1040: Intro to Data Analytics

Final Spring 2019

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Instructor: | Konstantinos Pelechrinis |  | Name: | Angela Tseng |

1. Read carefully each multiple-choice question and circle the correct answer. Every question is worth 3 points.
2. The degree of node A in the following network is:



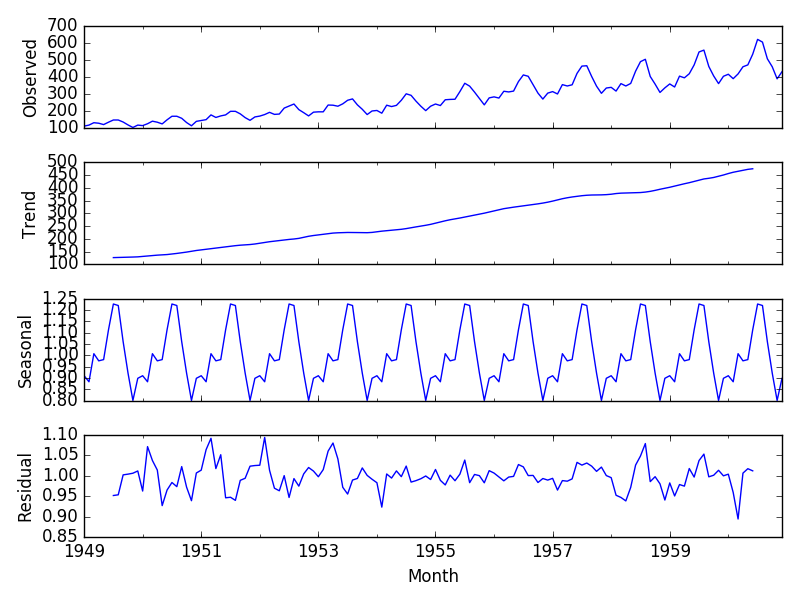
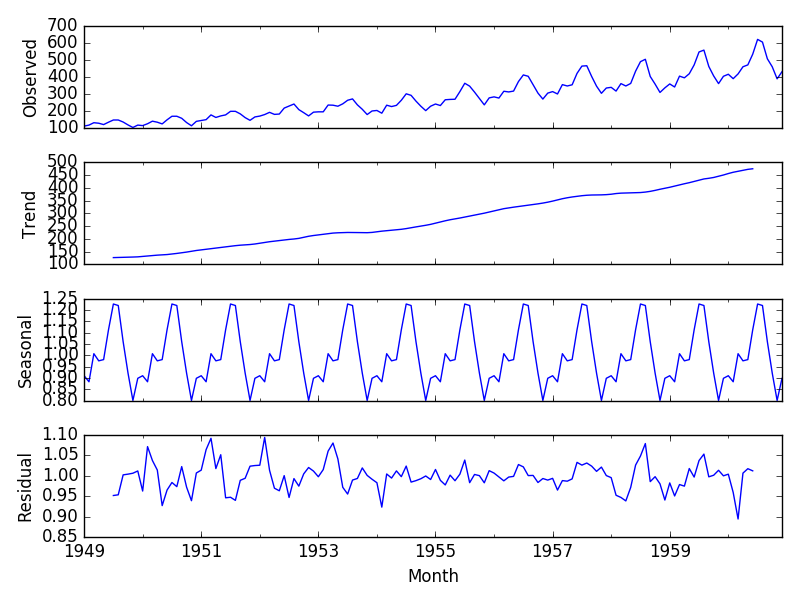
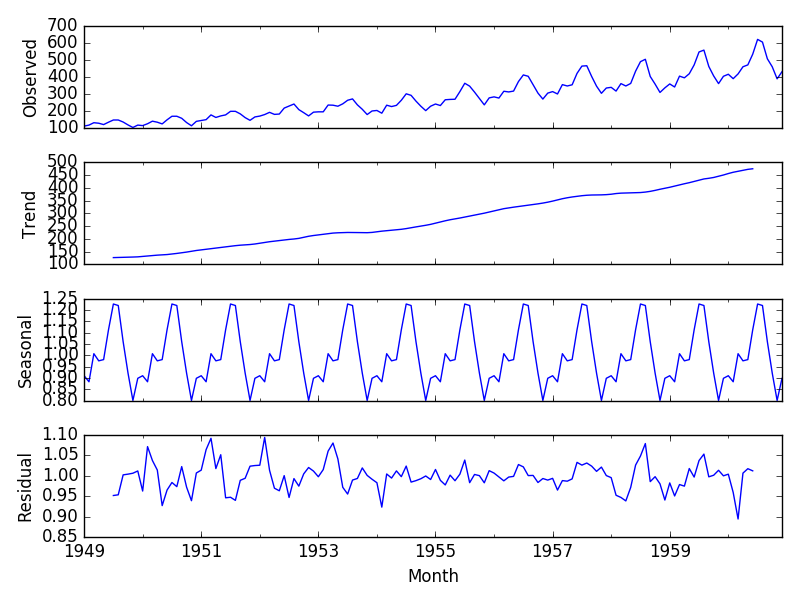
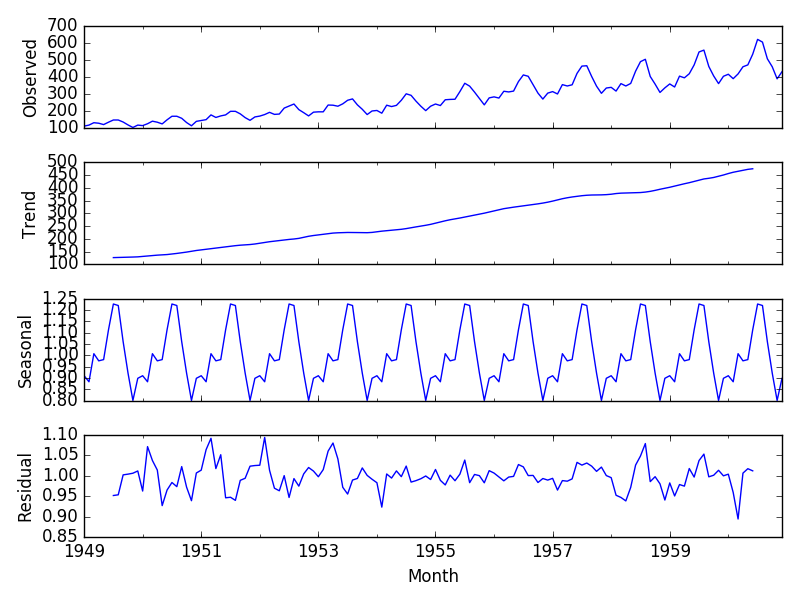
* + 2
  + 3
  + 4
  + 5

