Machine Learning Nanodegree Capstone Project

Spam Detection using Naive Bayes and Support Vector Machines

By

Sylvester Ranjith Francis

1.Introduction:

With the rapid growth of online information, text categorization has become one of the key techniques for handling and organizing text data. Text categorization techniques are used to classify news stories, to find interesting information on the WWW, and to guide a user's search through hypertext. Since building text classifiers by hand is difficult and time-consuming, it is advantageous to learn classifiers from examples. In this project we try to build such a Text classifier using Naive Bayes and Support Vector Machines to classify the incoming text messages as either Spam/Not Spam (Ham). The below figure explains the working of the Text Classifier.

A basic workflow of this project is as follows

- Step 1:Initial exploration of the dataset is performed
- Step 2:Once the data exploration was performed,We proceed to convert the raw messages into vectors.
- Step 3:The mapping was not 1 to 1 hence the **Bag of Words** approach is used where each unique word in a text is represented by a number
- Step 4: Converting the text into a vector involves three steps namely,Counting the frequency of a word occurring in each message (term frequency),Weighting the counts, so that frequent tokens get lower weight (inverse document frequency),Normalizing the vectors to unit length, to abstract from the original text length (L2 norm)
- Reference : http://scikit-learn.org/stable/modules/feature_extraction.html#the-bag-of-words-representation
- Step 5: The model has to trained using a naive bayes classifier to check if our model works
- Step 6: Split the data into testing and training data using Sklearn train_test_split

- Step 7:The testing data is fed into the designed Naive bayes classifier,to find out that the prediction engine gave an accuracy of 95%
- Step 8: To improve the performance of the prediction engine,I tried designing a Support Vector Machine after many different trials and errors to improve the accuracy
- Step 9: The resulting Support Vector Machine gave an accuracy of 98% which is a huge improvement compared to the Naive bayes classifier
- The result being that we were able to classify the messages successfully as spam and ham with an accuracy of 98%

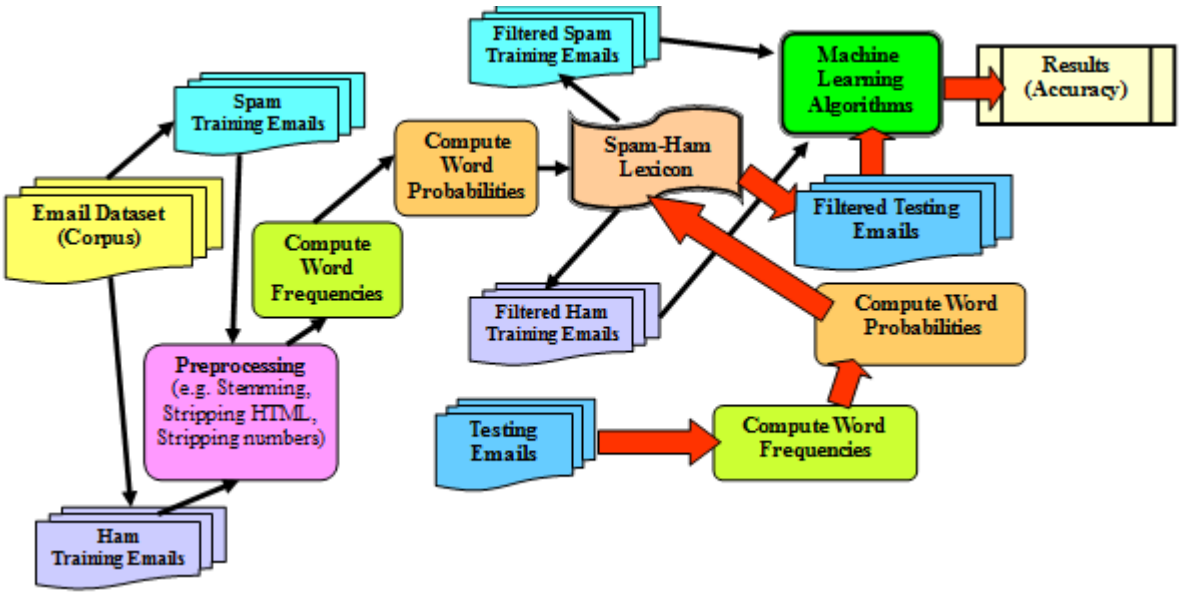


Fig 1 :Basic block diagram.

This block diagram will be explained in detail further as we proceed to the workings of the prediction engine.

