

Търсене и извличане на информация. Приложение на дълбоко машинно обучение

Стоян Михов



Лекция 5: Ентропия и кросентропия. Влагане на думи в многомерно векторно пространство.

План на лекцията

- 1. Формалности за курса (5 мин)**
2. Вероятностно очакване, вариация, ковариация. Емпирично разпределение (20 мин)
3. Ентропия, перплексия и оценяване на езиков модел (20 мин)
4. Семантично разширяване на заявката (10 мин)
5. Влагане на думи във многомерно разрежено векторно пространство (10 мин)
6. Влагане на термовете в контекстно пространство (15 мин)
7. Ранкиране на документи в контекстно пространство (5 мин)

План на лекцията

1. Формалности за курса (5 мин)
- 2. Вероятностно очакване, вариация, ковариация. Емпирично разпределение (20 мин)**
3. Ентропия, перплексия и оценяване на езиков модел (20 мин)
4. Семантично разширяване на заявката (10 мин)
5. Влагане на думи във многомерно разрежено векторно пространство (10 мин)
6. Влагане на термовете в контекстно пространство (15 мин)
7. Ранкиране на документи в контекстно пространство (5 мин)

Вероятностно очакване

- **Очакване на случайна величина** X означаваме с $\mathbb{E}[X]$ и дефинираме като

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} \Pr[\omega] X(\omega) = \sum_{x \in X(\Omega)} \Pr[X = x] x$$

- Пример: Очакване на честен зар: $\sum_{i=1}^6 \frac{1}{6} i = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$

- Петербургски парадокс: Казиното предлага хазартна игра, в която на всеки етап се хвърля честна монета. Първоначалният залог започва от 2 лева и се удвоява всеки път, когато се падне тура. Първият път, когато се появят ези, играта приключва и играчът печели какъвто е текущият залог. Така играчът печели 2 лева, ако при първото хвърляне се падне ези, 4 лева, ако първото хвърляне се падне тура, а второто ези, 8 лева, ако първите две хвърляния са тура, а третото е ези и т.н. Какво е очакването за печалбата?

Вероятностно очакване

- **Очакване на случайна величина** X означаваме с $\mathbb{E}[X]$ и дефинираме като

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} \Pr[\omega] X(\omega) = \sum_{x \in X(\Omega)} \Pr[X = x] x$$

- Пример: Очакване на честен зар: $\sum_{i=1}^6 \frac{1}{6} i = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$

- Петербургски парадокс: Казиното предлага хазартна игра, в която на всеки етап се хвърля честна монета. Първоначалният залог започва от 2 лева и се удвоява всеки път, когато се падне тура. Първият път, когато се появят ези, играта приключва и играчът печели какъвто е текущият залог. Така играчът печели 2 лева, ако при първото хвърляне се падне ези, 4 лева, ако първото хвърляне се падне тура, а второто ези, 8 лева, ако първите две хвърляния са тура, а третото е ези и т.н. Какво е очакването за печалбата?

- За някои случайни величини очакването може да е безкрайно: Петербургски парадокс:

С вероятност $\frac{1}{2^n}$ стойността е 2^n . Тогава за очакването е $\mathbb{E}[X] = \sum_{n=1}^{\infty} \frac{1}{2^n} 2^n = \infty$.

- **Очакване на функция на случайна величина** $Y = f(X)$ означаваме с $\mathbb{E}[Y] = \mathbb{E}[f(X)]$

• **Свойство:**
$$\mathbb{E}[f(X)] = \sum_{y \in Y(\Omega)} \Pr[Y = y] y = \sum_{x \in X(\Omega)} \Pr[X = x] f(x)$$

Докажете свойството!

