Spam ? Ham?

生科1303班　李姜

How to judge whether it is a spam or not

Data Processing

tm wordcloud

Machine Learning

caret e1071 pROC

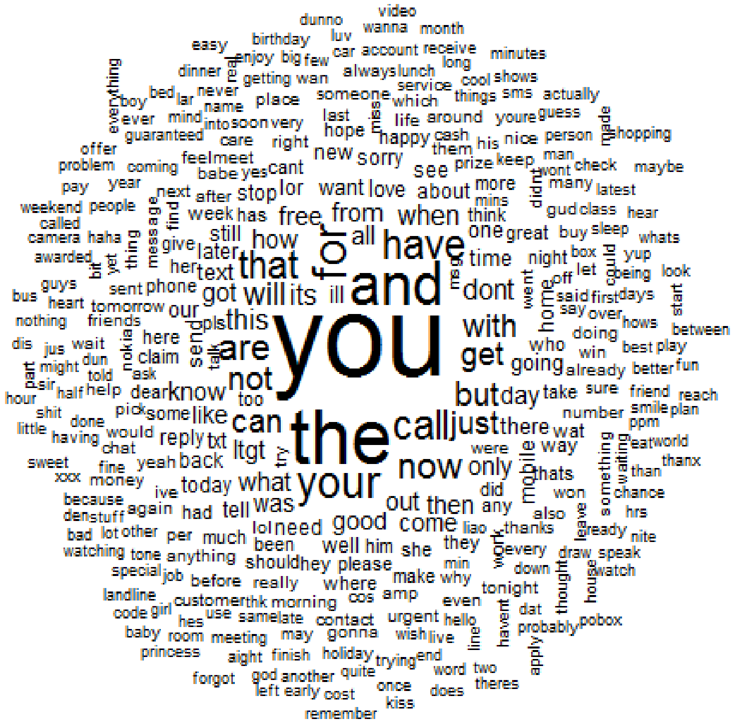
Data Processing

Import data

```
setwd("E:/deep_learing_R/Machine-Learning-with-R-datasets-master")
library(tm)
Sys.setlocale(category = "LC_ALL", locale = "us")
library(wordcloud)
sms_raw<-read.csv(file="sms_spam.csv",stringsAsFactors = F)
sms_raw$type<-factor(sms_raw$type)
```

