Notes on Poster

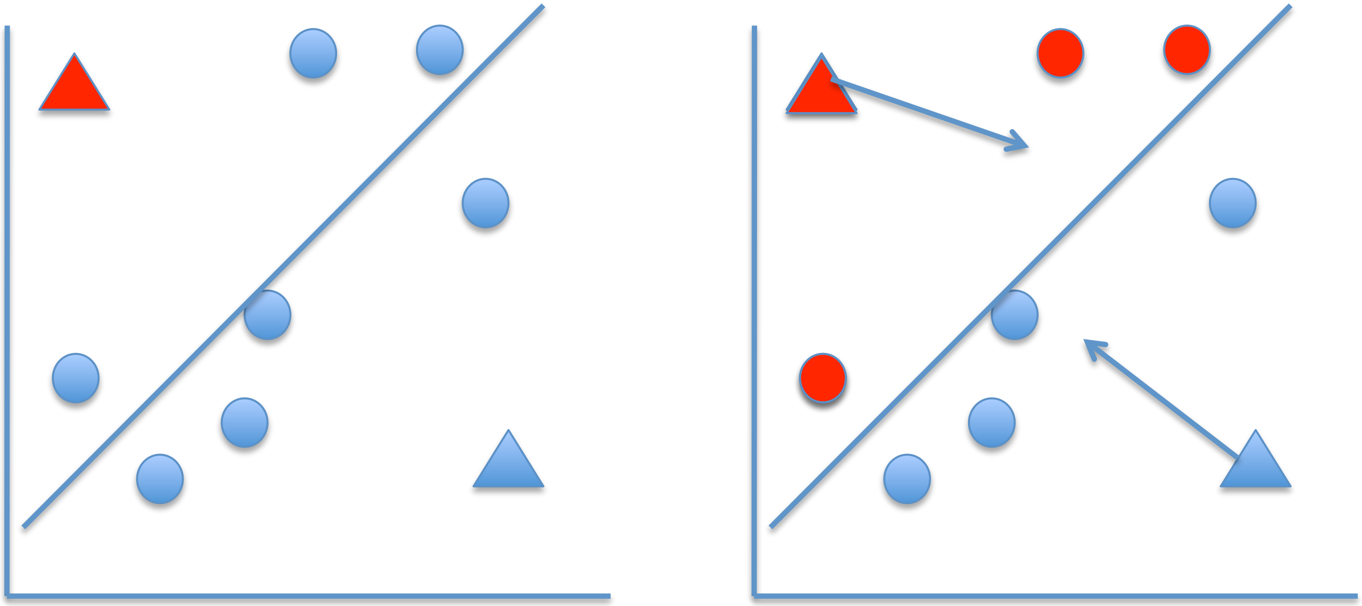
- I didn’t want to buy/print a poster, so I just printed out these slides in the appropriate size augmented with cut out colored paper shapes/backgrounds/labels/etc.
- Used lots of colored construction (primary red, but also yellow and blue) paper to make poster layout nicer and for more style overall, as much as my lackluster art skills allowed.
- (That’s why title isn’t here – used colored construction paper)
- Note some slides (e.g. k-means demo) are for poster (more targeted at lay audience) and will not be in final paper at all (but were useful several times at Poster Session!)
- Assumed would be accompanied by me talking about things… so awkward to turning it in. If you want something else, let me know...

