Support Vector Machines

Introduction

We are going to work with a total of 3175 DNA sequences, of which 2175 are training instances and 1000 are test instances. Each DNA sequence is of length 60. Our task is to identify whether or not a given DNA sequence would undergo splicing. Phenotype labels are S:spliced and N:not-spliced.

```
library(dplyr) # for manipulating tabular data
```

```
train_sequence <- read.table("data/train_sequence.txt", header=T)
test_sequence <- read.table("data/test_sequence.txt", header=T)
head(train_sequence)
```

```
##                                                          sequence
## 1 CCAGTAGTGTCGGGAGCAAGTTCAGCAAAACTCAAACCTTTCAGGGTTGTGGAATCTTGC
## 2 TGCCTCCTTTCACACTCCTCTTGGGGCTCGTGACATTACGAACCCTAACCCGGGCCCTGC
## 3 ATCTCACCGATCCGCCTTTTCGTTCTTTCTTTTTATTCTCTTTAGACGGAGTTTCACTCT
## 4 GCAGAGGCAGTGGGGGTGGGCAGCATTTACAGAAAATCTGTGATGAGACACCACAAAACC
## 5 ATGGCTCTTAATTATTATCTTTGGAATATTTGGGCTAACAGTGATGCTATTTGTATTCTT
## 6 CTCTTCCCCGCTTCCAACCTTCGGTTGCCGGTAACTACACCCCAGGGGTGGAACCCTAGC
```

Let's first compute the frequencies for 2-mers, 3-mers and 4-mers for each DNA sequence in our training and test data.

The set of all possible 2-mers is

```
base <- c("A", "T", "C", "G")

combns <- expand.grid(base, base)
combns <- within(combns, two_mers<-paste(Var1, Var2, sep=""))
two_mers <- combns$two_mers

two_mers
```

```
##  [1] "AA" "TA" "CA" "GA" "AT" "TT" "CT" "GT" "AC" "TC" "CC" "GC" "AG" "TG"
## [15] "CG" "GG"
```

The same can be done for 3-mers and 4-mers.

```
combns <- expand.grid(base, base, base)
combns <- within(combns, three_mers<-paste(Var1, Var2, Var3, sep=""))
three_mers <- combns$three_mers

combns <- expand.grid(base, base, base, base)
combns <- within(combns, four_mers<-paste(Var1, Var2, Var3, Var4, sep=""))
four_mers <- combns$four_mers
```

The total features for each DNA sequence is therefore 16 + 64 + 256 = 366 features (adding counts of `two_mers`, `three_mers`, and `four_mers`).