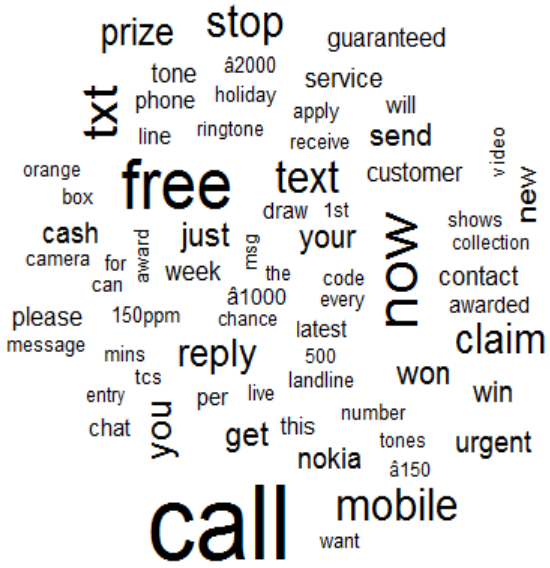
Data Processing

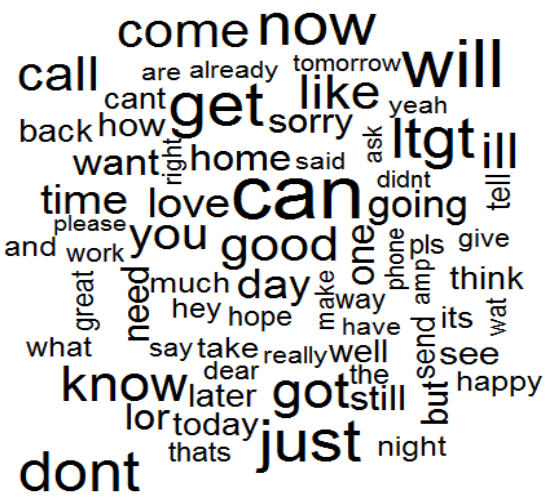
```
sms_corpus<-Corpus(VectorSource(sms_raw$text))

corpus_clean<-tm_map(corpus_clean,toupper)

corpus_clean<-tm_map(sms_corpus,removeNumbers)

corpus_clean<-tm_map(corpus_clean,removePunctuation)

corpus_clean<-tm_map(corpus_clean,stripWhitespace)

wordcloud(corpus_clean,min.freq = 30,random.order = F,random.color = F)
```



Data Processing

```
spam<-subset(sms_raw,sms_raw$type=="spam")

ham<-subset(sms_raw,sms_raw$type=="ham")

wordcloud(spam$text,max.words = 70)

wordcloud(ham$text,max.words = 70)
```





