STAT 391
Homework 1
Out April 5, 2022
Due April 12, 2022

©Marina Meilă
mmp@stat.washington.edu

[Problem 1 – Practice with sample spaces, events – NOT GRADED]

Professor Chance Gardner is in a hurry to get to the class she is teaching. She can drive to campus either on I5 (denote this event by I) or the shorter way over the University Bridge (denote this event by $\bar{I}$). The latter can be either up (denote this event by U) or down. After she gets to campus she may find a parking spot (denote this event by P) or not (in which case she will have to park in the Montlake area). Finally, denote by L the event "professor Gardner is late for class". Below are given the probabilities of all possible individual outcomes of this random experiment.

Outcome	**Probability**
IPL	0.05
$IP\bar{L}$	0.1
$I\bar{P}L$	0.15
$\bar{I}UPL$	0.18
$\bar{I}\bar{U}\bar{P}L$	0.07
$\bar{I}\bar{U}P\bar{L}$	0.35
$\bar{I}\bar{U}\bar{P}\bar{L}$	0.1

The outcomes not listed (like $I\bar{P}\bar{L}$="drives on I5, doesn't find a parking space and is not late for class") have probability 0. Some "individual outcomes" in the above list are really events (for example IPL which is the union of $IPLU$ and $IPL\bar{U}$). We still can call it an individual outcome assuming that if the route is over the I5 bridge then it doesn't matter what is the state of University bridge (or we may not find out what this state is).

a. Make a neatly labeled drawing of this sample space, showing all the possible outcomes and their probabilities.

b. What is the probability that Chance drives over the University Bridge ?

c. What is the probability that Chance makes it to class in time?

d. What is the probability that Chance doesn't find a parking spot and is late for class?

e. Which of the following events is more probable: A = "Chance drives over the University bridge and is not late for her class, or B = "Chance drives over the I5 bridge and is not late for her class" ?

[Problem 2 – Practice with repeated sampling – NOT GRADED]

The "Little Amazon" company sells books on the internet. "Little Amazon" has the following titles for sale: 0 – "War and Peace", 1 – "Harry Potter & the Deathly Hallows, 2 – "Winnie the Pooh", 3 – "Get rich NOW", 4 – "Probability". "Little Amazon" has collected data on the sales of each title over the last 3 months.

For all the following questions, please give the "literal" expression of the answer

a. Denote by θ_i the probability that a customer buys title i. Assume that each purchase of a book is independent of the other purchases by the same customer or by other customers. Estimate $\theta = (\theta_0 \dots \theta_4)$ from the data. What are the sufficient statistics?

b. A customer buys 3 books. What is the probability that he buys "War and Peace", "Harry Potter", "Probability" in this order?

c. A customer buys 4 books. What is the probability that she buys only non-fiction, that is, N=$\{3,4\}$?

d. A customer buys 2 "Probability" books and 3 fiction (i.e 0 or 1 or 2) books. What is the probability of this event?

e. A customer buys n books. What is the probability that he buys at least one "Probability" ?

Problem 3 – Geometric distribution

Let $S = \{0, 1, 2, \dots 9\}$ be the sample space of the 10 digits and $P(n)$ be an exponential distribution over this space, given by

$$P(n) \;=\; \frac{1}{Z}\gamma^n \tag{1}$$

with $\gamma = 1/2$.

a. What is the value of the normalization constant Z? Give either a numerical answer or a formula in terms of γ.

b. What is the probability that $n < 5$? Give either a numerical answer or a formula in terms of γ.

[c. – Not graded] What is the probability that n is odd? Give either a numerical answer or a formula in terms of γ.

d. Let $S_3 = S^3$ be the sample space of all sequences of 3 elements from S. The sequences in S_3 are obtained by sampling independently 3 times from S with the probability distribution P defined in (1). We denote the probability distribution over S_3 obtained this way also by P, with the understanding that $P(A)$ with $A \subseteq S$ refers to the original distribution over S, while $P(B)$ with $B \subseteq S_3$ refers to the distribution induced by P over S_3.

How many elements has S_3?

[e. – Not graded] Compute the probability of the sequences (1,2,3), (0,0,0) and (0,1,1).

f. What is the sequence $(n^{(1)}, n^{(2)}, n^{(3)})$ that has highest probability of occurrence? What is the sequence with the lowest probability?

g. For any outcome $(n^{(1)}, n^{(2)}, n^{(3)})$ define

$$s(n^{(1)}, n^{(2)}, n^{(3)}) = n^{(1)} + n^{(2)} + n^{(3)},$$

the sum of three digits drawn independently from P. For instance $s(2, 0, 6) = 8$.

Let $(n^{(1)}, n^{(2)}, n^{(3)})$ and $(\bar{n}^{(1)}, \bar{n}^{(2)}, \bar{n}^{(3)})$ be two sequences with $s(n^{(1)}, n^{(2)}, n^{(3)}) > s(\bar{n}^{(1)}, \bar{n}^{(2)}, \bar{n}^{(3)})$. Does this imply

$$P(n^{(1)}, n^{(2)}, n^{(3)}) > P(\bar{n}^{(1)}, \bar{n}^{(2)}, \bar{n}^{(3)})? \tag{2}$$

Prove or give a counterexample.

[h. Extra credit] Denote by S_s the outcome space of s, and by Q the probability distribution of s.

