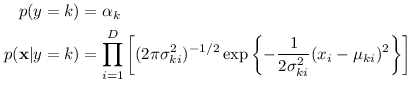
**PROBLEM STATEMENT**

Write a spam filter using discriminative and [generative classifiers](https://moodle.unive.it/mod/resource/view.php?id=160009). Use the [Spambase dataset](https://archive.ics.uci.edu/ml/datasets/spambase) which already represents spam/ham messages through a bag-of-words representations through a dictionary of 48 highly discriminative words and 6 characters. The first 54 features correspond to word/symbols frequencies; ignore features 55-57; feature 58 is the class label (1 spam/0 ham).

Perform SVM classification using linear, polynomial of degree 2, and RBF kernels over the TF/IDF representation.   
Can you transform the kernels to make use of angular information only (*i.e.*, no length)? Are they still positive definite kernels?

Classify the same data also through a Naive Bayes classifier for continuous inputs, modelling each feature with a Gaussian distribution, resulting in the following model:  
  
where α*k* is the frequency of class *k*, and *μki*, *σ*2*ki* are the means and variances of feature *i* given that the data is in class *k*.

Provide the code, the models on the training set, and the respective performances in 10 way cross validation.

Explain the differences between the two models

**SOLUTION**

Machine learning is subset of Artificial intelligence which aims to make systems able to learn automatically. It involves observing data instances and predicting models that will help take decisions in future. There are three categories in Machine learning.

* **Supervised learning**

In Supervised learning, data labels are given to learn the matchine explicitly.

* **Unsupervised learning:**

In unsupervsed learning, data is not give with labels but the matchine has to understand data and classify data with pattens/ structures.

* **Reinforcemnt learning**

It is reward based learning in which machine learns from positives and negatives. It learns how to act in certain environments.

Supervised learning is further divided into two categories i.e. Discriminative and Generative classification.

Discriminative model learns to build the decision boundary between the classes by using conditional probability distribution p(y/x). whereas a Generative Model ‌explicitly learns the distribution of individual class by using joint probability distribution p(x,y). Both approaches predicting conditional probabilities but the way the learn or classify is different. Some of the Examples of both the approaches is given below:

**Generative classifiers**

## Naïve Bayes

* Bayesian networks
* Markov random fields
* ‌Hidden Markov Models (HMM)

**Discriminative Classifiers**

* ‌Logistic regression
* Scalar Vector Machine
* ‌Traditional neural networks
* ‌Nearest neighbor
* Conditional Random Fields (CRF)s

**Application of ML In spam filtration**

In today’s era spam message are annoying as well as dangerous for the internet users. Spam contain virus, Trojan horses or other harmful content which is stressful. So, many machine learning techniques started working on automatic filtration procedures to give internet user relief from spam. The possible techniques for spam filtration are given bellow:

* **Content Based Filtering Technique:**

This technique filters emails by analysizg words, their occurenance in documents and measuring the distribution. Using machine learning approaches, Naïve Bayesian classification, Support Vector Machine, K Nearest Neighbor, Neural Networks.

* **Case Base Spam Filtering Method:**

In this technique all emails both spam and ham emails are collected by using collection model. Data is procedded and then classified into two vector sets. Then machine learning techniques are used to train datasets and then validates on test data. Which is then used to decide in future which is spam and which is ham.

* **Heuristic or Rule Based Spam Filtering Technique:**

This approach used the heuristics to recognise the patterns and classify messages into ham and spam.

* **Previous Likeness Based Spam Filtering Technique:**

This approach uses memory-based, or instance-based, machine learning methods to classify emails based to their resemblance to stored examples (e.g. training emails). k-nearest neighbor is used for such purpose.

**APPROACH WE ARE CONSIDERING**

In this assignment we are particularly working on content-based filtration using SVM and Naïve Bayes approach. We built a model that classifies between spam and ham. we are taking a feature vector formulated by bag of word model (finds term frequency) based on TF/IDF weights. TF-IDF stands for Term Frequency-Inverse Document Frequency. In addition to Term Frequency (tf) we compute Inverse document frequency (idf).

**Support vector machine (SVM), a Discriminative classifier**

SVM is used in classification and regression problems. This approach classifies different classes by finding hyperplane between them. When it comes to segregate classes by drawing hyperplanes there exit many possible hyperplanes, but we need to choose the best hyperplane, by best we mean that the one that classify classes better.

