1 Margin-based Text Classification

1.1 Margin-based Text Classification

generatic classifiers try to learn how the data is ‘generated’. It uses the distribution of the seen data to predict unseen data. For example the Naives-Bayes Network uses $P(Y|X) = \frac{P(X|Y)*P(Y)}{P(X)}$. It takes the propabilities of the learned data to calculate the propability of the new data.
Discriminative Classifiers just try to compare new data with the seen data. It makes way less assumptions and therefor relies on good data(quantity and representative).
An example would be SVM which plots the data and tries to calculate a function that seperates the data as good as possible into the different classes. With to classes the SVM would calculate the border f and than predicts $P(Y|X)$ with the help with a look at the distance of X to the Hyperplane f. a negatic distance means the first class positiv the second. The hyperplane of a linear svm can be written as for points x which satisfy the equation $< w * x + b = 0 >$ where w is a normal vector to the hyperplane and b the bias.

1.2 Language Identification with SVM

The C parameter tells the SVM how much you want to avoid missclassification. The SVM tries to find the Hyperplane with the maxiumum possible margin satisfying C. The gamma parameter sets the radius of the influence of each trained feature point. If gamma is to big only exact matches can lead to classification. In the other case the svm can not capture the complexity of the data.
Like you can see in the table RBF gets a little bit better for trigrams and the linear kernel is a little bit worse. This is caused by the fact that the number of possible trigrams is bigger than that of the bigrams, which lead to a more complex data distribution. RBF should be used for complex data (for example bigger n-grams where a linear kernel will fail to find a suitable hyperplane). linear kernel should be used for a linear problem since it is faster and in this case has a better accuracy.b

kernel	bigram	trigram
linear	76%	74.66%
RBF	63.66%	64.33%