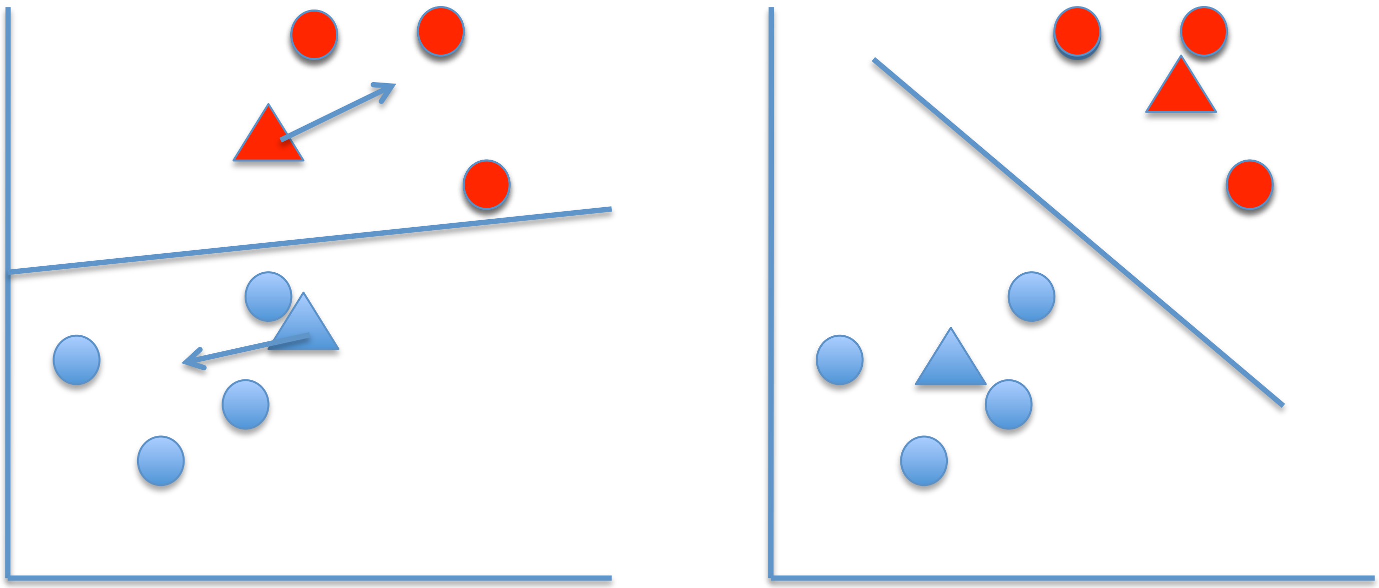
Motivation

- Policy initiatives promote particular factors such as education, home ownership, or marriage in the hope that they translate into well-being for citizens.
  - For the most part, these are promoted not as a goal in themselves, but as a means to other goals (e.g. reducing criminality or poverty)

- Therefore, finding the actual strong predictors of the desired outcomes would be exceptionally useful, if nothing else to verify that the factors we're promoting do actually promote the outcomes we want.
  - Or to gives us an idea what to prioritize when making trade offs (e.g. is education or family stability more important on the margin?)

Traditional K-Means clustering, 2-D





Data Set

- US Census Data (old but nicely formatted)
  - 3620 samples have incomplete data, removed
    - Alternately, could replace with e.g. mean (though most incomplete samples are irregular in other ways)
  - Leaves 45222, randomly split for training set
- First pass – Naïve Bayes
  - Training error (~15%) & Generalization error (~%20)

Sample Features, in Approximate Order of Importance

Feature	High Income	Low Income
Relationship	Married	Other
Age	Moderately-Older	Moderately-Younger
Hours per Week	Higher	Lower
Sex	Moderately-Male	Moderately-Female
Capital Gain	Higher	Lower
Education Level	More Educated	Less Educated
Race	Moderately-White	Other

K-means

- Since not after a particular prediction, just try to find any pattern in the data (i.e. unsupervised learning)
- Won’t discuss whole algorithm here, see diagram for simple example or ask (very intuitive graphically)
- Only works for continuous features – a problem because much of census data is categorical
  - Running on age, edu-num, cap-gain, cap-loss, and hours-worked. vary init and num clusters). also, normalize input.
  - Normalize Variance
  - Approximate Clusters (similar results for any number of clusters)
    - Young-ish, low-hours, low-education, low cap. Gains
    - Mid-age and average in every way
    - Mid-age with long hours, high education, and high capital gains

- Varied: cluster#, initialization pts, features used
  - Also varied distance measurement (traditionally L^2/ Euclidean norm), see notes below (k-modes)
- K-means is polythetic – though examples in the cluster are similar to each other, can't really put a finger on how they are similar (contrast monothetic)
  - This exact problem is why not particularly useful to display a 2-D graph along 2 features here – wouldn't display clear clustering. (Except a few rich people working long hours)
- Hurting for lack of Categorical data...

