Admin Notes

• Reminder: Group portion of In the News #2 due on Tuesday
• Reminder: Project proposal resubmissions due on March 9
• Reminder: Midterm 2 coming up on March 13
• Reminder: Final scheduled to be on Apr 17 @ 3:30
• Data mining exercises released on the website
Today we’re going to start on the group component of In The News call #2

• Today we’re going to spend some time in class having your groups work on the group component
• Make sure that you’ve read the grade rubric: https://www.ugrad.cs.ubc.ca/~cs100/2017W1/in-the-news.html#rubric
• Make sure you comment on the CT Building Block, Application, and/or Impact!
• Picking an article/topic and then looking for related articles is both okay and encouraged
• You can pick an article that one of the people in your group did or any other article that has been posted
• Make sure that you cite articles that you use – pick any citation style – see discussion of Plagiarism on project page: https://www.ugrad.cs.ubc.ca/~cs100/2017W1/project.html#plagiarism
Greed, for lack of a better word, is good

- The algorithm that we used to create the decision tree is a **greedy algorithm**
- In a greedy algorithm, you make a choice that’s the optimal choice for now and hope that it’s the optimal choice in the long run
- Sometimes it’s the best in the long run, sometimes it’s not.
- In building a decision tree, greedy will not always be optimal – but it’s pretty good, and it’s much faster than an optimal approach
- In some problems you can prove that greedy can find the best solution!
Computational thinking in your life: homework

In a group, discuss your algorithms for how you decide what order to do your homework in and why you choose that order.

- Due date
- How long it’ll take to complete
- Do things for the classes you like
- How many marks you’d lose
Which algorithm is best requires knowing what you’re trying to optimize (the “why”)

In a group, design a greedy algorithm to reduce the length of your homework todo list as fast as possible

Hint: your algorithm should look like “always do the [property] remaining assignment next”

Do the shortest one first
Clicker question: is it optimal?

Just guess: is a correctly-written greedy algorithm for minimizing the length of your todo list by doing the shortest one next optimal?

A. Yes
B. No
Sometimes greedy algorithms can make optimal choices

Minimize maximal lateness (minimize amount of time things are overdue – useful if you get progressively larger percentages off for lateness)

Example: Time   Deadline

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Lateness 1

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Lateness 3

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Greedily doing the assignment with the closest due date first will minimize maximal lateness… but if you want to have different priorities for different classes, greedy is no longer optimal
The second type of data mining that we will look at in detail involves putting similar items together in groups.
Exercise: Group this!

Given the list of items below, put items together into groups. You can have as many groups as you want. Groups do not need to have the same number of items.
Exercise: Group this!

What kind of groups did you get? What criteria did you use to form each group?

- Symbol vs. No symbol
  (flag, police car, ambulance) vs. (Green car, cup, water bottle)
- Handle vs. No handle (everything but flag in one group vs. Flag)
- Colours (red, green, blue)
Exercise: Group this! - Possible Solution 1

Group 1: Things I use when going to school
- Car
- Mug
- Water bottle

Group 2: People I can call 911 to get
- Police car
- Ambulance

Group 3: Flag
- Canadian flag
Exercise: Group this! - Possible Solution 2

Group 1: Things that are green

- Car
- Mug

Group 2: Things that are blue

- Police car
- Water bottle

Group 3: Things that are red and white

- Canadian flag
- Ambulance
What is clustering?

Clustering is partitioning a set of items into subgroups so as to ensure certain measures of quality (e.g., “similar” items are grouped together)
Why cluster?

Netflix movie recommendations

“There’s a mountain of data that we have at our disposal,” says Todd Yellin, Netflix’s VP of product innovation. “That mountain is composed of two things. Garbage is 99 percent of that mountain. Gold is one percent… .”

Group exercise: What information about customers do you think that Netflix uses when deciding what movies to recommend?

- Age
- Genre
- Demographics (geography)
- Cast
Why cluster? Netflix movie recommendations

“There’s a mountain of data that we have at our disposal,” says Todd Yellin, Netflix’s VP of product innovation. “That mountain is composed of two things. Garbage is 99 percent of that mountain. Gold is one percent… .

