The Analytics Edge

Test Your Knowledge of Text Analytics

Note to all. I have compiled the answers in the following format – for each question, the qualitative or "written" solutions will be provided together with their sub-questions. The R scripts (as well as the console outputs) will be provided *after* each whole question, followed by all the relevant plots. If I have missed anything in the solutions, or if you have any questions, you may email me at benjamin_tanwj@mymail.sutd.edu.sg. Thank you!

1. Nearly every email user has at some point encountered a "spam" email, which is an unsolicited message often advertising a product, containing links to malware, or attempting to scam the recipient. Roughly 80-90% of more than 100 billion emails sent each day are spam emails, most being sent from botnets of malware-infected computers. The remainder of emails are called "ham" emails.

   As a result of the huge number of spam emails being sent across the Internet each day, most email providers offer a spam filter that automatically flags likely spam messages and separates them from the ham. Though these filters use a number of techniques (e.g. looking up the sender in a so-called "Blackhole List" that contains IP addresses of likely spammers), most rely heavily on the analysis of the contents of an email via text analytics.

   In this problem, you will build and evaluate a spam filter using a publicly available dataset first described in the 2006 conference paper "Spam Filtering with Naive Bayes – Which Naive Bayes?" by V. Metsis, I. Androutsopoulos, and G. Paliouras. The "ham" messages in this dataset come from the inbox of former Enron Managing Director for Research Vincent Kaminski, one of the inboxes in the Enron Corpus. One source of spam messages in this dataset is the SpamAssassin corpus, which contains hand-labeled spam messages contributed by Internet users. The remaining spam was collected by Project Honey Pot, a project that collects spam messages and identifies spammers by publishing email address that humans would know not to contact but that bots might target with spam. The full dataset we will use was constructed as roughly a 75/25 mix of the ham and spam messages. The dataset contains just two fields:

   - **text:** The text of the email
   - **spam:** A binary variable, 1 indicating if the email was spam and 0 otherwise

   (a) Begin by loading the dataset **emails.csv** into a data frame called **emails**. Remember to pass the stringsAsFactors=FALSE option when loading the data. How many emails are in the dataset? How many of the emails are spam?

   Solution. There are 5728 emails in the dataset, 1368 of which are labeled as spam.

(b) Which word appears at the beginning of every email in the dataset?

Solution. The word is “Subject:”.

(c) Could a spam classifier potentially benefit from including the frequency of the word that appears in every email?

i. No - the word appears in every email so this variable would not help us differentiate spam from ham.

ii. Yes – the number of times the word appears might help us differentiate spam from ham.

Solution. Yes – since the number of times a word appears might be different in spam and ham email messages. For example, a long email might have the word “subject” occur more often, and this might be indicative of ham emails.

(d) The nchar() function counts the number of characters in a piece of text. How many characters are in the longest email in the dataset (where longest is measured in terms of the maximum number of characters)?

Solution. max(nchar(emails$text)) shows that the longest email message has 43952 characters.

(e) Which row contains the shortest email in the dataset? (Just like in the previous problem, shortest is measured in terms of the fewest number of characters). Write down the corresponding email.

Solution. which.min(nchar(emails$text)) shows that the 1992nd email message is the shortest – “Subject: fyi ”.

(f) Follow the standard steps to build and pre-process the corpus:

- Load the tm package.
- Build a new corpus variable called corpus.
- Using tm_map, convert the text to lowercase.
- Using tm_map, remove all punctuation from the corpus.
- Using tm_map, remove all English stopwords from the corpus.
- Using tm_map, stem the words in the corpus.
- Build a document term matrix from the corpus, called dtm

How many terms are in dtm?

Solution. There should be 28687 terms in the document-term matrix.

(g) To obtain a more reasonable number of terms, limit dtm to contain terms appearing in at least 5% of documents, and store this result as spdtm (don't overwrite dtm, because we will use it later). How many terms are in spdtm?

Solution. There should be 330 terms in the sparse document-term matrix.

(h) Build a data frame called emailsSparse from spdtm, and use the make.names function to make the variable names of emailsSparse valid. colSums() is an R function that returns the sum of values for each variable in our data frame. Our data frame contains the number of times each word stem (columns) appeared in each email (rows). Therefore, colSums(emailsSparse) returns the number of times a word stem appeared across all the emails in the dataset. What is the word stem that shows up most frequently across all the emails in the dataset?

Solution. "enron".

(i) Add a variable called "spam" to emailsSparse containing the email spam labels. How many word stems appear at least 5000 times in the ham emails in the dataset? Which word stems are these?

Solution.

hou	*will*	*vinc*	*subject*	*ect*	*enron*
5569	6802	8531	8625	11417	13388

These words appear at least 5000 times in the ham emails in the dataset.

(j) How many word stems appear at least 1000 times in the spam emails in the dataset? Which word stems are these?

Solution.

compani	*spam*	*will*	*subject*
1065	1368	1450	1577

These words appear at least 1000 times in the spam emails in the dataset.

(k) The lists of most common words are significantly different between the spam and ham emails. What does this likely imply?

i. The frequencies of these most common words are unlikely to help differentiate between spam and ham.

ii. The frequencies of these most common words are likely to help differentiate between spam and ham.

Solution. The frequencies of these most common words are likely to help differentiate between spam and ham. For example, “enron” appears very often in ham as compared to spam.

(l) Several of the most common word stems from the ham documents, such as “enron”, “hou” (short for Houston), “vinc” (the word stem of “Vince”) and “kaminski”, are likely specific to Vincent Kaminski’s inbox. What does this mean about the applicability of the text analytics models we will train for the spam filtering problem?

i. The models we build are still very general, and are likely to perform well as a spam filter for nearly any other person.

ii. The models we build are personalized, and would need to be further tested before being used as a spam filter for another person.

Solution. The models we build are personalised and would need to be further tested before being used as a spam filter for another person.

(m) First, convert the dependent variable to a factor with
> emailsSparse$spam <− as.factor(emailsSparse$spam)
Next, set the random seed to 123 and use the sample.split function to split emailsSparse 70-30 into a training set called “train” and a testing set called “test”. Make sure to perform this step on emailsSparse instead of emails. Using the training set, train the following three models. The models should predict the dependent variable “spam”, using all other available variables as independent variables. Please be patient, as these models may take a few minutes to train.

