STAT 391
Homework 2
Out April 21, 2020
Due April 28, 2020

©Marina Meilă
mmp@cs.washington.edu

Problem 1 – Estimation of small probabilities

Submit the code used for this problem through Canvas.
This problem requires you to use Maximum Likelihood estimation and the smoothing methods from Lecture 7 to estimate probabilities for the letters in the English alphabet.

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. You will estimate the parameters $\theta_{A:Z}$ of this simple model from the text below (also available in `hw3-mlk-letter-estimation.txt`).

```
To save man from the morass of propaganda, in my opinion, is one of the chief aims
of education.  Education must enable one to sift and weigh evidence, to discern
the true from the false, the real from the unreal, and the facts from the fiction.
[...]  The function of education, therefore, is to teach one to think intensively
and to think critically.
```

Martin Luther King, Jr., *The Purpose of Education*

First, preprocess this text: Turn all letters to lower (or upper) case, eliminate spaces and punctuation. Then proceed with the questions of the homework.

a. 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. Print out the counts $n_{a:j}$ only.

b. Let S be the sample space $\{\texttt{a},\texttt{b},\texttt{c},\ldots\texttt{z}\}$, with $m = |S| = 26$. Determine the sets $R_0, R_1, \ldots R_n$, where $R_k = \{j \in S,\, n_j = k\}$. Some of these sets will be empty; enumerate only those which are non-empty.

c. Let $r_k = |R_k|$ and r be the number of unique letters observed in thetext above. Verify that $r = \sum_{k=1}^n r_k$, $m = \sum_{k=0}^n r_k$, and $n = \sum_{k=1}^n kr_k$. What is the fingerprint r_k, $k = 0, \ldots$ of this data set?

For the estimation questions **d,e,f,g**, calculate the probability estimates for all $\theta_{a:z}$ in your code (you will need them for question **h**), but only show results for one letter (of your choice) from each type R_k, for $k = 0, 1, 2$ and K the largest k with $r_k > 0$ (i.e., for $k = 0, 1, 2, K$ choose a letter j for which $n_j = k$). For these θ_j estimates, give the formula, then the formula with numerical values replaced, and final numerical output. (E.g. $\theta_a^{ML} = n_a/n = 3/7 = 0.43$.)

d. Compute the ML estimates $\theta_{a:z}^{ML}$ of the letter probabilities.

e. Compute now the Laplace estimates $\theta_{a:z}^{Lap}$ of the same probabilities

f. Compute the simplified Good-Turing estimates $\theta_{a:z}^{GT}$ of the same probabilities.

g. Compute the Ney-Essen estimates $\theta_{a:z}^{NE}$ of the same probabilities, taking $\delta = 1$.

h. Now use the estimates you obtained to compute the (log-)probability of the text in either one of `hw3-test-letter-estimation.txt` or `hw3-test-letter-estimation-large.txt`. Also compute the log-probability of the *training data* `hw3-mlk-letter-estimation.txt`). Numerical results only for this question.

Which method gives the highest log-likelihood of the new data? Which method gives the highest log-likelihood of the training data?

[Problem 2 – CDF's and densities – Not graded]
Let

$$F(x) = \begin{cases} 0, & x \leq 0 \\ x^2, & 0 < x \leq 1 \\ 1, & 1 < x \end{cases} \tag{1}$$

and

$$G(x) = \begin{cases} 0, & x \leq 0 \\ 2x^2, & 0 < x \leq 0.5 \\ 1 - 2(1-x)^2 & 0.5 < x \leq 1 \\ 1, & 1 < x \end{cases} \tag{2}$$

be two cumulative distribution functions.

1. – **Not graded** Plot F, G (OK to do by hand, but make a neat drawing, labelling all the important coordinates).

2. Compute their corresponding densities f and g. Plot them on a graph (OK to do by hand, but make a neat drawing, labelling all the important coordinates).

3. Denote by P_F and P_G the probability distributions defined by F, G. Find a, a' such that $P_F(0, a) = P_F(a, 1)$ and $P_G(0, a') = P_G(a', 1)$. a, a' i.e. the *medians* of the distributions P_F, P_G. Represent x, x' on the graphs in questions **1., 2.**

4. – **Not graded** Find the probabilities of the following intervals $[0, 0.25]$, $[1, 1.75]$ under P_F, P_G.

5. – **Not graded** Find the shortest interval $[a_F, b_F]$ that has probability 0.1 under F. Find the shortest interval $[a_G, b_G]$ that has probability 0.1 under G. Motivate your answer. Show the intervals $[a_F, b_F]$, $[a_G, b_G]$ on the graphs of f, g respectively.

6. Calculate the means of f, g, denoted by $E_f[X]$, $E_g[X]$.

[Problem 3 – Probabilities of intervals – Not graded]

The exponential distribution with parameter $\gamma > 0$ is defined by the density $f_\gamma(x) = \gamma e^{-\gamma x}$ on $S = [0, \infty)$.

1. Denote by p_n the probability of the interval $[n-1, n)$ under the exponential distribution, i.e. $p_n = Pr[\, x \in [n-1, n)\,]$ for $n = 1, 2, \ldots$. What is the expression of p_n as a function of γ and n? What is this expression if $\gamma = \ln 2$?

2. What is the expression of $\frac{p_n}{p_{n+1}}$ as a function of γ and n? What is this expression if $\gamma = \ln 2$?

3. Plot on the same graph the densities $f_\gamma(x)$ for $\gamma = \ln 2, \ln 3, \ln 4$.

4. Let $g(x) = \frac{1}{Z} e^{\gamma(x+3)}$, $x \in S = [-3, \infty)$. Evaluate the normalization constant Z as a function of γ. Evaluate the expression of the CDF G of this distribution.

Problem 4 – Rayleigh distribution

The Rayleigh parametrized family of distributions is described by

$$f(r; a) = \frac{r}{a^2} e^{-\frac{r^2}{2a^2}} \quad r \geq 0 \tag{3}$$

If we shoot at a target centered at $(0, 0)$ and our bullets hit at (x, y), where each of x, y is normally distributed with mean $\mu = 0$, then the distance $r = \sqrt{x^2 + y^2}$ from the target center is distributed according to a Rayleigh distribution.

Assume you are given a data set $\mathcal{D} = \{r_1, r_2, \ldots r_n\}$ drawn independently from a Rayleigh distribution. The task is to determine the formula for the ML estimate of the parameter a as a function of the data.

1. Write the formula of the likelihood of a, $L(a)$.

2. Take the logarithm of $L(a)$ to obtain the log-likelihood $l(a) = \log L(a)$. Then compute the derivative

$$\frac{\partial l}{\partial a}$$

3. Now solve the equation $\frac{\partial l}{\partial a} = 0$ to obtain a formula for a^{ML} as a function of the data.

4. Does this problem have sufficient statistics? What are they (is it)?