Geography, age, and gender? We put that in the garbage heap. Where you live is not that important.”
Netflix group its tens of thousands of titles into a few thousand “clusters” based not on where people live, but what they like.

Netflix assigns each subscriber to a handful of these clusters, weighted by the degree to which each matches their taste. “When you have more than 75 million people around the world, you can get really specific about who’s your taste,” says Yellin.
Why cluster?

Netflix movie recommendations

The movies recommended to you are based on those that others in your clusters watch or recommend.

“We used to be more naive. We used to overexploit individual signals,” says Yellin. “If you watched a romantic comedy, years ago we would have overexploited that. The whole top of your screen would be more romantic comedies. Not a lot of variety. And that gets you into a quick cul-de-sac of too much content around one area.”
Why cluster?  

Netflix movie recommendations

A related problem: how to predict how users will rate a *new* movie?

Netflix has a competition with a 1 million dollar prize for algorithms that do this well. They provide training data: 100 million ratings generated by over 480 thousand users on over 17 thousand movies. Competitors use clustering (among other techniques) in their solutions.
Letters to Nature

*Nature* **415**, 530-536 (31 January 2002) | doi:10.1038/415530a; Received 24 August 2001; Accepted 22 November 2001

Gene expression profiling predicts clinical outcome of breast cancer

Laura J. van 't Veer\textsuperscript{1,2}, Hongyue Dai\textsuperscript{2,3}, Marc J. van de Vijver\textsuperscript{1,2}, Yudong D. He\textsuperscript{3}, Augustinus A. M. Hart\textsuperscript{1}, Mao Mao\textsuperscript{3}, Hans L. Peterse\textsuperscript{1}, Karin van der Kooy\textsuperscript{1}, Matthew J. Marton\textsuperscript{2}, Anke T. Witteveen\textsuperscript{1}, George J. Schreiber\textsuperscript{3}, Ron M. Kerkhoven\textsuperscript{1}, Chris Roberts\textsuperscript{3}, Peter S. Linsley\textsuperscript{3}, René Bernards\textsuperscript{1} & Stephen H. Friend\textsuperscript{3}
First, let’s define **Gene Expression**

When a gene is “on” and its protein or RNA product is being made, scientists say that the gene is being expressed.

The on and off states of all of a cell’s genes is known as a **gene expression profile**.

Each cell type has a unique gene expression profile.

**CELL X’s GENE EXPRESSION PROFILE:**

<table>
<thead>
<tr>
<th>Gene</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 90</td>
<td>OFF</td>
</tr>
<tr>
<td>Gene 91</td>
<td>OFF</td>
</tr>
<tr>
<td>Gene 92</td>
<td>ON</td>
</tr>
<tr>
<td>Gene 93</td>
<td>ON</td>
</tr>
<tr>
<td>Gene 94</td>
<td>OFF</td>
</tr>
</tbody>
</table>
“Breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome. [...] Chemotherapy or hormonal therapy reduces the risk of distant metastases by approximately one-third; however, 70–80% of patients receiving this treatment would have survived without it.”
“Here we applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastases ('poor prognosis' signature). Our findings provide a strategy to select patients who would benefit from adjuvant therapy.”

“An unsupervised, hierarchical clustering algorithm allowed us to cluster the 98 tumours on the basis of their similarities measured over [...] approximately 5,000 significant genes.”
Why cluster?

- A way to explore data for any hidden patterns or correlations
  - Once you see something, you can delve further but it is a good way to quickly try to see if there are any possible relationships you have missed
- Helps organize data
- Reduces the number of data points (e.g., you can reduce a cluster to a representative data point)
- Results might be fed into other data mining techniques
Clustering by numbers

- All of the examples we’ve seen can be framed as “clustering by numbers”
- What do we mean by that?
Clustering by numbers

- All of the examples we’ve seen can be framed as “clustering by numbers”
- What does that mean?
  - We cluster points, typically in a high-dimensional space
  - The example here is a 2-dimensional space
The goal in clustering data is to find points that are “near” each other

- For example, to form project groups, we might cluster students along the dimensions of “desired grade” and “procrastination tendency”
- Most of the time, there are many more dimensions
Clustering by numbers
Netflix example

Clustering task: cluster movies based on whether subscribers give them similar ratings

Data: tens of thousands of movies; for each movie, subscriber ratings (there are almost 100 million subscribers!)

