GUIDE TO ENSEMBLES OF MODELS (UNIT 8)

COURSE: MACHINE LEARNING I

MASTER IN BIG DATA ANALYTICS

Ricardo Aler Mur

- This lecture shows how models can be improved by making ensembles of basic models.
- There are basically two types of ensembles: bagging and boosting (also Stacking).

- **Bagging** creates ensembles of models by resampling with replacement the original training dataset. This is done many times and a model is trained with each resample.
- The final classifier is the ensemble of all models. Classification is done through majority voting. For regression, the average of the outputs of all models can be computed instead.

- **Boosting** creates models sequentially. Each model in the sequence focuses on the mistakes of the previous model

- There are some variants of Bagging and Boosting when trees are used as base models.

- Random forests is a variant of Bagging, where in addition to using resampling with replacement, different trees in the ensemble are created with different attributes. More specifically, the best attribute for each node in the tree is selected, not from the complete available set of attributes, but from a random set of them. In other words, the decision tree training process is randomized (i.e. a deterministic

algorithm is transformed into a stochastic one, by randomly sampling the attributes to be evaluated at each node). As a consequence, each tree uses a much smaller subset of attributes (compared to the original decision tree algorithm).

- Gradient Boosted Trees is a variant of Boosting, where trees are used as base models. It is shown how each new model is trained by learning the pseudo-residuals (the difference between the ensemble so far and the actual output). It is also shown that Boosting might be prone to overfitting, and care must be taken to avoid it by controlling some of the hyper-parameters.

Associated material: Lecture slides and some exercises are provided.