21 Anomaly Detection and Quantification

This lecture will help understand how to determine and quantify anomalies.

An anomaly from a data set X would be a grouping of data elements; a subset $S \subset X$ or a single element $s \in S$ which is somehow sufficiently different from what you would expect.

To follow this paradigm, you need to:

1. Define what you expect from your data

2. Measure how what you observe differs with respect to that expectation. If there is "shape" of anomalies we can consider, what are they? And how to find the one that differs the most.

3. Determine if that observed anomaly is sufficiently different from what you expected, if so we may consider it interesting or *anomalous*.

We'll next review how to think about and formalize these steps.

21.1 What to Expect from Data

Formally we need a distribution $D(\theta)$ with assumption that $X \sim D(\theta)$. Either X is drawn jointly from $D(\theta)$, or more simply, $X \sim_{iid} D(\theta)$ so each $x_i \in X$ is drawn independently $x_i \sim D(\theta)$.

Compute the likelihood of X as $L(X) = \arg\max_\theta(L(X;\theta))$ for some parameter set θ. One can formally define a *likelihood* as an (unnormalized) probability that something happens (under some modeling assumptions). For context here, we mostly need to understand it simply as a score of how likely something is to happen – the larger, the most likely.

However, if $S \subset X$ is anomalous, then it should be a different distribution than $D(\theta)$; e.g., $S \sim D(\theta')$, while $X \setminus S$ is drawn from some distribution $D(\theta'')$. Where in the baseline model we may have something simple like $\theta = p$ (a single scalar p as the parameters); to explain something more complicated (some regular points $X \setminus S$ and some anomalies S) we need more parameters e.g. $\theta' = (p', q')$. So the likelihood of such an anomaly would be:

$$L(S, X \setminus S) = \arg\max_{\theta'} L(S, X \setminus S; \theta') = \arg\max_{p,q} L(S, X \setminus S; p, q).$$

By default in the baseline case, perhaps $S = \emptyset$ and $q = 0$; but is ignored.

21.1.1 Log-Likelihood Ratio

$$LLR(X, S) = \log(\frac{L(S, X \setminus S)}{L(X)}) = \log(L(S, X \setminus S)) - \log(L(X))$$

The larger the $LLR(X, S)$ score, the more interesting the anomaly S is. But which potential anomaly S should we consider? The answer is: *the one that maximizes $LLR(X, S)$!* But that simplistic answer does not provide a complete explanation. We first will introduce a simple example, and then try to elaborate a bit.

21.1.2 Example: Change points

X is an ordered sequence of real values $x_1, x_2, \ldots, x_n \subset \mathbb{R}$, the index i (for point $x_i \in X$) can represent time of observation. These could be weather readings, prices of a stock, amount of electrical power used (or obtained e.g., via solar).

We will use a simple model with a *mean* parameter $\bar{p} = \frac{1}{n}\sum_{i=1} x_i$. With a normal noise assumption, the negative log-likelihood is $\sum_{i=1}^n (x_i - p)^2$ and is minimized at $p = \bar{p}$. (Note since we took the negative, we switch from maximizing to minimizing).

In a time series, a change point t is a value where the data distribution is different before and after t. So $S = \{x_1, \ldots, x_t\}$ and $X \setminus S = \{x_{t+1}, \ldots, x_n\}$. Given a choice of t and S, the negative log-likelihood is maximized by taking the average of each set.

Deriving sum of squared errors. Assume $X \sim_{iid} \mu$ and normal noise, so for a parameter p the probability of observing a given x_i is

$$\textbf{Pr}[x_i \mid p] = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-p)^2/\sigma^2)$$

for some unknown (but assumed constant) bandwidth parameter σ. Since x_i are independent, then the probability of seeing a set of them is

$$\begin{aligned}\textbf{Pr}[X \mid p] &= \prod_{i=1}^n \textbf{Pr}[x_i \mid p] = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-p)^2/\sigma^2) \\ &\propto \prod_{i=1}^n \exp(-(x-p)^2/\sigma^2)\end{aligned}$$

We only need to keep track of proportional solutions (using $\propto$) under a likelihood, as opposed to the probability.

