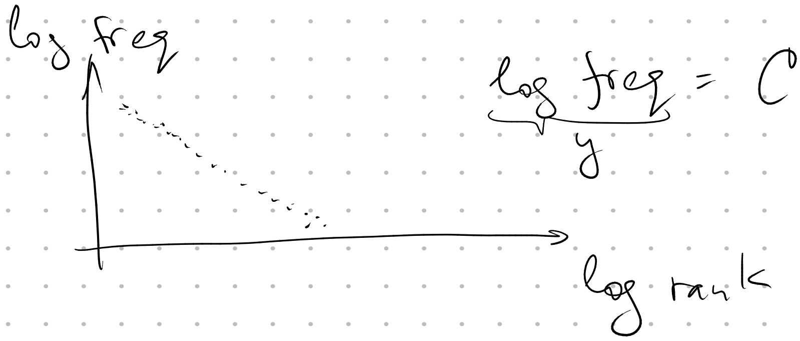
в Zipf



$$\underbrace{\log freq}_{y} = C - \underbrace{\log rank}_{x}$$

1) попробуйте "нормализовать" слова.

кот
кота
котом
...
→ кот

1 слово

стали ?? → сталь (сущ)
→ стать (гл)

норм. форма
→ лемма (сущ → ИП, ед.ч
гл → инфин.
...)
→ стем (отрезать окончание)

playing → play
player → play
dancing → danc (e)
⋮

Англ. Porter Stemmer
- алг. правил отрезания окончаний.

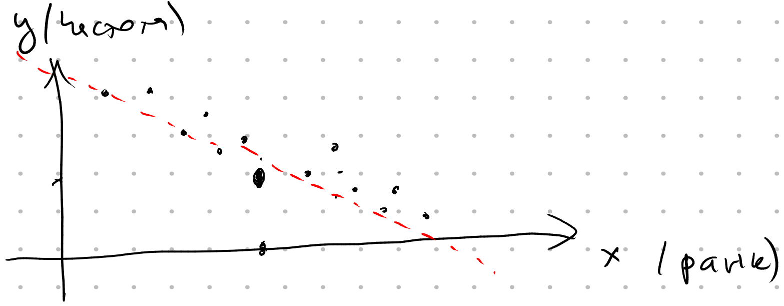
Доп. задание I
Нормализуем через морф. анализатор

- берём первый в-т анализа.

pymorphy 2
- для слова выдаёт варианты нач. форм, если не знает – угадывает.

Доп. задание II

Попытаемся подобрать прямую, проходящую ближе всего к нашим точкам.



Задача Давайте пытаемся предсказывать частоту по рангу

Есть примеры (обучающие)

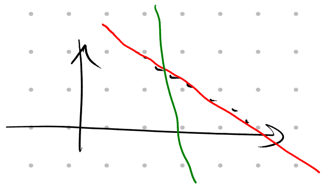
	ранк	частота	
много	100	10	
	20	50	
	120	?	← предсказать

Задача предсказания непрерывной величины – задача $\boxed{\text{регрессии}}$

(см. классификация)

Линейная регрессия предполагает.

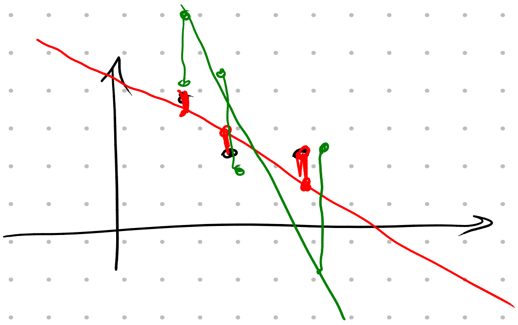
$$\underbrace{Y}_{\text{надо предсказать}} = b + k\underbrace{X}_{\text{дано}} + \boxed{Err} \sim N(0,\sigma)$$



Научиться предсказывать = подобрать b и k.

при подборе делаем так, чтоб $\sum_{i=1}^{N} err_i^2 \to \min$

сумма по примерам.



ош. зел >> ош. красн

Как подобрать прямую?

