DS201
Statistical Programming
Assignment 1

Ashutosh Kumar Jha
12340390
DSAI
2nd Year
Semester 4

----

1. Question 1: Analysis of text files and computation of probabilities and entropies

Introduction:

The problem statement includes the textual analysis of four files: fileA.txt, fileB.txt, fileC.txt, fileD.txt. The problem asks us to preprocess the text using the constraints given and compute the frequencies of the letters and their probabilities and also entropies for the letters using the files given. The same files are supposed to be used in computing the above mentioned parameters but instead of using letters, words are supposed to be used. The purpose of finding the probabilities is to report the top 10 words and letters present in the files and entropy computation would provide us with insights about the randomness of text in the files. Overall, the computation of the above mentioned parameters provides us with the insights about the diversities, uncertainties and randomness in the given text-based system.

Data:

The dataset contains 4 files: fileA.txt, fileB.txt, fileC.txt and fileD.txt, and the dataset was already provided to us. The files are text files and contain raw text. The text in the files are supposed to be preprocessed in order to apply computations on it and find the given parameters like frequencies, probabilities, entropy. The cleaned text was then analyzed for letter and word frequencies, and entropy was calculated to assess the randomness of their distributions. Different files are present in the dataset for computing different parameters like entropies and probabilities.

Methodology:

—---------------------

1. Cleaning and Preprocessing of data:
   The text was cleaned using a function named as “preprocess_text” which accepts the path of the text file which needs to be preprocessed and returns the cleaned content. During the cleaning process, the content of the file is first read using a file reader in python and then the special characters, whitespaces and punctuations are removed from the text and all the letters are then converted into lowercase in order to maintain uniformity in the content for feasible calculations to occur.

2. Calculation of alphabet probabilities:
   In order to find the probabilities of the letters, the frequencies of the alphabets are computed using the inbuilt Counter function present in python’s collections package. The Counter function returned us the dictionary containing the letters and their corresponding frequencies in the form of key value pairs. Since the frequencies are found, we can find the value of the probability for each letter by dividing the frequency with the total number of the letters in the text file. After finding the probabilities, sorting is done on the value of the probabilities for finding the top 10 letters existing.

3. Calculation of entropy:
   The entropy is a function of the probabilities and since we have already calculated the probabilities for the alphabets, we can calculate the entropy for the alphabet distribution very easily using the formula given.

4. Repetition of steps 2 and 3, for the words:
   The steps 2 and 3 are supposed to be repeated in fileC and also in fileD but the only change is that we would find the given parameters for the words and not for the letters. The preprocessing again takes place, but for the words this time, for cleaning the text. The word probabilities are then computed using the corresponding function and then the entropy is calculated using the same function since it is just a function of probabilities:

Results:

The results of the problem statement are:

1. Probabilities of letters in fileA.txt: The probabilities of the letters were computed with the top letter being:
   s: 0.0424
   z: 0.0413
   y: 0.0413
   f: 0.0408
   w: 0.0403
   t: 0.0401
   u: 0.0401
   x: 0.0400
   j: 0.0398
   k: 0.0396

2. Entropy of letters in fileB.txt: The entropy of the letters in the file fileB.txt is 4.1760

3. Probabilities of words in fileC.txt: The probabilities of the words were computed with the top words being:
   the: 0.0805
   of: 0.0356
   and: 0.0302
   to: 0.0272
   a: 0.0249
   in: 0.0244
   is: 0.0145
   be: 0.0122
   as: 0.0099
   that: 0.0092

4. Entropy of words in fileC.txt: The entropy of the words in the file fileC.txt is 8.9415

5. Probabilities of words in fileD.txt: The probabilities of the words were computed with the top words being:
   the: 0.0691
   and: 0.0369
   of : 0.0362
   i : 0.0340
   a: 0.0248
   to : 0.0216
   in : 0.0170
   was : 0.0157
   that : 0.0127
   my: 0.0124

6. Entropy of words in fileD.txt: The entropy of the words in the file fileC.txt is 9.1784

Discussions:

Letter Frequency:
In fileA.txt, the distribution of letters is somewhat unusual, with 's', 'z', and 'y' being among the most frequent letters. This could indicate a specific writing style, genre, or encoding. Typically, vowels and common consonants like 't', 'n', and 'e' dominate English texts, but the observed distribution suggests that the text may not adhere to typical English patterns.

Letter Entropy:
The entropy of 4.1760 in fileB.txt indicates that the letter usage is moderately unpredictable. While some letters dominate (such as 's', 'z', and 'y'), the overall letter distribution is relatively even, suggesting that the text might be a mix of different styles or types of content.

Word Frequency:
The top 10 words in fileC.txt and fileD.txt are dominated by common English stop words like “the”,”of”,”and”, which is expected in most unprocessed text data. These stop words occur frequently because they are fundamental to sentence structure, while more specific content words tend to be less frequent.