- A logistic regression model called spamLog. You may see a warning message here - we’ll discuss this more later.
- A CART model called spamCART, using the default parameters to train the model. Directly before training the CART model, set the random seed to 123.
- A random forest model called spamRF, using the default parameters to train the model. Directly before training the random forest model, set the random seed to 123 (even though we’ve already done this earlier in the problem, it’s important to set the seed right before training the model so we all obtain the same results. Keep in mind though that on certain operating systems, your results might still be slightly different).

For each model, obtain the predicted spam probabilities for the training set.
You may have noticed that training the logistic regression model yielded the messages

"algorithm did not converge" and "fitted probabilities numerically 0 or 1 occurred". Both of these messages often indicate overfitting and in some case corresponds to severe overfitting, often to the point that the training set observations are fit perfectly by the model. Let's investigate the predicted probabilities from the logistic regression model.
How many of the training set predicted probabilities from spamLog are less than 0.00001? How many of the training set predicted probabilities from spamLog are more than 0.99999? How many of the training set predicted probabilities from spamLog are between 0.00001 and 0.99999?

Solution. 3046, 954 and 10 respectively.

(n) How many variables are labeled as significant (at the p=0.05 level) in the logistic regression summary output?

Solution. None of the variables are significant at the $p = 0.05$ level. Note that there was also trouble for the logistic regression model to converge in this example.

(o) How many of the word stems "enron", "hou", "vinc", and "kaminski" appear in the CART tree? Recall that we suspect these word stems are specific to Vincent Kaminski and might affect the generalizability of a spam filter built with his ham data.

Solution. Plot of the classification tree is shown here. The words "vinc" and "enron" appear at the top of the CART model. The words "hou" and "kaminski" do not appear.

(p) What is the training set accuracy of spamLog, using a threshold of 0.5 for predictions? What is the training set AUC of spamLog?

Solution. The confusion matrix (predicted labels as rows, actuals as columns) is shown below:

	0	1
FALSE	3052	4
TRUE	0	954

and the accuracy is 0.9990025. The AUC on the training set is 0.9999959.

(q) What is the training set accuracy of spamCART, using a threshold of 0.5 for predictions?

Solution. The confusion matrix (predicted labels as rows, actuals as columns) is shown below:

	0	1
FALSE	2885	64
TRUE	167	894

and the accuracy is 0.942394.

(r) What is the training set AUC of spamCART? (Remember that you have to pass the prediction function predicted probabilities.)

Solution. The training set AUC for the CART model is 0.9696.

(s) What is the training set accuracy of spamRF, using a threshold of 0.5 for predictions? (Remember that your have to use type="prob" in your prediction for random forest.)

Solution. The confusion matrix (predicted labels as rows, actuals as columns) is shown below:

	0	1
FALSE	3046	0
TRUE	6	958

and the accuracy is 0.998503.

(t) What is the training set AUC of spamRF? (Remember to pass the argument type="prob" to the predict function to get predicted probabilities for a random forest model. The probabilities will be the second column of the output.)

Solution. spamRF has a training AUC of 0.9999959.

(u) Which of the models have the best training set performance, in terms of accuracy and AUC?

- Logistic regression
- CART
- Random forest

Solution. In this model, logistic regression and random forest have the best performances.

(v) Obtain predicted probabilities for the testing set for each of the models, again ensuring that probabilities instead of classes are obtained. What is the testing set accuracy of spamLog, using a threshold of 0.5 for predictions?

Solution. The confusion matrix (predicted labels as rows, actuals as columns) is shown below:

	0	1
FALSE	1257	34
TRUE	51	376

and the accuracy is 0.9505239.

(w) What is the testing set AUC of spamLog? What is the testing set accuracy of spamCART, using a threshold of 0.5 for predictions? What is the testing set AUC of spamCART? What is the testing set accuracy of spamRF, using a threshold of 0.5 for predictions? What is the testing set AUC of spamRF?

Solution. The test AUC for spamLog is 0.9627517.

For spamCART, the confusion matrix (predicted labels as rows, actuals as columns) is shown below:

	0	1
$FALSE$	1228	24
$TRUE$	80	386

with an accuracy of 0.9394645. The AUC on the test set is 0.963176.

For spamRF, the confusion matrix (predicted labels as rows, actuals as columns) is shown below:

	0	1
$FALSE$	1290	25
$TRUE$	18	388

with an accuracy of 0.9749709. The AUC on the test set is 0.997768.

(x) Which model had the best testing set performance, in terms of accuracy and AUC?

- Logistic regression
- CART
- Random forest

Solution. The random forest has the most impressive performance in the test set both in terms of accuracy and AUC.

(y) Which model demonstrated the greatest degree of overfitting?

- Logistic regression
- CART
- Random forest

Solution. Logistic regression – it had an almost perfect fit on the training set but not as good performance on the test set. On the other hand, CART and random forest models have similar accuracies in the training and test sets.

