FNLP Tutorial 2

Question 1. Evaluating data annotations

Imagine that this is your small corpus of named entities in a simple task, where we ignore a named entity's type and annotate the real named entities with square brackets:

> [Paris Hilton] stayed at the [Hilton] in [Paris] and
> [James Clerk Maxwell] was educated at [Edinburgh] and [Cambridge]

We formalise the annotation of a single sentence as a set A. Each element $a \in A$ represents a named entity as a span through an ordered pair of zero-based indices (a named entity $\langle a, b \rangle$ starts at position a and ends at b, including b). Assume a computational model predicts the following named entities: $\{\langle 0, 1 \rangle, \langle 5, 5 \rangle, \langle 9, 10 \rangle, \langle 15, 15 \rangle\}$

1. Compute the precision, recall, and F_1-score of this annotation.
2. Provide an annotation that would give a precision of more than 0.8 and a recall of less than 0.2, and use your answer to explain why the F_1-score uses the *harmonic* mean.
3. While the F_1-score is a better metric than precision and recall in isolation, there are other flaws all three metrics suffer from. What, specifically in the context of span identification, does it fail to capture about the model's predictions provided above?

Question 2. Text generation with a language model

We are given the following corpus, modified from J&M (Chapter 3 in the 3rd edition):

> <s> language is awesome </s>
> <s> language is awesome </s>
> <s> language and students </s>
> <s> awesome students like language </s>

1. Using a bigram language model without smoothing, generate 4 sentences, starting from '<s>'. To generate sentences manually, randomly choose a real number between 0 and 1, and use the cumulative probabilities of words in the vocabulary to select a word from the language model. For example, if there are two words in the vocabulary with non-zero conditional probabilities $P(\text{one}|\text{<s>}) = 0.25$ and $P(\text{two}|\text{<s>}) = 0.75$, then choosing a random number between 0 and 0.25 would result in 'one', whereas a random number between 0.25 and 1 would result in 'two'.
2. How does the lack of smoothing impact the novelty of the sentences generated?
3. How does the nature of the corpus impact the lengths of the generated sentences?

Question 3. Smoothing

1. Compute $P(\text{awesome}|\text{is})$ and $P(\text{like}|\text{is})$ for (a) the bigram language model without smoothing, (b) with add-1 smoothing, and (c) using interpolation.
2. Which values of λ_1 and λ_2 did you select, and why?
3. As discussed in J&M section 3.5, one would normally select the values for λ_1 and λ_2 based on a held-out corpus. What are the characteristic features of a corpus that would result in $\lambda_2 \ll \lambda_1$?
4. Mention a disadvantage of using interpolation to do smoothing, by referring to the bigrams 'is awesome' and 'is like'. Identify another type of smoothing that overcomes the shortcomings you described.

Question 4. Document classification

The following are features extracted from a set of short movie reviews, and the genre of the movie.

n	**Document**	**Class**
1	love, fast, romantic, couple	`romance`
2	romantic, fast, fun, fun	`comedy`
3	fast, violence, shoot, furious, fast	`action`
4	couple, fun, fast, fast, furious	`action`
5	fun, violence, fun, romantic	`comedy`
6	fast, fun	?

1. What is the MLE estimate of the prior probability $P(\texttt{action})$?
2. What is $P_{MLE}(\text{fast}|\texttt{action})$?
3. Apply add-α smoothing. Assuming that $\alpha = 0.7$, what is $P_{add-\alpha}(\text{fast}|\texttt{action})$?
4. Assuming d is the sixth document from the table above, compute the ratio $\frac{P(\texttt{comedy}|d)}{P(\texttt{action}|d)}$ with the same smoothing method applied. What does this ratio tell us about the document's classification?
5. Describe the type of classification model one would get for an extremely large value of α. Could changing α change the classification of our document number 6?

Bonus question

We said in the lecture that relative frequency estimation is a form of maximum likelihood estimation (MLE). Here we want you to *prove* that this is true for a categorical random variable.

Let X be a categorical random variable that can take values $1, \ldots, k$. Assume we have N independent samples $x_1, \ldots, x_N$ from X where each x_i takes one of the k values. We consider a family of distributions $\hat{P}_\theta(X = x) = \theta_x$ with parameters θ. The likelihood of the data as function of θ is:

$$L(\theta) = \prod_{i=1}^{N} \hat{P}_\theta(X = x_i) = \prod_{i=1}^{N} \theta_{x_i} \tag{1}$$

Show that $P_{MLE}(X = x) = \frac{C(x)}{N}$ where $C(x)$ is the count of x in our sample $x_1, \ldots, x_N$.

Hint you might want to use the following fact, known as Gibbs' inequality (see also Lecture 6b): If P and Q are distributions over the random variable X, then

$$-\sum_{i=1}^{k} P(X=i)\log P(X=i) \leq -\sum_{i=1}^{k} P(X=i)\log Q(X=i) \tag{2}$$

This holds with equality if and only if $P(X=i)=Q(X=i)$ for all i.