- Нека X_1, X_2, \dots, X_n са случайни величини над Ω . Тогава $\mathbb{E}[f(X_1, X_2, \dots, X_n)] =$

$$\sum_{x_1 \in X_1(\Omega), x_2 \in X_2(\Omega), \dots, x_n \in X_n(\Omega)} \Pr[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n] f(x_1, x_2, \dots, x_n)$$

- **Свойства:**

$$\sum_{x \in X(\Omega), y \in Y(\Omega)} \Pr[X = x, Y = y] f(x) = \sum_{x \in X(\Omega)} f(x) \sum_{y \in Y(\Omega)} \Pr[X = x, Y = y] =$$

•
$$= \sum_{x \in X(\Omega)} f(x) \Pr[X = x] = \mathbb{E}[f(X)]$$

•
$$\mathbb{E}[af(X) + bg(Y)] = a\mathbb{E}[f(X)] + b\mathbb{E}[g(Y)]$$

• Ако X и Y са независими случайни величини, то $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

Вариация

- **Вариацията (дисперсията) на случайна величина X** означаваме с $\text{Var}[X]$ и дефинираме като $\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$
- **Стандартна отклонение (девиация) на случайна величина X** означаваме с σ_X и дефинираме като $\sigma_X = \sqrt{\text{Var}[X]}$

Свойства:

- $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- $\text{Var}[aX] = a^2 \text{Var}[X]$
- Ако X и Y са независими то $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$

Ковариация

- Ковариацията на две случайни величини X и Y означаваме с $\text{Cov}(X, Y)$ и дефинираме като: $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$.

Свойства:

- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(X + X', Y) = \text{Cov}(X, Y) + \text{Cov}(X', Y)$, $\text{Cov}(aX, Y) = a \text{Cov}(X, Y)$
- $\text{Cov}(X, X) = \text{Var}[X] \geq 0$
- $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- Ако X и Y са независими, то $\text{Cov}(X, Y) = 0$

Емпирична функция на разпределение на вероятностна величина

- **Дефиниция:** Нека X_1, X_2, \dots, X_n са независими и идентично разпределени с X случайни величини. Нека сме наблюдавали (измерили) съответни стойности x_1, x_2, \dots, x_n за последователността от случайните величини X_1, X_2, \dots, X_n . Емпиричното разпределение на случайните величини наричаме функцията на разпределение $\Pr_n[X = x] : x \mapsto \frac{1}{n} \sum_{i=1}^n \delta_{X_i=x}$, където: $\delta_{X_i=x} = \begin{cases} 1 & \text{ако } X_i = x \\ 0 & \text{в противен случай} \end{cases}$.
- Емпирично очакване: $\mathbb{E}_n[X] = \sum_{x \in X(\Omega)} \Pr_n[X = x] x = \frac{1}{n} \sum_{x \in X(\Omega)} \sum_{i=1}^n \delta_{X_i=x} x = \frac{1}{n} \sum_{i=1}^n x_i$
- $\mathbb{E}_n[f(X)] = \sum_{x \in X(\Omega)} \Pr_n[X = x] f(x) = \frac{1}{n} \sum_{x \in X(\Omega)} \sum_{i=1}^n \delta_{X_i=x} f(x) = \frac{1}{n} \sum_{i=1}^n f(x_i)$
- **Закони за големите числа:** (Няма да доказваме)
 - $\lim_{n \rightarrow \infty} \Pr_n[X = x] = \Pr[X = x]$ (Закон за големите числа на Борел);
 - $\Pr[\lim_{n \rightarrow \infty} \mathbb{E}_n[X] = \mathbb{E}[X]] = 1$ (Закон за големите числа на Колмогоров).

Емпирични оценки

Емпирична вариация, ковариация и кросентропия на случайната величина:

- $$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] \approx \mathbb{E}_n[(X - \mathbb{E}_n[X])^2] = \frac{1}{n} \sum_{i=1}^n (x_i - \frac{1}{n} \sum_{j=1}^n x_j)^2$$

- $$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \approx \\ &\approx \mathbb{E}_n[(X - \mathbb{E}_n[X])(Y - \mathbb{E}_n[Y])] = \frac{1}{n} \sum_{i=1}^n (x_i - \frac{1}{n} \sum_{j=1}^n x_j)(y_i - \frac{1}{n} \sum_{j=1}^n y_j) \end{aligned}$$