SVM builds hyperplane by maximizing the distance between the nearest data points. The distance is called margin. The hyperplane margin is kept higher, the reason for this is to avoid possible misclassification. In case of outliers SVM ignore outlier and find plane that has maximum margin. So, one can say that it provides mechanism that is robust to outliers. Moreover, the best part of SVM is that, if classes are not linearly separable, then we can still classify them by adding new features. These additional features are added by kernel trick. The SVM kernel is function that takes low dimensional input space and transfer it to higher dimensional space thus converting nonlinear separable problems to separable ones. This way it can help to segregate that patterns which are not linearly separable.

In our assignment we are going to apply SVM over our dataset (spambase) that will classify spam and ham. We will be using linear, polynomial of base 2 and RBF kernel. For that purpose, I am using Python library sklearn because it is widely used library for such problem plus it also provides functions, for applying different kernels and implementation of cross validations, which is requirement of this assignment.

First of all, I need feature vector and label vector. My feature vector will be TFIDF weights and label vector will binary vectors which contain 0s and 1s. 1 meaning the spam and 0 meaning ham. After having my vectors, I converted into numpy array and labeled them. The next thing is creating a classifier which in our case is SVM. This is done by SVC function which is provided by sklearn library I also specify the kernel here. Linear, polynomial degree 2 and RBF.

As a requirement of assignment, I had to normalize the feature vector and use angular linear kernels because TFIDF is highly affected by length of document and occurrence of world and the kernel is also affected by length of vectors. So, I build new kernels that normalizes the vectors. I have got the results of two kernels over normalized and unnormalized data.

And then finally fitting the features to the labels. Now we are ready to predict the model function. Getting a hyperplane means getting coefficients. Coefficients are basically which can used to draw hyperplane. To get best hyperplane coefficients, I used cross validation. Finally testing model over test data .

**NAÏVE BASED APPROACH**

Naïve Bayes theorem is also a classification technique which is based on Bayes theorem. It assumes that the presence of one feature is independent of another feature. Meaning if “buy” occurs in spam messages that “credit” is independent. Every feature contributes to probability of class independently.

For implementation we need to Input a file(spambase) and feature and label vector. Then dividing the data into training and testing parts.

Naive Bayes comprised of summary of data in training set that each attribute by class value so, we can summarize dataset by summarizing features by class. we can study statistics for each class by creating map of each class to the list of instances belong to class and sorting dataset of instances into appropriate list.

Next, we need to calculate the mean of each attribute for class value because we use it as middle number in Gaussian distribution when calculating the probabilities. Moreover, we need standard deviation because it’s the second parameter of Gaussian distribution. We are ready to have summary of attributes and making predictions. We can do so by first pulling training dataset into instances grouped by class then calculating summaries for each attribute.

Now we are ready to make predictions using the summaries created by training data making predictions involved calculating the probability. The given data instance belongs to each class then selecting a class with the largest probability as the prediction.

We can divide this whole method into 4 tasks calculating Gaussian probability density function calculating class probability making prediction and then estimating the accuracy.

To calculate Gaussian probability density function. we use Gaussian function to estimate probability of given attribute value given the mean and standard deviation of the attribute estimated from the training data.

Next task is to calculate class probability. we can calculate the attribute belonging to the class. For that we combine probabilities of all attributes for a data instance and come up with the probability of the entire data instance belonging to the class. we can make predictions now. we can calculate the probability of instance belonging to each class value and we can look for largest probability and return the associated class. for that purpose, we use a function that will summaries and input vector which are probabilities which are being input for a particular label.

we can now estimate the accuracy of model by making predictions for each data instance in our test data. The methods are used to calculate predictions based upon test data and the summary of training data set. Lastly, the prediction values can be compared to the class values in our test data set and classification accuracy can be calculated.

**COMPARISION**

In this section I am going to compare both the approaches, as the requirement of this assignment. Firstly, within a supervised learning category, SVM comes under disseminative class which discriminate classes by building hyperplanes and the naive Bayes comes under generative approach which focuses in finding the probability distribution of data points by using joint probabilities.

Moreover, Naive Bayes assumes that the features are independent to each other whereas SVM considers the interactions between them. SVM is geometric in nature whereas probabilistic in nature.