The likelihood is then $L(X;p) = \prod_{i=1}^n \exp(-(x-p)^2/\sigma^2)$ and the log-likelihood is

$$\begin{aligned}\ln(L(X;p)) &= \ln(\prod_{i=1}^n \exp(-(x-p)^2/\sigma^2)) \\ &= \sum_{i=1}^n \ln(\exp(-(x-p)^2/\sigma^2)) \\ &= \sum_{i=1}^n [-(x-p)^2/\sigma^2] \\ &= -\frac{1}{\sigma^2}\sum_{i=1}^n (x-p)^2,\end{aligned}$$

and as a result we often *minimize* the *negative* log-likelihood.

Now note that this expression is minimized over p by setting p equal to the sample mean $\bar{p} = \frac{1}{n}\sum_{i=1}^n x_i$ from the observed data X. And the optimum value is the same regardless of the choice of σ.

Through a similar reduction (with the same constants, assuming same σ) when there is a change point at t then the optimal parameter values are $\bar{q} = \frac{1}{t}\sum_{i=1}^t x_i$ and $\bar{p} = \frac{1}{n-t}\sum_{i=t+1}^n x_i$.

21.2 Finding Anomalies

Now we can score how interesting an anomaly S is (using $LLR(X,S)$) we then need to determine how to find S.

In change point problems, this is simple. We only allow S defined as $S_t = X \cap (-\infty, t]$; so it is determined by a single parameter t. We can check each value $t \in [n]$, return the one that maximizes

$LLR(X, S)$. While it takes $O(n)$ time to compute $LLR(X, S_t)$, in updating from t to $t+1$, we can update the value $LLR(X, S_{t+1})$ in $O(1)$ time. Then the entire search takes linear $O(n)$ time (not the naive $O(n^2)$).

Thus, the most anomalous region of X (in terms of the change point model we consider) is

$$S^* = \arg\max_{S_t} LLR(X, S_t).$$

This change point model was chosen as an example because this optimization process is so simple and straight-forward. This course will show much more complex ideas of data, its structure, its distribution, and the form anomalies can take.

Here are two more general models, with a similar theme.

Subset anomalies. As with the change detection, this only considers anomalies that are defined by a subset S of X. But if we consider *any* subset $S \subset X$, this is too powerful: there are 2^n of them, which makes it hard to search over all, and easy to over-fit to values in a way that does not represent a real phenomenon.

As such, we typically define a *range space* $(X, \mathcal{R})$ which is a family of subsets of X. Typically these are defined to fit some structural (or geometric) property, like all one-sided ranges of the form $(-\infty, t]$ used in change detection. But it could also be two sided intervals $(s, t]$, or pairs of intervals... If X lies in a higher-dimensional space (e.g., $X \subset \mathbb{R}^2$), then we can define subsets with geometric shapes (like balls or rectangles or halfspaces). Typically, like with 1-sided intervals, there are a bounded number. One way to categorize this is VC-dimension (the same term that comes up in machine learning theory). If the VC-dimension of a range space is v, then there are at most $O(n^v)$ ranges to consider which induce different subsets $S \subset X$. For instance, for 2-sided intervals $v = 2$, for axis-aligned rectangles in 2 dimensions, then $v = 4$, and for balls in d-dimensions, then $v = d + 1$.

Functional anomalies. This is a more complicated form where we define a (parameterized function from class $\mathcal{F}$) $f : \mathcal{X} \to \mathbb{R}_+$ where $\mathcal{X}$ is the domain of the data X, and $\mathbb{R}_+$ is the set of non-negative real values. Then $f(x)$ somehow describes the contribution of a data point $x \in X$ to the anomaly. If we restrict the range of f to be $[0, 1]$ then it can describe the fractional membership of x in the anomaly ($f(x_i) = 0.8$ means x_i is 80% in the anomaly). If the range of f is $\{0, 1\}$ then it recovers the subset anomaly notion, since $f(x_i) = 1$ means $x_i \in S$ and $f(x_j) = 0$ means $x_j \notin S$.

