2024/17

The difference between row-wise and column-wise preprocessing

Importance of col-wise

In general, If we normalize a given sample's data across its feature (=across column = by rows),

We've tossed out a lot of information.

	height	weight	age
Person 1	180	65	50
2	140	45	50

Consider this data.

If we normalize a given sample's data across column, we'd loss informations like ratio, relationships between each feature.

⇒ P1 will seem to have different age than P2, but it isn't correct information.

Most cases we want to preprocess due to the need for reducing ratio difference between features.

So in general, we'd preprocess across samples rather than features.

	height	weight	age	Number of Cell
Person 1	180	65	50	21 trillion
2	140	45	50	35 trillion

If we analyze samples without preprocessing, it may yield disastrous result due to huge difference of scale between each features.

But in some cases, which have a needs for comparison between samples as a whole, it will be fine to use preprocessing across features.

Norm Normalization

If we want to analyze relationships between samples, we can preprocess each sample via Norm normalization.

Ref

One situation that might call for normalization by rows is when all of the features are of essentially the same type but their values can systematically differ among samples. For example, in RNA sequencing the absolute amount of RNA might differ from sample to sample, resulting in systematically different values among samples (rows) for the observed values of each of the 20,000 or so genes (columns) whose expression is being evaluated. In that type of case it might make sense to take those systematic differences among samples into account via normalizing by row. If the features are fundamentally different in type, however, your sense is correct.

Summary

• Normalization by samples ⇒ Comparison similarity between each samples
•   by features ⇒ Making every feature have same importance