Medians and blobs

Prof. Ramin Zabih

http://cs100r.cs.cornell.edu

Cornell University
Computer Science

Administrivia

- Assignment 2 is out, due in 2 pieces
  - Small amount is due this Friday
  - Most of it is due next Friday
- Quiz 2 on Tuesday 9/18
  - Coverage through today's lecture

♦ What is the complexity of:

- Finding the 2nd biggest element (>all but 1)? The 3rd biggest (> all but 2)?
  - What do you think?
  - It's actually O(n)
  - We do 2 (or 3) "find biggest" operations
    - Each of which takes O(n) time
- Finding the element bigger than all but 5%?
  - Assume we do this by repeated "find biggest"
  - What if we use modified quicksort?

Modified quicksort

- This change to quicksort gives us a very practical way to find a particular element without actually sorting the array!
  - It's actually much faster, as you will see
- The worst case is still bad
- Well beyond CS100R: for random input this method actually runs in linear time
  - You can try random input and see this

Putting it all together

- By modifying quicksort we can find the 5% largest (or smallest) element
  - This allows us to efficiently compute the trimmed mean
    - Significantly faster than sorting
- It's possible to select in linear time (1973)
  - Rev. Dodgson's problem
  - But the code is a little messy
    - And the analysis is messier

http://en.wikipedia.org/wiki/Selection_algorithm

What about the median?

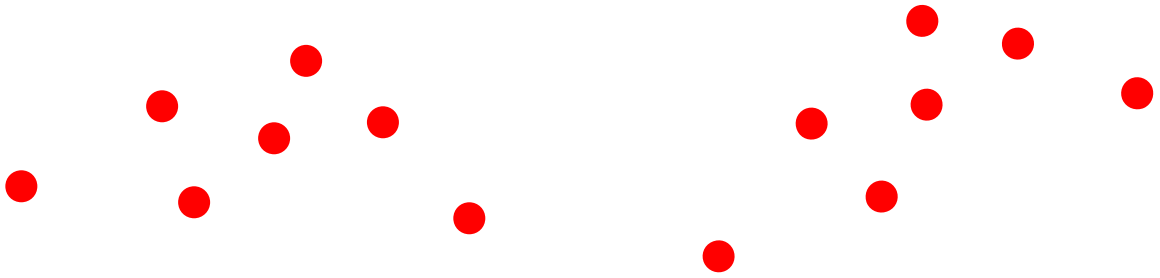
- Obvious way to avoid our bad data points:
  - Use the median instead of the mean
- The median of a set of 5 numbers is the 3rd largest (and thus the 3rd smallest)
- Mean, like median, was defined in 1D
  - For a 2D mean we used the centroid
    - I.e., we took the mean of the x coordinates and y coordinates "separately"
    - Call this the "mean vector"
    - Does this work for the median also?

What is the 2D median

- In 1900, statisticians wanted to find the "geographical center of the population"
  - In order to quantify the westward shift
- Why not the centroid?
  - Someone being born in San Francisco changes the centroid much more than someone being born in Indiana
- What about the "median vector"?
  - Take the median of the x coordinates and the median of the y coordinates separately

Median vector

- A little thought will show you that this doesn't really make a lot of sense
  - Nonetheless, it's a common solution, and we will implement it for CS100R
    - In situations like ours it works pretty well
- It's almost never an actual datapoint
- It depends upon rotations!



Can we do even better?

- None of what we described works that well if we have widely scattered red pixels
  - And we can't figure out lightstick orientation
- Is it possible to do even better?
  - Yes, and we'll spend a few more weeks on this
- In particular, we will focus on:
  - Finding "blobs" (connected red pixels)
  - Summarizing the shape of a blob
  - Computing orientation from this
- We'll need brand new tricks!

What is a blob?

1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	0	0	0	0	0	0

How to find blobs

- Pick a 1 to start with, where you don't know which blob it is in
  - You're done when there aren't any
- Give it a new blob number
- Give the same blob number to each pixel that is part of the same blob
  - But how do we figure this out?
  - You are part of blob N if you are next to someone who is part of blob N
    - But what exactly does "next to" mean?