The functional form is more powerful as it can capture the change point model and the LLR simultaneous by parameterizing f by the values t, p, and q. Then the effect of a point x_i depends on if $i \leq t$, and then the values p or q. We do not need to search over p and q in this setting, since we can set them optimally based on the choice of t (and values of data points in X).

We can then measure the distance between two data sets X and Z as follows. First define an extension of f applied to a data set (as opposed to one observation) as:

$$d_{\mathcal{F}}(X, Z) = \max_{f \in \mathcal{F}} \left| \mathbf{E}_{x_i \in X}[f(x_i)] - \mathbf{E}_{z_i \in Z} f(z_i) \right|.$$

Here $\mathbf{E}_{x \in X}[f(x)] = \frac{1}{|X|} \sum_{x_i \in X} f(x_i)$ is the (sample) expected value of f on the (discrete) distribution represented by X.

This is called an *integral probability measure*, and is a special class of distance functions between point sets and distributions (generalized to the integral for over distributions from which X and Z are drawn), and is always a pseudo-metric, and for all examples we discussed also a metric. We'll revisit these ideas later in the semester.

21.3 Quantifying Anomalies

How do we know if the anomaly is meaningful? We found the best one, but was it interesting enough to do something about? For every data set, there is some best anomaly.

For this, we need to see how unlikely it is to occur. But what does that mean?

Sometimes we can generate the precise probabilistic bounds, but this is rare (and often also has assumptions). For instance, in the change point model, we need to know the bandwidth parameter σ to calculate it. Estimating σ is possible, but has its own challenges.

Resampling. If we do know the distribution μ enough to draw a sample from it, then we can do that to create a new data set $X_1 \sim_{iid} \mu$ of the same size n as the original. Then we can repeat this process m times to create m *data sets*: $X_1, X_2, \ldots, X_m$ (that is in total $m \times n$ points).

Then for each X_j we can calculate $S_j^* = \arg\max_{[S_j]_t} LLR(X_j, [S_j]_t)$, and its $\gamma_j = LLR(X_j, S_j^*)$.

With this set $\Gamma = \{\gamma_1, \ldots, \gamma_m\}$ we can build an estimate of the distribution of the log-likelihood ratios.

Finally, we can measure our computed $\gamma = L(X, S^*)$ from our observed data against this distribution. If a β-fraction of the γ_j are larger than γ, then we say that (under this resampling model) that the p-value of our test is β. If β is sufficiently small (e.g., $\beta < 0.01$ or $\beta < 0.05$ or $\beta < 0.00001$, choose your favorite), we can say that the subset S^* we found is sufficiently anomalous.

Permutation testing. But what if we do not know μ well enough to sample from it?
A common trick is to use *permutation testing*, which works for structured subset anomalies.

This takes all of the values of our observation X, but randomly permutes their order. If the anomalies we consider depend on the order (e.g., the intervals in the change point example, or geometric shape examples), then it should be unlikely to be preserved under a random re-ordering of the data. Here we let the jth random re-order of X be the random data set X_j. Then we compute the rest (e.g., p-value β) the same we did if/when we could sample from μ directly.

Quick note on p-values. These are very very common in the sciences, and probably over-used – likely because that can directly provide a decision for people who do not really understand the underlying statistics!

An important observation here, is that *there are more than one reasonable way to generate p-values*. So what do we do if one crosses a threshold and other does not? The data scientist should think deeply about what this way of generating a baseline distribution means, which model they think most closely models reality, and if its unclear if something crosses a threshold – maybe they should be dubious!