mathbf{x}\in W} \|\mathbf{x}-\mathbf{y}\|^2 = \frac{\partial}{\partial \mathbf{y}}\left(2|W|\mathbf{y} - 2\sum_{\mathbf{x}\in W}\mathbf{x}\right) = 2|W|\mathbf{I}$$

Следователно Хесианът е положително дефинитна форма и следователно точката, в която се нулира градиента е глобален минимум.

Следствие: На всяка стъпка от алгоритъма RSS намалява или остава същия.

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. Логистична регресия (15 мин)
6. Обучение чрез спускане по градиента (15 мин)
7. Логистична регресия при много класове (15 мин)

Условия за край

- Функцията RSS е дискретна. 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 вероятностно: **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 $\Pr_l[\mathbf{x}] = \frac{D(\mathbf{x})^2}{\sum_{\mathbf{x}' \in \mathbf{X}} D(\mathbf{x}')^2}$. Избираме центроида μ_{l+1} от $\mathbf{X}$ със случайно разпределение $\Pr_l[\mathbf{x}]$.

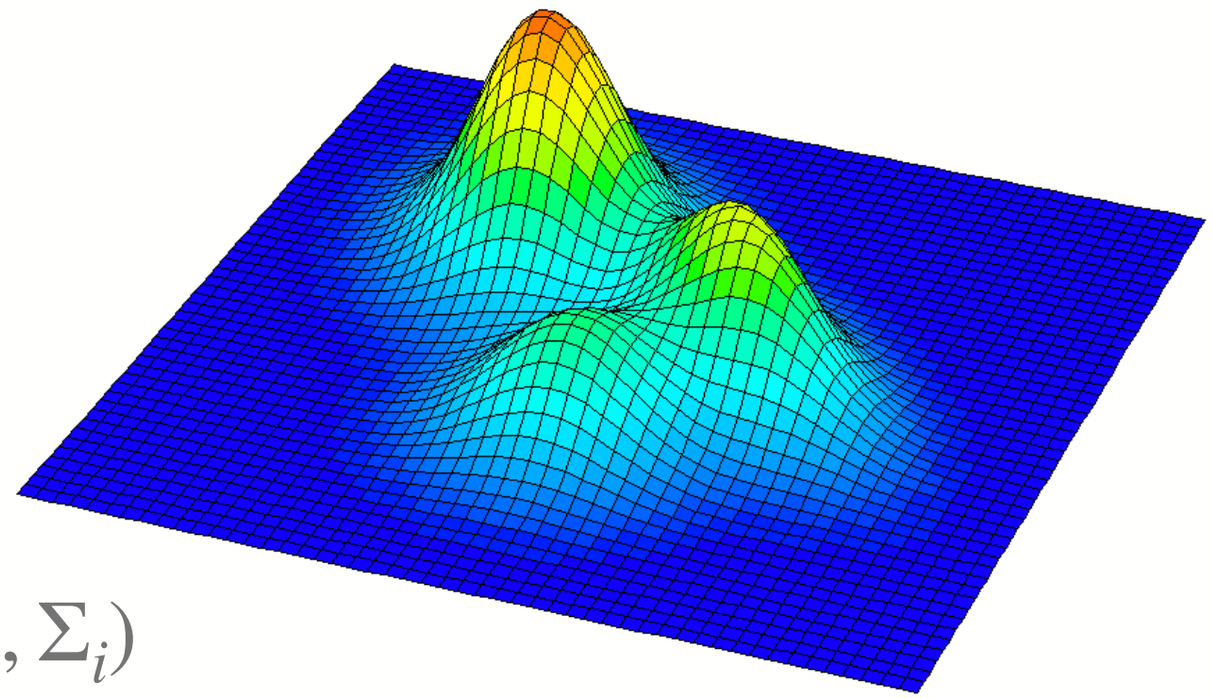
3. Повтаряме стъпки 2-3, докато изберем K центроида.

4. С избраните центроиди изпълняваме алгоритъма k-means

Доказва се, че $\mathbb{E}[\,\mathrm{RSS}_{\mathrm{k-means++}}\,] \leq 8\,(\ln k + 2)\;\mathrm{RSS}_{\min}$

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.