What is a blob?

1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	1	0	0	0	0	0	0	0

What is a neighbor?

- We need a notion of neighborhood
  - Sometimes called a neighborhood system
- One possibility, the standard one, only considers vertical and horizontal neighbors
  - Sometimes called "NEWS"
    - North, east, west, south
  - 4-connected, since you have 4 neighbors
- Other possibility includes diagonals
  - 8-connected neighborhood system

The long winding road to blobs

- We actually need to cover a surprising amount of material to get to blob finding
  - Some of which is not obviously relevant
  - But (trust me) it will all hang together!

An amazingly useful concept

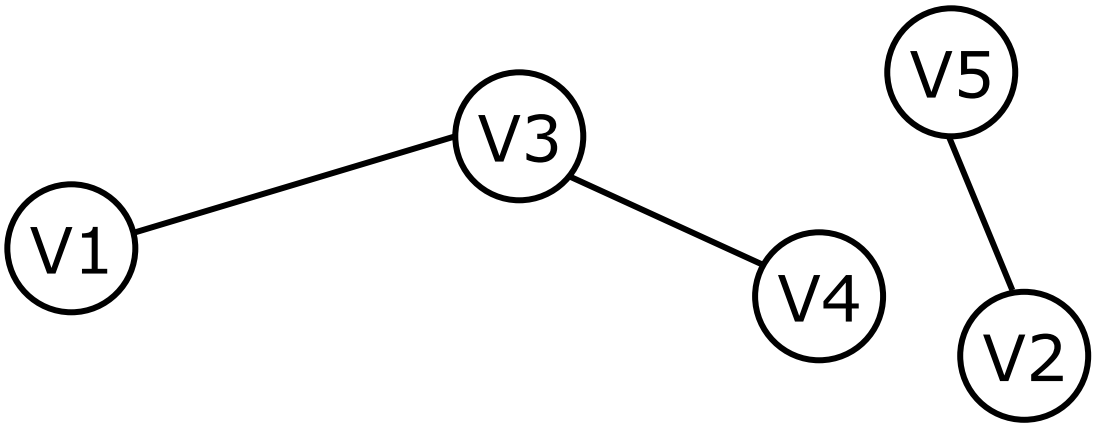
- A single idea can be used to think about:
  - Assigning frequencies to radio stations
  - Scheduling your classes so they don't conflict
  - Connecting computer chips on a motherboard
  - Figuring out if a chemical is already known
  - Finding groups in MySpace/Facebook
  - Searching for pages on the web
  - Determining how fragile the internet is
- ♦ Which one of these problems is related to finding blobs?

Graphs: always the answer

- We are going to look at an incredibly important concept called a graph
  - Note: not the same as a plot
- Nearly all CS professors do research that somehow involves graphs
  - Many math professors as well
- Most problems can be thought of in terms of graphs
  - But it may not be obvious, as with blobs
  - So, graph algorithms are very important

What is a graph?

- Loosely speaking, a set of things that are somehow paired up
- Precisely, a set of vertices V and edges E
  - Vertices sometimes called nodes
  - An edge (or link) connects a pair of vertices



Notes on graphs

- What can a graph represent?
  - Cities and direct flights
  - People and friendships
  - Web pages and hyperlinks
  - Rooms and doorways
  - IMAGES!!!
- A graph isn't changed by:
  - Drawing the edges differently
    - While preserving endpoints
  - Re-numbering the vertices

Problems, algorithms, programs

- A central distinction in CS
- Problem: what you want to compute
  - "Find the median"
  - Sometimes called a specification
- Algorithm: how to do it, in general
  - "Modified quicksort"
- Program: how to do it, in a particular programming language

```
function [med] = find_median[A]
...
```

Some major graph problems

- Graph coloring
  - Ensuring that radio stations don't clash
- Planarity testing
  - Connecting computer chips on a motherboard
- Graph isomorphism
  - Is a chemical structure already known?
- Graph cycles
  - Helping FedEx/UPS/DHL plan a route
- Graph connectivity
  - How fragile is the internet?