I. numpy

Данные с числами лучше хранить в numpy-матрицах.

numpy – создание и работа с матрицами из чисел. значительно эффективнее и по памяти, и по времени, по сравнению со списками и циклами.

```
import numpy as np

x = np.array([10, 20, 30])
y = np.array([[10,20],[30,40]])
```

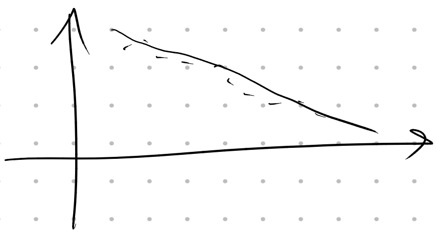
10	20	30

10	20
30	40

обучаем модель

$$\underset{X}{\begin{bmatrix} 0 & 7 \\ 1 & 2 \\ 1.5 & 4 \\ 2 & 3 \end{bmatrix}} \quad \underset{y}{\begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix}}$$

в Zipf



$$\log freq = C - \underbrace{\log rank}$$

$$\parallel$$

проверяем, что получилось ≈ −1

$$\log freq = b + \boxed{k} \cdot \log rank$$

$$k = -1$$

$$freq = \frac{C}{rank^{\text{аен}+1}}$$

N-граммы модели.

Строим модель языка.

∃ язык – это множество предложений.

Например, в русском "я иду спать" ∈ Русский

"я иду насос" ∉ Русский

"мы стали более лучшее одеваться" ?∈ Русский

Лучше иначе:

$$P(\text{предложение}) \longrightarrow \overset{\text{число от 0 до 1}}{[0, 1]}$$

вероятность встретить предлож.

P("я иду спать") >>> P("я иду насос")

Для вероятностой модели важно $\sum_{предл} P(предл) = 1$

Это вероятностная модель языка

Модель языка (P: предл → числа)

зачем

— при генерации текста, выбор лучшего варианта.

I sit by the table → Я сижу у стола (1)
Я сижу у таблицы (2)

$P(1) > P(2)$

выберем 1ый вариант.

— поиск ошибок. Я гляжу кита $^{(1)}$ — было
Я гляжу кота $^{(2)}$ — м.б.

$P(1) < P(2)$ ⇒ м.б. вы имели в виду "кота".

У нас будет N-грамм модель.

$P(w_1\ w_2\ w_3\ w_4\ ..\ w_n) =$ гипотеза: след слово зависит от N-1 предыдущих.

слово

N=2

$$= P(\overset{\text{начало}}{<s>}\ w_1\ w_2\ w_3\ ...\ \overset{\text{конец}}{w_4\ </s>}) :=$$

$$= \underbrace{P(w_1 | <s>)}_{\text{в-ть слова } w_1 \text{ после } <s>} \cdot P(w_2 | w_1) \cdot P(w_3 / w_2) \cdot ... \cdot (P(</s> | w_4).$$

вероятность слова w_2 после w_1

$$P(A|B) := \frac{P(AB)}{P(B)}$$

усл. в-ть

кубик

$$1\,2\,\boxed{3\,4\,5\,6}$$

$$\tfrac{1}{6}\ \tfrac{1}{6}\ \tfrac{1}{6}\ \tfrac{1}{6}\ \tfrac{1}{6}\ \tfrac{1}{6}$$

$$P(\text{чёт} \mid \geq 3) = \frac{P(\text{чёт} \geq 3)}{P(\geq 3)} = \frac{2/6}{4/6} = 1/2$$

$$P(\text{нечёт} \mid \geq 4) = \frac{P(\text{нечет} \geq 4)}{P(\geq 4)} = \frac{1/6}{3/6} = 1/3 \qquad 123\boxed{456}$$

$$P(\text{нечет}) = 3/6 = \frac{1}{2}$$

$N = 3$

$$P(\langle s\rangle\langle s\rangle\ w_1\ w_2\ w_3 \dots w_n\ \langle /s\rangle) =$$

$$= P(w_1 \mid \langle s\rangle\langle s\rangle) \cdot P(w_2 \mid \langle s\rangle w_1) \cdot P(w_3 \mid w_1 w_2) \cdot \dots$$

$$\cdot\ P(w_i \mid w_{i-2} w_{i-1}) \cdot \dots P(\langle /s\rangle \mid w_{n-1} w_n).$$

Как оценить $P(u \mid w)$? слово u после слова w?

Метод максимального правдоподобия (MLE)

ax ikelihood stimation

Дан корпус, где много текстов/предложений/слов.

можно посчитать $\underbrace{P(\text{корпус})}_{\text{правдоподобие}} = P(\text{предл } 1) \cdot P(\text{предл } 2) \cdot \dots \cdot P(\text{предл } n) =$

$$= P(w_1 \mid \langle s\rangle) \dots\dots \to \max.$$

Max достигается при

$$\boxed{P(u \mid w) = \frac{c(wu)}{c(w)}}$$

← сколько раз в корпусе было написано слово w, потом u

↑ сколько слов w в корпусе.

