Exercises

1. Derive a maximum likelihood estimator for casting a die.
2. Tell the difference between parameter types and parameter instances.
3. Which of the following are n-grams from the sentence "Tim bought a book for $1."
   (a) Tim bought
   (b) bought a
   (c) a book for
   (d) for $1.
   (e) Tim a for
   (f) Tim
   (g) Tim bought book
   (h) book a
4. Given three sentences "all models are wrong", "a model is wrong" and "some models are useful", and a vocabulary V ={¡s¿, ¡/s¿, a, all, are, model, models, some, useful, wrong}.
   (a) Calculate the probabilities of all bigrams without smoothing.
   (b) Calculate the probabilities of all bigrams and the unseen bigram "a models" with add-one smoothing.
   (c) Calculate the probabilities of all bigrams and the unseen bigram "a models" with add-α smoothing. Try $\alpha = 0.05$ and $\alpha = 0.15$.
   (d) Calculate the probabilities of all bigrams and the unseen bigram "a models" with back-off smoothing. Try $\lambda = 0.95$ and $\lambda = 0.75$.
5. Good-Turing smoothing gives an estimated count of *unseen* ngrams according to existing count of ngrams in a corpus. Denoting the count of n-grams that occur in the corpus *freq* times as N(*freq*), the estimated frequency ESITMATED(*freq*) is calculated as $(freq + 1)\frac{N_{freq+1}}{N_{freq}}$. Give Good-Turing smoothed value of all the bigrams in Ex 4 and the unseen bigram "a models".
6. Kneser-Ney smoothing calculates the bigram probability $P_{KN}(w|w')$ by considering both the bigram counts $\#(w'w)$ in a corpus and *diversity of histories* of the word w. Formally,

$$P_{KN}(w|w') = \frac{\max(\#(w'w) - \delta, 0)}{\sum_{w'} \#(w'w)} + \lambda_w P_{KN}(w), \tag{1}$$

   where δ is a hyper-parameter for discounting $\#(w'w)$, $0 < \delta < 1$, λ_w is a parameter to ensure $\sum_{w''} P_{KN}(w''|w') = 1$ and $P_{KN}(w)$ is the unigram

probability that depends on the the diversity of histories of the word w. Formally, the diversity of histories of a word w is defined as,

$$\textsc{DiversityOfHistories}(w) = \#(\{w' : w'w \in D\})$$

$\textsc{DiversityOfHistories}(w)$ can be used to replace $\#w$ in counting unigrams.

$$P_{KN}(w) = \frac{\textsc{DiversityOfHistories}(w)}{\sum_{w''} \textsc{DiversityOfHistories}(w'')}$$

Similarly, for bigrams, $\textsc{DiversityOfHistories}(w_1w_2) = \#(\{w' : w'w_1w_2 \in D\})$. Give Kneser-Ney smoothed value of all the bigrams in Ex 4 and the unseen bigram "a models" when $\delta = 0.1$.

7. True or False

   (a) If B is conditionally independent of A, then A is also conditionally independent of B.

   (b) If $P(B|A) = P(B)$ then $P(B|AC) = P(B|C)$.

   (c) $P(B|A)P(A) = P(A|B)P(B)$.

   (d) $P(ABC) = P(A|BC)P(B|AC)P(C|AB)$.

   (e) $P(ABC) = P(A)P(B)P(C)$ under iid assumption.

8. What are the differences between hyper-parameters and parameters?

9. Hyper-parameters are tuned over _____.
   (A) Training data (B) Development data (C) Test data.

10. The term "Naïve" in Naïve Bayes classification refers to the iid assumption. Extend the Naïve Bayes classifier using the concept of Bigram language modelling. The new model loses the "Naïve" attribute. Can you integrate bag-of-word features into this model by leveraging smoothing techniques?

11. Which if the following feature sets contain overlapping features?

    (a) bag-of-words and bag-of-bigrams for document modelling.

    (b) full word, prefix and suffix for word modelling.

    (c) a class label and a bag of words for document modelling.

    (d) The first letter of a word, and a binary feature indicating whether the word is capitalised.

    (e) number of words in a document, and bag of words.

    (f) word-class pairs and bag of words for document classification.