Mohammad Hosseinipour

Think about how range-type variables might be represented. Is it reasonable to represent them as any numeric variables, or is it better to have ad-hoc encoding?

The answer to this question depends on the case and data distribution. It Mostly depends on the effect of the distribution of that feature on the Goal. For example, let's compare two cases:

In the first case, the distribution of the feature is close to a uniform distribution. So there are no specific sub-ranges in which most of the training indexes occur.

In the second case, let us consider an example when the feature is mostly distributed on some finite specific K sub-ranges.

For the second case, we could consider each sub-range a Bin so that we can convert this feature into a K-bin feature. But we also have to consider the Goal, such that each bin can be related to a specific range or class of the Goal.

For the first case, it does depend on the Goal more than the feature's distribution. Consider this example that the feature has a uniform distribution, and the Goal is a binary classification that our feature(range [A, B]) can be represented into some sub-ranges (small finite # of sub-ranges) that each sub-range mainly points to a class of our Goal. So using K bin and one hot vector is reasonable, but if the number of sub-ranges is big, converting into a one-hot vector could increase our overhead and not worth it.

So overall, It depends on the case and data visualization to see if it is worth encoding the feature into a K bin and one hot vector. This task needs lots of experience in DATA MINING to preprocess the data correctly.