Ling 572 Reading 1

Daniel Campos dacampos@uw.edu

01/29/2019

1 Q1

Let P(X=i) be the probability of getting an i when rolling a dice (e.g., i=1, 2, . . . , 6). What is the value of P(X=i) with the maximum entropy if the following is true?

(a) P(X=1) + P(X=2) = 1/2:
Since we want to have the most uniform distribution our values for P(X=3)+ P(X=4) +P(X=5) + P(X=6) = 1/2 and they are all equal so P(X=1) = 1/4, P(X=2) = 1/4, P(X=3) = 1/8, P(X=4) = 1/8, P(X=5) = 1/8, P(X=6) = 1/8.

(b) P(X=1) + P(X=2) = 1/2 and P(X=6) = 1/3
Keeping with the logic mentioned above,P(X=3)+ P(X=4) +P(X=5) = 1-(1/2+1/3) = 1-5/6 = 1/6 so P(X=1) = 1/4, P(X=2) = 1/4, P(X=3) = 1/18, P(X=4) = 1/18, P(X=5) = 1/18, P(X=6) = 1/3.

2 Q2 How many feature functions?

If there are C classes and V features then there are —V—*—C— feautre functions since each feature function is a binary-valued function on events. That being said not all features may be used in the final Maxent model.

3 Q3 Similarities and differences

3.1 Modeling

Both methods are similair in the modeling of lingusitic task because they treach each word as a feature but Naive Bayes(NB) assumes feature independence while Maxent can model dependent feautres and assumes that the dataset is best representing by only modeling what is known. NB uses two parameters are the class priors(p(c)) and the conditional probability(P(F—c)) while Maxent parameters are a set of feautre functions of (feature, class)binary pairs. NB has —C— + —V—*—C— parameters(C being classes and V being the vocabulary of the training set, otherwise refered to as the features) while maxent has up to —V—*—C— parameters but will optimze for a set of those that maximize the entropy.

3.2 Training

This is where the methods differ most since NB calculates the class prior which is merely how often the class shows in the training data along with the conditional probability which is count(f, c)/counc(c) where f is feature of F and c is a classs of C. Maxent has a more traditional Training stage since training is done via either GIS or ISS algorithms which optimize feautre usage and weights to represent the data distribution of training. Basically Maxent compares the empiracal expectaions to the model and modify the weights of feature and use of features to optimize for maximum data entropy. Both models are prone to overfitting and poor generalization on test data which can be optimized by use of smoothing mechanism. Additionally, since Maxent has a more traditional training mechanism, it uses various regularizations and early stopping to improve training time and optimize model performance. In short training for Maxent may be able to model dependent features better but is much slower to NB since it may need to look at all possible feature combination and feature weights can be infinite and the iterative trainer can take a long time to reach those values.

3.3 Decoding

Decoding is done in a similair way for both models. Using the features(words) present in the candidate document the probability of all classes is measured and the class with the highest probability is selected. In NB this is done by adding the logprob of all features(1-P of feature if the feature t in V is not in the document) while Maxent uses the logprob of all features present in the document.