Tree, Bagging, Random Forests and Boosting

Only used for predictions

not CI / Testing / significant variables.

- Classification trees ← Leo Breiman
- Boosted Random Forests ← run a lot of trees and find one that is not in the forest but better than all existing trees in the forests

Non-parametric

Arm Bended Problem

Two-class Classifications

If tree is small ⇒ good prediction properties.

⋮ ⇒ how to determine when ⇒ in R package
to stop

huge ⇒ overfit

SPAM ⇒ R built-in

Sensitivity ⇒ proportion of true spam identified ↑ Type I ↑

Specificity ⇒ proportion of true email identified ↑ Power

Want both to be high

overfitting ⇒ could get high specificity & sensitivity

Freund → prof here

Decision Boundary : Tree

Model Averaging ⇒ Boosting > Random Forests > Bagging > Single Tree

Bagging

Bootstrap aggregation

⇒ bootstrap a thousand times

⇒ get a thousand trees

⇒ average the trees

Smoother decision boundaries

Random Forests:

⇒ Randomly choose m features (refinement)

then ↳ bagging ⇒ average the trees

⇒ high dimensional

Boosting

⇒ Bootstrap, take features

⇒ which data was predicted well ⇒ more likely ⇒ smaller weight

⇒ higher weight on misfit ⇒ weighted bootstrap ⇒ run trees.

⇒ which data predict well ⇒ reweight

↳ Loss function

pick the data you misclassified

⇒ won't be overfitting

run many forests

⇒ converged forests

Investigations

Big question:

What is the difference in weight between babies born to mothers who smoked during pregnancy and those who did not? Is this difference important to the health of the baby?

Instructions

- Summarize numerically the two distributions of birth weight for babies born to women who smoked during their pregnancy and for babies born to women who did not smoke during their pregnancy.
- Use graphical methods to compare the two distributions of birth weight.
- Compare the frequency, or incidence, of low-birth-weight babies for the two groups. How reliable do you think your estimates are? That is, how would the incidence of low birth weight change if a few more or fewer babies were classified as low birth weight?
- Asses the importance of the difference you found in your three types of comparisons (numerical, graphical, incidence). Summarize your findings and relate them to other studies.

qqplot, histogram

# of times something happened

Add noise to it? Don't change too much to check

How robust is the estimator?

How reliable the estimator is?

Discussion

↳ returns back

Simulate

↳ Illustrate whether this is a good estimator?

only use words

Note: If you make separate plots for smokers and nonsmokers, be sure to scale the axes identically for both graphs.

The objective of this study is to investigate the responses of the participants in the study with the intention of providing useful information about the students to the designers of the new computer lab?

* [Scenario 1:] Begin by providing an estimate for the fraction of students who played a video game in the week prior to the survey. Provide an interval estimate as well as a point estimate for this proportion.
* [Scenario 2:] Check to see how the amount of time spent playing vide games in the week prior to the survey compares to the reported frequency of play (daily, weekly, etc). How might the fact that there was an exam in the week prior to the survey affect your previous estimates and this comparison?
* [Scenario 3:] Consider making an internal estimate for the average amount of time spent playing video games in the week prior to the survey. Keep in mind the overall shape of the sample distribution. A simulation study may help determine the appropriateness of an interval estimate.
* [Scenario 4:] Next consider the "attitude" questions. In general, do you think the students enjoy playing video games? If you had to make a short list of the most important reasons why students like/dislike video games, what would you put on the list? Don't forget that those students who say that they have never played video games or do not at all like video games are asked to skip over some of these questions. So, there may be many nonrespondents to the questions as to whether they think video games are educational, where they play video games, etc.

The objective of this study is to investigate the responses of the participants in the study with the intention of providing useful information about the students to the designers of the new computer lab?

* [Scenario 1:] Begin by providing an estimate for the fraction of students who played a video game in the week prior to the survey. Provide an interval estimate as well as a point estimate for this proportion.
* [Scenario 2:] Check to see how the amount of time spent playing vide games in the week prior to the survey compares to the reported frequency of play (daily, weekly, etc). How might the fact that there was an exam in the week prior to the survey affect your previous estimates and this comparison?
* [Scenario 3:] Consider making an interval estimate for the average amount of time spent playing video games in the week prior to the survey. Keep in mind the overall shape of the sample distribution. A simulation study may help determine the appropriateness of an interval estimate.
* [Scenario 4:] Next consider the "attitude" questions. In general, do you think the students enjoy playing video games? If you had to make a short list of the most important reasons why students like/dislike video games, what would you put on the list? Don't forget that those students who say that they have never played video games or do not at all like video games are asked to skip over some of these questions. So, there may be many nonrespondents to the questions as to whether they think video games are educational, where they play video games, etc.

