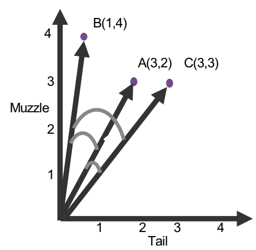
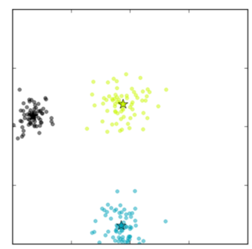
Clustering, *k*-means, Expectation-Maximization

AI and Ethics





Based partly on: M desJardins, T Oates, P Matuszek, RJ Mooney: www.cs.utexas.edu/~mooney/cs388/slides/TextClustering.ppt, and other sources as noted

1

Bookkeeping

- HW5 due 12/3
- Today: Clustering, EM (briefly), ethics of AI
- Next time: Applications - robotics
- Final exam: 12/13 (in class)

2

What is Clustering?

- Given some instances of data: group them such that
  - Examples within a group are similar
  - Examples in different groups are different
- These groups are **clusters**
- A kind of unsupervised learning – the instances do not include a class attribute.

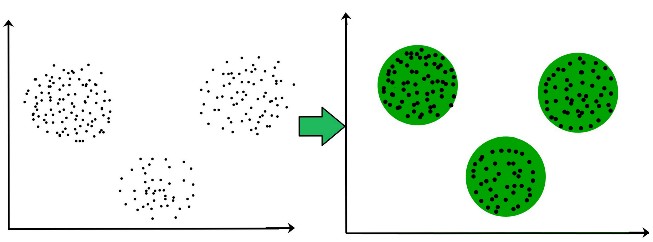
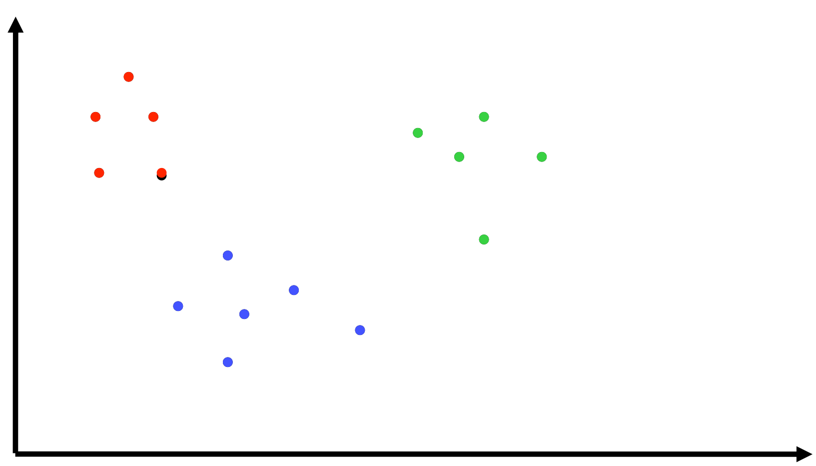


Image: geeksforgeeks.com

3

Clustering Example



4

A Different Example

- How would you group
  - ‘The price of crude oil has increased significantly’
  - ‘Demand for crude oil outstrips supply’
  - ‘Some people do not like the flavor of olive oil’
  - ‘The food was very oily’
  - ‘Crude oil is in short supply’
  - ‘Oil platforms extract oil’
  - ‘Canola oil is supposed to be healthy’
  - ‘Iraq has significant oil reserves’
  - ‘There are different types of cooking oil’

A note: you might or might not know how many clusters to look for.

5

A Different Example

- How would you group
  - ‘The price of crude oil has increased significantly’
  - ‘Demand for crude oil outstrips supply’
  - ‘Some people do not like the flavor of olive oil’
  - ‘The food was very oily’
  - ‘Crude oil is in short supply’
  - ‘Oil platforms extract oil’
  - ‘Canola oil is supposed to be healthy’
  - ‘Iraq has significant oil reserves’
  - ‘There are different types of cooking oil’

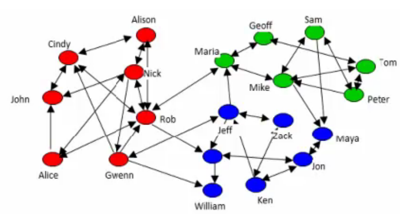
6

Another Example

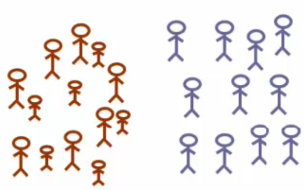
7

Some Example Uses

Organize computing clusters



Social network analysis



Market segmentation

Astronomical data analysis

8

Clustering Basics

- Collect examples
- Compute **similarity** among examples according to some metric
- Group examples together such that:
  - Examples within a cluster are similar
  - Examples in different clusters are different
- Summarize each cluster
- **Sometimes**: assign new instances to the most similar cluster

