Lab3assignment1

Alejo Perez Gomez

11/12/2020

1

Kernel methods

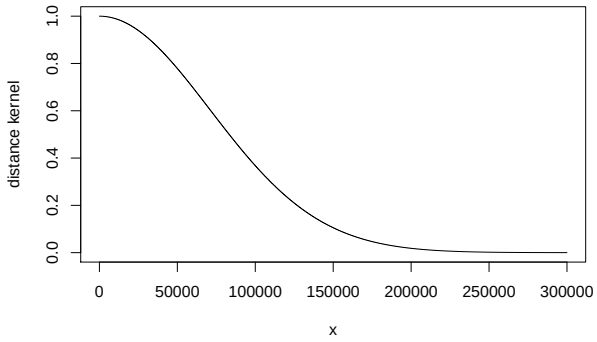
In this assignment we are required to use a Gaussian kernel-based algorithm to predict air temperatures based on Linköping meteorologic station Registers. Three Kernels operations will be computed:

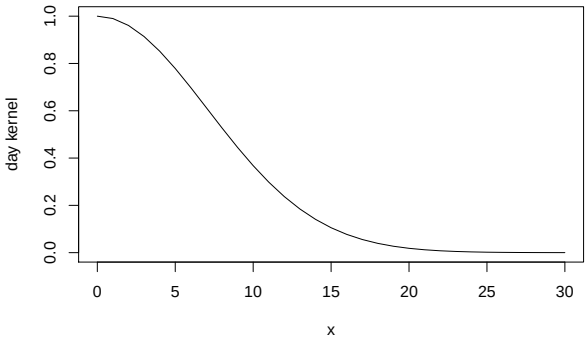
- 1. Based in Haversine Great Circle Distance using coordinates
- 2. Based in time distance in days
- 3. Based in time distance hours

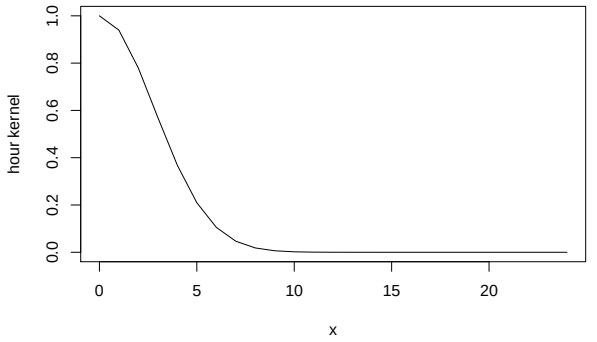
In order to choose the divisor constant h for each kernel we will plot the response of each over a reasonable support. The support of the Haversine kernel will be a sequence comprised between 1 and 300 km since that is the approximate span radius of Linköping city. The support the day-distance-kernel will be 30 days and for the hour-distance-kernel 24 hours. We tried several values h until getting plots which conferred us larger response for smaller distance values and less for bigger ones. Therefore we will choose the following values for h.

- 1. $h_{day} = 10$ To include distances within the third part of a month
- 2. $h_{hour} = 4$ Will allow us to account for time distances within a day with a strong variability of 3 hours span
- 3. $h_{Haversine} = 100$ So as we can account for distances smaller than the span of Linköping being the constant the third part of it.

Moreover, the response in the plots show $h_{day} = 10$ and $h_{hour} = 4$ show similar responses towards the variation of distances and $h_{Haversine} = 100$ has a lower cut-off value for distances, what means that it starts rejecting smaller distances than the other two.