Непрекъснато обобщение на клъстеризацията: Гаусови смески



- Гаусова смеска от ред K:

$$\Pr[\mathbf{X}=\mathbf{x};\bar{c},\bar{\mu},\bar{\Sigma}]=\sum_{i=1}^{K}c_i\mathcal{N}(\mathbf{x};\mu_i,\Sigma_i)$$

$$\mathcal{N}(\mathbf{x};\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^N|\Sigma|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$c_i\geq 0,\sum_{i=1}^{K}c_i=1$$

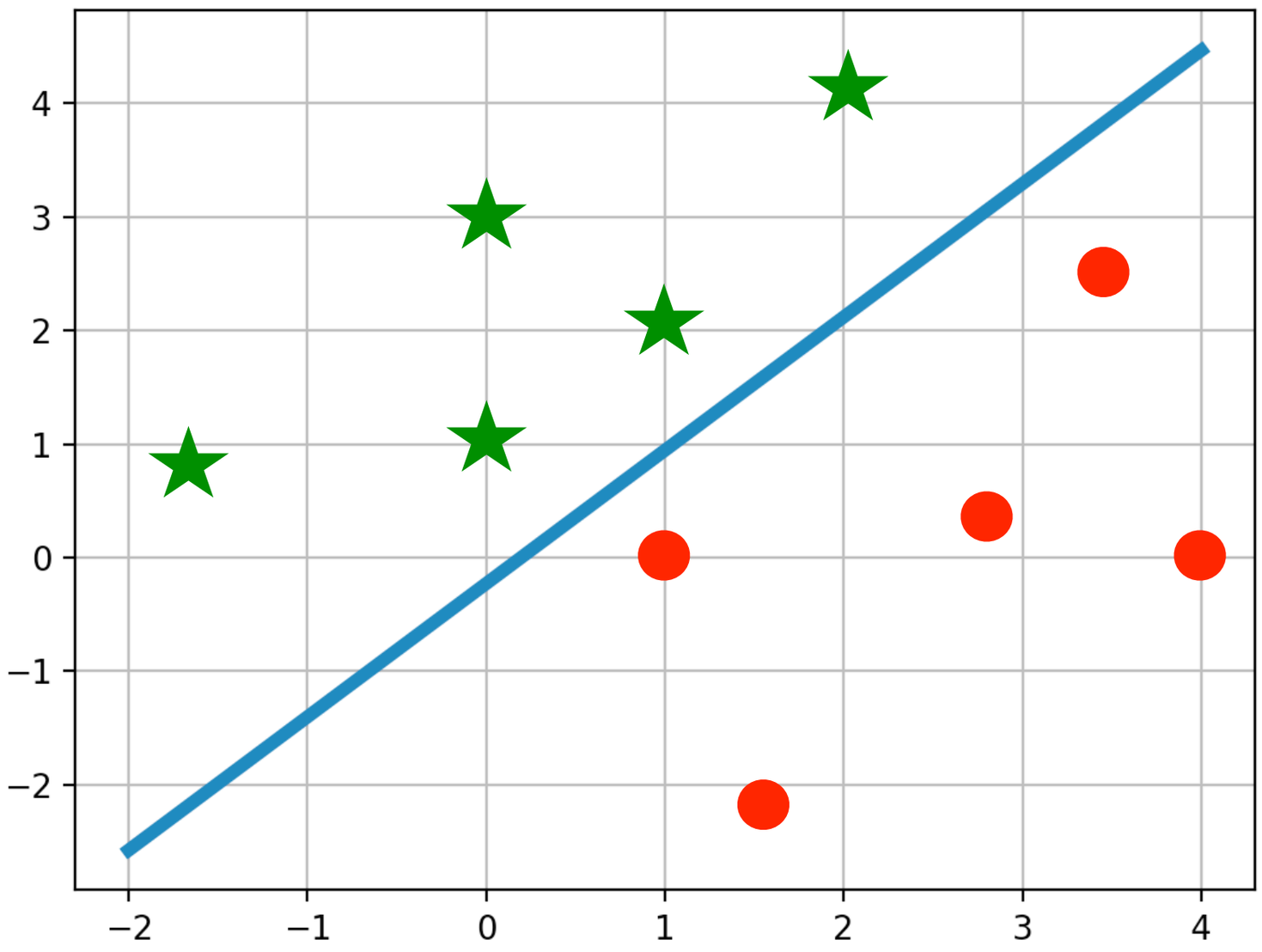
- Търсим: $\bar{c},\bar{\mu},\bar{\Sigma}=\arg\max_{\bar{c},\bar{\mu},\bar{\Sigma}}\prod_{j=1}^{S}\Pr[\mathbf{X}=\mathbf{x}_j;\bar{c},\bar{\mu},\bar{\Sigma}]$