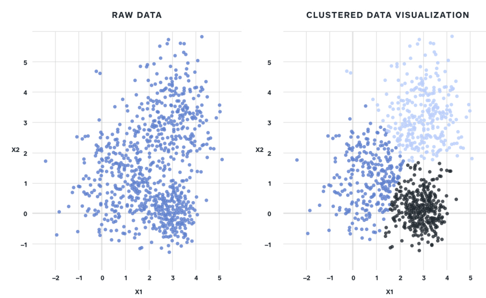


Image: developer.squareup.com/blog/so-you-have-some-clusters-now-what

9

Measures of Similarity

- To do clustering we need some measure of similarity.
- This is basically our "critic"
- Computed over a vector of values representing instances
- Types of values depend on domain:
  - Documents: bag of words, linguistic features
  - Purchases: cost, purchaser data, item data
  - Census data: most of what is collected
- Multiple different measures exist

10

Measures of Similarity

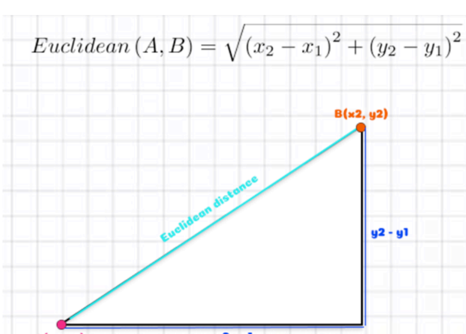
- Semantic similarity (but that's hard)
  - For example, olive oil/crude oil
- Similar attribute counts
  - Number of attributes with the same value
  - Appropriate for large, sparse vectors
  - Bag-of-Words: BoW
- More complex vector comparisons:
  - Euclidean Distance
  - Cosine Similarity

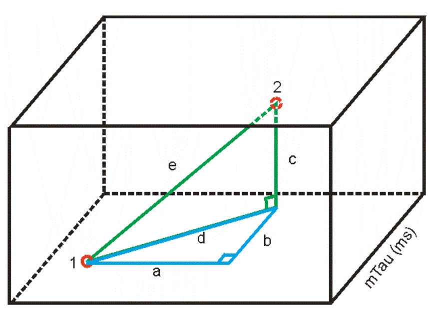
11

Euclidean Distance

- Euclidean distance: distance between two measures summed across each feature (between points in n-dimensional space)

$$\mathrm{dist}(x_i,\ x_j) = \mathrm{sqrt}((x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+..+(x_{in}-x_{jn})^2)$$





https://hlab.stanford.edu/brian/euclidean_distance_in.html
https://www.turing.com/kb/how-to-decide-perfect-distance-metric-for-machine-learning-model

12

Euclidean Distance

- Euclidean distance: distance between two measures summed across each feature (between points in n-dimensional space)

$$\mathrm{dist}(x_i,\ x_j) = \mathrm{sqrt}((x_{i1}\text{-}x_{j1})^2+(x_{i2}\text{-}x_{j2})^2+..+(x_{in}\text{-}x_{jn})^2)$$

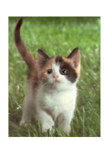
- Squared differences give more weight to larger differences
  - $\mathrm{dist}([1,2],[3,8]) = \mathrm{sqrt}((1\text{-}3)^2+(2\text{-}8)^2) =$
    $\mathrm{sqrt}((\text{-}2)^2+(\text{-}6)^2) =$
    $\mathrm{sqrt}(4+36) =$
    $\mathrm{sqrt}(40) = \sim 6.3$

13

Euclidean

- Calculate differences
  - Ears: pointy? [T/F → 0/1]
  - Muzzle: how many inches long? [1–4]
  - Tail: how many inches long? [2–6]


x_1


x_2


x_3


x_4

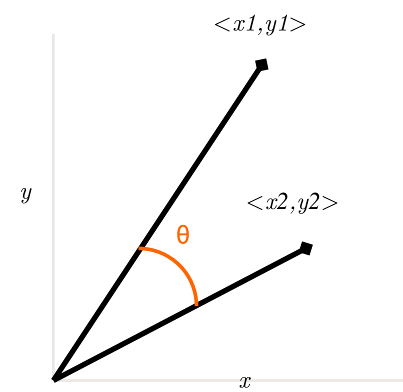
$$\mathrm{dist}(x_1,\ x_2) = \mathrm{sqrt}((0\text{-}1)^2+(3\text{-}1)^2+..+(2\text{-}4)^2)=\mathrm{sqrt}(9)=3$$

$$\mathrm{dist}(x_1,\ x_3) = \mathrm{sqrt}((0\text{-}0)^2+(3\text{-}3)^2+..+(2\text{-}3)^2)=\mathrm{sqrt}(1)=1$$