План на лекцията

1. Формалности за курса (5 мин)
2. Вероятностно очакване, вариация, ковариация. Емпирично разпределение (20 мин)
- 3. Ентропия, перплексия и оценяване на езиков модел (20 мин)**
4. Семантично разширяване на заявката (10 мин)
5. Влагане на думи във многомерно разрежено векторно пространство (10 мин)
6. Влагане на термовете в контекстно пространство (15 мин)
7. Ранкиране на документи в контекстно пространство (5 мин)

Пример: задача за компресиране

- Нека имаме 8 състезателни коня — A, B, C, D, E, F, G, H. Вероятността за печалба на даден кон е:

	A	B	C	D	E	F	G	H
Pr	1/2	1/4	1/8	1/16	1/64	1/64	1/64	1/64

- Нека са проведени n състезания между конете и сме записали резултатите от надбягванията. Колко най-малко памет ни е нужна?
- Наивен подход — за обозначаването на даден кон от 8 възможни са ни нужни 3 бита. Следователно ще ни трябват $3n$ бита.
- Можем ли да подобрим представянето на резултатите, като се възползваме, че кон A ще се среща много по-често от кон F. Можем ли да представим A с по-малък брой битове за сметка на другите коне?

Решение: Код на Хъфман

	A	B	C	D	E	F	G	H
Pr	1/2	1/4	1/8	1/16	1/64	1/64	1/64	1/64
Код	0	10	110	1110	111100	111101	111110	111111

- Код на Хъфман $h : \Sigma \rightarrow \{0,1\}^*$
 1. Префиксен код — никой код не е префикс на друг и следователно всяка последователност от кодове позволява еднозначно декодиране.
 2. За всеки символ σ е изпълнено
$$|h(\sigma)| = \lceil -\log_2 \text{Pr}[\sigma] \rceil$$
- **Твърдение:** Горните свойства могат да бъдат удовлетворени за всяко крайно разпределение

Очакване за размера на представянето с кодиране на Хъфман

- Очакваме, че при всеки n надбягвания, конят σ ще спечели средно $n \Pr[\sigma]$ надбягвания.
- Тогава очакването за размера на представянето е:

$$\begin{aligned}\sum_{\sigma \in \Sigma} n \Pr[\sigma] |h(\sigma)| &= -n \sum_{\sigma \in \Sigma} \Pr[\sigma] \log_2 \Pr[\sigma] = \\ &= -n \left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{4} \log_2 \frac{1}{4} + \frac{1}{8} \log_2 \frac{1}{8} + \frac{1}{16} \log_2 \frac{1}{16} + \frac{4}{64} \log_2 \frac{1}{64} \right) = \\ &= 2n\end{aligned}$$

- Доказва се, че при условие че наблюденията са независими и еднакво разпределени, кодирането на Хъфман е оптимално (показва се в курса ОСОЕ).
- Ако разпределението беше равномерно, то получаваме представяне с $3n$ бита, което изисква най-много памет — не може да се компресира.

Ентропия

- **Ентропията** на дадена случайна величина X наричаме:

$$H_X = - \sum_{x \in X(\Omega)} \text{Pr}[x] \log_2 \text{Pr}[x] = - \mathbb{E}[\log_2(\text{Pr}[X])]$$

- Ентропията е мярка за очакваното (средното) количество информация (брой битове) за представяне на резултат на случаен опит. Тя е мярка за неопределеността, хаоса или изненадата на дадена случайна величина.
- Очакваната памет (в брой битове) необходима за предаването на n резултата от случаен опит на случайна величина X е nH_X .
- Ентропията възниква естествено в различни области като теория на вероятностите, теория на информацията, термодинамиката, и други.