K-modes and custom dissimilarity measures

- Extending K-means to categorical data
  - Most economic data collected is categorical data (e.g. census questions), so would like to use
- Rather than use distance from means, find mode of cluster and determine distance by checking equality only (i.e. all non-matches to be treated as equidistant)
  - More general than dataset-specific distance measures
- Allows operation of (efficient) K-means-like algorithm on heterogeneous features.
  - Paper by Z. Huang demonstrates effectiveness with soybean disease and credit approval

Further Work

- Bayesian networks
  - Probabilistic Graphical model – directed acyclic graph with nodes representing random variables and edges representing conditional dependencies
  - Once Constructed, can be used to efficiently answer queries (e.g. prob(diseaseY|symptomsX))
  - Need to learn Structure, as well as parameters.
- Other Clustering Methods
  - (e.g. various hierarchical clustering methods)

- Better Datasets
  - The dataset used was well-formatted and fairly large (would never have been able to aggregate such data myself), but had relatively few features for prediction and imprecise data (income only given as > or <=thresh)

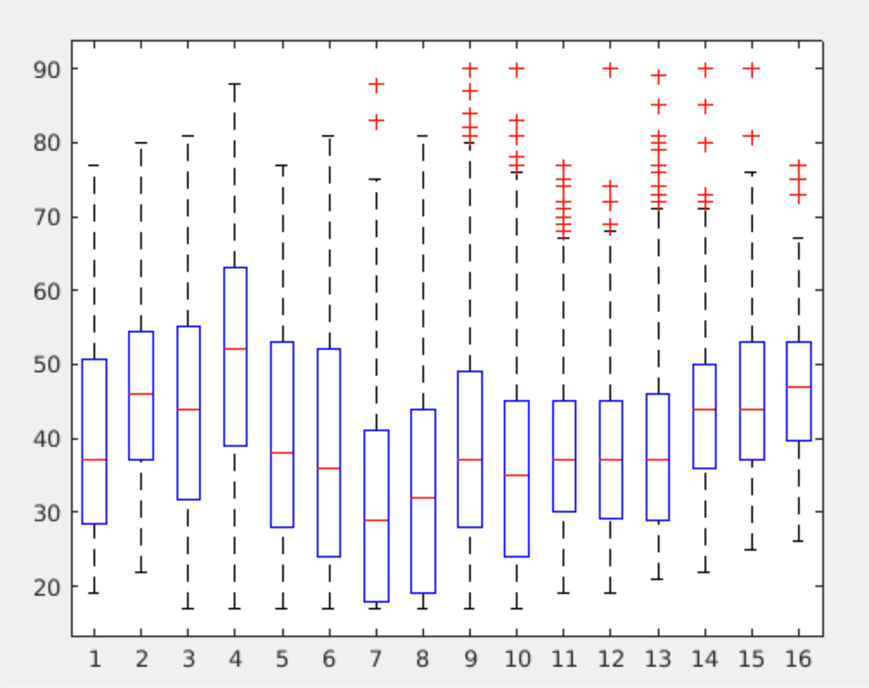
    (Goes with K-means section)
- Main unimproved K-means finding: If want to have lots of capital gains, work long hours

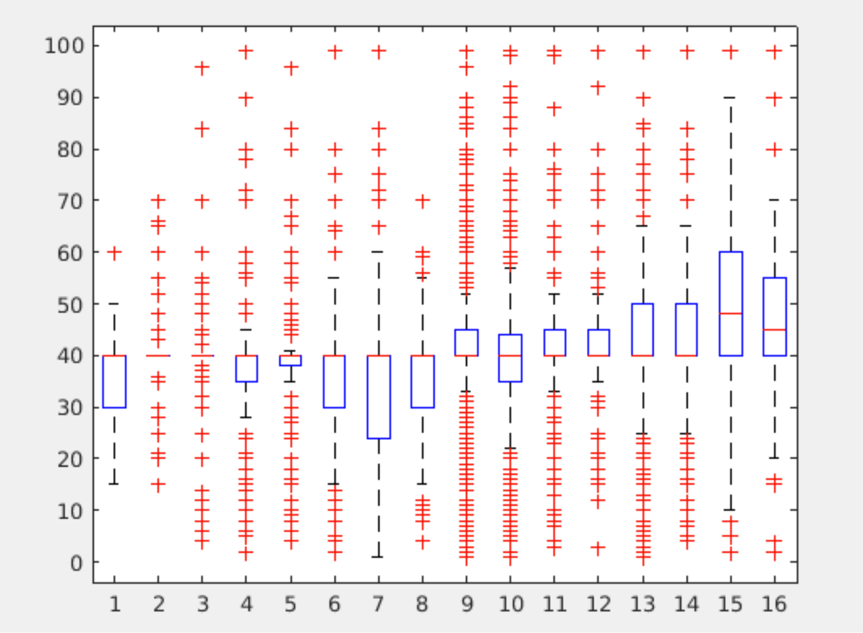
LABELS (to be cut out – see note at top)