The objective of this study is to investigate the responses of the participants in the study with the intention of providing useful information about the students to the designers of the new computer lab?

Compare two distributions → same tools as HW1

represent data in tables

2x2 maps
3x3

* [Scenario 5:] Look for the differences between those who like to play video games and those who don't. To do this, use the questions in the last part of the survey, and make comparisons between male and female students, those who work for pay and those who don't, those who own a computer and those who don't. Graphical display and cross-tabulations are particularly helpful in making these kinds of comparisons. Also, you may want to collapse the range of responses to a question down to two or three possibilities before making these comparisons.

Table ⇒ best way to represent

CI

Table

The objective of this study is to investigate the responses of the participants in the study with the intention of providing useful information about the students to the designers of the new computer lab?

Summary of HW1

* [Scenario 5:] Look for the differences between those who like to play video games and those who don't. To do this, use the questions in the last part of the survey, and make comparisons between male and female students, those who work for pay and those who don't, those who own a computer and those who don't. Graphical display and cross-tabulations are particularly helpful in making these kinds of comparisons. Also, you may want to collapse the range of responses to a question down to two or three possibilities before making these comparisons.

EC

Optional

* [Scenario 6:] Just for fun, further investigate the grade that students expect in the course. How will does it match the target distribution used in grade assignment of 20% A's, 30%B's,40% c's and 10%D's or lower? If the nonrespondents were failing students who no longer bothered to come to the discussion section, would this change the picture ?

Introduction
The data
Background
Investigations
Probability Theory
Goals
The Probability Model
Sample Statistics
Estimators for Standard Error
Population Total and Percentages
Normal Approximation and Confidence Intervals
An Example
An Example
The Bootstrap

Design stats lab

reason why we have data

from this finite population is for randomness

being in the sample ⇒ dependent

independence.
knowing info for one sample doesn't change the chance of other sample being selected in the data

because finite population ⇒ dependence
↳ sample of finite population

depends on the probability model. how good is the estimate

once you know one is in the sample, the chances of knowing others are in is reduced

chances of being in the sample

① how to estimate, quality of the estimate
② CI, why they work
③ how to correct them for finite population
④ simulations bootstrap ← data drawn from data
↑ finite population

survey data ⇒ independent?
↳ simple, random sample?
each column doesn't change other column ← independence
As a collection of individual
↳ independent from each other?

Jelena Bradic http://www.jelenabradic.net Assistant Professor, Department of Mathematics, University of California, San Diego jbradic@ucsd.edu

Goals

In this section we will use as out primary example the problem of estimating the average amount of time students in the class spent playing video games in the week prior to the survey.

* To determine the exact amount of time for the entire class we would need to interview all of the students (over 3000 of them).
* Alternatively, a subset of them can be interviewed, and the information collected from this subset could provide an approximation to the full group.
  * In this section we discuss one rule for selecting a subset of student to be surveyed, the simple random sample.
  * The simple random sample is a probability method for selecting the students..
  * Probability methods are useful because through chance we can make useful statements about the relation between the sample and the entire group.
  * With a probability method, we know the chance of each possible sample.

Goals (cont.)

Terminology

* Population units make up the group that we want to know more about
  * In this lab, the units are the students enrolled in the 1994 Fall semester class of Introductory Probability and Statistics.
* Population size, usually denoted by N, is the total number of units in the population. For very large population, often the exact size of the population is not known. Here we have 314 students in the class.
* Unit characteristic is a particular piece of information about each member of the population.
  * The characteristic that interests us in our example is the amount of time the student played video games in the week prior to the survey.
* Population parameter is a summary of the characteristic for all units in the population, such as the average value of the characteristic.
  * The population parameter of interests to us here is the average amount of time students in the class spent playing video games in the week prior to the survey.

Goals (cont.)

Terminology

* Sample units are those members of the population selected for the sample.
* Sample size usually denoted by n, is the number of units chosen for the sample. We will use 91 for our sample size, and ignore the four who did not respond.
* Sample statistic is a numerical summary of the characteristic of the units sampled. The statistic estimated the population parameter. Since the population parameter in our example is the average time spent playing video games by all students in the class in the week prior to the survey, a reasonable sample statistic is the average time spent playing video games by ll students in the sample.

Overall Questions

How do we find clusters of palindromes? How do we determine whether a cluster is just a chance occurrence or a potential replication site?

① Whether there is a cluster? → generate baseline → Uniform Distribution

② Whether that cluster is unusual?