Data points: one point per movie: \((\text{rating}_1, \text{rating}_2, \ldots, \text{rating}_n)\), where
\[
\text{rating}_k \text{ is the rating of subscriber } k \text{ (or "null" if no rating)}
\]

Dimension: the number of subscribers who provide ratings
Clustering by numbers
Breast cancer example

**Clustering task**: cluster breast cancer tumour samples, based on similarities between gene expression measurements

**Data**: 98 tumours; for each tumour, gene expression measurements of \(~5,000\) genes

Data points clicker question: How many data points? What is the data dimension?

A. 98 data points, dimension 5000

B. 5000 data points, dimension 98
Clustering task: cluster breast cancer tumour samples, based on similarities between gene expression levels

Data: 98 tumours; for each tumour, gene expression levels of ~5,000 genes

Data points: one point per tumour: \((\text{level}_1, \text{level}_2, \ldots, \text{level}_{5000})\), where
- \(\text{level}_k\) is the gene expression level of tumour \(k\)
- the data dimension is the number of genes
Measuring clustering quality

Super important is knowing which data to cluster on!

- Netflix does not use data pertaining to geography, gender, or age of subscribers when clustering movies, so there are no dimensions for that data.
- Libraries don’t cluster books by colour, rather by content.
- In what follows we'll assume that the data dimensions we’re clustering on are those that matter for quality.
Measuring clustering quality

• Suppose that we have two potential clusters of a bunch of points. Which is better?
• Let’s look at an example
Consider the following potential clusterings for assigning children to three different schools. What are the relative merits of each clustering?

A                                           B

Measuring clustering quality
Group exercise
Consider the following potential clusterings for assigning children to three different schools. What are the relative merits of each clustering?

Cluster A:
• By location (friends are in the same neighbourhood)
• Approximately even number of kids in each school

Cluster B:
• Meet new people
Measuring clustering quality
Which do you like better and why?

It depends
Measuring cluster quality
Possible criteria to use

• **Intra-class similarity**: points within a cluster contain are close to each other (or at least to their closest neighbours)

• **Inter-class dissimilarity**: points in two different clusters are far from each other (or at least to their closest neighbours in other clusters)

• **Size similarity**: clusters have similar size
Measuring cluster quality

Clicker question

In which graph are points 1 and 2 more similar?

A: Left graph
B: Right graph
C: Both are the same

Left graph

Right graph
K-means clustering algorithm

- This is a popular algorithm for clustering
- K is a number you choose; this is the number of clusters you want to end up with
- You can use this algorithm on any number of data points
- Depending on the data points, it is possible that there is no final answer so you should pick the max number of times you want to run it
K-means clustering algorithm

1. Choose k centroid points at random to act as the “centre” of your clusters

2. Repeat however many times you decide, or until the answer stabilizes (whichever comes first):
   a. Cluster assignment: For each point, determine which of the k centroids it’s closest to, and put it in the cluster of that centroid
   b. Move centroid: Average all the points inside each cluster to get a new centroid.

(The answer stabilizes when the assignment of points to clusters doesn’t change in two successive iterations)
K-means clustering example

Initial points – we’ll label them by letters so that we can refer to them more easily throughout the example.
K-means clustering example

Step 1: choose K centroid points at random to act as the “centre” of your clusters

We’ll let K = 2

We randomly choose the orange and pink x’s to be centroids of clusters O and P respectively

Note: the centroids do not have to be points that are being clustered
K-means clustering example

Step 2a: Cluster assignment: For each point, determine which of the k centroids it’s closest to, and put it in the cluster of that centroid.

