3.0 Naïve Bayes Classifier

Jonathan De Los Santos

2/3/2021

Bayes Overview

18th century mathematician Thomas Bayes developed principles for describing the probability of events and how to revise them based on new information - Bayesian classifiers use training data to claculate the probability of each outcome based on the evidence provided by freature values

Important concepts: - Event: potential outcome for which we measure the estimated likelihoods - Trials: opportunities for events to occur

Probability

The probability of an even is estimated by dividing the number of trials in which an even occurred by the total number of trials - Notation: $P(A)$ "Probability of event A" - Depends on mutually exclusive and exhaustive events, which cannot occur at the same time - They are the only possible outcomes - Events are mutually exclusive and exhaustive with their complement - Complement: the event comprising the outcomes in which the event of interest does not happen - Notation: The complement of A is denoted A^c or A'

Joint Probability

- What if we want to monitor non-mutually exclusive events?
- If some events occur concurrently with the event of interest, we can calculate that (think Venn Diagram)
- Joint probability: how the probability of one event is related to the probability of the other
  - Relies on events being dependent to be predictive
- Independent events: the events are totally unrelated
  - Impossible to predict one event by observing another

Notation

- Intersection: An event in which two or more events occur
  - Denoted with $\cap$ symbol
  - $A \cap B$ = an event where A and B occur
  - $P(A \cap B)$ = Probability that that both A and B occur

Bayes Theorem

The formulation for revising an estimate of the probability of one event given evidence provided by another. - Conditional probability: The probability of one event given another occurring - Denoted with a vertical bar |

The probability of A given B is estimated as the proportion of trials in which A occurred with B, divided by all trials in which B occurred.

Formulated as:

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

Rearranged algebraically to the more useful form:

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

Using an example of trying to predict whether an email is spam given that it contains the word "Viagra" we can replace our letters with the following A = spam B = Viagra

We are determining the probability of spam, given it contains "Viagra" or $P(spam|Viagra)$

Rewriting our formula:

$P(spam|Viagra) = \frac{P(Viagra|Spam)P(spam)}{P(Viagra)}$

Let's break down this formula into terminology components:

$P(spam|Viagra)$ is the **posterior probability** - How likely the message is to be spam - If this is greater than 50%, it's more likely to be spam

$P(Viagra|spam)$ is the **likelihood** - The probability that Viagra was used in previous spam messages

$P(spam)$ is the **prior probability** - The probability that any prior message was spam, a starting point of estimating whether an email is spam

$P(Viagra)$ is the **marginal likelihood** - The probability that Viagra appeared in any message at all

Calculating Bayes Theorem

See example pp. 96 of "Machine Learning with R" by Brett Lantz

- First construct a frequency table that records the number of times Viagra appeared in spam and non=spam messages
- Then convert this to a likelihood table to indicate the conditional probabilities for "Viagra" given that the email was spam or not spam

Naïve Bayes Classifier

Naive Bayes is a method to apply Bayes' theorem to classification problems - So named because it makes "naive" assumption about the data, namely that all of the features are **equally important and independent** - These assumptions are rarely true, but even when they're not it still performs well - One reason for this is that estimating probability precisely is not as important as making predictions - The difference between a 51% probability of spam and 99% probability of spam may not be important as long as its filtered

We perform the Bayes calculation as before, but with several features simultaneously - Besides Viagra we may have variables for other words like money, groceries, unsubscribe, or completely different variables like whether the sender is in your address book - **Class-conditional independence:** Events are independent so long as they are conditioned on the same class value - By making the naive assumption of **class-conditional independence** the math is simplified by multiplying the individual conditional probability rather than computing conditional joint probabilities

Naïve Bayes Classifier in R

e-1071() Package

Data Partioning with createDataPartition()

```
CreateDataPartition(
  y,
  times = 1,
  p = 0.5,
  list = TRUE,
  groups = min(5, length(y))
)
```

Arguments - y: a vector of outcomes. For createTimeSlices, these should be in chronological order. - times: the number of partitions to create - p: the percentage of data that goes to training - list: logical - should the results be in a list (TRUE) or a matrix with the number of rows equal to floor(p * length(y)) and times columns. - groups: for numeric y, the number of breaks in the quantiles

- Usage
  - Input 70% (`p=0.7`) of factored AHD column into `intrain`
  - List = FALSE for some reason
  - Add intrain to trainSet
  - Add all except intrain to testSet

```r
heart <- read.csv("Data Sets/3.0-Heart.csv")

# Save AHD as factor
heart$AHD.f <- as.factor(heart$AHD)


library(caret)
set.seed(1234)
intrain <- createDataPartition(y = heart$AHD.f, p = 0.7, list = FALSE)
trainSet <- heart[intrain,]
testSet <- heart[-intrain,]
#str(trainSet)
#summary(testSet)
```

naiveBayes() Model

> "Computes the conditional a-posterior probabilities of a categorical class variable given independent predictor variables using the Bayes rule."

