Practical session: String kernels

Jean-Philippe Vert

In this session you will

- Learn how to use string kernels with the R package kernlab
- Use string kernels to predict protein subcellular localization

1 Basic string kernels

kernlab implements several string kernels, including the spectrum and various substring kernels. They are created by the stringdot functions.

Question 1 Look at the string kernel implemented by help(stringdot). Check that you understand them.

To create a string kernel and test it on strings:

```
# We create a 2-spectrum kernel
sk <- stringdot(type="spectrum", length=2, normalized=FALSE)

# Compute the kernel between two words
sk('radar','abracadabra')
```

Question 2 Compute the kernel between two words (e.g., radar and abracadabra), for different types of kernels and different parameters. Check that it does what you want.

2 Application: text classification

As a toy application, let us show how string kernels can be used to manipulate texts. We load a small dataset of news from two newsgroupe or the Reuters dataset

```
data(reuters)
y <- rlabels
x <- reuters
```

We can then use the string kernels on x using the classical syntax of kernlab to run kernel methods.

Question 3 For different string kernels and different parameters, visualize the Reuters dataset by kernel PCA, and test the performance of a SVM to predict the newsgroup. Which kernel performs the best between spectrum, boundrange and exponential?

3 Application: Protein subcellular localization prediction

As a second application, we try to predict the subcellular localization of a protein from its aminoacid sequence.

Question 4 Download a set of protein sequence with subcellular localization information: http://www.psort.org/dataset/dataset1_0.txt .

Now we nead to read the aminoacid sequences and the subcellular localization in R. We use the read.fasta function of the seqinr package to read the file in FASTA format.

```
# Read data in FASTA format
protdata <- read.fasta("dataset1_0.txt",seqtype="AA",as.string=TRUE)

length(protdata)

# To speed up computation, we will only work on a subset of 100 randomly selected proteins
protdata <- protdata[sample(length(protdata),100)]

# Save it to a file for future use if needed
write.fasta(protdata,names=names(protdata),file.out="smalldataset.fa")


# Extract protein localization information
annotation <- getName(protdata)

# We get the location information by parsing the annotation as follows
extractlocationfromannotation <- function(s){strsplit(s,'|',fixed=TRUE)[[1]][3]}
loc <- unlist(lapply(annotation,extractlocationfromannotation))

# Extract the protein sequences
x <- unlist(getSequence(protdata,as.string=TRUE),recursive=FALSE)

# Focus on inner and outer membrane integral membrane proteins
y <- factor((loc=="Inner") | (loc=="Outer"))
```

Question 5 Test SVM with spectrum kernel ($k = 1, 2, 3$) and exponential kernel. Which works best?