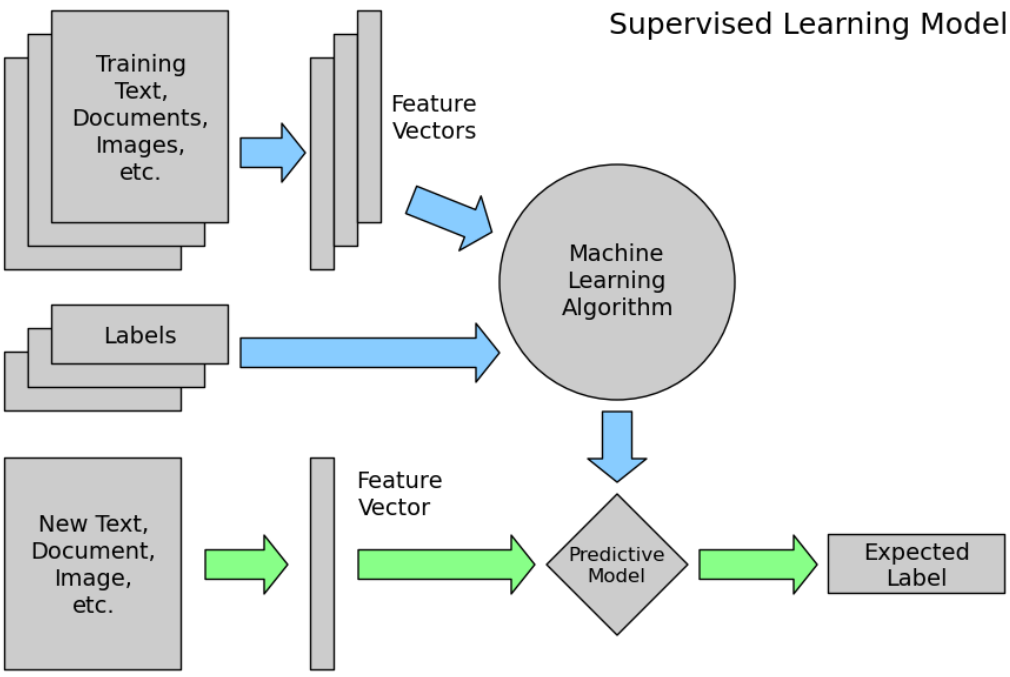
2.Dataset:

In this study, we have used the publicly available text corpus dataset of SMS spam messages found in the UCI Machine Learning Repository There are 5574 total messages in this dataset where 4459 is categorized for training and remaining 1115 messages is categorized for test phases. Figure 2 shows the first 10 messages of the text corpus dataset

```
In [11]: #printing the first 10 messages of the SMS CORPUS
         for message_no, message in enumerate(messages[:10]):
             print message_no, message

0 ham   Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wa
t...
1 ham   Ok lar... Joking wif u oni...
2 spam  Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry questio
n(std txt rate)T&C's apply 08452810075over18's
3 ham   U dun say so early hor... U c already then say...
4 ham   Nah I don't think he goes to usf, he lives around here though
5 spam  FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb
 ok! XxX std chgs to send, £1.50 to rcv
6 ham   Even my brother is not like to speak with me. They treat me like aids patent.
7 ham   As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for al
l Callers. Press *9 to copy your friends Callertune
8 spam  WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 090
61701461. Claim code KL341. Valid 12 hours only.
9 spam  Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free!
 Call The Mobile Update Co FREE on 08002986030
```

We can observe that these are Tab Separated Values where the first column is a label saying whether the given message is a normal message ("ham") or "spam". The second column is the message itself. This corpus will be our labeled training set. Using these ham/spam examples, we'll train a machine learning model to learn to discriminate between ham/spam automatically. Then, with a trained model, we'll be able to classify arbitrary unlabeled messages as ham or spam.



2.1 Data Exploration:

In this section we try to understand the ways of classifying the data on different parameters to classify the data better.By understanding if there are frequent patterns in the different messages such as “Length of a message”,”Addressed to”, “Sender info”,“Content of a message”. We try to correlate these different parameters to see if

there are any specific patterns in a message which may lead it to be classified under the label of spam.

```
In [13]: #Aggregating Statistics
         messages.groupby('label').describe()
```

Out[13]:

		message
label		
ham	count	4827
	unique	4518
	top	Sorry, I'll call later
	freq	30
spam	count	747
	unique	653
	top	Please call our customer service representativ...
	freq	4

By grouping the data by the label,We observe that there are 4827 messages which are classified as “Ham” and number of messages that were unique are 4518.The most repeated messages were with the text “Sorry,I’ll call later”.Similarly there are 747 “Spam” messages out of which 653 of them were unique and the frequently repeated text was “Please call our customer service representative...”. We don't necessarily infer much by classifying them by labels and describing it.