R Scripts.

```
> #a)
> emails <- read.csv("emails.csv", stringsAsFactors = FALSE)
> str(emails)
’data.frame’: 5728 obs. of  2 variables:
 $ text: chr  "Subject: naturally irresistible your corporate identity  lt is really hard to
recollect a company : the  market"| __truncated__ "Subject: the stock trading gunslinger
fanny is merrill but muzo not colza attainder and penultimate like esmar"| __truncated__
"Subject: unbelievable new homes made easy  im wanting to show you this  homeowner
you have been pre - approved"| __truncated__ "Subject: 4 color printing special  request
additional information now ! click here  click here for a printable "| __truncated__ ...
 $ spam: int  1 1 1 1 1 1 1 1 1 1 ...
> nrow(emails)
[1] 5728
> table(emails$spam)

   0    1
4360 1368



> #b)
> strwrap(emails[1,1])
 [1] "Subject: naturally irresistible your corporate identity lt is really hard to"
 [2] "recollect a company : the market is full of suqgestions and the information"
 [3] "isoverwhelminq ; but a good catchy logo , stylish statlonery and outstanding"
 [4] "website will make the task much easier .  we do not promise that havinq ordered"
 [5] "a iogo your company will automaticaily become a world ieader : it isguite ciear"
 [6] "that without good products , effective business organization and practicable aim"
 [7] "it will be hotat nowadays market ; but we do promise that your marketing efforts"
 [8] "will become much more effective . here is the list of clear benefits :"
 [9] "creativeness : hand - made , original logos , specially done to reflect your"
[10] "distinctive company image . convenience : logo and stationery are provided in"
[11] "all formats ; easy - to - use content management system letsyou change your"
[12] "website content and even its structure . promptness : you will see logo drafts"
[13] "within three business days . affordability : your marketing break - through"
[14] "shouldn ’ t make gaps in your budget . 100 % satisfaction guaranteed : we"
[15] "provide unlimited amount of changes with no extra fees for you to be surethat"
[16] "you will love the result of this collaboration . have a look at our portfolio _"
[17] "_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _"
[18] "_ _ _ _ _ _ _ _ _ _ _ not interested . . . _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _"
[19] "_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _"
> strwrap(emails[1000,1])
 [1] "Subject: 70 percent off your life insurance get a free quote instantly ."
 [2] "question : are you paying too much for life insurance ?  most likely the answer"
 [3] "is yes !  here ’ s why . fact . . . fierce , take no prisoner , insurance"
 [4] "industry price wars have driven down premiums - 30 - 40 - 50 - even 70 % from"
 [5] "where they were just a short time ago !  that ’ s why your insurance company"
 [6] "doesn ’ t want you to read this . . .  they will continue to take your money at"
 [7] "the price they are already charging you , while offering the new lower rates ("
 [8] "up to 50 % , even 70 % lower ) to their new buyers only .  but , don ’ t take"
 [9] "our word for it . . . click hereand request a free online quote . be prepared"
[10] "for a real shock when you see just how inexpensively you can buy term life"
[11] "insurance for today !  removal instructions : this message is sent in compliance"
[12] "with the proposed bill section 301 , paragraph ( a ) ( 2 ) ( c ) of s . 1618 ."
[13] "we obtain our list data from a variety of online sources , including opt - in"
[14] "lists . this email is sent by a direct email marketing firm on our behalf , and"
[15] "if you would rather not receive any further information from us , please click"
[16] "here . in this way , you can instantly opt - out from the list your email"
[17] "address was obtained from , whether this was an opt - in or otherwise . please"
[18] "accept our apologies if this message has reached you in error . please allow 5 -"
[19] "10 business days for your email address to be removed from all lists in our"
[20] "control . meanwhile , simply delete any duplicate emails that you may receive"
[21] "and rest assured that your request to be taken off this list will be honored ."
[22] "if you have previously requested to be taken off this list and are still"
[23] "receiving this message , you may call us at 1 - ( 888 ) 817 - 9902 , or write to"
[24] "us at : abuse control center , 7657 winnetka ave . , canoga park , ca 91306"



> #d)
> max(nchar(emails$text))
[1] 43952



> #e)
> which.min(nchar(emails$text))
[1] 1992
> emails$text[1992]
[1] "Subject: fyi "



> #f)
> library(tm)
> corpus <- Corpus(VectorSource(emails$text))
> corpus <- tm_map(corpus, content_transformer(tolower))
Warning message:
In tm_map.SimpleCorpus(corpus, content_transformer(tolower)) :
  transformation drops documents
> corpus <- tm_map(corpus, removePunctuation)
Warning message:
In tm_map.SimpleCorpus(corpus, removePunctuation) :
  transformation drops documents
> corpus <- tm_map(corpus, removeWords, stopwords("english"))
Warning message:
In tm_map.SimpleCorpus(corpus, removeWords, stopwords("english")) :
  transformation drops documents
> corpus <- tm_map(corpus, stemDocument)
Warning message:
In tm_map.SimpleCorpus(corpus, stemDocument) :
  transformation drops documents
> dtm <- DocumentTermMatrix(corpus)
> str(dtm)
List of 6
 $ i       : int [1:481719] 1 1 1 1 1 1 1 1 1 1 ...
 $ j       : int [1:481719] 1 2 3 4 5 6 7 8 9 10 ...
 $ v       : num [1:481719] 1 1 1 1 1 2 1 1 1 2 ...
 $ nrow    : int 5728
 $ ncol    : int 28687
 $ dimnames:List of 2
  ..$ Docs : chr [1:5728] "1" "2" "3" "4" ...
  ..$ Terms: chr [1:28687] "100" "afford" "aim" "amount" ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