Релативна ентропия и крос-ентропия

- **Крос-ентропия** на двете функции на разпределение \Pr и $\hat{\Pr}$ на случайната величина X дефинираме като: $H_X(\Pr \parallel \hat{\Pr}) = - \sum_x \Pr[x] \log_2 \hat{\Pr}[x] = - \mathbb{E}[\log_2 \hat{\Pr}[x]]$
- Крос-ентропията измерва очаквания брой битове, необходими за предаването на резултат от случаен опит, ако вместо действителната функция на разпределение на случайната величина \Pr , за кодиране се използва функцията на разпределение $\hat{\Pr}$.
- **Релативната ентропия** (разстояние на Кулбек-Лайблер) на двете функции на разпределение \Pr и $\hat{\Pr}$ на случайната величина X дефинираме като:
$$D(\Pr \parallel \hat{\Pr}) = H_X(\Pr \parallel \hat{\Pr}) - H_X = \sum_x \Pr[x] \log_2 \frac{\Pr[x]}{\hat{\Pr}[x]} = \mathbb{E} \left[\log_2 \frac{\Pr[x]}{\hat{\Pr}[x]} \right]$$
- Релативната ентропия измерва доколко функцията на разпределени \Pr се различава от $\hat{\Pr}$.
- **Теорема:** $D(\Pr \parallel \hat{\Pr}) \geq 0$, като равенство се достига т.с.т.к. $\Pr[x] = \hat{\Pr}[x]$ за всяко $x \in X(\Omega)$. (Доказателство в курса ОСОЕ)

Дефиниция на мярката за взаимна информация

- Нека X и Y са две случайни величини над вероятностно пространство Ω .

Тогава мярката за взаимна информация $I(X; Y)$ на X и Y дефинираме:

$$I(X; Y) = D(\Pr[x, y] \parallel \Pr[x] \Pr[y]) = \sum_{x \in X(\Omega), y \in Y(\Omega)} \Pr[x, y] \log_2 \frac{\Pr[x, y]}{\Pr[x] \Pr[y]}$$

За удобство ще предполагаме, че $0 \log 0 = 0$ и $0 \log \frac{0}{0} = 0$.

- Когато случайните величини X и Y са независими, тяхното съвместно разпределение $\Pr[x, y]$ е равно на произведението на $\Pr[x]$ и $\Pr[y]$.
Следователно взаимната информация е мярка за близостта на съвместното разпределение $\Pr[x, y]$ до неговата стойност, когато X и Y са независими, като близостта се измерва чрез релативната ентропията.
- По този начин $I(X; Y)$ може да се разглежда като мярка за количеството информация, която всяка една от величините може да предостави за другата.

Оценка на езиков модел

- Нека е даден езиков модел M зададен с точна фамилия от контекстни локални разпределения $\{\hat{\text{Pr}}[x | x_1x_2\dots x_n]\}_{x_1x_2\dots x_n \in V^*}$
- **Въпрос:** Как да оценим езиковия модел?
Идея: Да измерим крос-ентропията спрямо истинското разпределение.
- Тъй като действителното разпределение на езика Pr не ни е известно ние ще приближим действителната крос-ентропия с емпиричната крос-ентропия като използваме достатъчно голям корпус от текстове $\{\mathbf{x}^{(i)}\}_{i=1}^n$, където $\mathbf{x}^{(i)} \in V^*$ (често обемът на корпуса е в порядък от милиони или милиарди думи). Корпусът за оценяване не трябва да е използван за обучението на модела.

- Нека ни е даден езиков модел \hat{Pr} и корпус от последователности $\{\mathbf{x}^{(i)}\}_{i=1}^n$. Ще предполагаме, че $\mathbf{x}^{(i)} = x_1^{(i)} x_2^{(i)} \dots x_{k_i}^{(i)}$ и $x_{k_i+1}^{(i)} = \$$. Тогава емпиричната крос-ентропия е:

$$H_X(Pr_n | \hat{Pr}) = -\mathbb{E}_n[\log_2 \hat{Pr}[\mathbf{x}^{(i)}]] = -\frac{1}{n} \sum_{i=1}^n \log_2 \hat{Pr}[\mathbf{x}^{(i)}] =$$

$$= -\frac{1}{n} \sum_{i=1}^n \log_2 \prod_{j=1}^{k_i+1} \hat{Pr}[x_j^{(i)} | x_1^{(i)} x_2^{(i)} \dots x_{j-1}^{(i)}] =$$

$$= -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{k_i+1} \log_2 \hat{Pr}[x_j^{(i)} | x_1^{(i)} x_2^{(i)} \dots x_{j-1}^{(i)}]$$

Това ни дава очаквания брой битове за представяне на последователност от корпуса при използване на разпределението дадено от езиковия модел \hat{Pr} .

