Problem Set 1: Review of probability

Hanjun Dai

August 25, 2015

1 Zombie Bob (4 pts total)

1.1 Bayes rule (2 pts)

Answer:
$P(zombie|graagh) = P(graaph|zombie)P(zombie)/\left(P(graaph|zombie)P(zombie) + P(graaph|Bob)P(Bob)\right)$
So

$$P(\text{zombie}|\text{graagh}) = \frac{0.5 \times 10^{-6}}{0.5 * 10^{-6} + 10^{-5} \times (1 - 10^{-6})} = 4.7619\%$$

1.2 Expected utility (1 pt)

Answer:

$$\text{utility(stay)} = P(zombie) \times -20 + P(Bob) \times 0 = -0.9524$$
$$\text{utility(run)} = P(zombie) \times 3 + P(Bob) \times -1 = -0.8095$$

1.3 The chain rule and marginal probabilities (1 pt)

Answer:

$$P(survive|graagh) = P(zombie|graagh) \times 50\% + P(Bob|graagh) \times 100\% = 97.62\%$$

2 Necromantic Scrolls (4 pts total)

2.1 Bayes rule (1 pt)

Answer:

$$P(Anna|x = abracadabra) = P(x = abracadabra|Anna) \times P(Anna)/P(x = abracadabra)$$

Where

$$\begin{aligned} P(x = abracadabra) =& P(x = abracadabra|Anna)P(Anna) + P(x = abracadabra|Barry)(1 - P(Anna)) \\ =& 0.5\% \times 60\% + 1\% \times 40\% = 0.7\% \end{aligned}$$

So

$$P(Anna|x = abracadabra) = 0.5\% \times 60\%/0.7\% = 42.86\%$$

2.2 Breakeven point (1 pt)

Answer:
Let $P(Anna|x = abracadabra) = 50\%$, then we have

$$50\% = P(Y = Anna) \times 0.5\%/0.7\%$$

then we get $P(Y = Anna) = 70\%$

2.3 Multiple words (2 pts)

1. What is his posterior belief about the probability that Anna is author of the scroll? (1 pt)
   Answer: Since we have no prior, it is to say $P(Anna) = P(Barry)$, so

$$P(Anna|observation) \propto P(observation|Anna)P(Anna) \propto P(observation|Anna)$$

$$\begin{aligned} P(observation|Anna) =& P(\text{2 abracadabra, 1 gesundheit}; Param_{Anna}, N_w = 100) \\ =& P(\text{2 abracadabra}; Param_{Anna}, N_w = 100)P(\text{1 gesundheit}; Param_{Anna}, N_w = 98) \\ =& \binom{100}{2} \times 0.005^2 \times \binom{98}{1} \times 0.006 \\ =& 7.28\% \end{aligned}$$

   And also, $P(observation|Barry) = \binom{100}{2} \times 0.01^2 \times \binom{98}{1} \times 0.001 = 4.85\%$

   So $P(Anna|observation) = \frac{7.28\%}{7.28\%+4.85\%} = 60\%$

2. Does Dante need to consider the 97 words that were not abracadabra or gesundheit? Why or why not? (1 pt) Assume that he cannot obtain perauthor frequency statistics for any additional words that is, he cannot expand Table 1.
   Answer:

   No. Since we only observed the three words, we can only get the joint probability of these two, i.e., P(2 abracadabra, 1 gesundheit). Since the generating process of the 100-word corpus follows the multinomial distribution, we actually marginalize the combinations of other words.

3 Sentence lengths (5 pts total)

3.1 Maximum likelihood estimation (2 pts)

Answer:
For $n-$th sentence, the likelihood is

$$P(l_n|\lambda) = \lambda^{l_n}(1-\lambda)$$

The log-likelihood of the entire corpus is:

$$L = \sum_n \log P(l_n|\lambda) = \sum_n l_n \log\lambda + \log(1-\lambda) = N\log(1-\lambda) + \log\lambda \sum_n l_n$$

Take the derivative regarding λ and set it to zero:

$$\begin{aligned}\frac{\partial L}{\partial \lambda} =& -N\frac{1}{1-\lambda} + \frac{1}{\lambda}\sum_n l_n \\ =& 0\end{aligned}$$

Then we get $\lambda = \frac{\bar{l}}{1+\bar{l}}$, where $\bar{l} = \frac{1}{N}\sum_{i=1}^{N} l_i$ is the average length of sentences in the corpus.

3.2 Expectations (3 pts)

1. What is the expected sentence length, given a parameter λ?
   Answer:

$$\begin{aligned}E[l] =& \sum_{l=0}^{\infty} l\lambda^l(1-\lambda) \\ =& (1-\lambda)\sum_{l=0}^{\infty} l\lambda^l \\ =& (1-\lambda)\lambda\sum_{l=0}^{\infty} l\lambda^{l-1} \\ =& (1-\lambda)\lambda\left(\frac{d}{d\lambda}\sum_{l=0}^{\infty}\lambda^l\right) \\ =& (1-\lambda)\lambda\left(\frac{d}{d\lambda}\frac{1}{1-\lambda}\right) \\ =& (1-\lambda)\lambda\frac{1}{(1-\lambda)^2} \\ =& \frac{\lambda}{1-\lambda}\end{aligned}$$

2. What is the modal (most probable) sentence length, according to this model? (1 pt)
   Answer:

   Since $P(l|\lambda) = \lambda^l(1-\lambda)$, and $\lambda \in [0,1]$, so when $l = 0$, we get the maximum probability.

3. Extra credit:
   Answer:

   We can use the Poisson distribution, which parametrized by $\lambda > 0$, i.e.,

$$P(l|\lambda) = \frac{\lambda^l e^{-\lambda}}{l!}$$

   The mean value (expectation of l) is λ, while the mode is $\lceil\lambda\rceil - 1$, which makes the difference between mean and mode no more than 1.

4 Part-of-speech tagging accuracy (2 pts total)

1. Suppose a sentence contains n words, and that the chance of making an error on each word is independent and identically distributed (IID). What is the chance of tagging the entire sentence correctly? Give the answer for $n = 5$, rounding to two decimal places. (1 pt)
   Answer:

   $0.9^5 \simeq 59.05\%$

2. Felicias tagger makes errors that are IID. Gregory has also built a tagger that has a 10% per-word error rate, but his tagger makes all of its errors on verbs. Assume that Felicia and Gregory apply their taggers to the same corpus: whose tagger will get more sentences completely correct, and why? (1 pt; hint: you will have to apply a small amount of linguistic intuition to answer this question.)

   Answer:

   Since Gregory' tagger only makes errors on verbs, and the overall error is still 10%, suppose the fraction of verbs in English is $p \in (0, 1)$, then this tagger will have $\frac{10}{p}\%$ chance to make mistake on verbs.

   Typically there will only be one verb in a sentence, then the chance of Gregory to tag the sentence correctly is $1 - \frac{0.1}{p}$.

   Suppose the average sentence length is l, then the tagger of Felicia will have 0.9^l chance to correctly tag.

   So we are actually comparing $1 - \frac{0.1}{p}$ with 0.9^l. Suppose $l = 10$, and $p = 0.2$, then Gregory will have better chance to tag correctly.