

Dear members of the Decision Management Community,

Thank you for bringing your April / May 2015 predictive analytics problem to my attention. I was not aware of this challenge, presumably because I joined the corresponding LinkedIn group very recently. I started looking into this problem only today, after receiving your email. In any case, I found the challenge stimulating and I would like to present you the solution that I have found. This is not at all my field of research and it is moreover the first time I participate in such a challenge, so I'm far from being certain that the result is correct.

The prediction is based on a naïve Bayesian model. The code is written in R, and it relies on the e1071 package from the TU Wien. The input data is provided in an easily human-readable file.

I enclose the input file (patients.dat) and my R script (patients.r) in the attachment of this email.

According to this short code and the database that was provided, patient 21's illness is likely to be a cold - with a probability of somewhat above 83%, or an allergy, but the latter only with a probability of 12.4%. The probability of a flu is estimated to be very small in this case, with merely 3.8% .

I hope that this result is correct and that you will find this submission interesting. Thank you again for establishing this challenge and for bringing it to my attention.

Yours sincerely,
Riccardo Hertel

<https://www.linkedin.com/profile/view?id=242176070>
<https://plus.google.com/u/0/103922913049455254472>
www.riccardo-hertel.com

patients.r

```
# Challenge April-May 2015 "Decision Management Community"
#
# by Riccardo Hertel 2015-06-11
#
library(e1071)
patients <- read.csv('patients.dat', sep='\t')
# convert the logical values in the table [TRUE/FALSE] into numerical
ones:
fatigue <- 1 * patients[,1,drop=F]
s_nose <- 2 * patients[,2,drop=F]
sneeze <- 3 * patients[,3,drop=F]
s_throat <- 4 * patients[,4,drop=F]
# the numbers are unimportant as long as they are different for each
column.
```

```

# A shorthand notation of this would be, e.g., s_throat <- 4 *
patients[4]
# but I think it's less clear.
#
# Now let's reassemble the matrix:
num_patients <-
cbind(fatigue,s_nose,sneeze,s_throat,patients[,5,drop=F])
patients_symp_train <- num_patients[-21,-5]
illness <- factor(patients[-21,5])
patient_21_symp <- num_patients[21,-5]
# train the model:
patient_classifier <- naiveBayes(patients_symp_train,illness)
# make a prediction:
patient_prediction <- predict(patient_classifier,
patient_21_symp,type="raw")
print(patient_prediction)
#
# According to this naive Bayesian estimate, patient 21's illness is:
#
# with a probability of 83.7% a cold
# with a probability of 12.4% an allergy and
# with a probability of 3.8% a flu
#
# The a priori probabilities of the illnesses
# are 50% for cold, 30% for flu and 20% for allergies.

```

patients.dat

fatigue	s_nose	sneeze	s_throat	diag
F	F	T	T	Cold
T	T	T	F	Cold
T	T	F	T	Cold
T	F	F	T	Cold
T	T	F	T	Cold
F	T	T	T	Cold
T	F	F	T	Cold
T	T	F	F	Cold
F	T	F	F	Cold
F	F	T	T	Cold
T	F	F	T	Flu
T	F	F	T	Flu
F	T	T	F	Flu
T	T	F	F	Flu
T	F	T	F	Flu
T	F	F	F	Flu
T	F	F	T	Allergies
F	T	T	F	Allergies
F	F	T	F	Allergies
T	F	T	F	Allergies
F	T	T	T	Unknown