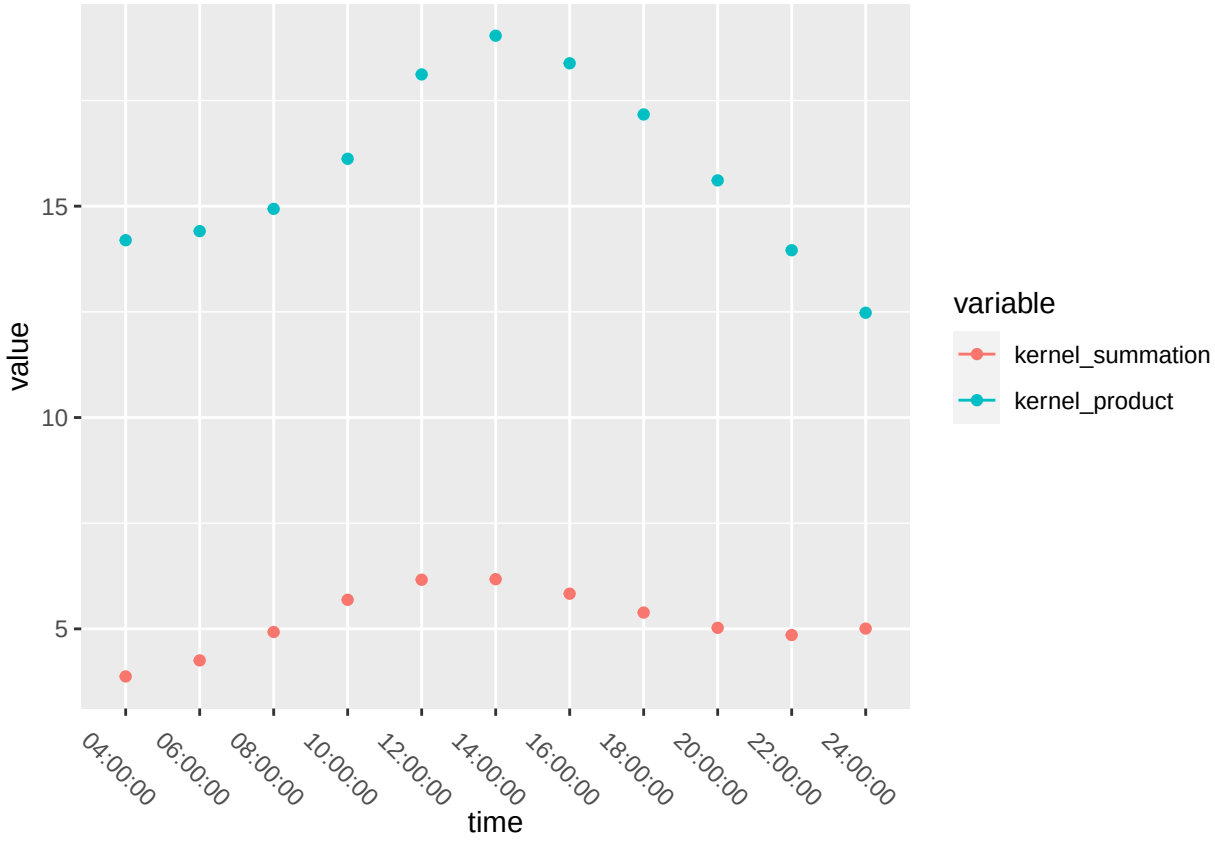
Here below we well initialize values, with the date to predict $day : 2013 - 8 - 15$ and coordinates $lon : 58.4274, lot : 14.826$. The algorithm will calculate the kernel operation for the *day distance* and *Haversine distance* for the dataset with earlier days than the mentioned. As for the *hour distance*, it will loop over the vector `times` in order to calculate a distance between each different time bucket in the vector and the rest of the filtered dataframe.

Afterwards, the predicted temperatures will be calculated out of the sum of kernels and their product separately.

```
### Initial values

set.seed(1234567890)
stations <- read.csv("stations.csv")
temps <- read.csv("temps50k.csv")
st <- merge(stations,temps,by="station_number")

h_distance <- 100
h_date <- 10
h_time <- 4

a <- 58.4274
b <- 14.826


date <- "2013-8-15" # The date to predict (up to the students)

times <- c("04:00:00", "06:00:00", "08:00:00", "10:00:00", "12:00:00",
```

```
           "14:00:00", "16:00:00", "18:00:00", "20:00:00", "22:00:00",
           "24:00:00")

temp <- list(kernel_summation=vector(length=length(times)), kernel_product= vector(length=length(times)))

input <- list(temp = temp,
              times=times,
              input_date=date,
              lat_lon=c(b,a),
              h_distance=h_distance,
              h_date=h_date,
              h_time=h_time)
```

```
## Warning: package 'reshape2' was built under R version 4.0.3
```



Up to this point some considerations should be taken regarding the result of the plot. At first glance, the result of the summation of Kernels does not make much sense based on the year season of the day we are predicting for. However, the kernel product has achieve another much more sensible result for a summer random day. The possible explanation of why the summation model is attaining smaller values than the product one might be the sum of three Gaussian exponential leads not as large values as the product of such.