```
> #g)
> spdtm <- removeSparseTerms(dtm, .95)
> str(spdtm)
List of 6
 $ i       : int [1:213551] 1 1 1 1 1 1 1 1 1 1 ...
 $ j       : int [1:213551] 1 2 3 4 5 6 7 8 9 10 ...
 $ v       : num [1:213551] 2 2 3 1 1 1 2 1 1 1 ...
 $ nrow    : int 5728
 $ ncol    : int 330
 $ dimnames:List of 2
  ..$ Docs : chr [1:5728] "1" "2" "3" "4" ...
  ..$ Terms: chr [1:330] "busi" "chang" "compani" "corpor" ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

```
> #h)
> emailsSparse <- as.data.frame(as.matrix(spdtm))
> colnames(emailsSparse) <- make.names(colnames(emailsSparse))
> which.max(colSums(emailsSparse))
enron
  324
```

```
> #i)
> emailsSparse$spam <- emails$spam
> sort(colSums(subset(emailsSparse,spam==0)))
      spam       life      remov      money      onlin    without     websit      click
         0         80        103        114        173        191        194        217
   special       wish      repli        buy        net       link     immedi       done
       226        229        239        243        243        247        249        254
      mean     design        lot     effect       info       read     either      write
       259        261        268        270        273        279        279        286
      line      begin    success      sorri     involv    softwar      creat     better
       289        291        293        293        294        299        299        301
    vkamin        say       keep      bring     believ       full    increas     realli
       301        305        306        311        313        317        320        324
   mention    thought     invest       idea      secur     specif      sever     experi
       325        325        327        327        337        338        340        346
     thing      allow        due      check       type      happi     return     expect
       347        348        351        351        352        354        355        356
     short     effort       open   internet     sincer     public     recent      anoth
       357        358        360        361        361        364        368        369
   alreadi       home       made    respond      given        etc        put     within
       372        375        380        382        383        385        385        386
     place    version      right      hello       sure       area        run     arrang
       388        390        390        395        396        397        398        399
   account       join       hour      locat     togeth     import      engin        per
       401        403        404        406        406        411        411        412
    corpor       high     result       hear      final       deal     applic       even
       414        416        418        420        422        423        428        429
       web     custom       soon       long       sinc      futur     member      event
       430        433        435        436        439        440        446        447
      X000        don       part       feel    tuesday  wednesday       unit      still
       447        450        450        453        454        456        457        457
      site       X853    continu understand    resourc     robert       form    analysi
       458        461        464        464        466        466        468        468
     point     assist    confirm     differ     intern      might       real       case
       474        475        485        489        489        490        490        492
     howev    comment    complet        abl       rate    appreci        tri       move
       496       505       515       515       516       518       521       526
     updat    approv   suggest      free  contract    detail      morn       end
       527       533       533       535       544       546       546       550
      mani    attend  thursday    direct    requir     cours    person     relat
       550       558       558       561       562       567       569       573
    depart     today     start       way      mark      valu   problem     peopl
       575       577       580       586       588       590       593       599
      note    school     invit    access      term      juli    monday    gibner
       600       607       614       617       625       630       630       633
      base  director     offer      cost     addit     kevin     great       set
       635       640       643       646       648       654       655       658
      file      find      much     order      oper     deriv       doc     april
       659       665       669       669       669       673       673       677
      book   address      copi   financi     month   student   respons   possibl
       680       693       700       702       709       710       711       712
    associ  particip       now     first  industri      dear   support      plan
       715       717       725       726       731       734       734       738
      back      name      come  opportun    report   product       two    origin
       739       745       748       760       772       776       787       796
       ask    credit     state    system   process      hope    london      just
       797       798       806       816       826       828       828       830
    receiv     chang    review   current     shall    friday      team     phone
       830       831       834       841       844       847       850       858
      issu      data     avail      last      good      give       www       gas
       865       868       872       874       876       883       897       905
      list     posit     visit    includ     resum      best     offic    servic
       907       917       920       924       928       933       935       942
      talk    number      well       fax    provid      sent     next.      send
       943       951       961       963       970       971       975       986
      http      john    univers    financ   stinson   schedul      take      date
      1009      1022      1025      1038      1051      1054      1057      1060
      want  question   program     think      X713  crenshaw    attach     trade
      1068      1069      1080      1084      1097      1115      1155      1167
      help     email    compani   request       see  communic    confer   discuss
      1168      1201      1225      1227      1238      1251      1264      1270
      make   contact    follow interview   project      mail   present      busi
      1281      1301      1308      1320      1328      1352      1397      1416
  interest    option       day      call       one      year      week    messag
      1429      1432      1440      1497      1516      1523      1527      1538
   houston      also      look       edu      corp   shirley   develop       get
      1577      1604      1607      1620      1643      1687      1691      1768
       new       use       let    regard    inform      need     power       may
      1777      1784      1856      1859      1883      1890      1972      1976
      like      risk    energi    market     model     price      work     manag
```

```
      1980       2097       2124       2150       2170       2191       2293       2334
      know      group       meet       time   research    forward      X2001        can
      2345       2474       2544       2552       2752       2952       3060       3426
     thank        com      pleas   kaminski      X2000        hou       will       vinc
      3558       4444       4494       4801       4935       5569       6802       8531
   subject        ect      enron
      8625      11417      13388
> # which(colSums(subset(emailsSparse, emailsSparse$spam == 0)) > 5000)


> #j)
> sort(colSums(subset(emailsSparse,spam==1)))
     enron   kaminski       X713   crenshaw     vkamin     gibner    stinson       vinc
         0          0          0          0          0          0          0          1
      X853      kevin        doc    shirley      deriv    houston      april      resum
         1          2          2          2          3          5          5          5
       edu     friday  wednesday        hou        ect     arrang  interview     london
         7          7          8          8         10         11         13         15
    attend     robert    student    schedul   thursday     monday       john    tuesday
        15         16         16         17         17         19         20         20
    attach    suggest    appreci       mark    comment      begin    analysi      X2001
        21         21         23         25         26         26         27         29
     model    mention       hope      X2000     togeth      invit     confer    univers
        29         30         30         32         32         33         33         34
    financ       talk     either        run      shall       morn      happi    thought
        35         38         39         39         40         40         42         42
    depart    confirm    respond     school       hear       corp        etc      howev
        46         47         48         48         49         49         49         49
     sorri       idea     energi    discuss       open     option understand       soon
        50         51         55         56         56         56         57         57
    experi      cours     associ      point      bring   director   particip       join
        59         59         62         62         63         65         65         66
     anoth      still   research      final        set     specif       case      given
        66         66         68         68         69         69         69         70
      juli    problem        put        ask    alreadi        fax        abl       deal
        71         73         73         74         74         75         75         75
      team       book      locat       issu       meet      updat        lot     sincer
        76         76         79         79         79         79         80         80
     short     better       sinc       done     recent   question    possibl        end
        82         82         82         83         83         83         84         85
  contract       move      might       data    continu       note    resourc      sever
        85         86         87         87         88         88         90         90
      feel       area   communic     realli        due     origin     direct       unit
        90         92         92         93         94         96         96         97
```