Arguments

- x: A numeric matrix, or a data frame of categorical and/or numeric variables.
- y: a class vector.
- formula: A formula of the form class ~ x1 + x2 + ... Interactions are not allowed.
- data: Either a data frame of predictors (categorical and/or numeric) or a contingency table.

```r
library(e1071)
nb.model <- naiveBayes(AHD.f~., data = trainSet)
nb.model
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##        No       Yes
## 0.5399061 0.4600939
##
## Conditional probabilities:
##      ID
## Y         [,1]     [,2]
##   No  151.3304 84.49569
##   Yes 148.0612 88.11027
##
##      Age
## Y         [,1]     [,2]
##   No  52.91304 9.712142
##   Yes 56.27551 7.979071
##
##      Sex
## Y          [,1]      [,2]
##   No  0.5652174 0.4978979
##   Yes 0.8367347 0.3715079
##
##      ChestPain
## Y     asymptomatic nonanginal nontypical    typical
##   No    0.26956522 0.41739130 0.21739130 0.09565217
##   Yes   0.82653061 0.09183673 0.05102041 0.03061224
##
##      RestBP
## Y         [,1]     [,2]
##   No  128.4783 17.12306
##   Yes 132.7143 17.09344
##
##      Chol
## Y         [,1]     [,2]
##   No  235.9217 54.04305
##   Yes 252.3061 51.62861
##
##      Fbs
## Y          [,1]      [,2]
##   No  0.1130435 0.3180317
##   Yes 0.1326531 0.3409434
##
##      RestECG
## Y         [,1]      [,2]
```

```
##   No  0.826087 0.9846263
##   Yes 1.214286 0.9765287
##
##      MaxHR
## Y         [,1]     [,2]
##   No  157.0000 19.88652
##   Yes 138.8061 22.81736
##
##      ExAng
## Y          [,1]      [,2]
##   No  0.1565217 0.3649394
##   Yes 0.5510204 0.4999474
##
##      Oldpeak
## Y          [,1]      [,2]
##   No  0.6286957 0.8122705
##   Yes 1.5836735 1.2905440
##
##      Slope
## Y         [,1]      [,2]
##   No  1.391304 0.5727075
##   Yes 1.816327 0.5438871
##
##      Ca
## Y          [,1]      [,2]
##   No  0.2767857 0.6466800
##   Yes 1.0824742 0.9646902
##
##      Thal
## Y          fixed     normal reversable
##   No  0.02631579 0.74561404 0.22807018
##   Yes 0.08247423 0.25773196 0.65979381
##
##      AHD
## Y     No Yes
##   No   1   0
##   Yes  0   1
```

predict() Function

```r
nb.model.pred <- predict(nb.model, testSet, type = 'class')
nb.model.pred
```

```
##  [1] No  Yes No  Yes No  No  No  No  Yes No  Yes No  No  No  No  No  No  Yes No
## [20] Yes No  Yes No  No  No  No  No  Yes Yes Yes Yes Yes No  Yes No  Yes No  No
## [39] No  Yes No  Yes No  No  No  Yes No  No  No  Yes Yes No  Yes No  No  Yes Yes
## [58] No  Yes Yes Yes Yes No  No  No  No  Yes Yes Yes No  No  Yes No  Yes No  No
## [77] No  No  No  Yes Yes Yes Yes No  No  Yes Yes Yes Yes Yes
## Levels: No Yes
```

Confusion Matrix

```
actual <- testSet$AHD.f
confusionMatrix(actual, nb.model.pred)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  49   0
##        Yes  1  40
##
##                Accuracy : 0.9889
##                  95% CI : (0.9396, 0.9997)
##     No Information Rate : 0.5556
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9776
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9800
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.9756
##              Prevalence : 0.5556
##          Detection Rate : 0.5444
##    Detection Prevalence : 0.5444
##       Balanced Accuracy : 0.9900
##
##        'Positive' Class : No
##
```

Compute Conditional Probability from R

From the above data we want to know:

$P(AHD = YES|ChestPain = typical, Thal = fixed)$

Needed Values

P(ChestPain=typical | AHD=YES) P(Thal=fixed | AHD=YES) P(AHD=YES)

P(ChestPain=typical | AHD=NO) P(Thal=fixed | AHD=NO) P(AHD=NO)

Finding Probabilities in `nb.model`

- Find ChestPain
- Find the probability at the intersection of "typical" and "YES"
- Repeat for Thal
- Probability for AHD = YES can be found under "A-Priori Probabilities"

Calculating Probability

For some reason these values differ from professor's, but I've been troubleshooting for too long. Here are the values from my model:

P(ChestPain=typical | AHD=YES): 0.031 **P(Thal=fixed | AHD=YES):** 0.082 **P(AHD=YES):** 0.460

AHD = YES $0.031 * 0.082 * 0.460 = 0.0012$

AHD = NO **P(ChestPain=typical | AHD=NO):** 0.096 **P(Thal=fixed | AHD=NO):** 0.0263 **P(AHD=NO):** 0.5399

$0.096 * 0.0263 * 0.5399 = 0.001$

Final Calculation $\frac{P(typical, fixed|AHD=YES)}{P(typical, fixed|AHD=YES)+P(typical, fixed|AHD=NO)}$

$\frac{0.0012}{0.0012+0.001} = 0.55$