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. **Логистична регресия (15 мин)**
6. Обучение чрез спускане по градиента (15 мин)
7. Логистична регресия при много класове (15 мин)

Линеен класификатор

$$\gamma(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$$



Проблеми при дискретната класификация

- В практиката рядко можем да класифицираме нещата в две крайности — черно или бяло.
- Ако наблюдението е близо до разделителната хиперравнина нашата увереност в класификацията би следвало да е по-ниска.
- Колкото по-далеч е наблюдението от разделителната хиперравнина, толкова по уверени можем да бъдем в правилността на класификацията.
- Желателно е класификатора да върне степен на увереност в класификацията.
- Вместо увереност е по-удобно да върнем вероятност.

Вероятностен линеен класификатор — логистична регресия

- Разглеждаме бинарен класификатор с класове $\mathcal{Y} = \{0,1\}$.
- **Задача:** Разстоянието на вектора $\mathbf{x}$ до разделителната хиперравнина $\mathbf{w}^\top \mathbf{u} + b = 0$ е $\dfrac{|\mathbf{w}^\top \mathbf{x} + b|}{\|\mathbf{w}\|}$ (докажете го).
- Нека с $y \in \mathcal{Y}$ означим класа на наблюдението $\mathbf{x}$. Дефинираме:

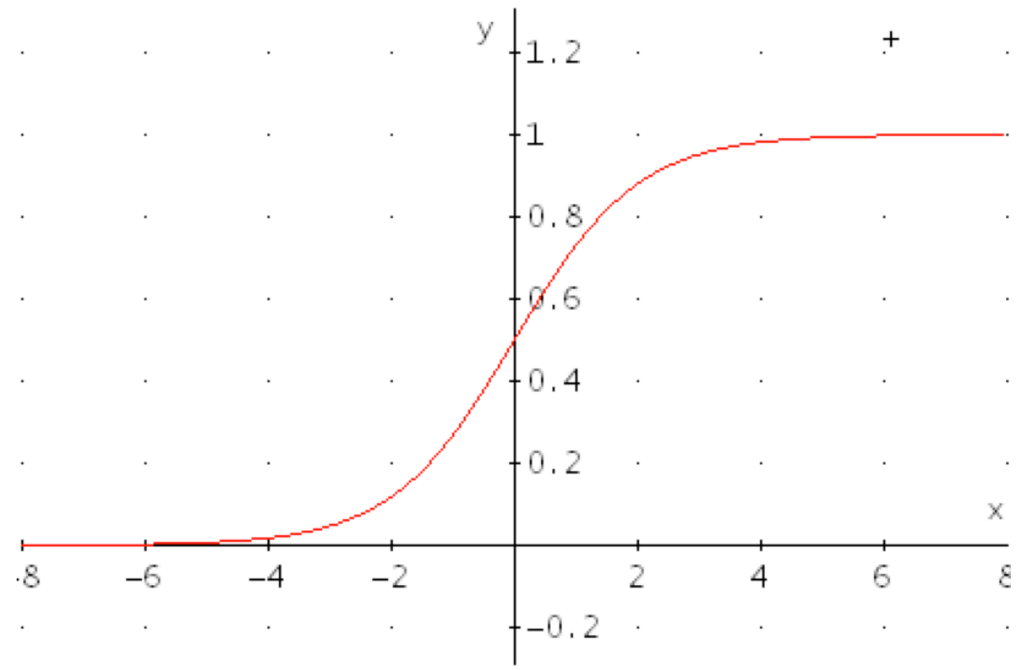
$$\Pr_{\mathbf{w},b}[y = 1 \mid \mathbf{x}] = \sigma(\mathbf{w}^\top \mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}}$$

$$\Pr_{\mathbf{w},b}[y = 0 \mid \mathbf{x}] = 1 - \sigma(\mathbf{w}^\top \mathbf{x} + b) = 1 - \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}}$$

- Използвайки логистичната функция (сигмоид) $\sigma(z) = \dfrac{1}{1 + e^{-z}}$, бинарните предсказания прекарани през лог-линейното преобразуване интерпретираме като вероятности за принадлежност.