* [**Random scatter**] To begin, pursue the point of view that structure in the data is indicated by departures from a uniform scatter of palindromes across the DNA.
  ** Of course, a random uniform scatter, does not mean that the palindromes will be equally spaced as milestones on a freeway. There will be some gaps on the DNA where no palindromes occur, and there will be some clumping together of palindromes.

  To look for structure examine the locations of the palindromes, the spacing between palindromes, and the counts of palindromes in non overlapping regions of the DNA. One starting place might be to see first how random scatter looks by using a computer to simulate it.

  ** A computer can simulate 296 palindrome sites chosen at random along a DNA sequence of 229,354 bases using a pseudo random number generator. When this is done several times,by making seller sets of simulated palindrome locations, then the real data can be compared to the simulated data.

How do we find clusters of palindromes? How do we determine whether a cluster is just a chance occurrence or a potential replication site?

* [Random scatter] To begin, pursue the point of view that structure in the data is indicated by departures from a uniform scatter of palindromes across the DNA.
  ** Of course, a random uniform scatter, does not mean that the palindromes will be equally spaced as milestones on a freeway. There will be some gaps on the DNA where no palindromes occur, and there will be some clumping together of palindromes.

  To look for structure examine the locations of the palindromes, the spacing between palindromes, and the counts of palindromes in non overlapping regions of the DNA. One starting place might be to see first how random scatter looks by using a computer to simulate it.

  ** A computer can simulate 296 palindrome sites chosen at random along a DNA sequence of 229,354 bases using a pseudo random number generator. When this is done several times,by making seller sets of simulated palindrome locations, then the real data can be compared to the simulated data.

Idea is to find the difference or a difference

Comparisons for two distributions (like HW1)

* [Locations and spacings] Use graphical methods to examine the spacings between consecutive palindromes and sum of consecutive pairs, triplets, etc, spacings. Compare what you find to what would you expect to find in a random scatter. Also, use graphical methods to compare locations of the palindromes.

How do we find clusters of palindromes? How do we determine whether a cluster is just a chance occurrence or a potential replication site?

try different length of interval ⇒ see if changing drastically

* [Counts] Use graphical methods and more formal statistical tests to examine the counts of palindromes in various regions of the DNA. Split the DNA into nonoverlaping regions of equal length to compare the number of palindromes in an interval to the number of that would you expect from uniform random scatter. The counts for shorter regions will be more variable than those for longer regions. Also consider classifying the regions according to their number of counts.
* [The biggest cluster] Does the interval with the greatest number of palindromes indicate a potential origin of replication?Be careful in making your intervals, for any small, but significant ,deviations from random scatter, such as a tight cluster of a few palindromes, could easily go undetected if the regions examined are too large. Also, if the regions are too small, a cluster of palindromes may be split between adjacent intervals and not appear as a high-count interval.

hypothesis testing.

is this bigger cluster statistically (significantly) difference than baseline (the one that is generated)

How would you advise biologist who is a bout to start experimentally searching for the origin of replication? Write your recommendations in the form of a report that a team members including biologist will read.

report ↘ to reduce budget

The aim of this lab is to provide a simple procedure for converting gain into density when the gauge is in operation. Keep in mind that the experiment was conducted by varying density and measuring the response in gain, but when the gauge is ultimately in use, the snow-pack density is to be estimated from the measured gain.

* [**Fitting**] Use the data to fit the gain, or a transformation of gain, to density. Try sketching the least squares line on a scatter plot.
  ** Do the residuals indicate any problems with the fit ?
  ** If the densities f the polyethylene blocks are not reported exactly, how might this affect the fit ?
  ** What if the blocs of polyethylene were not measured in random order?
* [**Predicting**] Ultimately we are interested in answering questions such as: Given a gain reading of 38.6, what is the density of the snow-pack ? or Given a gain reading of 426.7, what is the density of snow-pack? These two numeric values, 38.6 and 426.7, were chosen because they are the average gains for the 0.508 and 0.001 densities, respectively.
  ** Develop a procedure for adding bands around your least squares line that can be used to make interval estimates for the snow-pack density from gain measurements. Keep in mind how the data were collected: several measurements of gain were taken for polyenthylene blocks of known density.

How do we find clusters of palindromes? How do we determine whether a cluster is just a chance occurrence or a potential replication site?

* [**Cross-Validation**] To check how well your procedure works, omit the set of measurements corresponding to the block of density 0.508, apply your ”estimation”/calibration procedure to the remaining data, and provide an interval estimate for the density of a block with an average reading of 38.6. Where does the actual density fall in the interval? Try the same test, for the set f measurements at the 0.001 density.

Can you make your code available online ?

Question: How to connect the most donation requests to prospective donors? Which type of donation requests donors are most likely to respond to?