14

Cosine Similarity

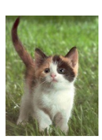
- A measure of similarity between vectors
  - Find **cosine of the angle** between them
  - Cosine = 1 when angle = 0
  - Cosine < 1 otherwise
- As angle between vectors shrinks, θ approaches 1
  - Meaning: the two vectors are getting closer
  - Meaning: the **similarity** of whatever is represented by the vectors **increases**
- Vectors can have any number of dimensions

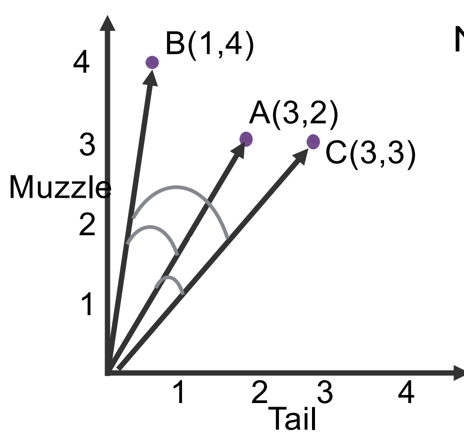


Based on home.iitk.ac.in/~mfelixor/Files/non-numeric-Clustering-seminar.ppt

15

Cosine Similarity





Most similar?

16

Euclidean Distance vs Cosine Similarity vs Other

- Cosine Similarity:
  - Measures **relative** proportions of various features
  - Ignores magnitude
  - When all the correlated dimensions between two vectors are in proportion, you get maximum similarity
- Euclidean Distance:
  - Measures **actual** distance between two points
  - More concerned with absolutes
- Either can deal with many dimensions
- Often similar in practice, especially on high dimensional data
- Consider meaning of features/feature vectors **for your domain**

Justin Washtell @ semanticvoid.com/blog/2007/02/23/similarity-measure-cosine-similarity-or-euclidean-distance-or-both/

17

Clustering Algorithms

- Flat:
  - K means
- Hierarchical:
  - Bottom up
  - Top down (not common)
- Probabilistic:
  - Expectation Maximization (E-M)

18

Partitioning (Flat) Algorithms

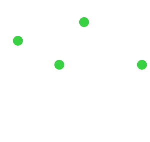
- Partitioning method
  - Construct a **partition** of *n* instances into a set of *k* clusters
- Given: a set of documents and the number *k*
- Find: a partition of *k* clusters that optimizes the chosen partitioning criterion
  - Globally optimal: exhaustively enumerate all partitions.
  - Usually too expensive.
  - Effective heuristic methods: k-means algorithm.

www.csee.umbc.edu/~nicholas/676/MRSslides/lecture17-clustering.ppt

19

k-means Clustering

- Simplest hierarchical method, widely used
- Create clusters based on a centroid; each instance is assigned to the closest centroid
  - Centroid means "approximate center"
- K is given as a parameter
- Heuristic and iterative