Сигмоид — логистичната функция

- $\sigma(z) = \frac{1}{1+e^{-z}} = \frac{e^{z}}{1+e^{z}},$



- $1 - \sigma(z) = \frac{e^{-z}}{1+e^{-z}} = \frac{1}{1+e^{z}}$

- $(\sigma(z))' = \frac{e^{-z}}{(1+e^{-z})^2} = e^{-z}\sigma^2(z) = (1-\sigma(z))\sigma(z)$

- $(\log \sigma(z))' = (\log 1 - log(1+e^{-z}))' = -\frac{-e^{-z}}{1+e^{-z}} = 1 - \sigma(z)$

- $(\log(1-\sigma(z)))' = (\log e^{-z} - log(1+e^{-z}))' = -1 + (1-\sigma(z)) = -\sigma(z)$

Целева функция

- Дадена е извадка от наблюдения заедно със съответните им етикети $X = ((\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})), (\mathbf{x}^{(i)}, y^{(i)}) \in \mathbb{R}^N \times \{0,1\}.$

- За дадена права $\mathbf{w}^\top \mathbf{x} + b = 0, \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}$, правдоподобието е:
  $L_{\mathbf{w},b}((\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})) = \prod_{i=1}^{m} \Pr_{\mathbf{w},b}[y = y^{(i)} \mid \mathbf{x}^{(i)}]$, където
  $\Pr_{\mathbf{w},b}[y = 1 \mid \mathbf{x}^{(i)}] = \sigma(\mathbf{w}^\top \mathbf{x}^{(i)} + b)$ и
  $\Pr_{\mathbf{w},b}[y = 0 \mid \mathbf{x}^{(i)}] = 1 - \sigma(\mathbf{w}^\top \mathbf{x}^{(i)} + b)$

- Целта е да намерим параметрите, при които се максимизира правдоподобието:
  $\hat{\mathbf{w}}, \hat{b} = \arg\max_{\mathbf{w},b} L_{\mathbf{w},b}((\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})) = \arg\max_{\mathbf{w},b} \prod_{i=1}^{m} \Pr_{\mathbf{w},b}[y = y^{(i)} \mid \mathbf{x}^{(i)}]$

Максимизиране на правдоподобието = минимизиране на емпиричната кросентропията

$$\hat{\mathbf{w}}, \hat{b} = \arg\max_{\mathbf{w},b} \prod_{i=1}^{m} \Pr_{\mathbf{w},b}[y = y^{(i)} \,|\, \mathbf{x}^{(i)}] =$$

$$= \arg\max_{\mathbf{w},b} \log \prod_{i=1}^{m} \Pr_{\mathbf{w},b}[y = y^{(i)} \,|\, \mathbf{x}^{(i)}] =$$

$$= \arg\min_{\mathbf{w},b} - \log \prod_{i=1}^{m} \Pr_{\mathbf{w},b}[y = y^{(i)} \,|\, \mathbf{x}^{(i)}] =$$

