In this question, you will train classifiers to predict whether sentences are about CS 188 or CS 186 (databases) using a bag-of-words Naive Bayes classifier. Each sentence is labeled with the class to which it pertains.

### Training set
- (188) agents need good models.
- (188) agents need data.
- (186) buffers need memory.
- (186) DBs need data models.

### Held-out set
- (188) agents need memory.
- (186) DBs have data.

### Test set
- (186) data have data models.

#### a) Write down all of the maximum likelihood (relative frequency) parameters for a bag-of-words naive Bayes classifier trained on the training set above. Let $Y$ be the class for a sentence, and $W$ be a word. You may omit any parameters equal to 0. Ignore punctuation. Note: Bag-of-words classifiers assume that the words at every sentence position are identically distributed. Repeated words affect both the likelihood of a word during estimation and sentence probabilities during inference.

| $Y$ | $P(Y)$ | $W$ | $P(W | Y = 188)$ | $W$ | $P(W | Y = 186)$ |
|-----|--------|-----|-----------------|-----|-----------------|
| 188 | $\frac{2}{3}$ | agents | $\frac{1}{3}$ | need | $\frac{2}{3}$ |
| 188 | $\frac{2}{3}$ | good | $\frac{1}{3}$ | buffers | $\frac{1}{3}$ |
| 186 | $\frac{1}{3}$ | models | $\frac{1}{3}$ | memory | $\frac{1}{3}$ |
| 186 | $\frac{1}{3}$ | classifiers | $\frac{1}{3}$ | DBs | $\frac{1}{3}$ |
| 186 | $\frac{1}{3}$ | data | $\frac{1}{3}$ | models | $\frac{1}{3}$ |

#### b) According to your classifier, what is the probability that the first held-out sentence “agents need memory” is about 188?

Since $P(W = \text{memory} | Y = 188) = 0$, the joint probability

$$P(Y = 188, W_1 = \text{agents}, W_2 = \text{need}, W_3 = \text{memory}) = 0$$

Likewise, since $P(W = \text{agents} | Y = 186) = 0$, the joint probability

$$P(Y = 186, W_1 = \text{agents}, W_2 = \text{need}, W_3 = \text{memory}) = 0$$

Adding these together, we find that according to our model, $P(W_1 = \text{agents}, W_2 = \text{need}, W_3 = \text{memory}) = 0$. Therefore, the posterior probability $P(Y = 188 | W_1 = \text{agents}, W_2 = \text{need}, W_3 = \text{memory})$ is undefined: $0/0$. 

---

CS188: Artificial Intelligence, Fall 2009

Naive Bayes and Smoothing
c) Using Laplace (i.e., add one) smoothing for all of your parameters, what is the probability of seeing the test sentence “data have data models”: \( P(W_1 = \text{data}, W_2 = \text{have}, W_3 = \text{data}, W_4 = \text{models}) \)? Assume that the only words you ever expect to see are those in your training and held-out sets. Hint: sum over \( Y \), using the new estimates after smoothing.

\[ W \text{ can take 9 different values, so we use the following smoothed parameters to compute the quantity desired:} \]

\[
\begin{array}{|c|c|c|}
\hline
Y & P(Y) & W & P(W|Y = 188) \\
\hline
188 & \frac{2}{9} & \text{data} & \frac{1}{5} \\
186 & \frac{2}{9} & \text{have} & \frac{1}{5} \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|c|}
\hline
Y & P(Y) & W & P(W|Y = 186) \\
\hline
188 & \frac{2}{9} & \text{have} & \frac{1}{5} \\
186 & \frac{2}{9} & \text{models} & \frac{1}{5} \\
\hline
\end{array}
\]

\[ P(W_1 = \text{data}, W_2 = \text{have}, W_3 = \text{data}, W_4 = \text{models}) \]

\[ = P(Y = 188, \text{data, have, data, models}) + P(Y = 186, \text{data, have, data, models}) \]

\[ = P(188)P(\text{data}|188)^2P(\text{have}|188)P(\text{models}|188) + P(186)P(\text{data}|186)^2P(\text{have}|186)P(\text{models}|186) \]

\[ = \frac{1}{2^8} \frac{1}{2^8} \frac{1}{16} \frac{1}{16} + \frac{1}{2^8} \frac{1}{2^8} \frac{1}{16} \frac{1}{16} \]

\[ = \frac{1}{2^{14}} + \frac{1}{2^{14}} = \frac{1}{2^{13}} \]

\[ \text{d) Using Laplace smoothing, what is the probability according to your classifier that the test sentence} \]

\[ \text{“data have data models” is about 186?} \]

\[ \text{From the previous question, we have:} \]

\[ P(Y = 186, W_1 = \text{data}, W_2 = \text{have}, W_3 = \text{data}, W_4 = \text{models}) = \frac{1}{2^7} \]

\[ P(W_1 = \text{data}, W_2 = \text{have}, W_3 = \text{data}, W_4 = \text{models}) = \frac{1}{2^7} \]

Hence, the conditional probability

\[ P(Y = 186|W_1 = \text{data}, W_2 = \text{have}, W_3 = \text{data}, W_4 = \text{models}) = \frac{1}{2} \]

\[ \text{, which will only classify the example correctly if we happen to break ties correctly (not a method we want} \]

\[ \text{to rely on).} \]

\[ \text{e) Suggest an additional feature that would allow the classifier to correctly conclude that} \]

\[ \text{“data have data models” is about 186 when trained on this training set.} \]

\[ \text{The bigram feature “data models” would suffice, as it only appears in data labeled 186. Several other} \]

\[ \text{features are also acceptable, like a feature for the absence of the word “agents”.} \]

\[ \text{In general, any feature that favors 186 over 188 in the training data that is also relevant} \]

\[ \text{for the test datum we are trying to satisfy would work.} \]