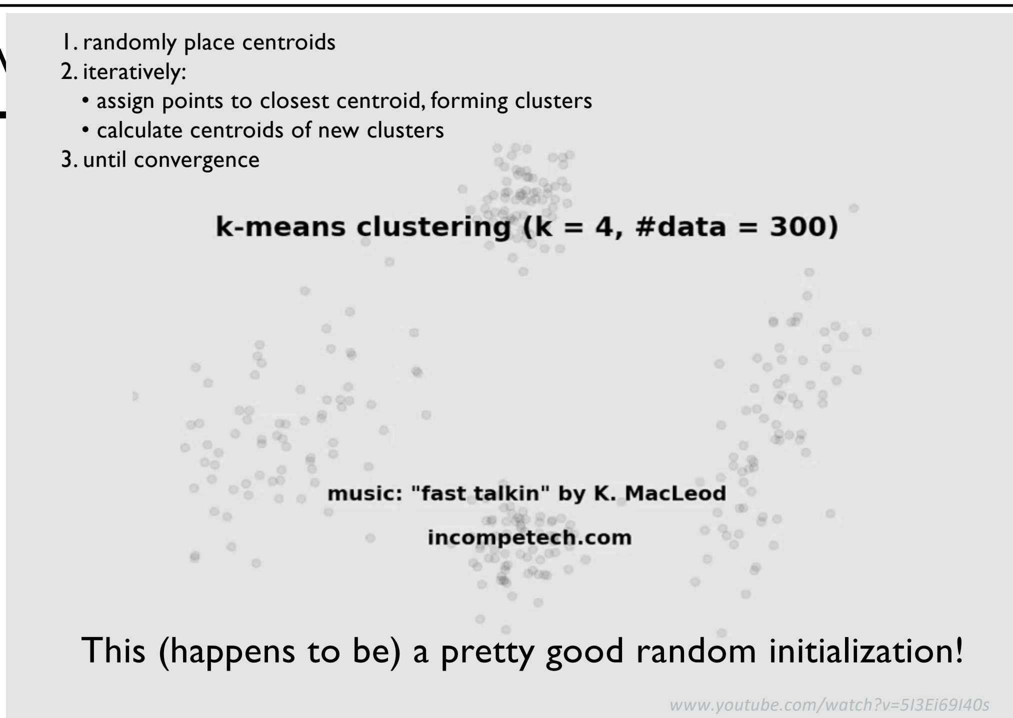
20

k-means Algorithm

- Choose k (the number of clusters)
- Randomly choose k instances to center clusters on
- Assign each point to the centroid it’s closest to, forming clusters
- Recalculate centroids of new clusters
- Reassign points based on new centroids
- Iterate until…
- Convergence (no point is reassigned) or after a fixed number of iterations.

21

K M



24

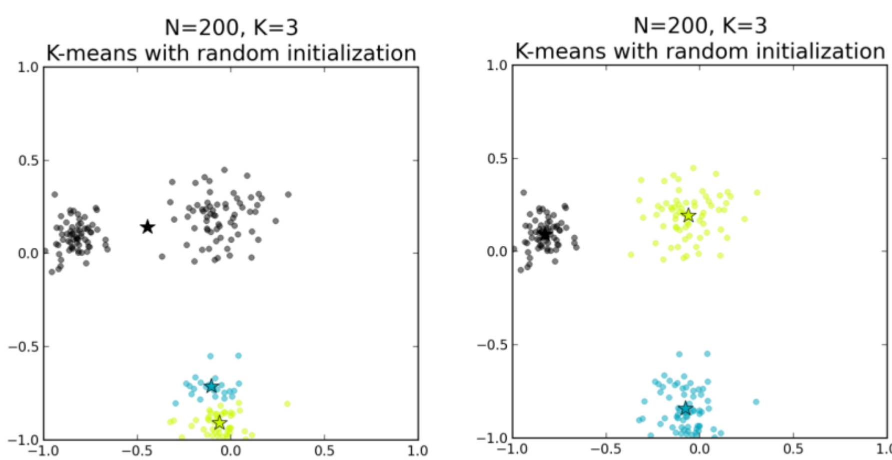
k-means

- Tradeoff between having more clusters (better focus within each cluster) and having too many clusters.
  - Overfitting is a possibility with too many!
- Results depend on random seed selection.
  - Some seeds can result in slow convergence or convergence to poor clusters
- Algorithm is sensitive to outliers
  - Data points that are very far from other data points
  - Could be errors, special cases, …

www.csee.umbc.edu/~nicholas/676/MRSslides/lecture17-clustering.ppt

25

Problem: Bad Initial Seeds



datasciencelab.wordpress.com/2014/01/15/improved-seeding-for-clustering-with-k-means/

26

Strengths of k-means

- Strengths:
  - Simple: easy to understand and to implement
  - Efficient: Time complexity: O(tkn),
    - where n is the number of data points,
    - k is the number of clusters, and
    - t is the number of iterations.
  - Since both k and t are small. k-means is considered a linear algorithm.
- K-means is most popular clustering algorithm.
- In practice, performs well, especially on text.

28

k-means Weaknesses

- Must choose k
  - Poor k → poor clusters
  - But sometimes we don't know
- Clusters may differ in size or density
- All attributes are weighted
- Heuristic, based on initial random seeds; clusters may differ from run to run

www.cs.uic.edu/~liub/teach/cs583-fall-05/CS583-unsupervised-learning.ppt

29

Expectation Maximization Clustering

- Expectation-Maximization is a core ML algorithm
  - Not just for clustering!
- Basic idea: assign instances to clusters **probabilistically** rather than **absolutely**
  - Instead of assigning membership in a group, learn a probability function for each group
- Instead of absolute assignments, output is probability of **each instance** being in **each cluster**
- **Like K-means with soft assignment.**
  - **Assign point partly to all clusters based on probability it belongs to each**
  - **Compute weighted averages (centers, and covariances)**

30

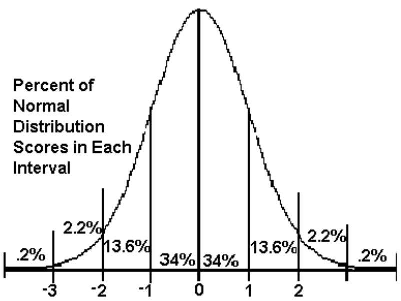
Expectation Maximization (EM)

