数据分析时采用的“一维数值”建议

我们之前在做可视化或者拟合插值等数据的处理和分析时，通常会使用 AQI 数值。我们先来看 AQI 的计算规则如下：

对 PM2.5, PM10, SO2, NO2, O3, CO （其中PM2.5和PM10为 24小时平均浓度值) 分别计算:

$$IAQI_p = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}}(C_p - BP_{Lo}) + IAQI_{Lo}$$

式中：

$\mathrm{IAQI_P}$——污染物项目P的空气质量分指数；
C_P——污染物项目P的质量浓度值；
$\mathrm{BP_{H}}_i$——相应地区的空气质量分指数及对应的污染物项目浓度指数表中与C_P相近的污染物浓度限值的高位值；
$\mathrm{BP_{L}}_o$——相应地区的空气质量分指数及对应的污染物项目浓度指数表中与C_P相近的污染物浓度限值的低位值；
$\mathrm{IAQI_{H}}_i$——相应地区的空气质量分指数及对应的污染物项目浓度指数表中与$\mathrm{BP_{H}}_i$对应的空气质量分指数；
$\mathrm{IAQI_{L}}_o$——相应地区的空气质量分指数及对应的污染物项目浓度指数表中与$\mathrm{BP_{L}}_o$对应的空气质量分指数。

$$AQI = max\{IAQI_1, IAQI_2, \ldots, IAQI_n\}$$

式中：

n——污染物项目。

可见 AQI 的取值为改点最大污染物的 IAQI 取值。不同点的最大污染物很可能不同，所以其不适合作为实时变化数据模型的简化“一维数值”。故此我们提出另一个数值。

不同污染物的 IAQI 比起浓度值 C 有一个优势是对于每一种污染物它们的取值范围，平均值，最小最大值更一致。所以我们选择 IAQI 而不是 C 来计算。

选择一定范围的样本数值后，令 $x^{(i,j)}$ 表示第 i 时刻第 j 监测点的 $[IAQI_1, IAQI_2, \ldots, IAQI_n]$

令：

$$Sigma = \frac{1}{m \times l}\sum_{i=1}^{m}\sum_{j=1}^{l}(x^{(i,j)})(x^{(i,j)})^T$$

式中：

m——样本的时间范围数量。
l——样本的监测点点数量。

易知 Sigma 为 $n \times n$ 的矩阵，求

$$[U, S, V] = svd(Sigma)$$

式中：

svd——奇异值分解函数（Singular Value Decomposition）

U 也为 $n \times n$ 的矩阵，取 U 的第一个列向量 $u^{(1)} \in R^{n\times 1}$，$x^{(i,j)} \in R^{1\times n}$ 令：

$$z^{(i,j)} = x^{(i,j)} u^{(1)}$$

则 $z^{(i,j)} \in R$ 即为我们所需要的一维数值。

评估

上述的一维数值能在多大程度上代表 n 维的不同污染物，需要实际评估。评估的方法如下：

先将得到的一维数值重建成 n 维向量：

$$x_{rec}^{(i,j)} = z^{(i,j)} u^{(1)T}$$

然后计算

$$V = \frac{\frac{1}{m\times l} \sum_{i=1}^{m} \sum_{j=1}^{l} \left\| x^{(i,j)} - x_{rec}^{(i,j)} \right\|^2}{\frac{1}{m\times l} \sum_{i=1}^{m} \sum_{j=1}^{l} \left\| x^{(i,j)} \right\|^2}$$

即可得到 $z^{(i,j)}$ 能在什么程度上代表 n 维数值的变化。

这个算法同时也可以用在机器学习中提高训练速度，如当网格用 32×32 时，有 1024 个维度，如果运算时间比较慢，可以在满足 V < 0.01 ~ V < 0.1 的前提下，选择一个合适的 k，将 1024 维缩减到 k 维，且保留了 90 ~ 99% 的变化量。以上简述的是 k 取 1 的情况。