- На практика е по-удобно използването на очакваният брой битове за представяне на символ от

корпус: $-\frac{1}{m} \sum_{i=1}^n \sum_{j=1}^{k_i+1} \log_2 \hat{Pr}[x_j^{(i)} | x_1^{(i)} x_2^{(i)} \dots x_{j-1}^{(i)}]$, където $m = \sum_{i=1}^n k_i + 1$ (cross-entropy rate).

- За по-задълбочено изучаване на теория на информацията:
 - Курсът “Основи на статистическата обработка на естествен език. Теория на информацията”.
 - Elements of Information Theory, Thomas M. Cover, Joy A. Thomas, John Wiley & Sons, 2012

Минимизиране на емпиричната крос-ентропия

- Емпиричната скорост на крос-ентропията е най-често използваната целева функция, която се минимизира при дълбокото машинно обучение.
- **Перплексията** на езиковия модел M дефинираме като $2^{H_x(\mathbf{Pr} || \hat{\mathbf{Pr}})}$, където $H_x(\mathbf{Pr} || \hat{\mathbf{Pr}})$ е крос-ентропията между действителното разпределение на езика \mathbf{Pr} и разпределението дадено от езиковия модел $\hat{\mathbf{Pr}}$.
- Добрият езиков модел следва да даде високи вероятности на наблюденията в корпуса, което води до по-ниска крос-ентропия и съответно перплексия.
- **ЗАДАЧА:** Има ли връзка между принципите за минимизиране на крос-ентропията и максимизиране на правдоподобие?

План на лекцията

1. Формалности за курса (5 мин)
2. Вероятностно очакване, вариация, ковариация. Емпирично разпределение (20 мин)
3. Ентропия, перплексия и оценяване на езиков модел (20 мин)
4. **Семантично разширяване на заявката (10 мин)**
5. Влагане на думи във многомерно разрежено векторно пространство (10 мин)
6. Влагане на термовете в контекстно пространство (15 мин)
7. Ранкиране на документи в контекстно пространство (5 мин)

Недостатъци при търсене базирано на съвпадение на ключови думи

- На упражнението видяхме пример за нерелевантно ранкиране при използване на $tf \cdot idf$ тегла (*“Румъния вирус”*).
- Задачата е да се удовлетвори информационната потребност, а не да се броят съвпадения на срещания на ключови думи между заявката и документа.

Пример:

Заявка: *“добра застраховка за кола”*

Релевантен документ: *“идеалното автомобилно каско”*,

Нерелевантен документ: *“добра застраховка живот покрива инциденти с кола”*

- Търсене базирано само на съвпадение на ключови думи в много случай връща нерелевантни и изпуска релевантни резултати.
- Следващата цел е да се реализира търсене по смисъл — т.е. по семантична близост.

Семантично разширяване на заявката

- Използване на семантичен речник.
 - Експертно съставен семантичен речник
Пример: WordNet съдържа синонимни, антонимни, меронимни и хипонимни и други семантични връзки между думите.
 - Автоматично съставен семантичен речник — на базата на съвместно срещане на думите в документите:
Пример: *Отидохме да берем ябълки и круши.* ===> термовете *ябълки* и *круши* са близки.
- Ключовите думи от заявката се разширяват със семантично свързани термове — също като при толерантното търсене.

Проблеми при семантично разширяване на заявката

- Експертно съставените речници са непълни, трудно се поддържат и не обхващат новите термини.
- Автоматично съставените семантични речници съдържат много шум (нерелевантност и неточност).
- Много от релациите са валидни само в определени контексти, което води до нерелевантно разширяване; (*маса \approx тегло, маса от дърво \neq тегло от дърво*).
- Ще разгледаме алтернативно решение чрез влагане на думите в “семантично” векторно пространство.

