STAT 391
Homework 1
Out March 29, 2018
Due April 5, 2018

©Marina Meilă
mmp@cs.washington.edu

[Problem 1 – Practice with probability – 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. This data is collected in the file `hw2-little-amazon.dat` which is available on the assignments web page.

For all the following questions, please give the "literal" expression of the answer as well as the numeric value.

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 2 – Practice with probability – NOT GRADED]

A man claims to have extrasensory perception. As a test, a fair coin is flipped 10 times, and the man is asked to predict the outcome in advance. He gets 7 out of 10 correct.

a. Let Y refer to the number of correct tests, and denote the outcomes of the 10 individual tests with the random variables $X_1, X_2, ..., X_{10}$. What are the distributions of each X_i? What is the relationship between $X_{1:10}$ and Y? (No proof required).

b. What is the probability that he would have done at least this well if he had no ESP, i.e. if his guesses were essentially random?

c. Suppose the test is changed – now, the coin is flipped until the man makes an incorrect guess. He guesses the first two correctly, but guesses the third wrong. What is the probability of this experimental outcome (again, assuming no ESP)?

d. Assume both tests were planned beforehand. What is the probability that both of these tests turned out the way they did? In other words, the plan was to first flip the coin 10 times and count how many times the man is correct (Y) then to continue flipping until the man makes the next mistake, at flip $10 + Z$. You are asked the probability that $Y = 7$ and $Z = 3$.

Problem 3

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. `La verite vaut bien qu'on passe quelques annees sans la trouver.''` (Truth is worth taking a few years to find.)

– Renard

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 po, non vo; chi vo, non po; chi sa, non fa; chi fa, non sa; e cosi, male il mondo va."` ("Who can do, don't want to; Who wants to, can't do; Who knows how to do, won't do it; Who does it, doesn't know how to; and, so, badly goes the world.")

–Italian Proverb

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

For each sentence, do the following:

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

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

- 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` or see and alternative template (and data file) at `spring13/homework03.hmtl`.

Note: 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.

Problem 4 – ML Estimation

Estimate a model for "Lincoln-English" using the following texts:

"What constitutes the bulwark of our own liberty and independence? It is not our frowning battlements, our bristling sea coasts, the guns of our war steamers, or the strength of our gallant and disciplined army. These are not our reliance against a resumption of tyranny in our fair land. All of them may be turned against our liberties, without making us stronger or weaker for the struggle. Our reliance is in the love of liberty which God has planted in our bosoms. Our defense is in the preservation of the spirit which prizes liberty as the heritage of all men, in all lands, every where. Destroy this spirit, and you have planted the seeds of despotism around your own doors."
–From the September 13, 1858 Speech at Edwardsville, IL

‘‘It is not merely for to-day, but for all time to come that we should perpetuate for our children's children this great and free government, which we have enjoyed all our lives. I beg you to remember this, not merely for my sake, but for yours. I happen temporarily to occupy this big White House. I am a living witness that any one of your children may look to come here as my father's child has. It is in order that each of you may have through this free government which we have enjoyed, an open field and a fair chance for your industry, enterprise and intelligence; that you may all have equal privileges in the race of life, with all its desirable human aspirations. It is for this the struggle should be maintained, that we may not lose our birthright--not only for one, but for two or three years. The nation is worth fighting for, to secure such an inestimable jewel.''
–August 22, 1864 Speech to the 166th Ohio Regiment

a Estimate $\hat{\theta}_a, \hat{\theta}_b, \ldots \hat{\theta}_z$ the parameters of the Lincoln-English language model from the above texts (available in the files `lincoln_text.txt`).

b Decide if the text

‘‘Fourscore and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty and dedicated to the proposition that all men are created equal.’’

– Lincoln’s Gettysburg Address on November 19, 1863.

is written in English, Spanish, German, French or Lincoln-English by the Maximum Likelihood method used in **Problem 3**.

Note that for Lincoln-English **some ML parameters may be 0** because not all letters of the alphabet are present in this text. If this happens, *do something reasonable* to get around this problem and explain what you did.