SCI 199 Y Mar 1

Service Learning Updates

- Where are you volunteering
- Have you…
  - Made contact?
    - **** If you have not made contact with the organization yet you will need to contact Prof. Borodin ASAP. ****
  - Volunteered?
- What has your experience been so far?

Machine learning

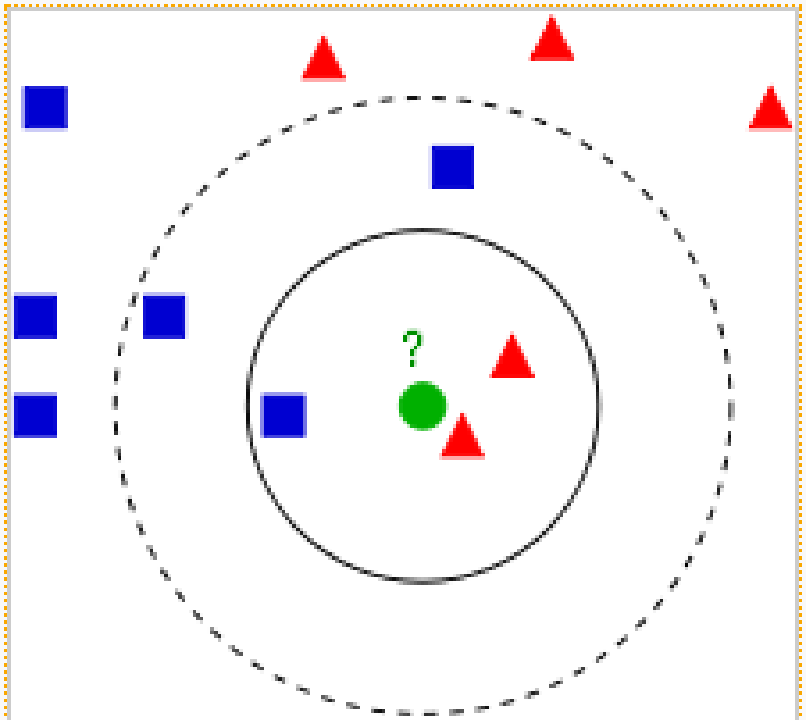
- A definition:
  - *Machine learning refers to a system capable of the autonomous acquisition and integration of knowledge. This capacity to **learn from experience**, analytical observation, and other means, results in a system that can continuously self-improve and thereby offer increased efficiency and effectiveness.*
  - http://www.aaai.org/AITopics/pmwiki/pmwiki.php/AITopics/MachineLearning

Algorithm types

- Supervised learning:
  - Algorithm takes in a labeled training set
  - Example: training set: apple->red; banana->yellow
  - If we input “apple” to the algorithm, it should be able to tell us it is “red” (usually more complex algorithms than this)
- Unsupervised learning:
  - Does not need labeled training data
  - Usually builds statistical model of data

K-nearest neighbors

- Supervised learning algorithm
- Classifies objects based on their **k** nearest neighbors in feature space
- What class should the green circle belong to? (k=3)



How does a computer do this?

- Each element in training set is associated with a **"feature vector"**
- Features here are x,y
- E.g., 1 -> (2,5)
- How to compute distance?
  - Many choices
  - E.g., Euclidean distance

num	x	y	class
1	2	5	Bad
2	3	5	Bad
3	4	5	Bad
4	3	4	Bad
5	1	2	Good
6	2	2	Good
7	1	1	Good

$$d(1->3) = \sqrt{(2-4)^2 + (5-5)^2} = \sqrt{(-2)^2} = 2$$

How does a computer do this?

- Using k=3
- How to classify (2,3)
- What are its k nearest neighbors?

num	x	y	class
1	2	5	Bad
2	3	5	Bad
3	4	5	Bad
4	3	4	Bad
5	1	2	Good
6	2	2	Good
7	1	1	Good

How does a computer do this?

- Using k=3
- How to classify (2,3)
- What are its k nearest neighbors?
  - 4 (Bad) ; 5 (Good);
  - 6 (Good)
- What class do we assign it?

num	x	y	class
1	2	5	Bad
2	3	5	Bad
3	4	5	Bad
4	3	4	Bad
5	1	2	Good
6	2	2	Good
7	1	1	Good

How does a computer do this?

