Другие сглаживания: $u_1 u_2 \dots u_{n-1} \underset{?}{w}$

$$P(\underbrace{w \mid u_1 u_2 \dots u_{n-1}}_{n \text{ грамм}}) = \frac{c(u_1 \dots u_{n-1} w)}{c(u_1 \dots u_{n-1})}$$

max likelyhood

Корпус

3гр. иногда я сильно — 0

2гр я сильно — 10

1гр. сильно — 20

Интерполяция для сглаживания

$$p(w \mid u_1 u_2 \dots u_{n-1}) = \lambda_{n-1} \frac{c(u_1 \dots u_{n-1} w)}{c(u_1 \dots u_{n-1})} + \lambda_{n-2} \frac{c(u_2 \dots u_{n-1} w)}{c(u_2 \dots u_{n-1})} + \dots + \lambda_1 \frac{c(u_{n-1} w)}{c(u_{n-1})} + \lambda_0 \underbrace{c(w)}_{\text{длина корпуса}}$$

λ надо подбирать, $\sum \lambda_i = 1$

Для биграм $p(w \mid u) = \lambda_1 \frac{c(uw)}{c(u)} + \lambda_0 \frac{c(w)}{N}$

$\sum_{w \in V} p(w \mid u) = 1$

$1 = \sum \left(\lambda_1 \frac{c(uw)}{c(u)} + \lambda_2 \frac{c(w)}{N} \right) =$

$\lambda_1 \underbrace{\sum_w \frac{c(uw)}{c(u)}}_{=1} + \lambda_2 \sum_w \underbrace{\frac{c(w)}{N}}_{=1} = \lambda_1 + \lambda_2$

$\lambda_1 + \lambda_0 = 1$

0.5 0.5

0.8 0.2

0.2 0.8

подбор гиперпараметров.

Лучший вариант: Kneyser-Ney.

только если $c(uw) > 0$

$$P(w \mid u) = \frac{c(uw) - \delta}{c(u)} + \lambda \frac{|\{u : c(uw) > 0\}|}{|\{(u' u'') : c(u' u'') > 0\}|} \quad (1)$$

— всего разных (2)

	$w^{(1)}$	$w^{(2)}$	$w^{(i)}$	$w^{(N)}$
u	0	0	~~17~~ 16.5 / 200	~~18~~ 17.5 / 200

← все слова корпуса

$0.5/200 + 0.5/200$

Больше вероятности отдаётся тем словам, которые часто встречаются в новых контекстах.
(у которых разнообразнее контекст).

Пример: New York

1 вариант контекста слова York

big
small
gray
...
cat

100 вариантов контекста

+ мало в-ти

$P(York|...)$

+ много в-ти

$P(cat|...)$

(1) – кол-во разных контекстов, которые встречались перед w

(2) – кол-во разных пар слов, которые вообще встречались в корпусе.

λ – подбирается, чтоб $\sum_w p(w|u) = 1$.

Схема подбора гипернараметров в машинном обучении

Хотим выбрать лучшую модель: ~~1. MLE $\left(p = \frac{c(uw)}{c(w)}\right)$~~ – 0.

гипернараметры: метод / λ, λ_i, δ

2. Лаплас $\frac{c(uw)+\lambda}{c(w)+\lambda V}$

λ – разная

3. интерполяция, подобрать λ_i

4. Kneyser-Ney $\delta = ?$

корпус

80% обучение	10%	10%
training set	validation set	test set

Как делить:
all_sentences = [, ,]
shuffle – ф-ия из модуля случайных чисел.

после shuffle → all_sentences[:0.8*len]
→ all_sentences[0.8*len : 0.9*len]
→ all_sentences[0.9*len :]

Процесс подбора.
Берем 1ое значение гипер параметров (Лаплас, $\lambda=1$)
обучаем на train_set, оцениваем качество на validation_set
Берем 2ое значение ... (Лаплас, $\lambda=2$)
обучаем на train_set оцениваем качество на validation_set
Берем 3ье значение (Кнезер_Ней, $\delta=0.1$)
обучаем на train_set оцениваем качество на validation_set
и т.д.

Побеждает тот, у кого качество на validation_set максимально.

Ответ: победившая модель + качество на test_set.

Оцениваем качество модели на корпусе |test_set|

$P\{\text{текста}\}$ ↑ больше — лучше

$P\{\text{текста}\} = P\{\text{предл 1}\} \cdot P\{\text{предл 2}\} \cdot \ldots \cdot P\{\text{предл N}\} =$

$= P(w_1^{(1)} | <s>) \cdot P(w_2^{(1)} | w_1^{(1)}) \cdot \ldots \cdot P(</s> | w_{n_1}^{(1)}) \cdot$
$P(w_1^{(2)} | <s>) \cdot P(w_2^{(2)} | w_1^{(2)}) \cdot \ldots P(</s> | w_{n_2}^{(2)}) \cdot$ ← underflow
$\ldots$

$\log P\{\text{текста}\} = \log P(|) + \log P(|) + \ldots + \log P(|)$

— нет underflow

Значение теста — вероятность меньше.

поэтому берут не P текста, а среднюю вероятность n-граммы.

$\sqrt[N]{P(\text{текст})}$ $\qquad \frac{1}{N}\left(\log P(\text{текста})\right)$

сколько N-грамм

$$E = -\frac{\log P(w_1^{(1)}|\langle s\rangle) + \log P(\,|\,) + \ldots + \log P(\langle /s\rangle \mid w_n^{(n)})}{N} > 0$$

N = всего n-грамм.

$\log P(\,.\,) < 0$

Perplexity $= 2^E$ $\quad$ 1/средняя вероятность одной n-граммы.

perplexity $= 20 \Rightarrow$ средняя вероятность $= 1/20$

perplexity $= 2000 \Rightarrow$ средняя вероятность $1/2000$ $\quad \uparrow$

perplexity должно быть как можно меньше.