copi	long	member	sure	allow	dear	write	public
97	98	99	99	102	104	104	104
event	let	differ	file	involv	respons	creat	type
105	107	109	111	111	113	114	114
effort	approv	detail	request	intern	say	import	support
115	115	115	117	117	118	119	120
relat	part	assist	two	last	back	keep	addit
121	121	123	124	124	125	125	126
date	place	group	mean	valu	think	offic	read
127	128	130	131	131	132	133	134
immedi	check	hello	applic	tri	review	phone	believ
136	137	139	139	140	142	143	143
hour	power	present	process	corpor	oper	full	return
144	145	146	149	151	151	152	154
sent	come	opportun	real	repli	line	engin	term
155	155	158	158	158	159	160	161
credit	well	gas	info	plan	risk	next.	increas
162	164	165	165	166	170	170	171
access	give	thank	version	requir	link	cost	great
172	172	172	174	174	174	175	182
wish	regard	posit	thing	call	develop	much	complet
185	186	187	188	190	191	192	192
even	project	form	design	without	expect	person	trade
193	194	196	196	198	198	198	199
buy	effect	rate	base	find	current	first	chang
199	201	201	202	202	203	203	204
visit	financi	high	mani	forward	good	special	success
206	207	208	208	209	221	225	226
don	per	number	week	result	web	industri	made
226	230	231	231	237	238	239	242
contact	follow	month	right	today	also	internet	help
242	244	249	249	251	260	262	262
manag	know	way	state	avail	futur	home	start
266	269	278	280	280	282	285	300
system	take	net	includ	life	see	name	onlin
302	304	305	314	320	329	344	345
within	remov	best	program	peopl	custom	year	like
346	357	358	358	359	363	367	372
interest	send	servic	look	work	day	want	product
385	393	395	396	415	420	420	421
www	account	provid	need	softwar	messag	site	address
426	428	435	438	440	445	455	461
may	list	price	new	websit	report	secur	just
489	503	503	504	506	507	520	524

```
     offer     invest      order        use      click       X000        now        one
       528        540        541        546        552        560        575        592
      time     market       http       make       free      pleas      money        get
       593        600        600        603        606        619        662        694
    receiv     inform        can      email       busi       mail        com    compani
       727        818        831        865        897        917        999       1065
      spam       will    subject
      1368       1450       1577
> #which(colSums(subset(emailsSparse, emailsSparse$spam == 1)) > 1000)


> #m)
> emailsSparse$spam <- as.factor(emailsSparse$spam)
> library(caTools)
> set.seed(123)
> spl <- sample.split(emailsSparse$spam, .7)
> train <- subset(emailsSparse, spl == TRUE)
> test <- subset(emailsSparse, spl == FALSE)
> spamLog <- glm(spam ~ ., data = train, family = "binomial")
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
> library(rpart)
> set.seed(123)
> spamCART <- rpart(spam ~ ., data = train)
> library(randomForest)
> set.seed(123)
> spamRF <- randomForest(spam ~ ., data = train)
>
> predictLog <- predict(spamLog, newdata = train, type = "response")
> predictCART <- predict(spamCART, newdata = train)
> predictRF <- predict(spamRF, newdata = train, type = "prob")
> #predictNB <- predict(spamNB, newdata = train, type = "class")
> predictCART <- predictCART[,2]
> predictRF <- predictRF[,2]
> table(predictLog < 0.0001)

FALSE  TRUE
  964  3046

> table(predictLog > .9999)

FALSE  TRUE
 3056   954
```

```
> table(predictLog >= 0.0001 & predictLog <= .9999)

FALSE  TRUE
 4000    10


> #n)
> summary(spamLog)

Call:
glm(formula = spam ˜ ., family = "binomial", data = train)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.011   0.000   0.000   0.000   1.354

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.082e+01  1.055e+04  -0.003    0.998
busi        -4.803e+00  1.000e+04   0.000    1.000
chang       -2.717e+01  2.215e+04  -0.001    0.999
compani      4.781e+00  9.186e+03   0.001    1.000
corpor      -8.286e-01  2.818e+04   0.000    1.000
day         -6.100e+00  5.866e+03  -0.001    0.999
done         6.828e+00  1.882e+04   0.000    1.000
effect       1.948e+01  2.100e+04   0.001    0.999
effort       1.606e+01  5.670e+04   0.000    1.000
even        -1.654e+01  2.289e+04  -0.001    0.999
full         2.125e+01  2.190e+04   0.001    0.999
good         5.399e+00  1.619e+04   0.000    1.000
inform       2.078e+01  8.549e+03   0.002    0.998
interest     2.698e+01  1.159e+04   0.002    0.998
list        -8.692e+00  2.149e+03  -0.004    0.997
look        -7.031e+00  1.563e+04   0.000    1.000
made         2.820e+00  2.743e+04   0.000    1.000
make         2.901e+01  1.528e+04   0.002    0.998
manag        6.014e+00  1.445e+04   0.000    1.000
market       7.895e+00  8.012e+03   0.001    0.999
much         3.775e-01  1.392e+04   0.000    1.000
order        6.533e+00  1.242e+04   0.001    1.000
origin       3.226e+01  3.818e+04   0.001    0.999
product      1.016e+01  1.345e+04   0.001    0.999
provid       2.422e-01  1.859e+04   0.000    1.000
```

```
realli      -2.667e+01  4.640e+04  -0.001    1.000
result      -5.002e-01  3.140e+04   0.000    1.000
see         -1.120e+01  1.293e+04  -0.001    0.999
special      1.777e+01  2.755e+04   0.001    0.999
subject      3.041e+01  1.055e+04   0.003    0.998
system       3.778e+00  9.149e+03   0.000    1.000
use         -1.385e+01  9.382e+03  -0.001    0.999
websit      -2.563e+01  1.848e+04  -0.001    0.999
will        -1.119e+01  5.980e+03  -0.002    0.999
within       2.900e+01  2.163e+04   0.001    0.999
without      1.942e+01  1.763e+04   0.001    0.999
continu      1.487e+01  1.535e+04   0.001    0.999
group        5.264e-01  1.037e+04   0.000    1.000