- Using k=3
- How to classify (2,3)
- What are its k nearest neighbors?
  - 4 (Bad) ; 5 (Good);
  - 6 (Good)
- What class do we assign it?
  - 2 neighbors = Good
  - 1 neighbor = Bad
  - (2,3) -> Good

num	x	y	class
1	2	5	Bad
2	3	5	Bad
3	4	5	Bad
4	3	4	Bad
5	1	2	Good
6	2	2	Good
7	1	1	Good

How to represent non-numerical features?

- Example: Text Classification
- Say we want to find all news articles about the Olympics
- What are the features?
- What distance function?

Text classification:

- Feature vector: frequency count of terms
- Example headline:
  - "Vancouver Olympics: Was it worth the cost?"
  - Feature vector (referred to as "term vector")

Term	Vancouver	Olympics	Was	it	worth	The	cost
Frequency	1	1	1	1	1	1	1

Text classification:

- How to compute distance between 2 term vectors?
- Cosine measure http://en.wikipedia.org/wiki/Cosine_similarity

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

- Review: Dot product & Magnitude

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

$$\|\mathbf{x}\| := \sqrt{x_1^2 + \cdots + x_n^2}$$

Cosine measure example:

- Phrase 1: the cat ran fast
- Phrase 2: the cat drank milk
- Vectors (note: elements are unique terms in both phrases)

Term	The	Cat	Ran	Fast	Drank	Milk
Freq. (Phrase 1)	1	1	1	1	0	0
Freq. (Phrase 2)	1	1	0	0	1	1

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

Cosine measure example:

Term	The	Cat	Ran	Fast	Drank	Milk
Freq. (Phrase 1)	1	1	1	1	0	0
Freq. (Phrase 2)	1	1	0	0	1	1

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

Cosine measure example:

Term	The	Cat	Ran	Fast	Drank	Milk
Freq. (Phrase 1)	1	1	1	1	0	0
Freq. (Phrase 2)	1	1	0	0	1	1

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

$$V_1 \cdot V_2 = (1\times1) + (1\times1) + (1\times0) + (1\times0) + (1\times0) + (1\times0) = 2$$

Cosine measure example:

Term	The	Cat	Ran	Fast	Drank	Milk
Freq. (Phrase 1)	1	1	1	1	0	0
Freq. (Phrase 2)	1	1	0	0	1	1

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

$$V_1 \cdot V_2 = (1\times1)+(1\times1)+(1\times0)+(1\times0)+(1\times0)+(1\times0) = 2$$

$$\| V_1 \| = \sqrt{1^2+1^2+1^2+1^2+0^2+0^2} = 2$$

Cosine measure example:

Term	The	Cat	Ran	Fast	Drank	Milk
Freq. (Phrase 1)	1	1	1	1	0	0
Freq. (Phrase 2)	1	1	0	0	1	1

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

$$V_1 \cdot V_2 = (1\times1)+(1\times1)+(1\times0)+(1\times0)+(1\times0)+(1\times0) = 2$$

$$\| V_1 \| = \sqrt{1^2+1^2+1^2+1^2+0^2+0^2} = 2 \qquad \| V_2 \| = \sqrt{1^2+1^2+0^2+0^2+1^2+1^2} = 2$$

Cosine measure example:

Term	The	Cat	Ran	Fast	Drank	Milk
Freq. (Phrase 1)	1	1	1	1	0	0
Freq. (Phrase 2)	1	1	0	0	1	1

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

$$V_1 \cdot V_2 = (1\times1)+(1\times1)+(1\times0)+(1\times0)+(1\times0)+(1\times0) = 2$$

$$\| V_1 \| = \sqrt{1^2+1^2+1^2+1^2+0^2+0^2} = 2 \qquad \| V_2 \| = \sqrt{1^2+1^2+0^2+0^2+1^2+1^2} = 2$$

$$similarity = \frac{V_1 \cdot V_2}{\| V_1 \| \| V_2 \|} = \frac{2}{2\times2} = \frac{1}{2} = 0.5$$

Putting it all together

- Training set:

Headline	About Olympics?
Vancouver Olympics: Was it worth the cost?	Yes
Final day of the 2010 Vancouver winter Olympics	Yes
Canadian stocks up on GDP data, copper prices	No
Canada's economic recovery picks up speed	No

- Using kNN (k=3); is the following about the Olympics?
  - "Games over: Winter Olympics end on Sunday in Vancouver"

Challenges to kNN

- Complexity
  - Need to compute distance to every member of training set to find k nearest neighbors
- Need a training set!
- Text classification challenges:
  - Stop-words (e.g., the, is, a, as etc.)
  - Stemming words (e.g., runs vs. running; Olympic vs. Olympics)