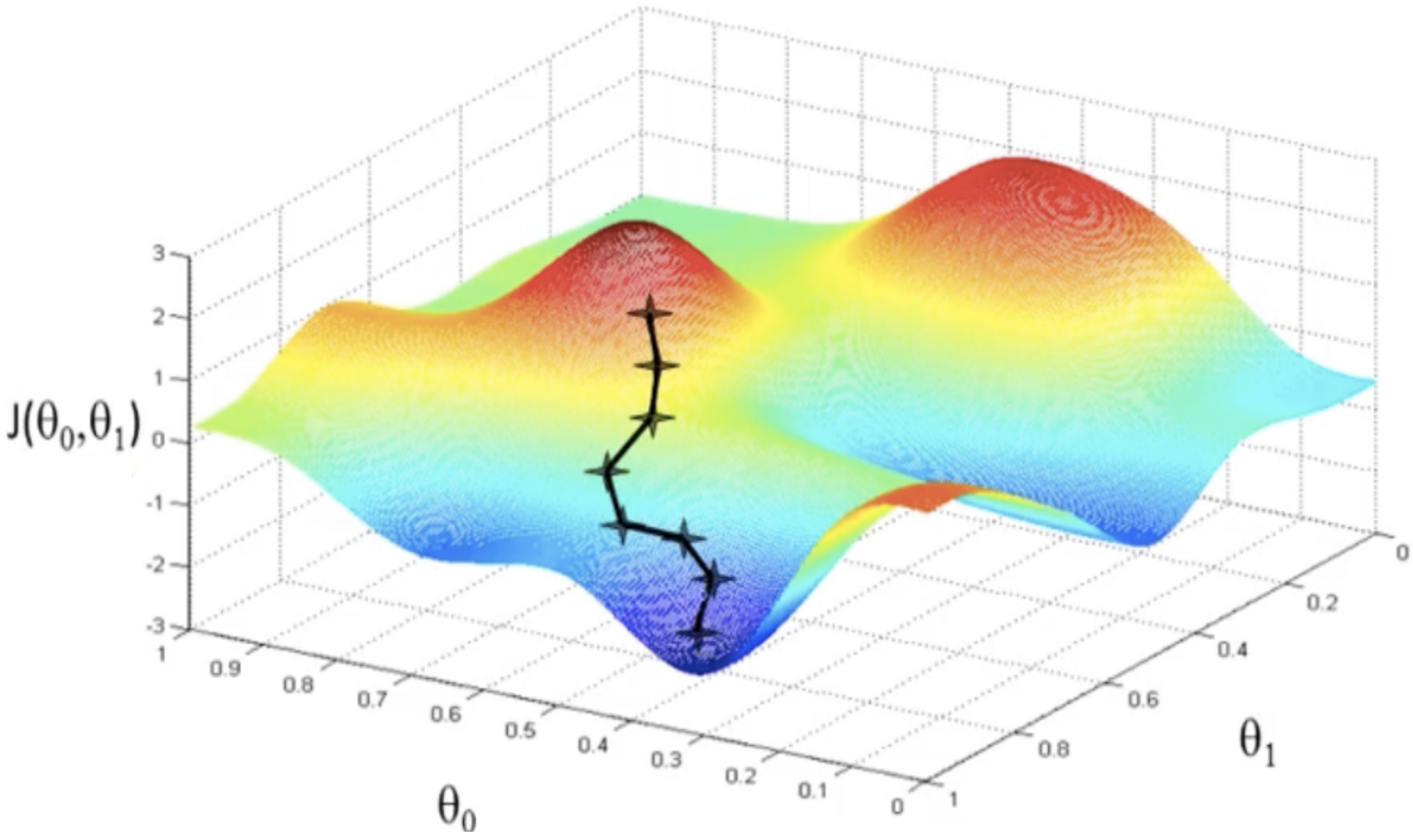
$$= \arg\min_{\mathbf{w},b} - \frac{1}{m} \sum_{i=1}^{m} \log \Pr_{\mathbf{w},b}[y = y^{(i)} \,|\, \mathbf{x}^{(i)}] =$$

$$= \arg\min_{\mathbf{w},b} H_X(\Pr \,\|\Pr_{\mathbf{w},b})$$

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. Логистична регресия (15 мин)
6. **Обучение чрез спускане по градиента (15 мин)**
7. Логистична регресия при много класове (15 мин)

Интуиция за спускането по градиента



Обучение чрез спускане по градиента

- При по-сложни функции аналитичното намиране на параметрите, при които се минимизира кросентропията, е трудно или дори невъзможно.

- Нека е дадена целева функция $J(\theta)$, където параметрите, по които минимизираме, са θ, която е частично-диференцируема относно θ.

- **Спускането по градиента** е следния итеративен алгоритъм:

    1. Започваме с начална стойност на параметрите θ_0.

    2. На стъпка $i + 1$ намираме: $\theta_{i+1} = \theta_i - \alpha \frac{\partial}{\partial \theta} J(\theta_i)$.