like         5.649e+00  7.660e+03   0.001    0.999
trade       -1.755e+01  1.483e+04  -0.001    0.999
tri          9.278e-01  1.282e+04   0.000    1.000
approv      -1.302e+00  1.589e+04   0.000    1.000
ask         -7.746e+00  1.976e+04   0.000    1.000
complet     -1.363e+01  2.024e+04  -0.001    0.999
credit       2.617e+01  1.314e+04   0.002    0.998
form         8.483e+00  1.674e+04   0.001    1.000
hear         2.887e+01  2.281e+04   0.001    0.999
home         5.973e+00  8.965e+03   0.001    0.999
new          1.003e+00  1.009e+04   0.000    1.000
offer        1.174e+01  1.084e+04   0.001    0.999
opportun    -4.131e+00  1.918e+04   0.000    1.000
rate        -3.112e+00  1.319e+04   0.000    1.000
take         5.731e+00  1.716e+04   0.000    1.000
time        -5.921e+00  8.335e+03  -0.001    0.999
visit        2.585e+01  1.170e+04   0.002    0.998
want        -2.555e+00  1.106e+04   0.000    1.000
way          1.339e+01  1.138e+04   0.001    0.999
addit        1.463e+00  2.703e+04   0.000    1.000
click        1.376e+01  7.077e+03   0.002    0.998
com          1.936e+00  4.039e+03   0.000    1.000
fax          3.537e+00  3.386e+04   0.000    1.000
mail         7.584e+00  1.021e+04   0.001    0.999
messag       1.716e+01  2.562e+03   0.007    0.995
now          3.790e+01  1.219e+04   0.003    0.998
phone       -6.957e+00  1.172e+04  -0.001    1.000
request     -1.232e+01  1.167e+04  -0.001    0.999
version     -3.606e+01  2.939e+04  -0.001    0.999
best        -8.201e+00  1.333e+03  -0.006    0.995
end         -1.311e+01  2.938e+04   0.000    1.000
```

```
get          5.154e+00  9.737e+03   0.001    1.000
great        1.222e+01  1.090e+04   0.001    0.999
money        3.264e+01  1.321e+04   0.002    0.998
softwar      2.575e+01  1.059e+04   0.002    0.998
custom       1.829e+01  1.008e+04   0.002    0.999
hello        2.166e+01  1.361e+04   0.002    0.999
one          1.241e+01  6.652e+03   0.002    0.999
onlin        3.589e+01  1.665e+04   0.002    0.998
pleas       -7.961e+00  9.484e+03  -0.001    0.999
access      -1.480e+01  1.335e+04  -0.001    0.999
account      2.488e+01  8.165e+03   0.003    0.998
allow        1.899e+01  6.436e+03   0.003    0.998
alreadi     -2.407e+01  3.319e+04  -0.001    0.999
also         2.990e+01  1.378e+04   0.002    0.998
applic      -2.649e+00  1.674e+04   0.000    1.000
area         2.041e+01  2.266e+04   0.001    0.999
assist      -1.128e+01  2.490e+04   0.000    1.000
base        -1.354e+01  2.122e+04  -0.001    0.999
believ       3.233e+01  2.136e+04   0.002    0.999
buy          4.170e+01  3.892e+04   0.001    0.999
can          3.762e+00  7.674e+03   0.000    1.000
cost        -1.938e+00  1.833e+04   0.000    1.000
creat        1.338e+01  3.946e+04   0.000    1.000
current      3.629e+00  1.707e+04   0.000    1.000
design      -7.923e+00  2.939e+04   0.000    1.000
develop      5.976e+00  9.455e+03   0.001    0.999
differ      -2.293e+00  1.075e+04   0.000    1.000
director    -1.770e+01  1.793e+04  -0.001    0.999
discuss     -1.051e+01  1.915e+04  -0.001    1.000
due         -4.163e+00  3.532e+04   0.000    1.000
email        3.833e+00  1.186e+04   0.000    1.000
event        1.694e+01  1.851e+04   0.001    0.999
expect      -1.179e+01  1.914e+04  -0.001    1.000
file        -2.943e+01  2.165e+04  -0.001    0.999
forward     -3.484e+00  1.864e+04   0.000    1.000
futur        4.146e+01  1.439e+04   0.003    0.998
gas         -3.901e+00  4.160e+03  -0.001    0.999
give        -2.518e+01  2.130e+04  -0.001    0.999
given       -2.186e+01  5.426e+04   0.000    1.000
high        -1.982e+00  2.554e+04   0.000    1.000
import      -1.859e+00  2.236e+04   0.000    1.000
includ      -3.454e+00  1.799e+04   0.000    1.000
increas      6.476e+00  2.329e+04   0.000    1.000
industri    -3.160e+01  2.373e+04  -0.001    0.999
```

```
invest       3.201e+01  2.393e+04   0.001    0.999
involv       3.815e+01  3.315e+04   0.001    0.999
just        -1.021e+01  1.114e+04  -0.001    0.999
know         1.277e+01  1.526e+04   0.001    0.999
locat        2.073e+01  1.597e+04   0.001    0.999
mani         1.885e+01  1.442e+04   0.001    0.999
may         -9.434e+00  1.397e+04  -0.001    0.999
mean         6.078e-01  2.952e+04   0.000    1.000
mention     -2.279e+01  2.714e+04  -0.001    0.999
might        1.244e+01  1.753e+04   0.001    0.999
month       -3.727e+00  1.112e+04   0.000    1.000
need         8.437e-01  1.221e+04   0.000    1.000
note         1.446e+01  2.294e+04   0.001    0.999
number      -9.622e+00  1.591e+04  -0.001    1.000
offic       -1.344e+01  2.311e+04  -0.001    1.000
oper        -1.696e+01  2.757e+04  -0.001    1.000
person       1.870e+01  9.575e+03   0.002    0.998
posit       -1.543e+01  2.316e+04  -0.001    0.999
possibl     -1.366e+01  2.492e+04  -0.001    1.000
present     -6.163e+00  1.278e+04   0.000    1.000
price        3.428e+00  7.850e+03   0.000    1.000
problem      1.262e+01  9.763e+03   0.001    0.999
process     -2.957e-01  1.191e+04   0.000    1.000
project      2.173e+00  1.497e+04   0.000    1.000
read        -1.527e+01  2.145e+04  -0.001    0.999
relat       -5.114e+01  1.793e+04  -0.003    0.998
report      -1.482e+01  1.477e+04  -0.001    0.999
requir       5.004e-01  2.937e+04   0.000    1.000
research    -2.826e+01  1.553e+04  -0.002    0.999
resourc     -2.735e+01  3.522e+04  -0.001    0.999
return       1.745e+01  1.844e+04   0.001    0.999
review      -4.825e+00  1.013e+04   0.000    1.000
risk        -4.001e+00  1.718e+04   0.000    1.000
secur       -1.604e+01  2.201e+03  -0.007    0.994