- **Probabilistic method for soft clustering**
- Idea: learn k classifications from **unlabeled** data
- Assumes k clusters:$\{c_1, c_2, \ldots c_k\}$
- "Soft" version of k-means
- Assumes a probabilistic model of categories (such as Naive Bayes)
- Allows computing $P(c_i \mid I)$ for each category, c_i, for a given instance I

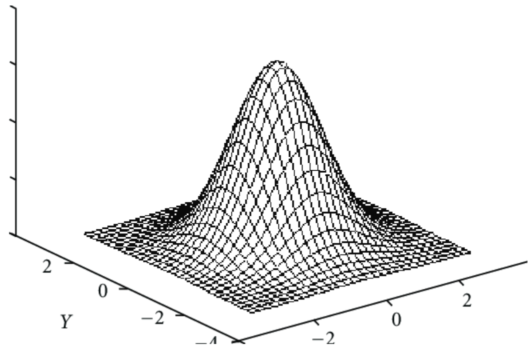
31

Notation: Normal distributions

N(μ , σ) is a 1D normal (Gaussian) distribution with mean μ and standard deviation σ (so the variance is σ^2).

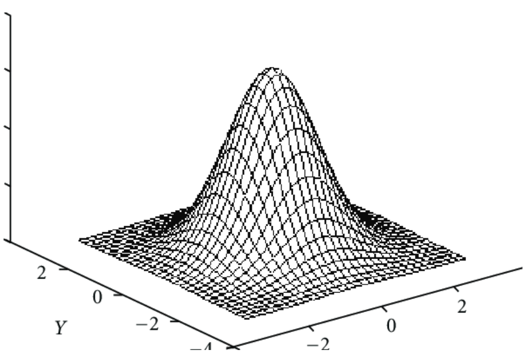


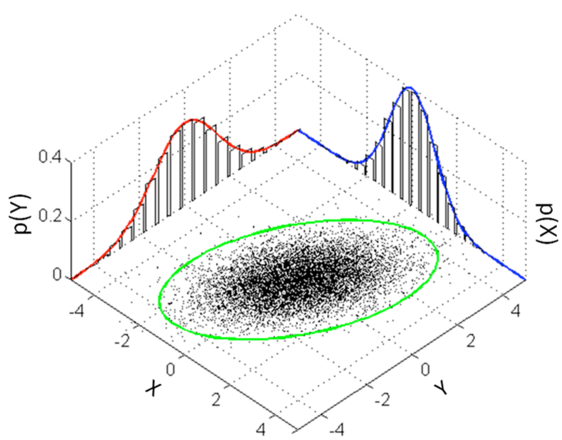
2D (Gaussian) distribution



32

2D Gaussian





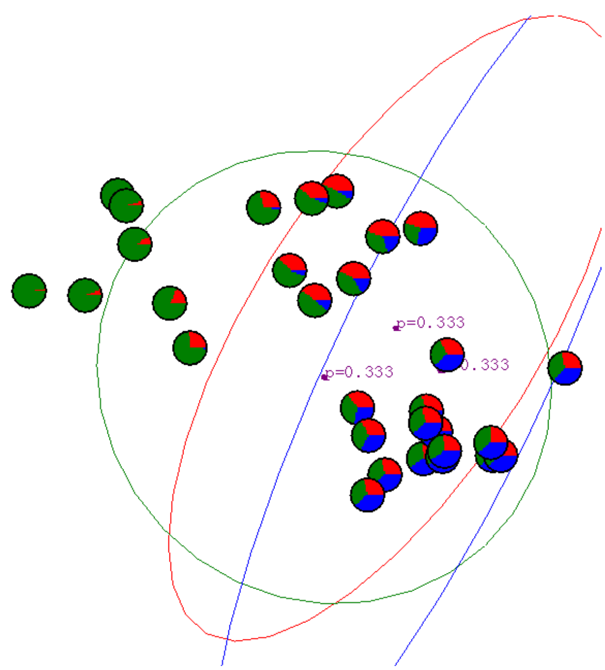
33

K-means vs. EM

	K-means	EM
Cluster Representation	mean	mean, variance, and weight
Cluster Initialization	randomly select K means	initialize K Gaussian distributions
Expectation	assign each point to closest mean	soft-assign each point to each distribution
Maximization	compute means of current clusters	compute new params of each distribution