    3. Повтаряме стъпки 2-3 докато не удовлетворим условие за край.

- Параметърът α наричаме **скорост на обучение**. Той оказва съществено значение за броя на итерациите и намирането на минимум.

Намиране на градиента при логистичната регресия

$$\frac{\partial}{\partial b} \log \Pr_{\mathbf{w},b}[y = 1 \,|\, \mathbf{x}] = \frac{\partial}{\partial b} \log \sigma(\mathbf{w}^\top \mathbf{x} + b) =$$

- $$= (1 - \sigma(\mathbf{w}^\top \mathbf{x} + b)) \frac{\partial}{\partial b} (\mathbf{w}^\top \mathbf{x} + b) = 1 - \sigma(\mathbf{w}^\top \mathbf{x} + b)$$

$$\frac{\partial}{\partial b} \log \Pr_{\mathbf{w},b}[y = 0 \,|\, \mathbf{x}] = \frac{\partial}{\partial b} \log(1 - \sigma(\mathbf{w}^\top \mathbf{x} + b)) =$$

- $$= - \sigma(\mathbf{w}^\top \mathbf{x} + b) \frac{\partial}{\partial b} (\mathbf{w}^\top \mathbf{x} + b) = 0 - \sigma(\mathbf{w}^\top \mathbf{x} + b)$$

- Следодателно: $\frac{\partial}{\partial b} \log \Pr_{\mathbf{w},b}[y = y^{(i)} \,|\, \mathbf{x}^{(i)}] = y^{(i)} - \sigma(\mathbf{w}^\top \mathbf{x}^{(i)} + b)$

$$\frac{\partial}{\partial \mathbf{w}} \log \Pr_{\mathbf{w},b}[y = 1 \,|\, \mathbf{x}] = \frac{\partial}{\partial \mathbf{w}} \log \sigma(\mathbf{w}^\top \mathbf{x} + b) =$$

- $$= (1 - \sigma(\mathbf{w}^\top \mathbf{x} + b)) \frac{\partial}{\partial \mathbf{w}} (\mathbf{w}^\top \mathbf{x} + b) = (1 - \sigma(\mathbf{w}^\top \mathbf{x} + b))\mathbf{x}$$

$$\frac{\partial}{\partial \mathbf{w}} \log \Pr_{\mathbf{w},b}[y = 0 \,|\, \mathbf{x}] = \frac{\partial}{\partial \mathbf{w}} \log(1 - \sigma(\mathbf{w}^\top \mathbf{x} + b)) =$$

- $$= -\sigma(\mathbf{w}^\top \mathbf{x} + b) \frac{\partial}{\partial \mathbf{w}} (\mathbf{w}^\top \mathbf{x} + b) = 0 - \sigma(\mathbf{w}^\top \mathbf{x} + b)\mathbf{x}$$

- Следодателно:
  $$\frac{\partial}{\partial \mathbf{w}} \log \Pr_{\mathbf{w},b}[y = y^{(i)} \,|\, \mathbf{x}^{(i)}] = (y^{(i)} - \sigma(\mathbf{w}^\top \mathbf{x}^{(i)} + b))\mathbf{x}^{(i)}$$

Градиент на логистична регресия

- $$\frac{\partial}{\partial b} H_X(\mathrm{Pr} \,\|\mathrm{Pr}_{\mathbf{w},b}) = -\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \sigma(\mathbf{w}^\top \mathbf{x}^{(i)} + b))$$
- $$\frac{\partial}{\partial \mathbf{w}} H_X(\mathrm{Pr} \,\|\mathrm{Pr}_{\mathbf{w},b}) = -\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \sigma(\mathbf{w}^\top \mathbf{x}^{(i)} + b))\mathbf{x}^{(i)}$$

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. Логистична регресия (15 мин)
6. Обучение чрез спускане по градиента (15 мин)
7. **Логистична регресия при много класове (15 мин)**

