Asmt 6A: Noise, Outliers, and Anomalies

Turn in through GradeScope by 1pm:
Wednesday, November 19
50 points

Overview

In this assignment you will explore different approaches for finding Anomalous regions, and Outliers.

You will use one data sets for this assignment:

- `http://www.cs.utah.edu/~jeffp/teaching/DM/A6/T.csv`

As usual, it is recommended that you use LaTeX or another method which can properly display mathematical notation for this assignment. If you do not, you may lose points if your assignment is difficult to read or hard to follow. Find a sample form in this directory: `http://www.cs.utah.edu/~jeffp/teaching/latex/`

You could also utilize an LaTeX template specifically created for this assignment. Click here.

1 Anomalies (50 points)

Consider the sequence of 100 values in data array `T`.

A (20 points): Find the index s (start counting at 0) so that all values less than this index (a set $S = \{T[0], T[1], T[2], \ldots, T[s-1]\}$) define the most anomalous set. Use the change point detection model (a 1-sided subset anomaly) discussed in class (Lecture 21). Assume the value in each set S and $T \setminus S$ is constant with normal noise; you can use $\sigma = 1$.

Explain why your chosen index s is the right change point to define S.

B (10 points): Compute the log-likelihood ratio $LLR(T, S)$ where S is your 1-sided subset anomaly, and T is the full array. That is, let

$$L(T) = \max_{\mu_T} \prod_{j \in [100]} \frac{1}{\sqrt{2\pi}} \exp\left(-(T[j] - \mu_T)^2/2\right)$$

be the maximum likelihood of all $x \in T$ fit with the maximum likelihood estimate. Then let $L(S, T \setminus S) = L(S) \cdot L(T \setminus S)$ where $L(S)$ is the same as $L(T)$ except fit in the subset S, and again for $L(T \setminus S)$. Finally the log-likelihood ratio is

$$LLR(T, S) = \ln\left(\frac{L(S, T \setminus S)}{L(T)}\right).$$

C (20 points): Find and report the outliers in this data set. Hint, there are less than 10 of them.

Report how you did this, and why the set you report are the anomalous points.

2 BONUS: Bad Anomaly (1 point)

Change the value of the point at index 3 so that the likelihood ratio for your same choice of s and S is 10. Report this updated value of point at index 3, and explain how you arrived at it.