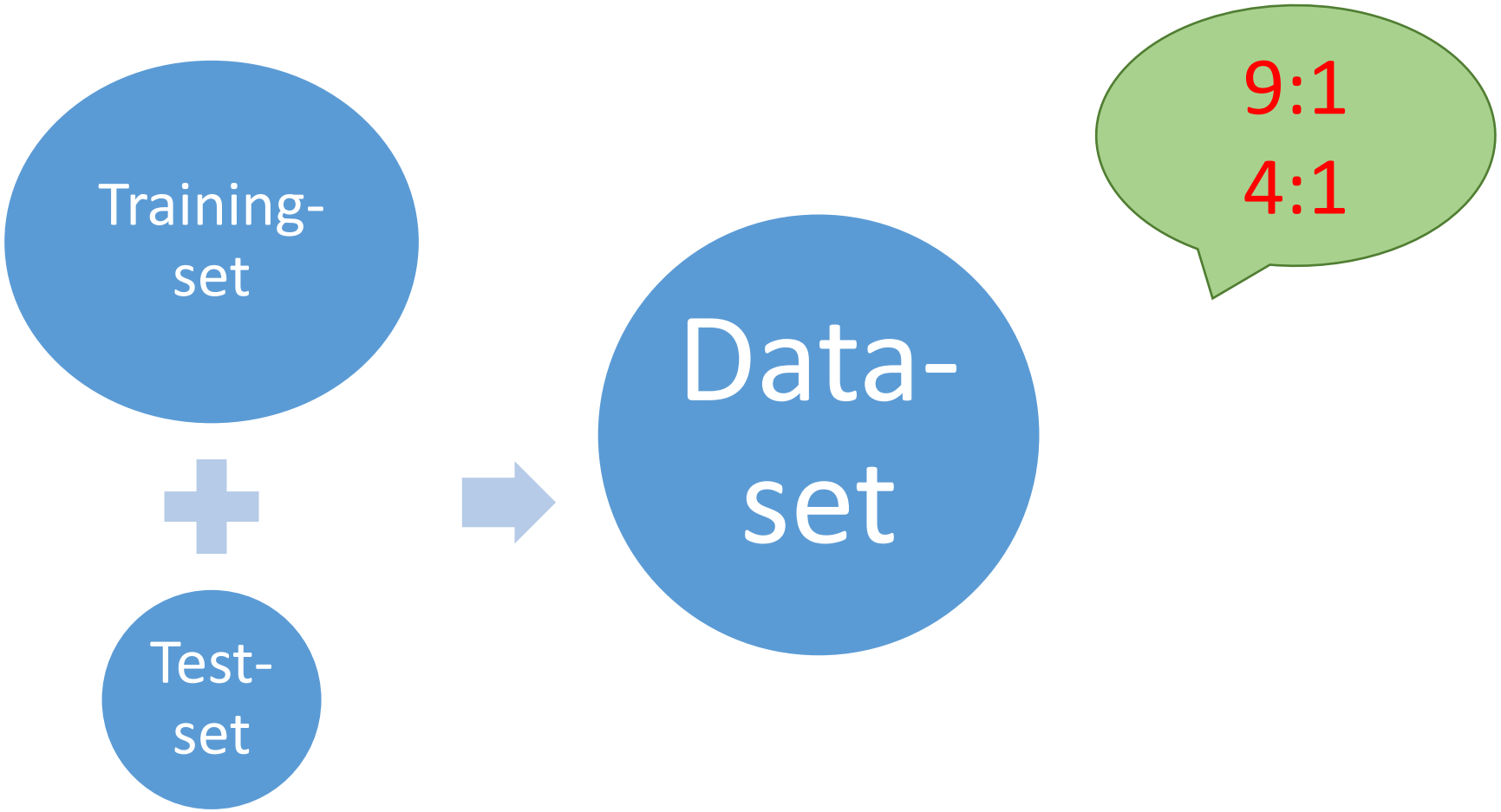
Machine Learning

Na ïve Bayes algorithm

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

1. Collect the data you need
2. Data cleaning , primary data mining
3. Train the model based on your data
4. Test the performance of the model
5. Optimize the model

Machine Learning



Machine Learning

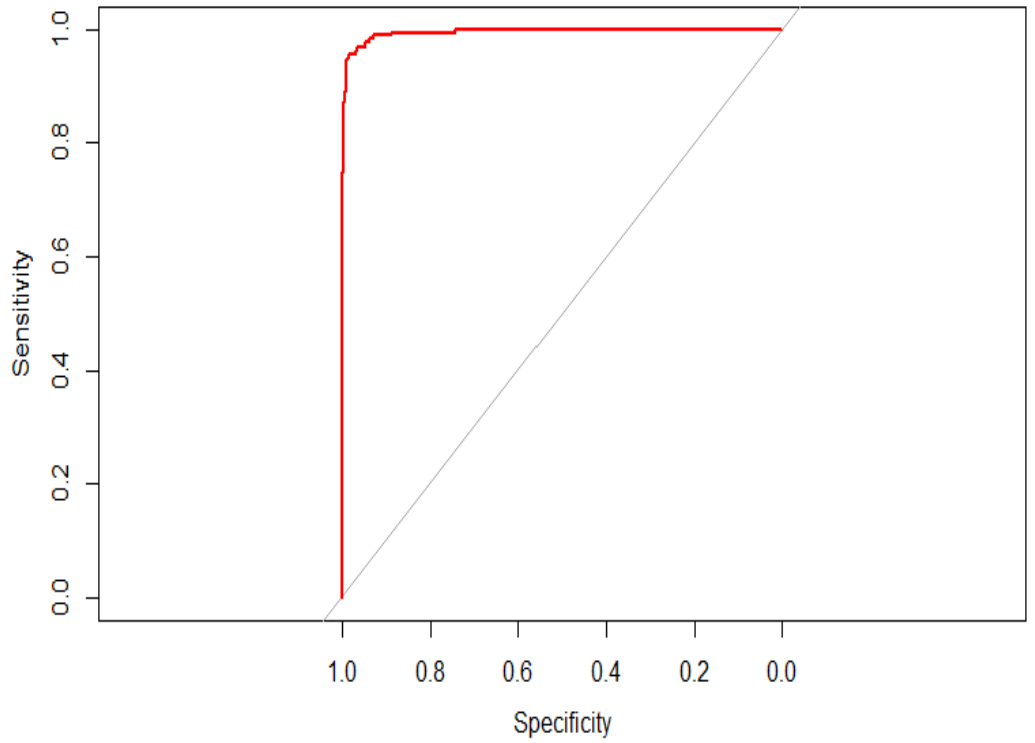
```
model<-naiveBayes(sms_train,sms_raw_train$type,laplace = 1)
pred<-predict(model,sms_test,type = "raw")
confusionMatrix(pred,sms_raw_test$type)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  ham spam
      ham  1206   24
      spam    5  158

               Accuracy : 0.9792
                 95% CI : (0.9702, 0.986)
    No Information Rate : 0.8693
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9041
 Mcnemar's Test P-Value : 0.0008302

            Sensitivity : 0.9959
            Specificity : 0.8681
         Pos Pred Value : 0.9805
         Neg Pred Value : 0.9693
             Prevalence : 0.8693
         Detection Rate : 0.8658
   Detection Prevalence : 0.8830
      Balanced Accuracy : 0.9320

       'Positive' Class : ham
```

