Is it essential to do normalization for SVM and Random Forest?

My features' every dimension has different range of value. I want to know if it is essential to normalize this dataset. Thanks

machine-learning

asked Apr 24 '13 at 0:15

user22062

131 1 7

2 Answers

The answer to your question depends on what similarity/distance function you plan to use (in SVMs). If it's simple (unweighted) Euclidean distance, then if you don't normalize your data you are unwittingly giving some features more importance than others.

For example, if your first dimension ranges from 0-10, and second dimension from 0-1, a difference of 1 in the first dimension (just a tenth of the range) contributes as much in the distance computation as two wildly different values in the second dimension (0 and 1). So by doing this, you're exaggerating small differences in the first dimension. You could of course come up with a custom distance function or weight your dimensions by an expert's estimate, but this will lead to a lot of tunable parameters depending on dimensionality of your data. In this case, normalization is an easier path (although not necessarily ideal) because you can at least get started.

Finally, still for SVMs, another thing you can do is come up with a similarity function rather than a

distance function and plug it in as a kernel (technically this function must generate positive-definite matrices). This function can be constructed any way you like and can take into account the disparity in ranges of features.

For random forests on the other hand, since one feature is never compared in magnitude to other features, the ranges don't matter. It's only the range of one feature that is split at each stage.

edited Apr 24 '13 at 1:56

answered Apr 24 '13 at 1:44

Ansari

258 2 7

Random Forest is invariant to monotonic transformations of individual features. Translations or per feature scalings will not change anything for the Random Forest. SVM will probably do better if your features have roughly the same magnitude, unless you know apriori that some feature is much more important than others, in which case it's okay for it to have a larger magnitude.

answered Apr 24 '13 at 1:43

rrenaud

613 3 8