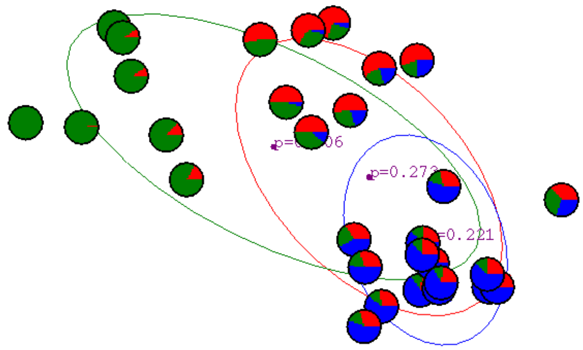
34

Gaussian Mixture Example: Start



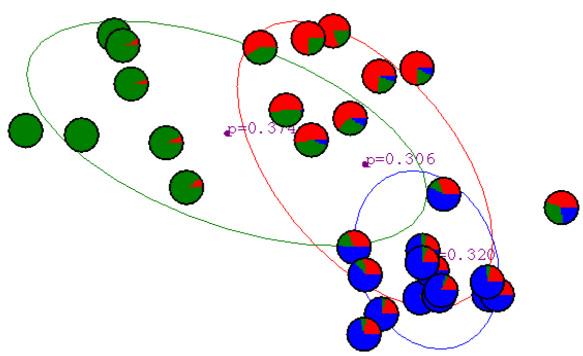
35

After first iteration



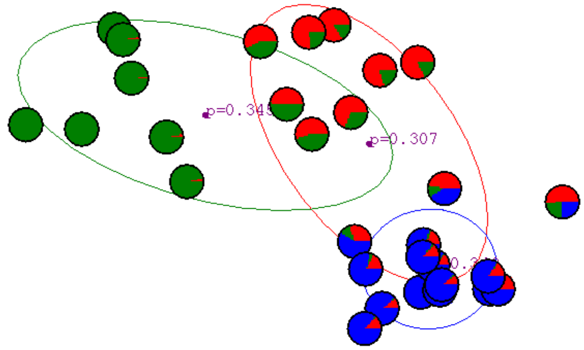
36

After 2nd iteration



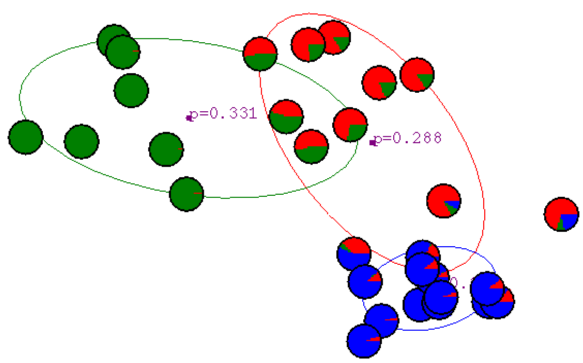
37

After 3rd iteration



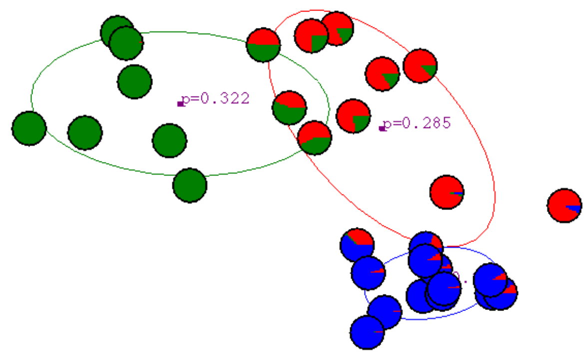
38

After 4th iteration



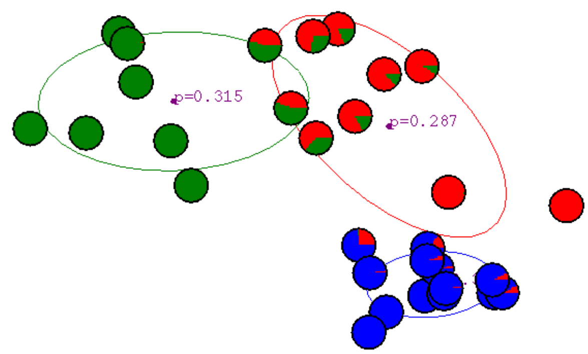
39

After 5th iteration



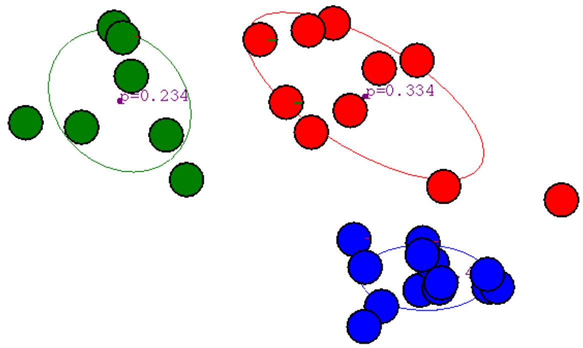
40

After 6th iteration



41

After 20th iteration



42

K –Means -> EM with GMM : The Intuition (1)