Word Entropy:
The entropy value for words in fileC.txt and fileD.txt are 8.9415 and 9.1784, which is quite high and therefore indicates a very varied and diverse vocabulary. This suggests that the text is rich in content, with a relatively even distribution of word usage. This could point to a more sophisticated or technical piece of writing, or it could indicate a text that is designed to use a wide range of vocabulary

Conclusion:

This analysis of letter and word frequencies, as well as entropy values, has provided insights into the structure and unpredictability of the text in the four analyzed files. The results suggest that the texts vary in their letter and word usage patterns, with some files showing highly predictable distributions and others exhibiting more randomness and variety.

- File A: Letter frequency shows some unusual patterns with certain letters (like 's', 'z', 'y') being dominant.
- File B: The entropy value of the letters indicates a moderately unpredictable letter distribution.
- File C: The top words are typical stopwords and also the entropy of words is quite high, which is standard for most English text and also show a diverse and varied vocabulary

● File D: The word entropy is very high with the most frequent words being the stop words, indicating again a very diverse and varied vocabulary.

These findings highlight the importance of entropy as a tool for understanding the complexity and randomness of language in different contexts.

2. Question 2: Analysis of mean and variance for samples of randomly generated numbers, from a uniform distribution

1. Introduction:

This report explores the convergence behavior of the sample mean and sample variance for a uniform distribution between 0 and 1. Through randomly generated samples of numbers following the uniform distribution, we examine how the sample mean and variance converge to their theoretical values as the sample size increases. The uniform distribution has a known mean of μ=0.5 and variance σ(sq)=1/12. The goal is to observe the behavior of these sample statistics and verify their convergence to the true values

2. Data

In this process of generation, random samples are drawn from a uniform distribution U(0,1). For each sample size n from 1 to 10,000 with the step size of 100, the sample mean and sample variance are computed. The following theoretical values are used for comparison:

● True Mean: μ=0.5
● True Variance: σ(sq)=1/12
● We simulate 100 trials for each sample size to calculate the sample mean and sample variance, and these values are then plotted to observe their convergence

3. Methodology

3.1. Simulation Process:

The simulation process is as follows:
1. Generating Random Samples: For each sample size n (from 1 to 10,000), n random numbers are generated from a uniform distribution between 0 and 1 using the NumPy function np.random.uniform(0, 1, size=n).
2. Calculating Sample Mean: For each set of samples, the sample mean is computed using stats.mean(samples).
3. Calculating Sample Variance: The sample variance is calculated using stats.variance(samples).

The stats refers to the statistics library in python which is used here for calculating the statistical data about the generated samples.

3.2. Plotting:
Two key statistics are plotted as functions of sample size:
1. Sample Mean: The convergence of the sample mean is plotted against the sample size.
2. Sample Variance: The convergence of the sample variance is plotted against the sample size

3.3. True Values:
For comparison, the following true values for the uniform distribution are plotted:
- The true mean of the uniform distribution is 0.5.
- The true variance of the uniform distribution is 1/12≈0.0833

4. Results

Part (a) - Convergence of Sample Mean:
The plot of the sample mean as a function of sample size shows how the sample mean approaches the true mean of 0.5 as the sample size increases.

Observation:
- At small sample sizes, the sample mean fluctuates significantly, reflecting the inherent variability in smaller samples.
- As the sample size increases, the sample mean becomes increasingly stable and converges towards the true mean of 0.5.

Sample Mean Convergence Plot:
- The blue curve represents the sample mean at each sample size, while the red dashed line at y=0.5 represents the true mean.
- As the sample size increases, the blue curve closely follows the true mean line, demonstrating convergence

Part (b) - Convergence of Sample Variance:
The plot of the sample variance as a function of sample size shows how the sample variance approaches the true variance of 1/12≈0.0833 as the sample size increases.

Observation:

- At small sample sizes, the sample variance shows high variability due to the small number of data points.
- As the sample size increases, the sample variance stabilizes and converges towards the true variance of the uniform distribution.

Sample Variance Convergence Plot:

- The green curve represents the sample variance at each sample size, while the red dashed line at y=1/12 represents the true variance.
- As the sample size increases, the green curve stabilizes around the true variance, demonstrating convergence

5. Discussion

Sample Mean:
The sample mean converges to the true mean of 0.5 as expected from the law of large numbers. Initially, for small sample sizes, the sample mean fluctuates due to the limited data, but as the number of samples increases, the sample mean becomes increasingly accurate and stabilizes around the true mean. This behavior illustrates the concept that larger sample sizes provide better estimates of the population mean.

Sample Variance:
The sample variance also converges to the true variance of 1/12 with increasing sample size. In the beginning, the variance is more volatile due to the small sample size, but as the number of samples increases, the sample variance stabilizes around the true value. This reflects the fact that larger samples better represent the variability of the entire population.

Convergence Behavior:
Both the sample mean and sample variance demonstrate typical convergence behavior, where the fluctuations decrease as the sample size increases. This is consistent with the central limit theorem and the law of large numbers, which guarantee that the sample mean and sample variance converge to their true values as the sample size grows

6. Conclusion

This simulation demonstrates the convergence of the sample mean and sample variance for a uniform distribution between 0 and 1. As expected, both statistics converge to their true values with increasing sample size. Specifically:

- Sample Mean converges to 0.5.
- Sample Variance converges to 1/12≈0.0833

These results highlight the power of large sample sizes in obtaining accurate estimates of population parameters. The simulation also provides a clear illustration of the law of large numbers in practice, where increasing the sample size leads to more reliable estimates of the true mean and variance.