1. Adding a dropout layer in a neural network helps:
   * Improve its generalization performance
   * Speed up the convergence of the model
   * Speed up the inference
   * Does not help in general
2. Consider the following dataset representing the height (in cm) ofs the students in 8th grade of your local high school; {165, 161, 168, 163, 153, 171, 168, 165}. Which of the following samples ***can*** be a bootstrapped sample obtained from the dataset:
   * {165, 163, 153, 153, 168, 171, 165, 161}
   * {168, 168, 171, 171, 171, 163, 168, 168}
   * {161, 161, 161, 161, 165, 161, 161, 161}
   * All of the above
3. For a clustering problem you obtain the following Silhouette values for different number of clusters. The final choice of clusters based on the Silhouette heuristic is:



* + 2
  + 10
  + 1
  + 6

1. Consider the time series shown at the top of the following figure (within the blue box). This time series has been decomposed using additive decomposition. The components observed are shown in the other three plots of the figure. Which figure corresponds to each of the components?
   * A: random/residuals, B: trend, C: seasonal
   * A: trend, B: random/residuals, C: seasonal
   * A: seasonal, B: trend, C: random/residuals
   * A: seasonal, B: random/residuals, C: trend



A

B

C

1. Each of the following questions is worth 2 points. Read carefully and answer choose whether the sentence is true or false.

1. Every time you apply k-means on the same dataset with the same number of clusters (k value) you will obtain the same clusters. T

2. In a directed network the average in-degree is equal to the average out-degree. T

3. Bootstrap works in most of the cases because it assumes the sampled variable follows a normal distribution that can essentially model any distribution with the right parameters. F

4. Hierarchical clustering provides *better* results compared to k-means. F

5. Adjacent observations in time-series data (excluding the noise term) are independent and identically distributed. F

1. The following question is worth 5 points and it is for **extra credit**.

Your roommates are getting into data analysis this semester and are enthused of clustering algorithms and in particular, affinity propagation (this is a clustering algorithm that we have not seen in class, but the inner workings of the algorithm are not important). They claim that *affinity propagation is the current state-of-the-art in clustering, in that it provides the optimal clusters for a dataset*. How would you respond to this claim?

I would break down the benefits of silhouette analysis and what it does before asking them to compare the clustering provided by their affinity propagation algorithm to the recommended amount by the silhouette score. Since the silhouette score measures how similar an object is to its own cluster, it stands as a more objective analysis of what the optimal clusters for a dataset should be.