Модель построена.

$$\begin{tikzcd} & & \text{предложение} \arrow[d] \\ \text{Дан корпус} \arrow[r] & \text{считает } P(u|w) \arrow[r] & \boxed{\text{N-гр. модель } P(u|w)} \arrow[d] \\ & & p(\text{предложения}) \end{tikzcd}$$

$P(\text{я иду спать}) = P(\text{я}|\langle s\rangle) \cdot P(\text{иду}|\text{я}) \cdot P(\text{спал}|\text{иду}) \cdot P(\langle /s\rangle|\text{спать})$

— $P(u|w) = 0$, если в корпусе нет подряд wu.

$P(\ldots\ldots wu\ldots) = 0$. неудобно иметь 0 вероятност., они будут часто встречаться.

Сглаживание — $P(u|w)$ = чуть исправить значение, чтобы всегда $\neq 0$.

P должна остаться вероятностью.

$$\sum_{u \in V} P(u|w) = 1 \qquad V \text{ - весь словарь}$$

например

$$\underbrace{p(\text{я}|\langle s\rangle) + p(\text{он}|\langle s\rangle) + p(\text{они}|\langle s\rangle) + p(\text{стол}|\langle s\rangle) + \ldots + p(\text{ящер}|\langle s\rangle)}_{\text{все слова}} = 1$$

— как сравнить качество разных сглаживаний? в маш. обучении — подбор гиперпараметров

— оценить качество модели

перерыв до 10:45

Пример: Корпус $N=2$

<s> я вижу стол </s>
<s> я вижу стул </s>
<s> я вижу сыр </s>
<s> я ем сыр </s>
<s> я стол </s>

Результат обучения

$$\frac{c(\langle s\rangle я)}{c(\langle s\rangle)} = \frac{c(я\ вижу) = 3}{c(я) = 5}$$

$\Sigma = 1$ →

$$\frac{c(вижу\ стол)}{c(вижу)} = \frac{1}{3}$$

w \ u	<s>	я	вижу	ем	стол	стул	сыр	</s>
<s>	0	1	0	0	0	0	0	0
я	0	0	3/5	1/5	1/5	0	0	0
вижу	0	0	0	0	1/3	1/3	1/3	0
ем	0						1	
стол	0							1
стул	0							1
сыр	0							1
</s>	0	0	0	0	0	0	0	0

$$P(я\ вижу) = p(я \mid \langle s\rangle) \cdot P(вижу \mid я) \cdot P(\langle/s\rangle \mid вижу) =$$

$$= 1 \cdot 3/5 \cdot 0 = 0.$$

Избавление от $P(u|w) = 0$.

гиперпараметр.
1 0.1 0.01

$$P(u|w) = \frac{c(u) + \lambda}{c(wu) + \lambda|V|}$$

весь словарь

$\lambda = 1$

w \ u	<s>	я	вижу	ем	стол	стул	сыр	</s>
<s>	0	1	0	0	0	0	0	0
я	1/13	1/13	4/13	2/13	2/13	1/13	1/13	1/13
вижу	0	0	0	0	1/3	1/3	1/3	0
ем	0						1	
стол	0							1
стул	0							1
сыр	0							1
</s>	0	0	0	0	0	0	0	0

$\Sigma = 1$

$$\frac{0}{5} \rightarrow \frac{1}{5+1 \cdot 8}$$

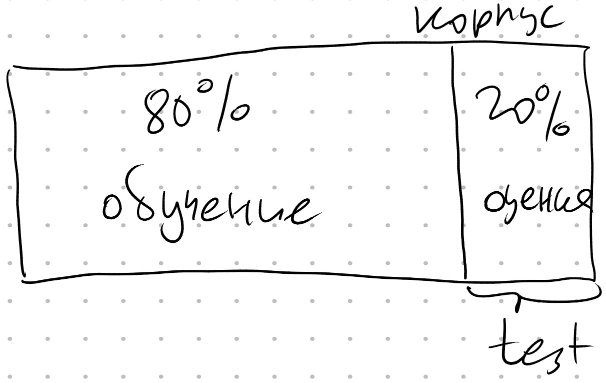
$$\frac{1}{5} \rightarrow \frac{2}{5+1 \cdot 8}$$

$$\frac{3}{5} \rightarrow \frac{4}{5+1 \cdot 8}$$

При сглаживании все неизвестные слова (не из корпуса) заменяются на <u>UNKN.</u>

<u>Качество модели.</u> Оцениваем на корпусе.

обычно делим исходный корпус на части



$$P(\text{тест. корпуса}) =$$
$$= P(\text{предл } 1) \cdot P(\text{предл } 2) \cdot P(\text{предл } 3) \cdot \ldots\ldots \to \max$$

Perplexity — примерно как P, но → min