Next we try to see if we can infer anything about classifying the messages based on the length of the text

```
In [14]: #length of messages
         messages['length'] = messages['message'].map(lambda text: len(text))
         print messages.head()
```

```
  label                                            message  length
0   ham  Go until jurong point, crazy.. Available only ...     111
1   ham                      Ok lar... Joking wif u oni...      29
2  spam  Free entry in 2 a wkly comp to win FA Cup fina...     155
3   ham  U dun say so early hor... U c already then say...      49
4   ham  Nah I don't think he goes to usf, he lives aro...      61
```

```
In [16]: messages.length.describe()
```

```
Out[16]: count    5574.000000
         mean       80.604593
         std        59.919970
         min         2.000000
         25%        36.000000
         50%        62.000000
         75%       122.000000
         max       910.000000
         Name: length, dtype: float64
```

The longest message is 910 characters long and it is found that is a ham message and not spam. Representing them graphically in figure 2.1. We infer that spam messages generally have a lesser character count than the ham messages.

```
In [18]: #difference in length between Spam and Ham messages
         messages.hist(column='length', by='label', bins=50)
```

```
Out[18]: array([<matplotlib.axes._subplots.AxesSubplot object at 0x10a335690>,
                <matplotlib.axes._subplots.AxesSubplot object at 0x1173fc1d0>], dtype=object)
```

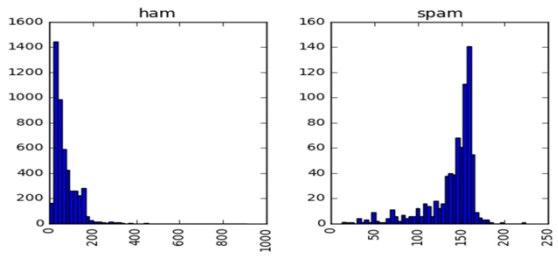


Fig 2.1 Ham vs Spam (message length)

So through this initial data exploration we find out that “length of the messages” is a factor by which we can classify the messages.

In section 3 let us get into detail about the working of the classifier and how and why two different approaches were used to implement the classifier

3. Implementation :

There are two ways that the Ham/Spam classifier has been implemented in this project namely

1. Naive bayes classification
2. Using Support Vector Machines

We shall discuss in detail about these two approaches in this section

3.1 Data Preprocessing:

Before proceeding with either approach preprocessing of the data is necessary.
Initially the raw messages which is a sequence of characters is converted into vectors by utilising a bag of words approach,since the mapping is not one to one.The bag of words approach assigns a different number to each unique word in the sentence.We will convert each message, represented as a list of tokens into a vector that machine learning models can understand. It involves three steps in the bag of words approach

1.Counting the frequency of a word occurring in each message (term frequency)

2.Weighting the counts, so that frequent tokens get lower weight (inverse document frequency)

3.Normalizing the vectors to unit length, to abstract from the original text length (L2 norm)

Each vector has as many dimensions as there are unique words in the SMS corpus

Text Analysis is a major application field for machine learning algorithms. However the raw data, a sequence of symbols cannot be fed directly to the algorithms themselves as most of them expect numerical feature vectors with a fixed size rather

than the raw text documents with variable length. In order to address this, scikit-learn provides utilities for the most common ways to extract numerical features from text content, namely: tokenizing strings and giving an integer id for each possible token, for instance by using white-spaces and punctuation as token separators. counting the occurrences of tokens in each document. normalizing and weighting with diminishing importance tokens that occur in the majority of samples / documents. In this scheme, features and samples are defined as follows: each individual token occurrence frequency (normalized or not) is treated as a feature. the vector of all the token frequencies for a given document is considered a multivariate sample. A corpus of documents can thus be represented by a matrix with one row per document and one column per token (e.g. word) occurring in the corpus. We call vectorization the general process of turning a collection of text documents into numerical feature vectors. This specific strategy (tokenization, counting and normalization) is called the Bag of Words or “Bag of n-grams” representation. Documents are described by word occurrences while completely ignoring the relative position information of the words in the document.

3.1.1 Bag of Words:

The **bag-of-words** model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the **bag** (multiset) of its **words**, disregarding grammar and even **word** order but keeping multiplicity.

Bag of Words (BoW) is an algorithm that counts how many times a word appears in a document. Those word counts allow us to compare documents and gauge their similarities for applications like search, document classification and topic modeling. BoW is a method for preparing text for input in a deep-learning net.

BoW lists words with their word counts per document. In the table where the words and documents effectively become vectors are stored, each row is a word, each column is a document and each cell is a word count. Each of the documents in the corpus is represented by columns of equal length. Those are word count vectors, an output stripped of context.