Instead of making a “hard” decision on to which class a sample belongs to, we use probability theory and assign samples to classes probabilistically

Likelihood

Prior

Using Bayes rule
$$p(C_l|x) = \frac{p(x|C_l)p(C_l)}{\sum_{i=1}^{l} p(x|C_i)p(C_i)}$$

Posterior

We want to maximize the Posterior

43

K-Means $\rightarrow$ EM with Gaussian Mixture Models (GMM)

- Boot Step:
  - Initialize K clusters: $C_1, \ldots, C_K$
  - (μ_j, Σ_j) and $P(C_j)$ for each cluster j
- Iteration Step:
  - Estimate the (soft) cluster assignment of each datum ➡ Expectation

$$p(C_j \mid x_i)$$

  - Re-estimate the cluster parameters ➡ Maximization

$$(\mu_j, \Sigma_j), p(C_j) \quad \text{For each cluster } j$$

46

EM Clustering Algorithm

- **Goal:** maximize overall probability of data
- Iterate between:
  - Expectation: **estimate probability** that each instance belongs to each cluster
  - Maximization: **recalculate parameters** of probability distribution for each cluster
- Until convergence or iteration limit.

47

(Slightly) More Formally

- Iteratively learn **probabilistic categorization model** from **unsupervised data**
- Initially assume random assignment of examples to categories
  - "Randomly label" data
- Learn initial probabilistic model by estimating **model parameters θ** from randomly labeled data
- Iterate until convergence:
  - **Expectation (E-step):**
    - Compute $P(ci \mid I)$ for each instance (example) given the current model
    - Probabilistically re-label the examples based on these posterior probability estimates
  - **Maximization (M-step):** Re-estimate model parameters, θ, from re-labeled data

48

EM Summary

- Basically a probabilistic k-means.
- Has many of same advantages and disadvantages
  - Results are easy to understand
  - Have to choose k ahead of time
- Useful in domains when we want likelihood that an instance belongs to more than one cluster
  - Natural language processing for instance
  - "Oil is a valuable commodity" – 80% crude, 20% food?

54

Ethics in AI

Some interesting questions from 20,000 feet

55

Meta-Questions

- Questions we will not answer today:
  - What do “right” and “wrong” mean?
  - Who gets to decide what’s right and wrong?
  - How do/should those decisions be made?
  - What should we do about things that are wrong?
- We’ll use commonly understood ideas of wrong:
  - It’s wrong to **harm** people
    - Physically, emotionally, financially…
  - It’s wrong to **discriminate** against people
  - It’s wrong to **steal** from people
  - It’s wrong to **invade people’s privacy**
  - It’s wrong to be **unfair** to people

“Without extenuating circumstances,” and understanding that sometimes there’s no “right” alternative

56

Big Questions

- Can computers "hurt" people? **Absolutely.**
- What about robots? **Yes.**
- Can a machine be "unfair"? An algorithm? **Sort of. There's a GIGO aspect.**
- Why do we, as computing professionals, care? **Ethics and morals, legal liability**
- What are some ways in which AI is doing wrong, right now? **Let us count the ways…**

57

Topics

- We will drive the discussion with current examples:
  - Self-driving cars (and other robots)
  - Discrimination and machine learning
  - Privacy, machine learning, and big data
- …but we will try to generalize from that

58

Self-Driving Cars

- Cars can hurt or kill people.
  - How many fatalities is acceptable?
  - Is it enough to not cause accidents?
- People cause accidents!
  - ~38,000 deaths per year in the U.S.
  - Lately it's been going up
  - How many of you text and drive?
- Do cars have to be perfect? Just better than humans? Somewhere in between?

59

Harder Questions

- What about naked self-driving cars?
  - No control mechanisms inside at all
- Should it be legal for a person to drive?
  - Even if cars are demonstrably better at it?
- Why?
  - Because I wanna?
  - Because we dislike giving up control?
- Even if you accept the risks, what about my rights?
- Who's legally liability? ← this is a big question that will affect the future

60

The Hardest One

- When an accident is inevitable…
  - Should the car occupants get hurt?
  - That is, the person who paid for it?
  - If it's not their fault?

- Would you buy a car that could hurt or kill you?
  - If it could be avoided by hurting or killing someone else?
- But consider:
  - Would you swerve to avoid a kid in the road?
  - What about a baby stroller?
- Who should be deciding these things? Uber?