Логистична регресия при много класове

- Разглеждаме класификатор при класове $\mathcal{Y} = \{1,2,\ldots,K\}$.
- Можем да разглеждаме K ,разделителни хиперравнини $\mathbf{w}_c^\top \mathbf{x} + b_c = 0$.
- Дадена е извадка от наблюдения заедно със съответните им етикети $X = ((\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})), (\mathbf{x}^{(i)}, y^{(i)}) \in \mathbb{R}^N \times \{1,2,\ldots,K\}$
- Дефинираме $W \in \mathbb{R}^{K \times N}$, $W = \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_K \end{bmatrix}$, $\mathbf{b} \in \mathbb{R}^K$, $\mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_K \end{bmatrix}$
- $\Pr_{W,\mathbf{b}}[y = c \,|\, \mathbf{x}] = \text{softmax}(W\mathbf{x} + \mathbf{b})_c = \dfrac{e^{(W\mathbf{x}+\mathbf{b})_c}}{\sum_{j=1}^{K} e^{(W\mathbf{x}+\mathbf{b})_j}}$
- Обучаваме модела, като минимизираме кросентропията $H_X[\Pr \,||\Pr_{W,\mathbf{b}}]$.
- **Задача**: Покажете аналитично, че при модел с 2 класа softmax е еквивалентен на сигмоид.

Верижно правило за диференциране на композиция на функции на много променливи

- Нека $f : \mathbb{R}^m \to \mathbb{R}$, $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^n$
- Тогава:

$$\frac{\partial}{\partial x_k} f(\mathbf{g}(\mathbf{x})) = \frac{\partial}{\partial x_k} f(g_1(x_1, \ldots, x_n), \ldots, g_m(x_1, \ldots, x_n)) = \sum_{j=1}^{m} \frac{\partial f}{\partial g_j} \frac{\partial g_j}{\partial x_k}$$

- Векторен запис с използване на **якобияни**:

$$\frac{\partial}{\partial x_k} f(\mathbf{g}(\mathbf{x})) = \left(\frac{\partial f}{\partial \mathbf{g}} \right)^{\mathsf{T}} \frac{\partial \mathbf{g}}{\partial x_k}, \text{ тук } \frac{\partial f}{\partial \mathbf{g}}, \frac{\partial \mathbf{g}}{\partial x_k} \in \mathbb{R}^m,$$

$$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{g}(\mathbf{x})) = \left(\frac{\partial f}{\partial \mathbf{g}} \right)^{\mathsf{T}} \frac{\partial \mathbf{g}}{\partial \mathbf{x}}, \text{ тук } \frac{\partial \mathbf{g}}{\partial \mathbf{x}} \in \mathbb{R}^{m \times n}.$$

$$\arg\min_{W,\mathbf{b}} H_X(\Pr \| \Pr_{W,\mathbf{b}}) = \arg\min_{W,\mathbf{b}} -\frac{1}{m}\sum_{i=1}^{m} \log \Pr_{W,\mathbf{b}}[y = y^{(i)} \mid \mathbf{x}^{(i)}] =$$

$$= \arg\min_{W,\mathbf{b}} -\frac{1}{m}\sum_{i=1}^{m} \log \operatorname{softmax}(W\mathbf{x}^{(i)} + \mathbf{b})_{y^{(i)}} =$$

$$= \arg\min_{W,\mathbf{b}} -\frac{1}{m}\sum_{i=1}^{m} \log \frac{e^{(W\mathbf{x}^{(i)}+\mathbf{b})_{y^{(i)}}}}{\sum_{j=1}^{K} e^{(W\mathbf{x}^{(i)}+\mathbf{b})_j}}$$

$$\frac{\partial}{\partial u_k} \log \operatorname{softmax}(\mathbf{u})_y = \frac{\partial}{\partial u_k} \log \frac{e^{u_y}}{\sum_{j=1}^{K} e^{u_j}} = \frac{\partial}{\partial u_k} u_y - \frac{\partial}{\partial u_k} \log \sum_{j=1}^{K} e^{u_j} =$$

$$= \delta_{k=y} - \frac{1}{\sum_{j=1}^{K} e^{u_j}} \frac{\partial}{\partial u_k} \sum_{j=1}^{K} e^{u_j} = \delta_{k=y} - \frac{e^{u_k}}{\sum_{j=1}^{K} e^{u_j}} =$$