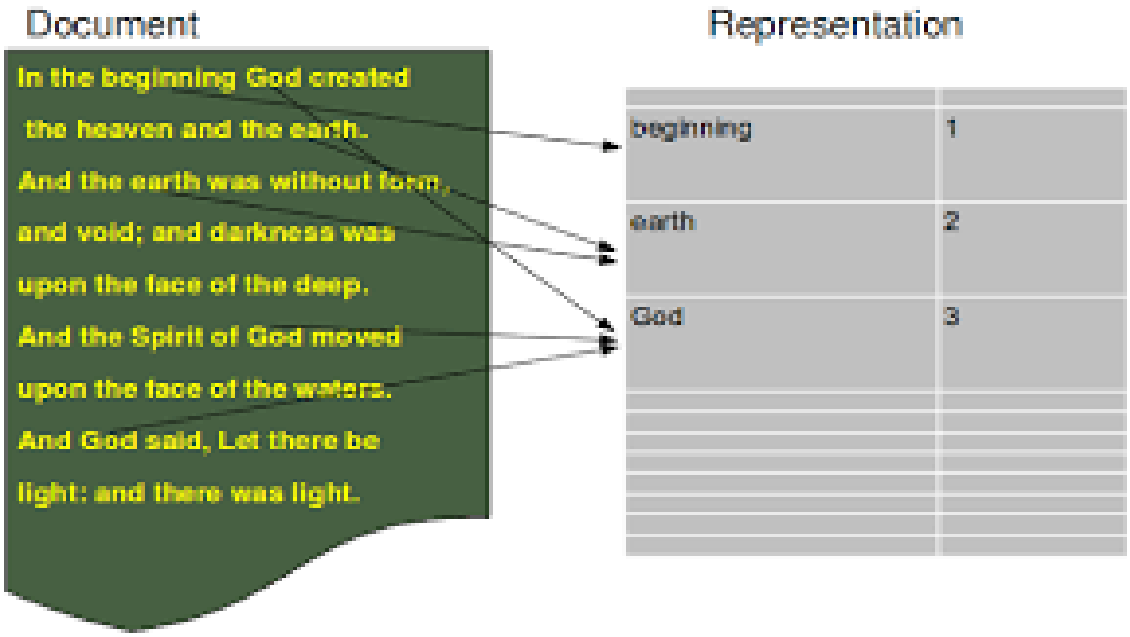


Fig 3.1 Bag of Words - Representation example

After counting, the term weighting and normalization can be done with TF-IDF, using scikit-learn's TfidfTransformer

3.1.2 TF-IDF Algorithm:

Term Frequency - Inverse Document Frequency is a method to judge the topic of an article by the words it contains, With this method words are given weight – TF-IDF measures relevance, not frequency. That is, word counts are replaced with TF-IDF scores across the whole dataset.

First, TF-IDF measures the number of times that words appear in a given document But because words such as “and” or “the” appear frequently in all documents, those are systematically discounted. That’s the inverse-document frequency part. The more documents a word appears in, the less valuable that word is as a signal. That’s intended to leave only the frequent AND distinctive words as markers. Each word’s TF-IDF relevance is a normalized data format that also adds up to one.