3. Question 3: Analysis of mean and variance for samples of randomly generated numbers, from a uniform distribution

1. Introduction

This report investigates the convergence behavior of the sample mean and sample variance for a Gaussian distribution with a specified mean and standard deviation. The Gaussian (or normal) distribution is commonly used in statistics and has the following parameters:

- True Mean: μ=4
- True Variance: σ(sq)=9

The goal here is to observe how the sample mean and variance converge to the true values as the sample size increases.

By running simulations for increasing sample sizes, we can demonstrate the law of large numbers and the central limit theorem.

2. Data

The data used for the simulation comes from a Gaussian distribution with the following parameters:

- Mean: μ=4
- Standard Deviation: σ=3

- Sample Sizes: From 1 to 10,000 samples with the step_size of 100

For each sample size, the sample mean and variance are computed. The results are averaged over 100 trials to reduce variability and show the trend as sample size increases.
Parameters:

- Max Sample Size: 10,000
- Number of Trials: 100
- Gaussian Distribution: N(μ=4,σ=3)

3. Methodology

3.1. Simulation Process:
The simulation process is outlined as follows:
1. Generate Random Samples: For each sample size n (from 1 to 10,000), n random samples are drawn from a Gaussian distribution with mean 4 and standard deviation 3 using NumPy's np.random.normal(mean, std_dev, size=n).
2. Calculate Sample Mean: For each sample set, the sample mean is calculated using stats.mean(samples)
3. Calculate Sample Variance: The sample variance is computed using stats.variance(samples)

The stats here is referred to the statistics library in python which is used to find the statistical parameters needed.

3.2. Plotting:
Two main plots are generated:
1. Convergence of Sample Mean: The plot shows how the sample mean converges to the true mean of 4 as the sample size increases.
2. Convergence of Sample Variance: The plot shows how the sample variance converges to the true variance of 9 as the sample size increases.

3.3. True Values:
For comparison, the following true values for the Gaussian distribution are plotted:

- True Mean: μ=4
- True Variance: σ(sq)=9

4. Results

Part (a) - Convergence of Sample Mean:

The plot of the sample mean as a function of sample size illustrates how the sample mean approaches the true mean of 4 as the sample size increases.

Observations:
- At small sample sizes, the sample mean fluctuates significantly due to the random nature of the samples.
- As the sample size increases, the sample mean becomes more stable and converges to the true mean of 4.

Sample Mean Convergence Plot:
- The blue curve represents the sample mean at each sample size, and the red dashed line at y=4 represents the true mean.
- As the sample size increases, the blue curve converges to the true mean of 4, demonstrating convergence.

Part (b) - Convergence of Sample Variance:
The plot of the sample variance as a function of sample size shows how the sample variance approaches the true variance of 9 as the sample size increases.

Observation:
- At small sample sizes, the sample variance fluctuates more due to the small number of data points.
- As the sample size increases, the sample variance stabilizes and converges toward the true variance of 9.

Sample Variance Convergence Plot:
- The green curve represents the sample variance at each sample size, and the red dashed line at y=9 represents the true variance.
- As the sample size increases, the green curve approaches the true variance value, showing the expected convergence

5. Discussion:

Sample Mean:
The sample mean exhibits the typical behavior described by the law of large numbers. Initially, with smaller sample sizes, the sample mean fluctuates because of the random variation in the data. However, as the number of samples increases, the sample mean becomes more accurate and converges towards the true mean of the Gaussian distribution, which is 4.

Sample Variance:

The sample variance converges to the true variance of 9 in a similar manner. For smaller sample sizes, the variance shows large fluctuations because the data points are sparse. As the sample size grows, the variance stabilizes and becomes a more reliable estimate of the true variance. This behavior is also consistent with the law of large numbers, which ensures that with larger sample sizes, sample statistics such as variance converge to their true population values.

Convergence Behavior:
Both the sample mean and variance converge as expected with increasing sample size. The initial fluctuations at small sample sizes gradually decrease, illustrating that larger samples provide more accurate estimates of the population parameters. This is a key feature of statistical inference: with sufficient data, sample statistics converge to their true values.

6. Conclusion

This simulation demonstrates the convergence of the sample mean and sample variance for a Gaussian distribution with mean 4 and standard deviation 3.

The following observations were made:
- Sample Mean converges to the true mean of 4.
- Sample Variance converges to the true variance of 9.

These results confirm the expected behavior outlined by the law of large numbers. As the sample size increases, the sample mean and variance become increasingly reliable estimates of the true population parameters.

This experiment illustrates the importance of large sample sizes in obtaining accurate estimates and can be applied in various statistical analyses, where large datasets help ensure that the sample statistics closely match the population parameters.

Code: 12340390 Ashutosh Asg1.ipynb