$$= \delta_{k=y} - \operatorname{softmax}(\mathbf{u})_k$$

$$\frac{\partial}{\partial \mathbf{u}} \log \operatorname{softmax}(\mathbf{u})_y = \bar{\delta}_y - \operatorname{softmax}(\mathbf{u}), \text{ където } \bar{\delta}_y \in \mathbb{R}^K, (\bar{\delta}_y)_k = \delta_{k=y}$$

$$\frac{\partial}{\partial \mathbf{b}} \log \frac{e^{(W\mathbf{x}+\mathbf{b})_y}}{\sum_{j=1}^{K} e^{(W\mathbf{x}+\mathbf{b})_j}} = (\bar{\delta}_y - \text{softmax}(W\mathbf{x}+\mathbf{b}))^\top \left(\frac{\partial}{\partial \mathbf{b}} (W\mathbf{x}+\mathbf{b}) \right) =$$

$$= (\bar{\delta}_y - \text{softmax}(W\mathbf{x}+\mathbf{b}))^\top \mathbf{I} =$$

$$= \bar{\delta}_y - \text{softmax}(W\mathbf{x}+\mathbf{b})$$

Разглеждаме функцията: $u : \mathbb{R}^{KN} \to \mathbb{R}^K, u(W) = W\mathbf{x} + \mathbf{b}$. Якобиянът $\frac{\partial \mathbf{u}}{\partial W}$ е матрица $\mathbb{R}^{K \times KN}$.

$$\left(\frac{\partial \mathbf{u}}{\partial W_{p,q}} \right)_k = \frac{\partial \mathbf{u}_k}{\partial W_{p,q}} = \frac{\partial \sum_{l=1}^{N} W_{k,l} x_l}{\partial W_{p,q}} = \begin{cases} 0 & \text{if } k \neq p \\ x_q & \text{if } k = p \end{cases} = \delta_{p=k} x_q$$

$$\frac{\partial}{\partial W} \log \frac{e^{(W\mathbf{x}+\mathbf{b})_y}}{\sum_{j=1}^{K} e^{(W\mathbf{x}+\mathbf{b})_j}} = (\bar{\delta}_y - \text{softmax}(W\mathbf{x}+\mathbf{b}))^\top \left(\frac{\partial}{\partial W} (W\mathbf{x}+\mathbf{b}) \right) =$$

$$= (\bar{\delta}_y - \text{softmax}(W\mathbf{x}+\mathbf{b}))^\top \frac{\partial \mathbf{u}}{\partial W} =$$

$$= (\bar{\delta}_y - \text{softmax}(W\mathbf{x}+\mathbf{b})) \otimes \mathbf{x}$$

Защото, ако $\mathbf{v} = \bar{\delta}_y - \text{softmax}(W\mathbf{x}+\mathbf{b})$, то:

$$\left(\mathbf{v}^\top \frac{\partial \mathbf{u}}{\partial W} \right)_{p,q} = \mathbf{v}^\top \frac{\partial \mathbf{u}}{\partial W_{p,q}} = \sum_{k=1}^{K} \mathbf{v}_k \left(\frac{\partial \mathbf{u}}{\partial W_{p,q}} \right)_k = \sum_{k=1}^{K} \mathbf{v}_k \delta_{p=k} \mathbf{x}_q = \mathbf{v}_p \mathbf{x}_q$$

Градиент на логистична регресия при много класове

- $$\frac{\partial}{\partial \mathbf{b}} H_X(\Pr \| \Pr_{W,\mathbf{b}}) = -\frac{1}{m} \sum_{i=1}^{m} (\bar{\delta}_{y^{(i)}} - \text{softmax}(W\mathbf{x}^{(i)} + \mathbf{b}))$$
- $$\frac{\partial}{\partial W} H_X(\Pr \| \Pr_{W,\mathbf{b}}) = -\frac{1}{m} \sum_{i=1}^{m} (\bar{\delta}_{y^{(i)}} - \text{softmax}(W\mathbf{x}^{(i)} + \mathbf{b})) \otimes \mathbf{x}^{(i)}$$

Заключение

- Клъстеризацията k-means
  - Сравнително прост метод за научаване на общи закономерности от наблюденията (чрез групиране) без учител.
- Логистичната регресия
  - Обобщение на линеен класификатор при използване на логистично вероятностно разпределение
- Спускането по градиента
  - Общ метод за намиране на локален минимум на произволно сложни гладки функции