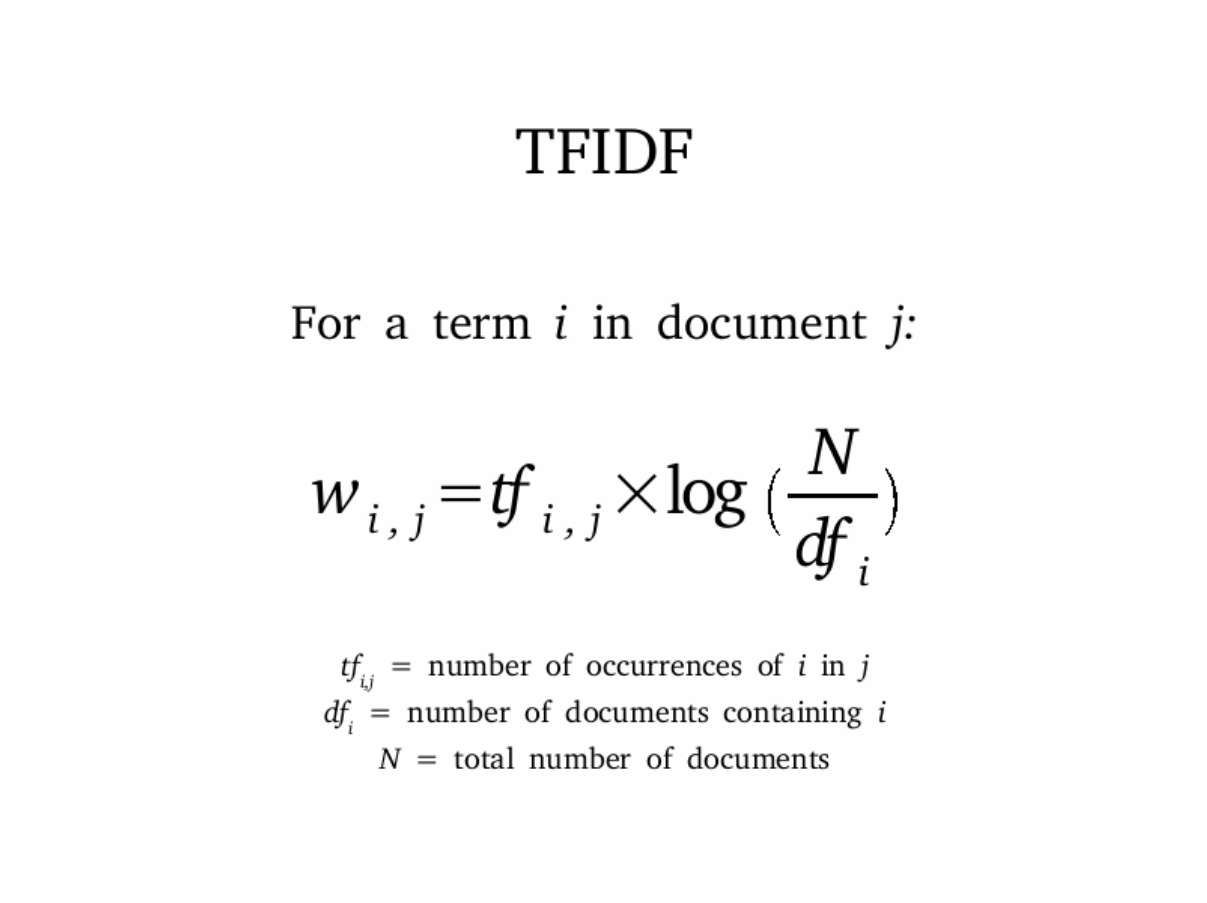


Fig 3.1.2 TFIDF

Now we can proceed into the explanation of the Naive Bayes approach in the implementation of the classifier

3.2 Naive Bayes Approach:

Naive Bayes is a collection of classification algorithms based on Bayes Theorem. It is not a single algorithm but a family of algorithms that all share a common principle, that every feature being classified is independent of the value of any other feature.

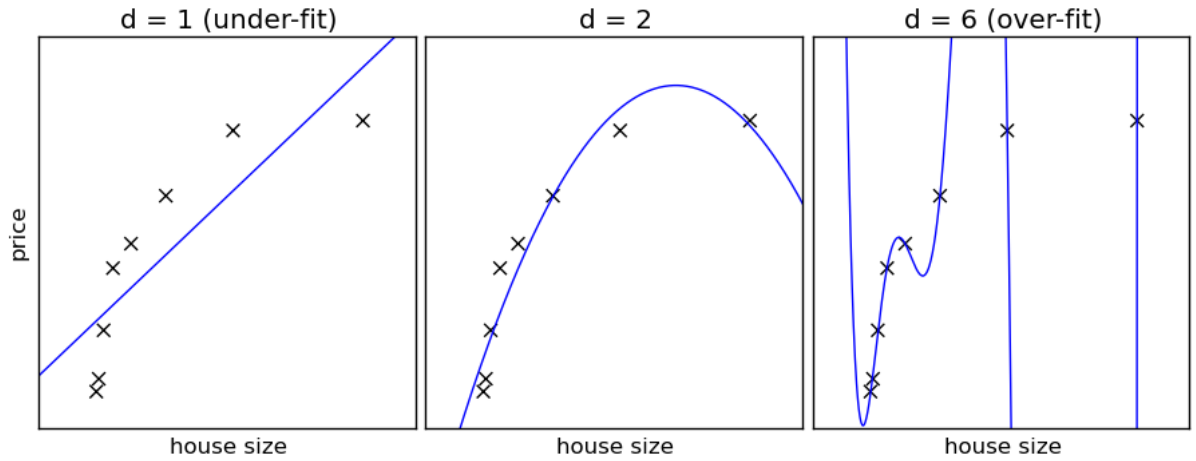
MultinomialNB implements the naive Bayes algorithm for multinomial distributed data, and is one of the two classic naive Bayes variants used in text classification (where the data are typically represented as word vector counts, although tf-idf vectors are also known to work well in practice). The distribution is parametrized by

vectors $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ for each class y, where n is the number of features (in text classification, the size of the vocabulary) and θ_{yi} is the probability $P(x_i \mid y)$ of feature i appearing in a sample belonging to class y.

Naive Bayes is an example of a high bias - low variance classifier (aka simple and stable, not prone to overfitting). An example from the opposite side of the spectrum would be Nearest Neighbour (kNN) classifiers, or Decision Trees, with their low bias but high variance (easy to overfit). Bagging (Random Forests) as a way to lower variance, by training many (high-variance) models and averaging.

High bias = classifier: is opinionated. Not as much room to change its mind with data, it has its own ideas. On the other hand, not as much room it can fool itself into overfitting either (picture on the left).

Low bias = classifier: more obedient, but also more neurotic. Will do exactly what you ask it to do, which, as everybody knows, can be a real nuisance (picture on the right).



Advantages of Naive Bayes Classifier:

- It is very simple to use,If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data
- In the case of the SMS classifier this model gives an accuracy of 95% which is a good result

- It’s not sensitive to irrelevant features

Disadvantages of Naive Bayes Classifier:

- It assumes independence of all features
- We cannot decide with utilising one algorithm that it returns the best result,The accuracy may change with another algorithm. This is the reason we utilise another approach to implement the SMS classifier

In [39]:
```
pipeline = Pipeline([
    ('bow', CountVectorizer(analyzer=split_into_lemmas)),  # strings to token integer counts
    ('tfidf', TfidfTransformer()),  # integer counts to weighted TF-IDF scores
    ('classifier', MultinomialNB()),  # train on TF-IDF vectors w/ Naive Bayes classifier
])
```