61

Discrimination and ML

- Machine learning is only as good as its training data
- **GIGO: Garbage In, Garbage Out.**
- If we're drawing training data from some source, we perpetuate any bias in that source
- So a "fair" **algorithm** can yield biased **results**
  - Depends on source of training data
  - Depends on representation choices
  - Depends on chosen application

62

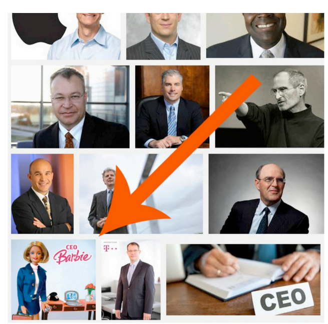
Case 1: Predictive Policing

- Predict where more/more serious crimes will occur and concentrate police presence there
  - People there are more likely to be caught/arrested
- "But it works!"
  - Because… more people are arrested in those places?
    - Where you have more police? What about all of them?
  - Studies: it doesn't work better than existing best practices
- Sending someone to jail is one of the few known things that causes subsequent criminal behavior
  - Causes, not correlates with

63

CEO Barbie

- A study of image search results for professions (e.g., CEO)
- Compare gender of results to ground truth from BLS
- Results of study:

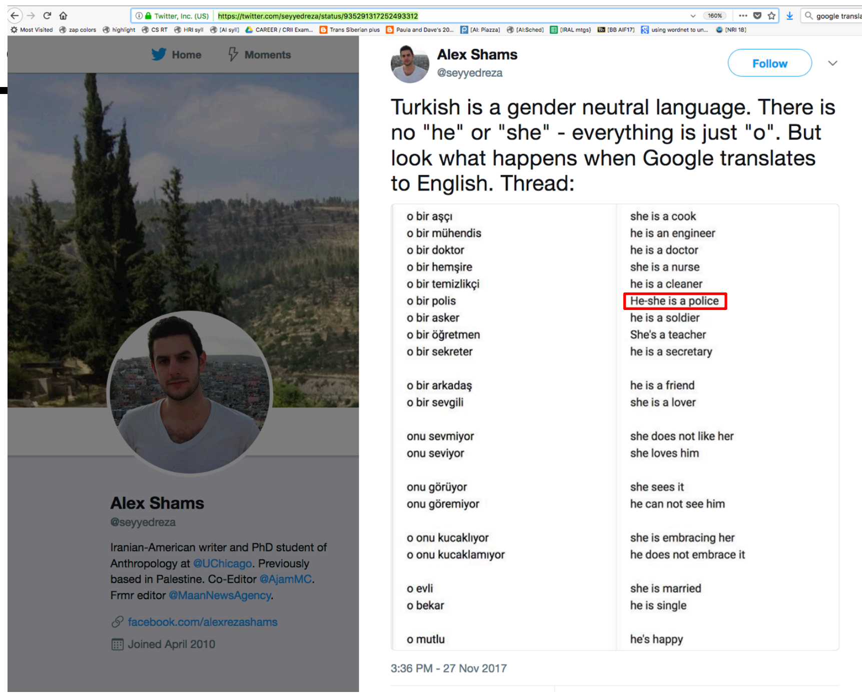


the only woman returned in a GIS for "CEO"

1. Women are under-represented in higher-paid fields, over-represented in lower-paid ones
2. People's guess as to the percentage split **is affected by** images viewed – there are real-world consequences

64

Transl



65

How Did This Happen?

- Google Translate is not a “translation” algorithm.
  - It is a pattern-matching, predictive algorithm
- It **reproduces patterns**, whether or not they are good/appropriate translations
  - Mostly they are, and translations come out
  - Sometimes they are not!
- Why not just hardcode gender-neutrality?
  - Very little of it is hardcoded – or even seen by human eyes

66

(Why) Is It a Problem?

* completely made-up fake example

- Some translations are wrong
  - Consider: "President's Erdogan's cook travels with him; ← her advice is indispensible"*
  - This may be importantly wrong.
- It's self-reinforcing
  - Once published, text becomes part of Google's statistical model
- It affects people's ideas of who can/should do what
  - As mentioned in the CEO Barbie study and others
  - These results and representations do affect minds
    - Think they don't affect yours? Let's look at those survey results.

67

Government and Privacy

- AI makes it possible to collect more data, correlate it better, analyze it better (clustering, anyone?)
  - Often framed as a dichotomy: "Privacy or safety"
  - We can disagree on the appropriate balance, but…
  - Only if loss of privacy **actually** leads to improved security
- "Nothing to hide* is, ethically speaking, nonsense
  - You can want to have privacy for many reasons
  - AKA: "I have nothing to hide (*that I think is actually bad, and that could be found out*) and (*I think*) nobody would ever target me for harassment."

68

Commerce and Privacy

- Read this terrifying longform:
  - http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html
- Google vs. Privacy
  - https://techcrunch.com/2013/04/02/google-unified-privacy-policy-vs-european-data-protection-regulators
- Short summary: Target knows everything.

69