Questions and Answers for SVMs and Ensembles: Boosting

Bingzhang Zhu

November 15, 2021

1 Where you will use an SVM over a RandomForest Machine Learning algorithm?

The main reason to use an SVM instead is that the problem might not be linearly separable. In that case, we will have to use an SVM with a non-linear kernel (e.g. RBF).

Another related reason to use SVMs is if you are in a higher-dimensional space. For example, SVMs have been reported to work better for text classification.

2 Can we apply the kernel trick to logistic regression? Why is it not used in practice then?

Logistic Regression is computationally more expensive than SVM — O(N^3) vs O(N^2k) where k is the number of support vectors.

The classifier in SVM is designed such that it is defined only in terms of the support vectors, whereas in Logistic Regression, the classifier is defined over all the points and not just the support vectors. This allows SVMs to enjoy some natural speed-ups (in terms of efficient code-writing) that is hard to achieve for Logistic Regression.

3 Is AdaBoost less or more prone to over-fitting?

In practical experience AdaBoost is quite robust to over-fitting, and LPBoost (Linear Programming Boosting) even more so (because the objective function requires a sparse combination of weak learners, which is a form of capacity control). The main factors that influence it are:

The “strength” of the “weak” learners: If you use very simple weak learners, such as decision stumps (1-level decision trees), then the algorithms are much less prone to over-fitting. Whenever you try using more complicated weak learners (such as decision trees or even hyper-planes) You’ll find that over-fitting occurs much more rapidly.

The noise level in the data: AdaBoost is particularly prone to over-fitting on noisy datasets. In this setting the regularized forms (RegBoost, AdaBoostReg, LPBoost, QPBoost) are preferable.

The dimensionality of the data: We know that in general, we experience over-fitting more in high dimensional spaces ("the curse of dimensionality"), and AdaBoost can also suffer in that respect, as it is simply a linear combination of classifiers which themselves suffer from the problem. Whether it is as prone as other classifiers is hard to determine.

Of course we can use heuristic methods such as validation sets or k-fold cross-validation to set the stopping parameter (or other parameters in the different variants) as you would for any other classifier.