In [40]:
```
scores = cross_val_score(pipeline,  # steps to convert raw messages into models
                         msg_train,  # training data
                         label_train,  # training labels
                         cv=10,  # split data randomly into 10 parts: 9 for training, 1 for scoring
                         scoring='accuracy',  # which scoring metric?
                         n_jobs=-1,  # -1 = use all cores = faster
                         )
print scores
```

```
[ 0.93736018  0.95302013  0.94630872  0.9573991   0.95515695  0.96412556
  0.96404494  0.9505618   0.93258427  0.95505618]
```

In [41]:
```
print scores.mean()
print scores.std()
```

```
0.951561784339
0.00982568625738
```

Thus we get an accuracy of 95% using the naive bayes classifier

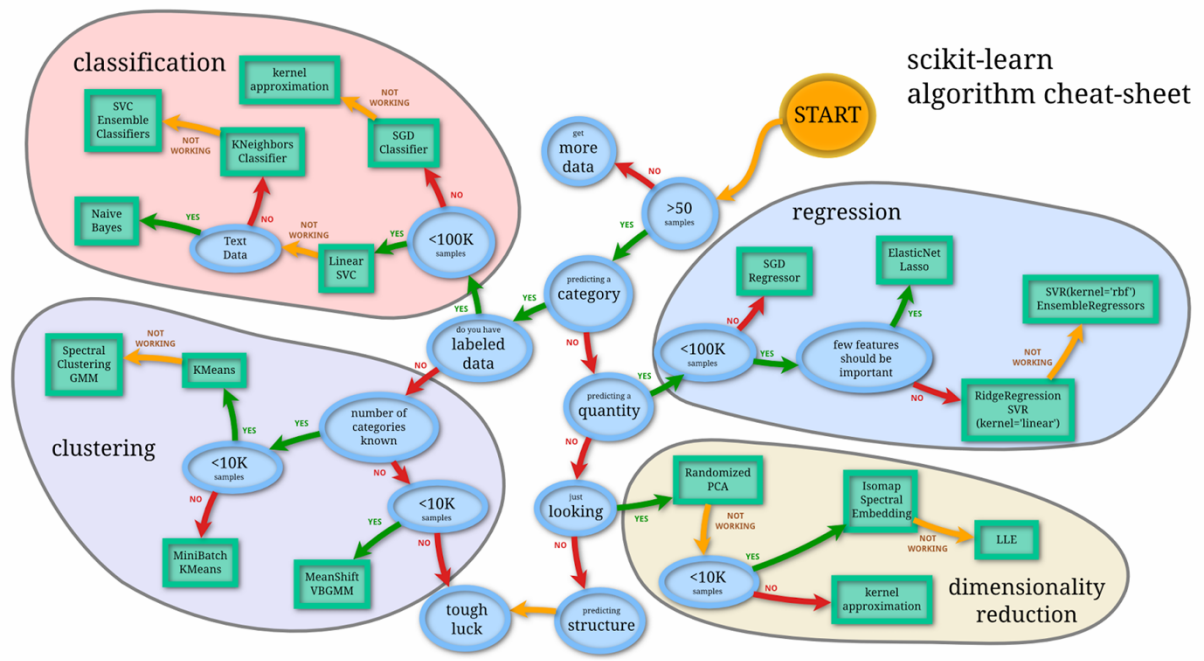


Fig 3.2 Sklearn Algorithm Cheat-Sheet

Let us discuss the second approach that was used to implement this project. One may think, Does IDF weighting have an effect on accuracy? Do we require extra processing cost of lemmatization (vs. just plain words)?

3.3 Support Vector Machine Approach:

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges.it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.

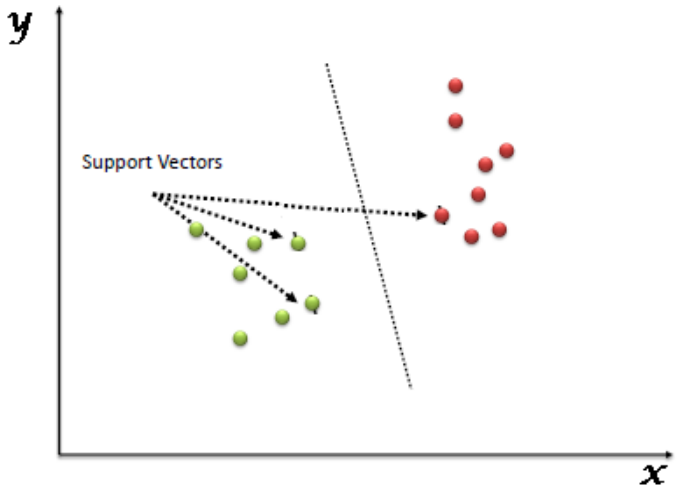


Fig 3.3 Support Vector Machines - SVM

Advantages of Support Vector Machines:

- It works really well with clear margin of separation
- It is effective in high dimensional spaces.
- It is effective in cases where number of dimensions is greater than the number of samples.
- It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- It gives an accuracy of 98% in our Ham/Spam classifier which is higher than an accuracy of 95% while using Naive Bayes.