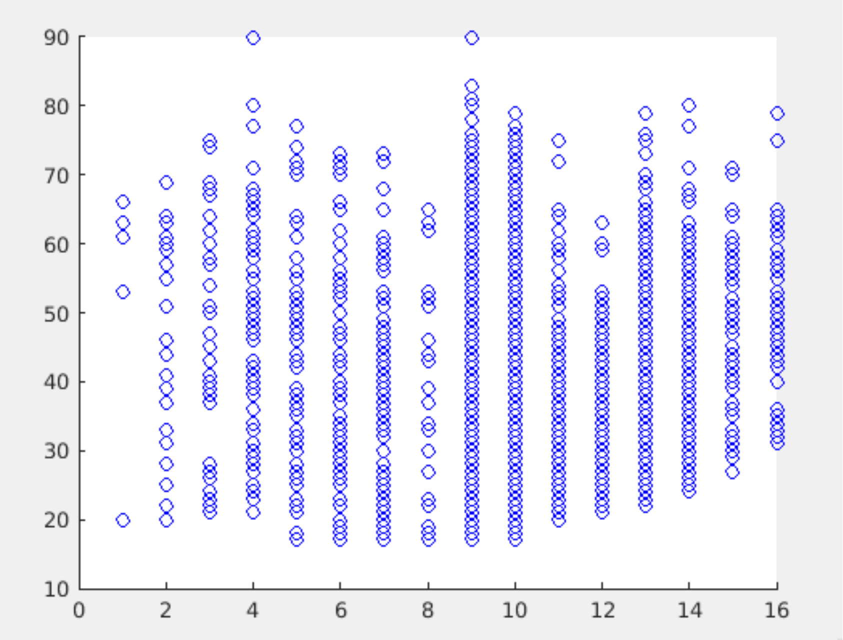
Age Age

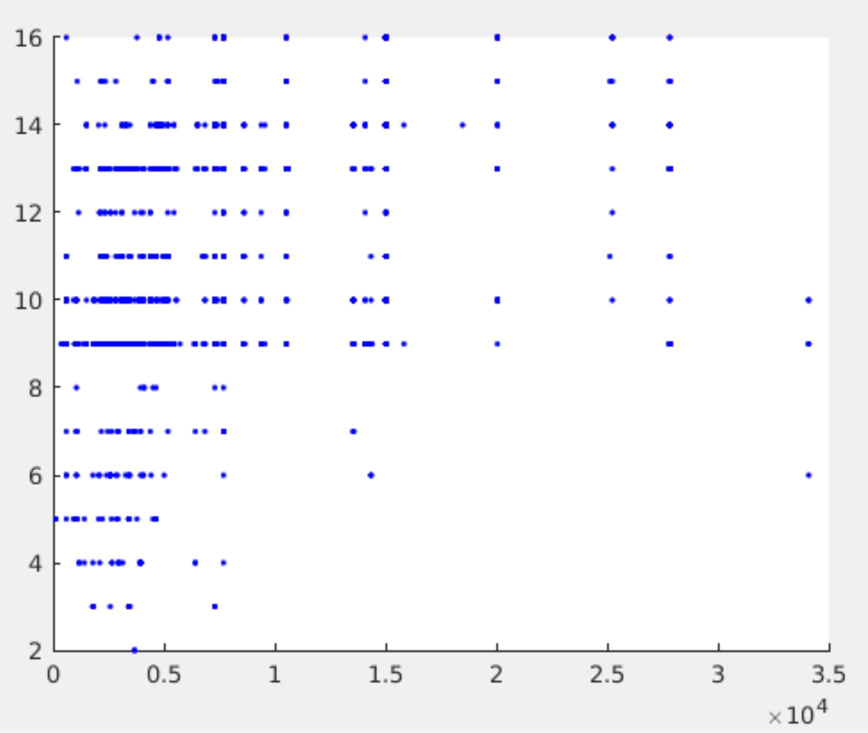
Education-Level Hours-Worked

Capital-Gains Hours-Worked






90
80
70
60
50
40
30
20
10
0
2
4
6
8
10
12
14
16


16
14
12
10
8
6
4
2
0
0.5
1
1.5
2
2.5
3
3.5
$\times 10^4$