ADA BOOST:

Consider a text classification task, such that the document X can be expressed as a binary

feature vector of the words. More formally X = [X1, X2, X3, ..., Xm], where Xj = 1 if word j

is present in document X, and zero otherwise. Consider using the AdaBoost algorithm with

a simple weak learner, namely

h(X; θ) = yXj

θ = {j, y} j is the word selector ; y is the associated class

y ∈ {−1, 1}

More intuitively, each weak learner is a word associated with a class label. For example

if we had a word football, and classes {sports,non-sports}, then we will have two weak

learners from this word, namely

• Predict sports if document has word football

• Predict non-sports if document has word football.

1. [2 points] How many weak learners are there ?

8

⋆ SOLUTION: Two weak learners for each word, i.e. 2m weak learners.

2. This boosting algorithm can be used for feature selection. We run the algorithm and

select the features in the order in which they were identified by the algorithm.

(a) [4 points] Can this boosting algorithm select the same weak classifier more than

once? Explain.

⋆ SOLUTION: The boosting algorithm optimizes each new α by assuming that

all the previous votes remain fixed. It therefore does not optimize these coefficients

jointly. The only way to correct the votes assigned to a weak learner later on is to

introduce the same weak learner again. Since we only have a discrete set of possible

weak learners here, it also makes sense to talk about selecting the exact same weak

learner again.

(b) [4 points] Consider ranking the features based on their individual mutual information with the class variable y, i.e. ˆI(y; Xj). Will this ranking be more informative

than the ranking returned by AdaBoost ? Explain.

⋆ SOLUTION: The boosting algorithm generates a linear combination of weak

classifiers (here features). The algorithm therefore evaluates each new weak classifier

(feature) relative to a linear prediction based on those already included. The mutual

information criterion considers each feature individually and is therefore unable to

recognize how multiple features might interact to benefit linear prediction.