<table>
<thead>
<tr>
<th>Point</th>
<th>Distance to O</th>
<th>Distance to P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1.4</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>2.2</td>
</tr>
<tr>
<td>C</td>
<td>1.4</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>3.2</td>
<td>2.8</td>
</tr>
<tr>
<td>E</td>
<td>4.5</td>
<td>4.2</td>
</tr>
</tbody>
</table>
K-means clustering example

Step 2a: Cluster assignment: For each point, determine which of the k centroids it’s closest to, and put it in the cluster of that centroid.

<table>
<thead>
<tr>
<th>Point</th>
<th>Distance to O</th>
<th>Distance to P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1.4</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>2.2</td>
</tr>
<tr>
<td>C</td>
<td>1.4</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>3.2</td>
<td>2.8</td>
</tr>
<tr>
<td>E</td>
<td>4.5</td>
<td>4.2</td>
</tr>
</tbody>
</table>
K-means clustering example

Step 2b: move centroid: Average all the points inside each cluster to get a new centroid.

Average x for “O” cluster = (1+1)/2 = 1
Average y for “O” cluster = (1+0)/2 = .5
Average x for “P” cluster = (0+2+3)/3 = 1.7
Average y for “P” cluster = (2+4+5)/3 = 3.7
Step 2b: move centroid: Average all the points inside each cluster to get a new centroid.

Average x for “O” cluster = (1+1)/2 = 1
Average y for “O” cluster = (1+0)/2 = .5
Average x for “P” cluster = (0+2+3)/3 = 1.7
Average y for “P” cluster = (2+4+5)/3 = 3.7

New centroid for O: (1, .5)
New centroid for P: (1.7, 3.7)
K-means clustering example

Back to the beginning! We need to calculate the distance to the new centroids
K-means clustering example

Step 2a: Cluster assignment: For each point, determine which of the k centroids it’s closest to, and put it in the cluster of that centroid.

<table>
<thead>
<tr>
<th>Point</th>
<th>Distance to O</th>
<th>Distance to P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.5</td>
<td>2.7</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
<td>3.7</td>
</tr>
<tr>
<td>C</td>
<td>1.8</td>
<td>2.4</td>
</tr>
<tr>
<td>D</td>
<td>3.6</td>
<td>0.5</td>
</tr>
<tr>
<td>E</td>
<td>4.9</td>
<td>1.9</td>
</tr>
</tbody>
</table>
K-means clustering example

Step 2a: Cluster assignment: For each point, determine which of the k centroids it’s closest to, and put it in the cluster of that centroid

<table>
<thead>
<tr>
<th>Point</th>
<th>Distance to O</th>
<th>Distance to P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.5</td>
<td>2.7</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
<td>3.7</td>
</tr>
<tr>
<td>C</td>
<td>1.8</td>
<td>2.4</td>
</tr>
<tr>
<td>D</td>
<td>3.6</td>
<td>0.5</td>
</tr>
<tr>
<td>E</td>
<td>4.9</td>
<td>1.9</td>
</tr>
</tbody>
</table>
K-means clustering example

Step 2b: move centroid: Average all the points inside each cluster to get a new centroid.

Average x for “O” cluster = (1+1+0)/3 = .7
Average y for “O” cluster = (1+0+2)/3 = 1
Average x for “P” cluster = (2+3)/2 = 2.5
Average y for “P” cluster = (4+5)/2 = 4.5
K-means clustering example

Step 2b: move centroid: Average all the points inside each cluster to get a new centroid.

Average x for “O” cluster = (1+1+0)/3 = 0.7
Average y for “O” cluster = (1+0+2)/3 = 1
Average x for “P” cluster = (2+3)/2 = 2.5
Average y for “P” cluster = (4+5)/2 = 4.5

New centroid for O: (0.7, 1)
New centroid for P: (2.5, 4.5)
What happens next?

A. We re-calculate distances between points and the two centroids
B. We re-calculate the centroid positions
C. We stop
Back to the beginning (again)! We need to calculate the distance to the new centroids.
Step 2a: Cluster assignment: For each point, determine which of the k centroids it’s closest to, and put it in the cluster of that centroid.