servic      -7.164e+00  1.235e+04  -0.001    1.000
set         -9.353e+00  2.627e+04   0.000    1.000
short       -8.974e+00  1.721e+04  -0.001    1.000
specif      -2.337e+01  3.083e+04  -0.001    0.999
state        1.221e+01  1.677e+04   0.001    0.999
term         2.013e+01  2.303e+04   0.001    0.999
thing        2.579e+01  1.341e+04   0.002    0.998
today       -1.762e+01  1.965e+04  -0.001    0.999
two         -2.573e+01  1.844e+04  -0.001    0.999
understand   9.307e+00  2.342e+04   0.000    1.000
```

```
unit        -4.020e+00  3.008e+04   0.000    1.000
well        -2.222e+01  9.713e+03  -0.002    0.998
work        -1.099e+01  1.160e+04  -0.001    0.999
hour         2.478e+00  1.333e+04   0.000    1.000
lot         -1.964e+01  1.321e+04  -0.001    0.999
real         2.046e+01  2.358e+04   0.001    0.999
right        2.312e+01  1.590e+04   0.001    0.999
start        1.437e+01  1.897e+04   0.001    0.999
X000         1.474e+01  1.058e+04   0.001    0.999
X2001       -3.215e+01  1.318e+04  -0.002    0.998
follow       1.766e+01  3.080e+03   0.006    0.995
name         1.672e+01  1.322e+04   0.001    0.999
sent        -1.488e+01  2.195e+04  -0.001    0.999
last         1.046e+00  1.372e+04   0.000    1.000
avail        8.651e+00  1.709e+04   0.001    1.000
first       -4.666e-01  2.043e+04   0.000    1.000
http         2.528e+01  2.107e+04   0.001    0.999
join        -3.824e+01  2.334e+04  -0.002    0.999
line         8.743e+00  1.236e+04   0.001    0.999
next.        1.492e+01  1.724e+04   0.001    0.999
remov        2.325e+01  2.484e+04   0.001    0.999
repli        1.538e+01  2.916e+04   0.001    1.000
wish         1.173e+01  3.175e+04   0.000    1.000
www         -7.867e+00  2.224e+04   0.000    1.000
year        -1.010e+01  1.039e+04  -0.001    0.999
back        -1.323e+01  2.272e+04  -0.001    1.000
internet     8.749e+00  1.100e+04   0.001    0.999
member       1.381e+01  2.343e+04   0.001    1.000
receiv       5.765e-01  1.585e+04   0.000    1.000
site         8.689e+00  1.496e+04   0.001    1.000
anoth       -8.744e+00  2.032e+04   0.000    1.000
associ       9.049e+00  1.909e+04   0.000    1.000
comment     -3.251e+00  3.387e+04   0.000    1.000
corp         1.606e+01  2.708e+04   0.001    1.000
date        -2.786e+00  1.699e+04   0.000    1.000
find        -2.623e+00  9.727e+03   0.000    1.000
free         6.113e+00  8.121e+03   0.001    0.999
issu        -3.708e+01  3.396e+04  -0.001    0.999
long        -1.489e+01  1.934e+04  -0.001    0.999
move        -3.834e+01  3.011e+04  -0.001    0.999
particip    -1.154e+01  1.738e+04  -0.001    0.999
recent      -2.067e+00  1.780e+04   0.000    1.000
respons     -1.960e+01  3.667e+04  -0.001    1.000
say          7.366e+00  2.217e+04   0.000    1.000
```

```
week        -6.795e+00  1.046e+04  -0.001    0.999
dear        -2.313e+00  2.306e+04   0.000    1.000
regard      -3.668e+00  1.511e+04   0.000    1.000
thank       -3.890e+01  1.059e+04  -0.004    0.997
address     -4.613e+00  1.113e+04   0.000    1.000
contact      1.530e+00  1.262e+04   0.000    1.000
engin        2.664e+01  2.394e+04   0.001    0.999
etc          9.470e-01  1.569e+04   0.000    1.000
immedi       6.285e+01  3.346e+04   0.002    0.999
net          1.256e+01  2.197e+04   0.001    1.000
per          1.367e+01  1.273e+04   0.001    0.999
place        9.005e+00  3.661e+04   0.000    1.000
respond      2.974e+01  3.888e+04   0.001    0.999
sincer      -2.073e+01  3.515e+04  -0.001    1.000
type        -1.447e+01  2.755e+04  -0.001    1.000
come        -1.166e+00  1.511e+04   0.000    1.000
confirm     -1.300e+01  1.514e+04  -0.001    0.999
analysi     -2.405e+01  3.860e+04  -0.001    1.000
bring        1.607e+01  6.767e+04   0.000    1.000
call        -1.145e+00  1.111e+04   0.000    1.000
data        -2.609e+01  2.271e+04  -0.001    0.999
detail       1.197e+01  2.301e+04   0.001    1.000
happi        1.939e-02  1.202e+04   0.000    1.000
idea        -1.845e+01  3.892e+04   0.000    1.000
info        -1.255e+00  4.857e+03   0.000    1.000
send        -2.427e+01  1.222e+04  -0.002    0.998
success      4.344e+00  2.783e+04   0.000    1.000
sure        -5.503e+00  2.078e+04   0.000    1.000
team         7.940e+00  2.570e+04   0.000    1.000
web          2.791e+00  1.686e+04   0.000    1.000
don          2.129e+01  1.456e+04   0.001    0.999
copi        -4.274e+01  3.070e+04  -0.001    0.999
help         1.731e+01  2.791e+03   0.006    0.995
part         4.594e+00  3.483e+04   0.000    1.000
life         5.812e+01  3.864e+04   0.002    0.999
meet        -1.063e+00  1.263e+04   0.000    1.000
sever        2.041e+01  3.093e+04   0.001    0.999
question    -3.467e+01  1.859e+04  -0.002    0.999
write        4.406e+01  2.825e+04   0.002    0.999
think       -1.218e+01  2.077e+04  -0.001    1.000
point        5.498e+00  3.403e+04   0.000    1.000
let         -2.763e+01  1.462e+04  -0.002    0.998
link        -6.929e+00  1.345e+04  -0.001    1.000
communic     1.580e+01  8.958e+03   0.002    0.999
```

```
contract    -1.295e+01  1.498e+04  -0.001    0.999
either      -2.744e+01  4.000e+04  -0.001    0.999
final        8.075e+00  5.008e+04   0.000    1.000
howev       -3.449e+01  3.562e+04  -0.001    0.999
peopl       -1.864e+01  1.439e+04  -0.001    0.999
 [ reached getOption("max.print") -- omitted 81 rows ]

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4409.49  on 4009  degrees of freedom
Residual deviance:   13.46  on 3679  degrees of freedom
AIC: 675.46

Number of Fisher Scoring iterations: 25
```

```
> #o)
> library(rpart.plot)
> prp(spamCART)
```