План на лекцията

1. Формалности за курса (5 мин)
2. Вероятностно очакване, вариация, ковариация. Емпирично разпределение (20 мин)
3. Ентропия, перплексия и оценяване на езиков модел (20 мин)
4. Семантично разширяване на заявката (10 мин)
5. **Влагане на думи във многомерно разрежено векторно пространство (10 мин)**
6. Влагане на терموвете в контекстно пространство (15 мин)
7. Ранкиране на документи в контекстно пространство (5 мин)

Документно представяне във многомерно векторно пространство с “one hot” вектори

- В мултиномния документен модел на всеки документ съпоставяме $|V|$ -мерен вектор d , в който на позиция t записваме броя на срещанията на съответния терм.
- Дефинираме за всеки терм с индекс t съответен “**one hot**” $|V|$ -мерен вектор, $\chi_t \in \{0,1\}^{|V|}$, който се състои от $|V| - 1$ нули и една единица на позиция k . Т.е.

$$(\chi_t)_i = \begin{cases} 1 & \text{ако } i = t \\ 0 & \text{в противен случай} \end{cases}$$

$$\chi_t = (\quad 0, \quad 0, \quad \dots, \quad 0, \quad \underset{\substack{\uparrow \\ t}}{1}, \quad 0, \quad \dots, \quad 0 \quad)^T$$

- В такъв случай получаваме: $d = \sum_{j=1}^{L_d} \chi_{t_j}$

- Други документни представяния също могат да се разглеждат като получени от влагане на думи в многомерно векторно пространство. Например при някои варианти на **tf • idf** теглата.
- Във тези случаи векторите, които съпоставяме на термовете съдържат ненулева стойност единствено на позицията, която съответства на терма.
- Документното представяне получаваме като сумираме (или акумулираме по друг начин) векторите, съответстващи на термовете от документа.
- Векторите, съответстващи на различни думи са ортогонални.
- Векторите, съответстващи на документите са силно разреждени — размерността им е $|V|$, често над 100000, а броя на ненулевите стойности е по-малък от L_d — около 1000.
- **Проблем**: Няма никаква връзка между семантичната близост между термовете и техните векторни представяния. Следователно, векторното представяне на документа изцяло зависи от това, какви точно термове са използвани.

План на лекцията

1. Формалности за курса (5 мин)
2. Вероятностно очакване, вариация, ковариация. Емпирично разпределение (20 мин)
3. Ентропия, перплексия и оценяване на езиков модел (20 мин)
4. Семантично разширяване на заявката (10 мин)
5. Влагане на думи във многомерно разрежено векторно пространство (10 мин)
- 6. Влагане на термовете в контекстно пространство (15 мин)**
7. Ранкиране на документи в контекстно пространство (5 мин)

Алтернативен подход за векторно представяне на думи и документи

- **Дистрибутивна семантика**: Значението на дадена дума се определя от думите, които често се срещат около нея.
 - *“You shall know a word by the company it keeps” (Firth 1957)*
- В тълковните речници значението се определя с примери за използването на думата.
- Пример: ***Зад храста се показва малък космат пирентил с вирната опашка.***

- Контекстът на дадена дума са думите, които са около нея — в рамките на параграф, изречение или фиксиран по размер прозорец.
- Две думи ще считаме за семантично свързани, ако често се срещат в един и същ контекст.
- Чрез статистически анализ върху контекстите на срещанията на думите определяме тяхната семантична близост.

Пример: Матрица на срещанията на терм в контекст

K1: Иван кара кола. Иван купи кола.

K2: Мария купи колело. Мария кара колело.

K3: Иван обича кола. Мария обича колело.

K4: Иван обича Мария.

Брой срещания на терма в
съответния контекст

	K1	K2	K3	K4
Иван	2	0	1	1
Мария	0	2	1	1
кара	1	1	0	0
купи	1	1	0	0
обича	0	0	2	1
кола	2	0	1	0
колело	0	2	1	0

Пример: Матрица на съвместните срещания

Иван кара кола . Иван купи кола .
Мария купи колело . Мария кара колело .
Иван обича кола . Мария обича колело .
Иван обича Мария .