<table>
<thead>
<tr>
<th>Point</th>
<th>Distance to O</th>
<th>Distance to P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.3</td>
<td>3.81</td>
</tr>
<tr>
<td>B</td>
<td>1.04</td>
<td>4.74</td>
</tr>
<tr>
<td>C</td>
<td>1.22</td>
<td>2.4</td>
</tr>
<tr>
<td>D</td>
<td>3.27</td>
<td>0.71</td>
</tr>
<tr>
<td>E</td>
<td>4.61</td>
<td>0.71</td>
</tr>
</tbody>
</table>

No points changed clusters. We’re done!
What clusters does step 2a yield?

Group exercise

The example and a tool to calculate distance between points is on today’s lecture page under “resources”
Use the first step of k-means to create two clusters for these points: Group exercise.
What happens next in the algorithm

A. We re-calculate distances between points and the two centroids

B. We re-calculate the centroid positions

C. We stop
What happens next in the algorithm? We move the centroids.
How do we move the centroids?

Move centroid step:

x-coordinate of pink centroid is at position 
\[
\frac{1+1+2+2+3+4+5+6}{8}
\]

y-coordinate of purple centroid is at position 
\[
\frac{1+1+2+3+3+3+4+5}{8}
\]

(Similar for blue)
Next: calculate the distances again

- Using the same strategy as before, calculate the distances from the centroids to each point
- The clusters do not change → the centroids do not change → the algorithm is done!
Downsides of k-means clustering

- The algorithm may give different cluster solutions depending on how the initial centroids are chosen.
- It’s not always clear how to choose k, the number of clusters.
  - If the size of the data set is small, different values of k can be chosen.
  - Or, a large value of k can be chosen, and then clusters can be merged to yield a hierarchical cluster structure.
Dirty data

- One catch with clustering, and data mining in general, is “dirty” data
- Unless the data is clean, the results aren’t meaningful
- Example: “smoking information is very hard to parse... If you read the records, you understand right away what the doctor meant. But good luck trying to make a computer understand. There’s ‘never smoked’ and ‘smoking = 0.’ How many cigarettes does a patient smoke? That’s impossible to figure out.”

http://fortune.com/2014/06/30/big-data-dirty-problem/
Coming full circle: back to privacy issues

Massachusetts released anonymized medical records for state employees. They removed all identifiers but left birthdate (including year), gender, and zip code.

Group discussion: what percentage of people in the US could likely be uniquely identified by this information? (Note: there are ~7,500 people per zip code)

A. 0-19%
B. 20-39%
C. 40-59%
D. 60-79%
E. 80-100%
Group exercise

Is it a problem that we can tell that in one database one individual (we don’t know the name, but we know the age, gender, and zip code) has a set of medical conditions?
Group exercise

Is it a problem that we can tell that in one database one individual (we don’t know the name, but we know the age, gender, and zip code) has a set of medical conditions?

Do you think this is problematic?
A: Yes
B: No
Group exercise

Why?

- You might not want people to know/not ready to let people know of your medical condition
- Your employer could use it against you in some way
Well…

- Okay, so we can uniquely determine that there exists some person with some medical visits. We still don’t who they are.
- But there are other data sources, too. Publically available voting records include name, zip code, birthdate and gender of voters.
- So if you put the two together, you now have names and health records together
- Security researcher (and graduate student) Latanya Sweeney sent the Governor’s full health records to his office.

Learning Goals Revisited

- [CT Building Block] Students will be able to create English language descriptions of algorithms to analyze data and show how their algorithms would work on an input data set.
- [CT Application] Students will be able to use computing to examine datasets and facilitate exploration in order to gain insight and knowledge (data and information).
- [CT Impact] Students will be able to give examples of privacy and security issues that arise as a result of data mining.
- [CT Building Block] Students will be able to describe what a greedy algorithm is.
- [CT Building Block] Students will be able to describe what the general decisions are in building a decision tree.
- [CT Building Block] Students will be able to build a simple decision tree.
- [CT Building Block] Students will be able to describe what considerations are important in building a decision tree.
- [CT Building Block] Students will be able to give examples showing why clustering is useful, describe how clustering tasks can be formulated in terms of high-dimensional numerical data, and infer what is the output of the k-means clustering algorithm on a small input.