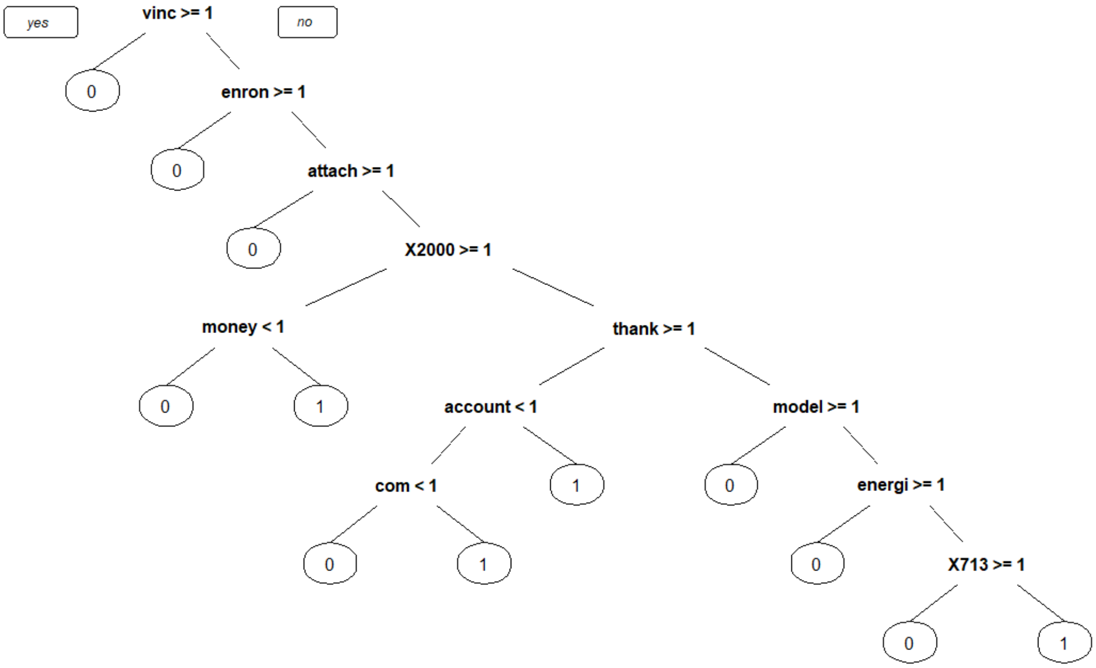
```
> #p)
> table(predictLog >= .5, train$spam)

           0    1
  FALSE 3052    4
  TRUE     0  954

> library(ROCR)
> predictLog1 <- prediction(predictLog, train$spam)
> perfLog1 <- performance(predictLog1, measure = "auc")
> perfLog1
An object of class "performance"
Slot "x.name":
[1] "None"

Slot "y.name":
[1] "Area under the ROC curve"

Slot "alpha.name":
[1] "none"

Slot "x.values":
```

```
list()

Slot "y.values":
[[1]]
[1] 0.9999959


Slot "alpha.values":
list()



> #q)
> table(predictCART >= .5, train$spam)

           0    1
  FALSE 2885   64
  TRUE   167  894


> #r)
> predictCART1 <- prediction(predictCART, train$spam)
> perfCART1 <- performance(predictCART1, measure = "auc")
> perfCART1
An object of class "performance"
Slot "x.name":
[1] "None"

Slot "y.name":
[1] "Area under the ROC curve"

Slot "alpha.name":
[1] "none"

Slot "x.values":
list()

Slot "y.values":
[[1]]
[1] 0.9696044


Slot "alpha.values":
list()
```

```
> #s)
> table(predictRF >= .5, train$spam)

           0    1
  FALSE 3046    0
  TRUE     6  958


> #t)
> predictRF1 <- prediction(predictRF, train$spam)
> perfRF1 <- performance(predictRF1, measure = "auc")
> perfRF1
An object of class "performance"
Slot "x.name":
[1] "None"

Slot "y.name":
[1] "Area under the ROC curve"

Slot "alpha.name":
[1] "none"

Slot "x.values":
list()

Slot "y.values":
[[1]]
[1] 0.9999959


Slot "alpha.values":
list()



> #v)
> predLogtest <- predict(spamLog, newdata = test, type = "response")
> predCARTtest <- predict(spamCART, newdata = test)
> predRFtest <- predict(spamRF, newdata = test, type = "prob")
> predCARTtest <- predCARTtest[,2]
> predRFtest <- predRFtest[,2]
> table(predLogtest >= .5, test$spam)

           0    1
  FALSE 1257   34
  TRUE    51  376


> #w)
> predictLog2 <- prediction(predLogtest, test$spam)
> perfLog2 <- performance(predictLog2, measure = "auc")
> perfLog2
An object of class "performance"
Slot "x.name":
[1] "None"

Slot "y.name":
[1] "Area under the ROC curve"

Slot "alpha.name":
[1] "none"

Slot "x.values":
list()

Slot "y.values":
[[1]]
[1] 0.9627517


Slot "alpha.values":
list()

>
> table(predCARTtest >= .5, test$spam)

           0    1
  FALSE 1228   24
  TRUE    80  386

> predictCART2 <- prediction(predCARTtest, test$spam)
> perfCART2 <- performance(predictCART2, measure = "auc")
> perfCART2
An object of class "performance"
Slot "x.name":
```

```
[1] "None"

Slot "y.name":
[1] "Area under the ROC curve"

Slot "alpha.name":
[1] "none"

Slot "x.values":
list()

Slot "y.values":
[[1]]
[1] 0.963176


Slot "alpha.values":
list()

>
> table(predRFtest >= .5, test$spam)

          0    1
  FALSE 1290   25
  TRUE    18  385

> predictRF2 <- prediction(predRFtest, test$spam)
> perfRF2 <- performance(predictRF2, measure = "auc")
> perfRF2
An object of class "performance"
Slot "x.name":
[1] "None"

Slot "y.name":
[1] "Area under the ROC curve"

Slot "alpha.name":
[1] "none"

Slot "x.values":
list()

Slot "y.values":
[[1]]
```

```
[1] 0.997768


Slot "alpha.values":
list()
```



Plot for Q1o. Click here to go back to the question.