

QA Session

(SNLP tutorial)

Vilém Zouhar

21st July 2021

Perplexity

- How can we define it?

Perplexity

- How can we define it?
- Suppose that we generate a random sequence of digits S in the decimal system, where each digit in S is drawn from a uniform distribution. What is the perplexity of S in the following cases:

Perplexity

- How can we define it?
- Suppose that we generate a random sequence of digits S in the decimal system, where each digit in S is drawn from a uniform distribution. What is the perplexity of S in the following cases:

Perplexity

- How can we define it?
- Suppose that we generate a random sequence of digits S in the decimal system, where each digit in S is drawn from a uniform distribution. What is the perplexity of S in the following cases:
 - ▶ S is a sequence of length 5

Perplexity

- How can we define it?
- Suppose that we generate a random sequence of digits S in the decimal system, where each digit in S is drawn from a uniform distribution. What is the perplexity of S in the following cases:
 - ▶ S is a sequence of length 5

Perplexity

- How can we define it?
- Suppose that we generate a random sequence of digits S in the decimal system, where each digit in S is drawn from a uniform distribution. What is the perplexity of S in the following cases:
 - ▶ S is a sequence of length 5
 - ▶ S is a sequence of length 25

Word Sense Disambiguation

- Formal definition of the problem

Word Sense Disambiguation

- Formal definition of the problem
- Naive Bayes for WSD

Word Sense Disambiguation

- Formal definition of the problem
- Naive Bayes for WSD
- Flip-Flop for clustering (???)

Conditional Random Fields

- First and second-order HMM

Conditional Random Fields

- First and second-order HMM
- Suppose you have a sequence of length M and tagset of size T , what would be the complexity of the normalization factor $Z(x)$ in this case?

Conditional Random Fields

- First and second-order HMM
- Suppose you have a sequence of length M and tagset of size T , what would be the complexity of the normalization factor $Z(x)$ in this case?
- Bayesian network, cliques

Naive Bayes

- Naive Bayes pseudocode

Naive Bayes

- Naive Bayes pseudocode
- Suppose you have a Naive Bayes classifier for topic categorization that is defined over a vocabulary of size V and a category label set of size C . How many parameters does this model have?

Naive Bayes

- Naive Bayes pseudocode
- Suppose you have a Naive Bayes classifier for topic categorization that is defined over a vocabulary of size V and a category label set of size C . How many parameters does this model have?
- Exercise 12 from 2020 exam

Naive Bayes - Code

```
for each class  $c \in C$                 # Calculate  $P(c)$  terms
     $N_{doc}$  = number of documents in D
     $N_c$  = number of documents from D in class  $c$ 
     $logprior[c] \leftarrow \log \frac{N_c}{N_{doc}}$ 
     $V \leftarrow$  vocabulary of D
     $bigdoc[c] \leftarrow$  append(d) for  $d \in D$  with class  $c$ 
    for each word  $w$  in  $V$                 # Calculate  $P(w|c)$  terms
         $count(w,c) \leftarrow$  # of occurrences of  $w$  in  $bigdoc[c]$ 
         $loglikelihood[w,c] \leftarrow \log \frac{count(w,c) + 1}{\sum_{w' \text{ in } V} (count(w',c) + 1)}$ 
return  $logprior, loglikelihood, V$ 
```

Naïve Bayes

- Algorithm: Continuous-valued Features
 - Numberless values for a feature
 - Conditional probability often modeled with the normal distribution

$$\hat{P}(X_j | C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2}\right)$$

μ_{ji} : mean (average) of feature values X_j of examples for which $C = c_i$

σ_{ji} : standard deviation of feature values X_j of examples for which $C = c_i$

- **Learning Phase:** for $\mathbf{X} = (X_1, \dots, X_n)$, $C = c_1, \dots, c_L$
 Output: $n \times L$ normal distributions and $P(C = c_i) \quad i = 1, \dots, L$

- **Test Phase:** Given an unknown instance $\mathbf{X}' = (a'_1, \dots, a'_n)$
 - Instead of looking-up tables, calculate conditional probabilities with all the normal distributions achieved in the learning phase
 - Apply the MAP rule to make a decision

Significance Testing

- How to use Chi-Square?

Compression

- Kraft's inequality and trees
- Optimal code length: $-\log_D p(w_i)$
- Encoding using a tree

Vector-Space Model

- Representation
- Retrieval (scoring vs bayes) + decision rule
- Classification

Jensen's Inequality

- For convex functions, opposite holds for concave ones

Jensen's Inequality

- For convex functions, opposite holds for concave ones
- Such as: x^2 , \log

Jensen's Inequality

- For convex functions, opposite holds for concave ones
- Such as: x^2 , \log
- $WA(f(x_i)) \geq f(WA(x_i))$

Resources

- https://miro.medium.com/max/1400/1*neaBooRXSloZAz6A2XfHJA.png
- <https://image.slidesharecdn.com/naive-bayes-150514165844-lva1-app6891/95/naive-bayes-15-638.jpg?cb=1431622795>