Disadvantages of Support Vector Machines:

- It doesn't perform well, when we have large data set because the required training time is higher
- It also doesn't perform very well, when the data set has more noise i.e. target classes are overlapping

- SVM doesn't directly provide probability estimates, these are calculated using an expensive five-fold cross-validation. It is related SVC method of Python scikit-learn library.

```
In [88]: pipeline_svm = Pipeline([
             ('bow', CountVectorizer(analyzer=split_into_lemmas)),
             ('tfidf', TfidfTransformer()),
             ('classifier', SVC()),  # <== change here
         ])

         # pipeline parameters to automatically explore and tune
         param_svm = [
           {'classifier__C': [1, 10, 100, 1000], 'classifier__kernel': ['linear']},
           {'classifier__C': [1, 10, 100, 1000], 'classifier__gamma': [0.001, 0.0001], 'classifier__kernel': ['rbf']},
         ]

         grid_svm = GridSearchCV(
             pipeline_svm,  # pipeline from above
             param_grid=param_svm,  # parameters to tune via cross validation
             refit=True,  # fit using all data, on the best detected classifier
             n_jobs=-1,  # number of cores to use for parallelization; -1 for "all cores"
             scoring='accuracy',  # what score are we optimizing?
             cv=StratifiedKFold(label_train, n_folds=5),  # what type of cross validation to use
         )

In [89]: %time svm_detector = grid_svm.fit(msg_train, label_train) # find the best combination from param_svm
         print svm_detector.grid_scores_
```

```
In [90]: print svm_detector.predict(["Hi mom, how are you?"])[0]
         print svm_detector.predict(["WINNER! Credit for free!"])[0]
         print svm_detector.predict(["Hey!!! you won 25000$ on a lottery click link to claim your reward"])[0]

         ham
         spam
         spam

In [91]: print confusion_matrix(label_test, svm_detector.predict(msg_test))
         print classification_report(label_test, svm_detector.predict(msg_test))

         [[975   2]
          [ 13 125]]
                      precision    recall  f1-score   support

                 ham       0.99      1.00      0.99       977
                spam       0.98      0.91      0.94       138

         avg / total       0.99      0.99      0.99      1115
```

Thus we get an accuracy of 98% using Support Vector Machines.

4. Performance Metrics:

To test the performance of our model we utilise the following performance metrics:

Precision: The precision is the ratio `tp / (tp + fp)` where `tp` is the number of true positives and `fp` the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

Recall: The recall is the ratio `tp / (tp + fn)` where `tp` is the number of true positives and `fn` the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

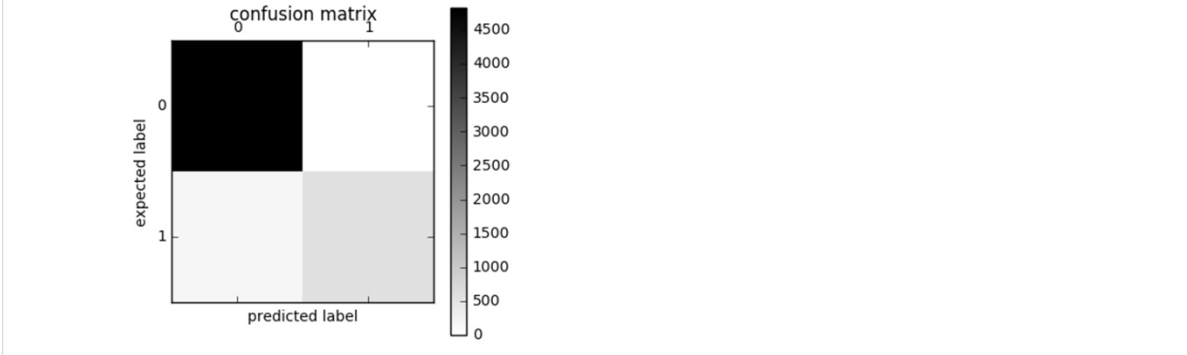
F1-Score: The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.

The F-beta score weights recall more than precision by a factor of `beta`. `beta == 1.0` means recall and precision are equally important.

Support: The support is the number of occurrences of each class in `y_true`.

Confusion Matrix:A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

```
In [75]: plt.matshow(confusion_matrix(messages['label'], all_predictions), cmap=plt.cm.binary, interpolation='nearest')
         plt.title('confusion matrix')
         plt.colorbar()
         plt.ylabel('expected label')
         plt.xlabel('predicted label')
Out[75]: <matplotlib.text.Text at 0x121ee74d0>
```



In [76]:
```
#From this confusion matrix, we can compute precision and recall, or their combination (harmonic mean) F1:
print classification_report(messages['label'], all_predictions)
```

```
             precision    recall  f1-score   support

        ham       0.97      1.00      0.98      4827
       spam       1.00      0.77      0.87       747

avg / total       0.97      0.97      0.97      5574
```

Cross Validation: In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data. **K-fold cross validation** is one way to improve over the holdout method. The data set is divided into *k* subsets, and the holdout method is repeated *k* times. Each time, one of the *k* subsets is used as the test set and the other *k-1* subsets are put together to form a training set. Then the average error across all *k* trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set *k-1* times. The variance of the resulting estimate is reduced as *k* is increased. The disadvantage of this method is that the training algorithm has to be rerun from scratch *k* times, which means it takes *k* times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set *k* different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over.