We would like to know if $Q[s]$ decreases with s. First, set $\gamma = \frac{1}{2}$. Is it true that $Q(s)$ is monotonically decreasing in s? Prove or give a counterexample.

Find the values of $\gamma \in (0, 1]$ for which the probability of $Q[s]$ decreases with s when $s \leq 10$.

Fact 1 The number of ways to write s as a sum of 3 non-negative integers is equal to $\binom{s+2}{2}$.

Fact 2 The number of ways to write s as a sum of 3 integers from S is equal to

$$\begin{cases} \binom{s+2}{2} & \text{for } s = 0, \dots 9 \\ \binom{s+2}{2} - 3\binom{s-8}{2} & \text{for } s = 10, \dots 19 \\ \binom{s+2}{2} - 3\binom{s-8}{2} + 3\binom{s-18}{2} & \text{for } s = 20, \dots 27 \end{cases} \tag{3}$$

[**Extra credit:** Prove Fact 1 or Fact 2]

FYI: The sum of the geometric progression

$$a + ax + ax^2 + ax^3 + \ldots + ax^{m-1} \;=\; a\frac{1-x^m}{1-x}$$

Problem 4 – A toy language classification program

This problem is a language classification experiment (demoed in class).

We assume that sentences in a language are generated by sampling letters independently from the alphabet { a, b, c, ... z }. Spaces and punctuation are ignored. For instance, the probability of the sentence ``who's on first?'' is

$$\theta_w\theta_h\theta_o^2\theta_s^2\theta_n\theta_f\theta_i\theta_r\theta_t$$

because the sentence contains (w, h, o, s, o, n, ...t) in this order. The parameters $\theta_{a:z}$ of this simple model depend on the language. The files english.dat, french.dat, german.dat, spanish.dat are ASCII files containing the probabilities of the letters a–z in each of the languages, multiplied by 1000[1] For example, below is the beginning of english.dat:

a 81.51
b 14.40
c 27.58
...

The data mean that for the English language $\theta_a = 0.08151$, $\theta_b = 0.0144$, $\theta_c = 0.02758$. These estimates are obtained by taking a long text, eliminating all the spaces and punctuation (and other non-literals like numbers), turning everything to lower case, and treating the obtained sequence as the outcome of a series of independent trials.

a. Use the above *language models* to decide on the language of the following sentences by the *Maximum Likelihood (ML)* method. The sentences are:

1. "As far as the laws of mathematics refer to reality, they are not certain, as far as they are certain, they do not refer to reality."

–Albert Einstein

2. ``Chi trova un amico, trova un tesoro.'' (He who finds a friend finds a treasure.) – Italian proverb http://en.wikiquote.org/wiki/Italian_proverbs

This sentence is in Italian, so none of the models you have is true. However, your program will still output a "best guess".

3. ``Las cuentas, claras, y el chocolate, espeso'' (Keep accounts (or relationships) transparent, and the chocolate, opaque)

–Spanish proverb

4. `` Wir finden in den Buchern immer nur uns selbst. Komisch, dass dann allemal die Freude gross ist und wir den Autor zum Genie erklaren.'' (We find in books always only ourselves. Funny how great the joy is, and how we think the author a genius.)

– Thomas Mann

5. ``Darkness cannot drive out darkness; only light can do that. Hate cannot drive out hate; only love can do that.''

– M. L. King, Jr.

For each sentence, do the following:

- Preprocess: Turn all letters to lower case, eliminate spaces and punctuation.

[1]The source of this data is http://www.santacruzpl.org/readyref/files/g-l/ltfrqeng.shtml, ltfrqger.shtml, ltfrqsp.shtml, ltfrqfr.shtml.

- Get the sufficient statistics: Count the number of times each letter appears in the sentence. These are the counts n_a, n_b, ... n_z.
- For each language, compute the log-likelihood of the sentence in that language $l_{E,G,S,F}(sentence) = \log_2 P_{E,G,S,F}(sentence)$. Print out these log-likelihoods. Make sure to convert into base 2 logarithms or to indicate the basis of the logarithm if it is not base 2.
- Output the best guess according to the ML method, i.e the language that gives highest likelihood to the data.
- Comment on what you observe: are the guesses correct? If not, why do you think not? How does the likelihood of the best guess depend on the length of the sentence? How does the difference in log-likelihoods between the best guess and the second best guess depend on the length of the sentence? What do you think of the probability models defined here as description of how language is produced?

Here is a short matlab code that computes the statistics of a sentence, typed all in lower case. It ignores all characters different from "a–z".

```
alphabet='abcdefghijklmnopqrstuvwxyz';
sentence = input('Type a sentence (lower case only):  ');
for ii = 1:26;
counts( ii ) = length( find( sentence == alphabet( ii )));
end;
```

For python, find sample code for reading the data in `hw1-language-template.py`.

A note on computing likelihoods Numbers like $\theta_i^{n_i}$ decrease exponentially with n_i, and for longer sequences produce numeric underflow. On the other hand, expressions like $n!$ easily produce numeric overflow. This is why all but the simplest calculations of likelihood should be done with logarithms. To make the results intuitive, we will use $\log_2$ or $\log_{10}$.

Note: this is classification The task that you just performed, deciding which of a given set of sources has generated an observation (in this case a sentence) is called **classification** or **pattern recognition**. Classification is very important both in Artificial Intelligence and in Statistics. We will talk more about classification later in this course.