Machine Learning

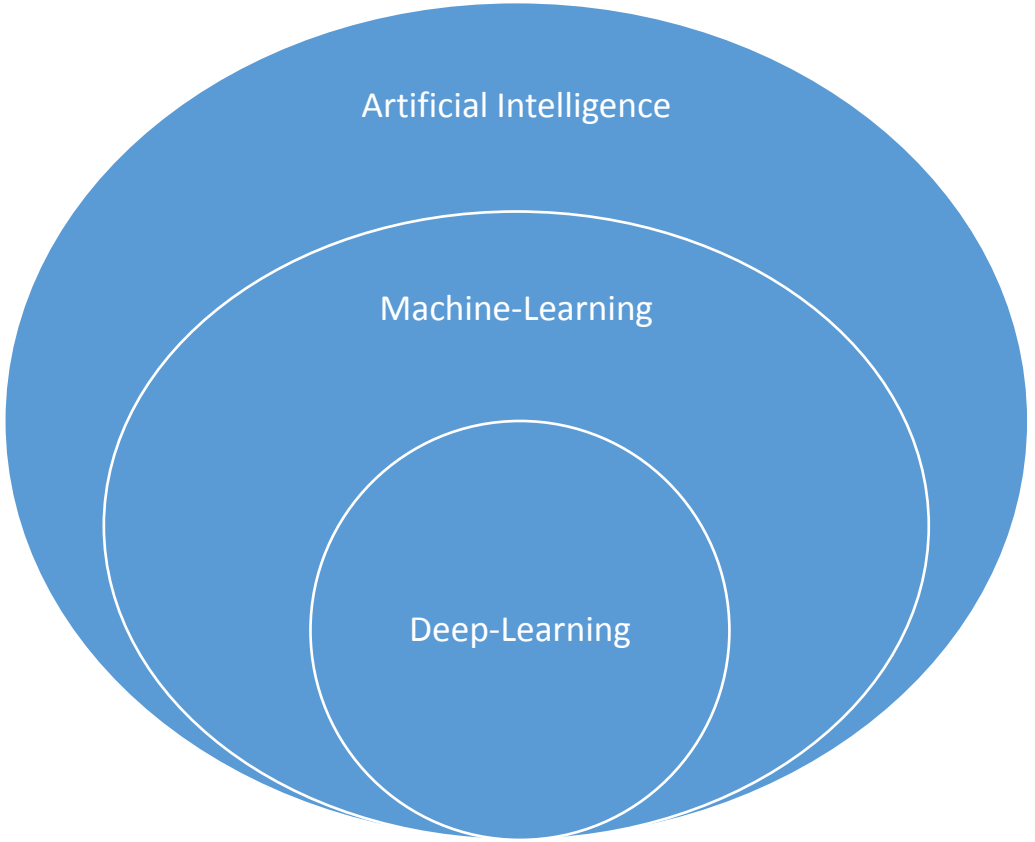
```
plot.nb<-
roc(response=sms_raw_test$type,predictor=pred[,2],levels=le
vels(sms_raw_test$type))

plot(plot.nb,type = "S",col="red")
```

Area under the curve(AUC):0.9953



Still Learning

SVM,NB,RandomForest,GLM,C5.0,KNN,K-means,ANN,Apriori



Still Learning

https://zhuanlan.zhihu.com/p/25432634?utm_source=qq&utm_medium=social

http://openclassroom.stanford.edu/MainFolder/VideoPage.php?course=MachineLearning&video=01.4-Introduction-UnsupervisedLearning&speed=100

http://v.youku.com/v_show/id_XMTM1MzQ1NDk5Ng==.html?from=y1.7-1.2

Still Learning

My GitHub Home Page： https://github.com/Ronlee12355

Thank you for listening and look forward to high scores

2017年3月2日