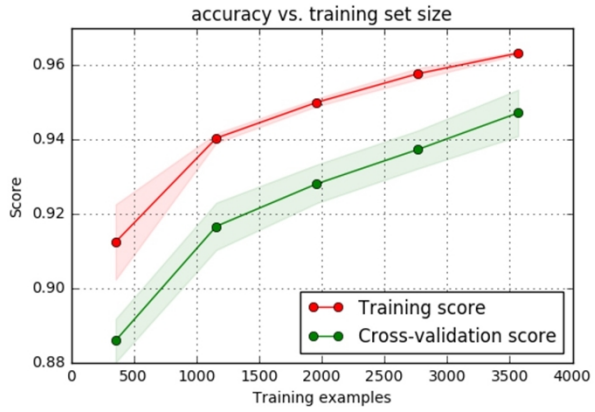
In [79]:
```
scores = cross_val_score(pipeline,  # steps to convert raw messages into models
                         msg_train,  # training data
                         label_train,  # training labels
                         cv=10,  # split data randomly into 10 parts: 9 for training, 1 for scoring
                         scoring='accuracy',  # which scoring metric?
                         n_jobs=-1,  # -1 = use all cores = faster
                         )
print scores
```

```
[ 0.94170404  0.9573991   0.96412556  0.9529148   0.94170404  0.94618834
  0.9529148   0.94843049  0.94394619  0.94831461]
```

In [80]:
```
print scores.mean()
print scores.std()
```

```
0.9497641961
0.00681792912038
```

```
Out[82]: <module 'matplotlib.pyplot' from '/Users/sylvester/anaconda/lib/python2.7/site-packages/matplotlib/pyplot.pyc'>
```



(We're effectively training on 64% of all available data: we reserved 20% for the test set above, and the 5-fold cross validation reserves another 20% for validation sets => 0.8*0.8*5574=3567 training examples left.) Since performance keeps growing, both for training and cross validation scores, we see our model is not complex/flexible enough to capture all nuance, given little data. In this particular case, it's not very pronounced, since the accuracies are high anyway. At this point, we have two options:

- Use more training data, to overcome low model complexity
- Use a more complex (lower bias) model to start with, to get more out of the existing data

Over the last years, as massive training data collections become more available, and as machines get faster, approach 1. is becoming more and more popular (simpler algorithms, more data). Straightforward algorithms, such as Naive Bayes, also have the added benefit of being easier to interpret (compared to some more complex, black-box models, like neural networks).

5.Results:

In the end we have designed a Ham/Spam Classifier using two different approaches

1.**Naive Bayes Classifier with TF-IDF** which results in an accuracy of 95%

2.**Support Vector Machine** which results in an accuracy of 98%

References:

Text Feature Extraction- Bag of Words Representation:
http://scikit-learn.org/stable/modules/feature_extraction.html#the-bag-of-words-representation
http://scikit-learn.org/stable/auto_examples/model_selection/grid_search_text_feature_extraction.html
http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html
Naive Bayes - Sklearn Documentation:
http://scikit-learn.org/stable/modules/naive_bayes.html
Support Vector Machines -Sklearn Documentation:
http://scikit-learn.org/stable/modules/svm.html
http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
Pipelines-Sklearn Documentation:
http://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html
Cross validation-Sklearn Documentation:
http://scikit-learn.org/stable/modules/cross_validation.html
http://scikit-learn.org/0.17/modules/generated/sklearn.grid_search.GridSearchCV.html
http://scikit-learn.org/0.17/modules/generated/sklearn.grid_search.GridSearchCV.html
Model Selection-Sklearn Documentation:
http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html#sklearn.model_selection.StratifiedKFold
http://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html
Accuracy score/Metrics-Sklearn Documentation:
http://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html#sklearn.metrics.accuracy_score

Decision Tree Classifier-Sklearn Documentation:
http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
Other References:
https://www.thinkful.com/projects/building-a-text-classifier-using-naive-bayes-499/
http://blog.fliptop.com/blog/2015/03/02/bias-variance-and-overfitting-machine-learning-overview/
https://deeplearning4j.org/bagofwords-tf-idf
http://blog.aylien.com/naive-bayes-for-dummies-a-simple-explanation/
https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code/
http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
https://www.cs.cmu.edu/~schneide/tut5/node42.html
http://zacstewart.com/2015/04/28/document-classification-with-scikit-learn.html

Affirmation:

I hereby confirm that this submission is my work. I have cited above the origins of any parts of the submission that were taken from Websites, books, forums, blog posts, github repositories, etc.

Sylvester Ranjith Francis