Брой срещания на терма
в прозорец около
съответната дума

		К О Н Т Е К С Т						
Д У М А		Иван	Мария	кара	купи	обича	кола	колел
	Иван	0	0	1	1	2	0	0
	Мария	0	0	1	1	2	0	0
	кара	1	1	0	0	0	1	1
	купи	1	1	0	0	0	1	1
	обича	2	2	0	0	0	1	1
	кола	0	0	1	1	1	0	0
	колело	0	0	1	1	1	0	0

Терм / контекст матрица

- Нека V е наредено множество от термове и C е наредено множество от контексти. Нека функцията $f: V \times C \rightarrow \mathbb{R}$ е мярка за свързването на даден терм с даден контекст. Тогава дефинираме матрицата терм / контекст $M^f \in \mathbb{R}^{|V| \times |C|}$ като $M_{i,j}^f = f(t_i, c_j)$, където $t_i \in V$ е i -тия терм в V и $c_j \in C$ е j -тия контекст в C .

- На терм t_i съпоставяме съответния вектор ред на матрица: $t_i \mapsto M_{i,\bullet}^f$.

- Близост или подобие между термовете t_i, t_k дефинираме:

$$\text{sim}_{\cos}(t_i, t_k) = \cos(M_{i,\bullet}^f, M_{k,\bullet}^f) = \frac{M_{i,\bullet}^f \cdot M_{k,\bullet}^f}{|M_{i,\bullet}^f| |M_{k,\bullet}^f|}$$

- Възможни са и други мярки за подобие но косинусовата близост е най-често и най-успешно използваната.

Мярка за свързването на терм с контекст

- Най-простата мярка е броя на срещанията:

$f(t_i, c_j) = \#(t_i, c_j)$, където с $\#(t_i, c_j)$ означаваме броя на срещанията на терма t_i в контекста c_j .

- Често се използва честотата на срещанията — броя нормализиран към сумата от всички срещания:

$$f(t_i, c_j) = \frac{\#(t_i, c_j)}{|D|}, \text{ където с } |D| = \sum_{t \in V, c \in C} \#(t, c).$$

В такъв случай имаме, че $f(t_i, c_j) = \text{Pr}_{MLE}[t_i, c_j]$.

- Недостатък на броя на срещанията е, че се получават много високи стойности за често срещани термове като предлози, определителни думи и други.

- Най-добри резултати се получават с използване на поточкова мярка за взаимна информация:

$$\text{PMI}(t; c) = \log \frac{\text{Pr}[t, c]}{\text{Pr}[t] \text{Pr}[c]} = \log \frac{\#(t, c) |D|}{\#(t, \bullet) \#(\bullet, c)}.$$

(Предполагаме, че ако $\#(t, c) = 0$ то $\text{PMI}(t; c) = 0$.)

- Недостатък на поточкова мярка за взаимна информация е, че ако двете явления се срещнат само веднъж и то заедно, то мярката ще е много висока. Затова често се прилага праг, за да се избегнат редките явления.

План на лекцията

1. Формалности за курса (5 мин)
2. Вероятностно очакване, вариация, ковариация. Емпирично разпределение (20 мин)
3. Ентропия, перплексия и оценяване на езиков модел (20 мин)
4. Семантично разширяване на заявката (10 мин)
5. Влагане на думи във многомерно разрежено векторно пространство (10 мин)
6. Влагане на термовете в контекстно пространство (15 мин)
- 7. Ранкиране на документи в контекстно пространство (5 мин)**

Влагане на документите в контекстно пространство

- Представянето на документите може да получим например като просто сумираме контекстните вектори съответстващи термовете, които се срещат в документа (BOW).
- Ранкирането може да се извърши според косинусовата близост спрямо вектора, получен за заявката.
- Съществуват значително по-релевантни методи за получаване на контекстния вектор съответстващ на документа — ще разглеждаме по-нататък в курса.
- Каква е гъстотата на векторите при това представяне?

Проблеми при влагането в контекстното пространство

- Размерността на контекстното пространство е броя на контекстите — може да бъде огромно.
- Даден терм може да се среща в стотици хиляди контексти. Документните вектори могат да съдържат милиони ненулеви елементи.
- Ранкирането в такова пространство е абсолютно невъзможно на практика.
- Следващата лекция ще разгледаме решение на този проблем