**Asignments 3**

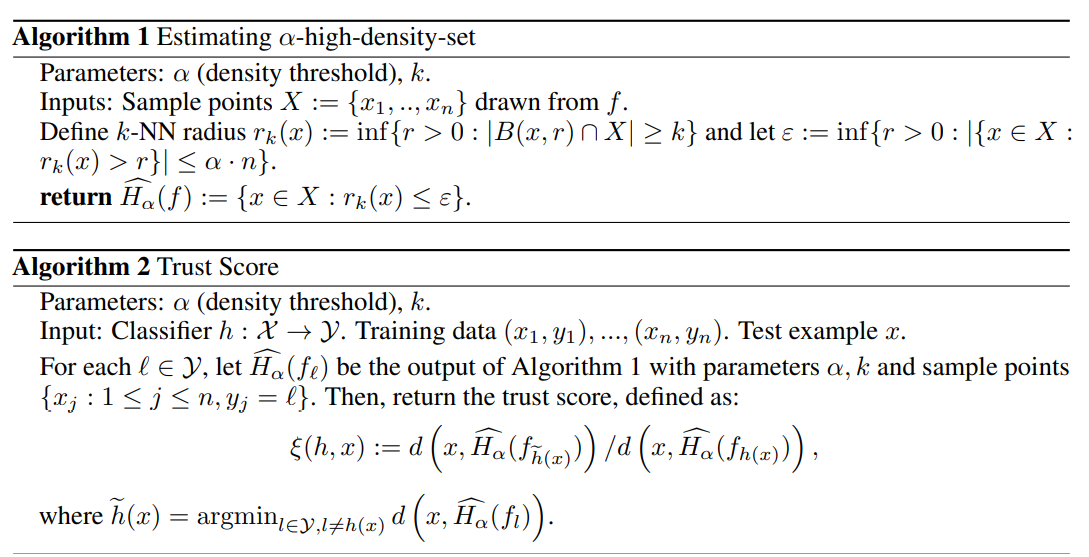
1. quantitative analysis methods

**To Trust Or Not To Trust A Classifier**

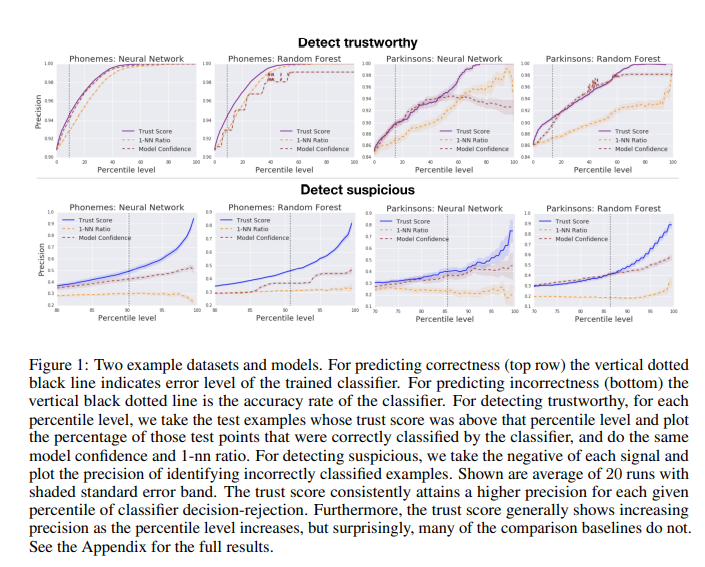
**Heinrich Jiang∗ Google Research heinrichj@google.com Been Kim Google Brain beenkim@google.com Melody Y. Guan† Stanford University** [**mguan@stanford.edu**](mailto:mguan@stanford.edu)

Method : The Trust Score

The approach proceeds in two steps outlined in Algorithm 1 and 2. first pre-process the training data, as described in Algorithm 1, to find the α-high-density-set of each class, which is defined as the training samples within that class after filtering out α-fraction of the samples with lowest density (which may be outliers)



The method has two hyperparameters: k (the number of neighbors, such as in k-NN) and α (fraction of data to filter) to compute the empirical densities. It show in theory that k can lie in a wide range 3 and still give us the desired consistency guarantees. Throughout our experiments, we fix k = 10, and use cross-validation to select α as it is data-dependent. Remark 2. Observed that the procedure was not very sensitive to the choice of k and α. As will be shown in the experimental section, for efficiency on larger datasets, we skipped the initial filtering step of Algorithm 1 (leading to a hyperparameter-free procedure) and obtained reasonable results. This initial filtering step can also be replaced by other strategies. One such example is filtering examples whose labels have high disagreement amongst its neighbors, which is implemented in the open-source code release but not experimented with here.

****

1. qualitative analysis methods

**Julius Adebayo∗ , Justin Gilmer] , Michael Muelly] , Ian Goodfellow] , Moritz Hardt]† , Been Kim] juliusad@mit.edu, {gilmer,muelly,goodfellow,mrtz,beenkim}@google.com ]Google Brain †University of California Berkeley**

Method : The Trust Score