Midterm Review

Data Mining & Analytics

Prof. Zach Pardos

INFO254/154: Spring ’19

Midterm

- Thursday, March 14th in-class, regular time
- 15 main questions + extra credit question(s)
- Two sided page of notes allowed (printed or handwritten)
- Otherwise closed computer/book

Don’t sit next to someone you have worked with on a Lab

Example Exam Question

3. You are training a decision tree on the below dataset consisting of two classes (True and False) and one continuously-valued numeric feature (A). Using the Gini Index, answer the following questions. *(4pts)*

Class Label	Feature A
True	0
False	2
True	4
False	8
False	10

(a) What is the best split point of Feature A? *(2pts)*

(b) Choose **two** of the following decision tree over fitting reduction methods and describe them: minimum leaf, max depth, pre-pruning, post-pruning. *(2pts)*

Topics

- Data transformation (pre-processing)
- Clustering (k-means)
- Classification (trees, neural nets)
- Combining classifiers (ensembles)
- Testing generalizability of models (cross-validation)

Data transformation (pre-processing)

Subtopics

- Feature engineering (pandas)
- Representing data to fit the prediction task
- Normalization

  e.g., Z-score:

Formula to find population mean

$$\mu = \frac{\sum x}{n}$$

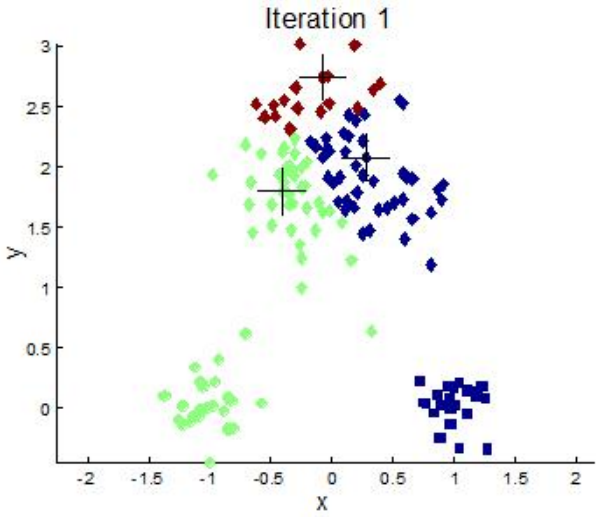
Formula to find population standard deviation

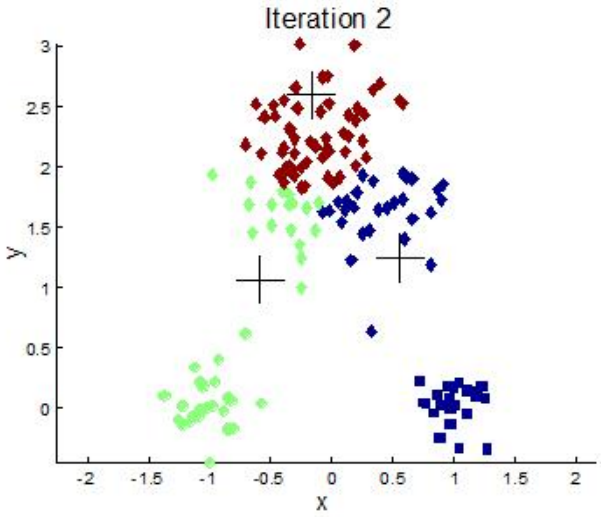
$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

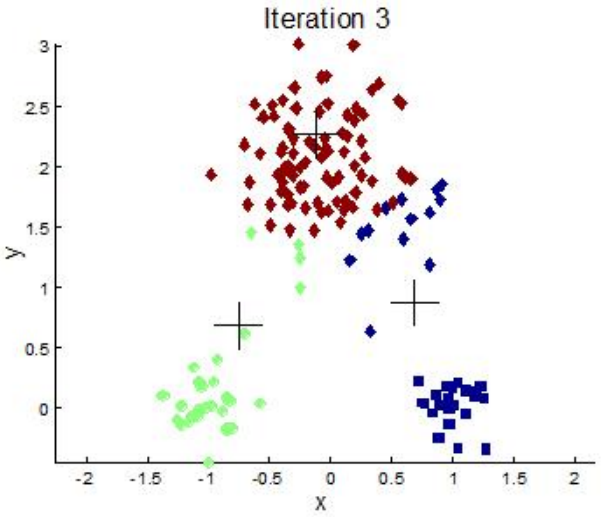
Formula to find the **z-score**

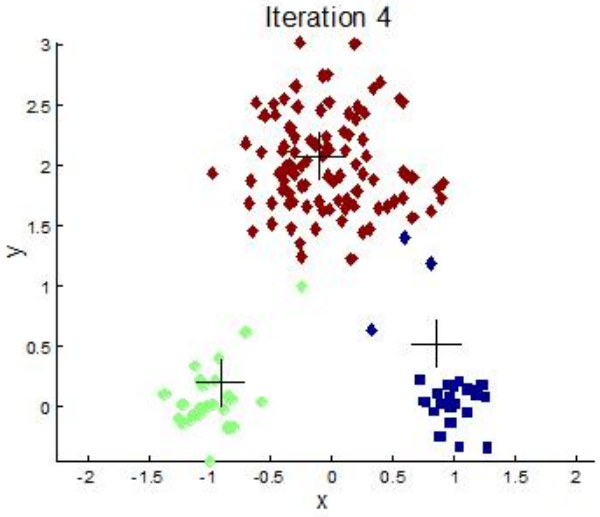
$$z\ score = \frac{(x - \mu)}{\sigma}$$

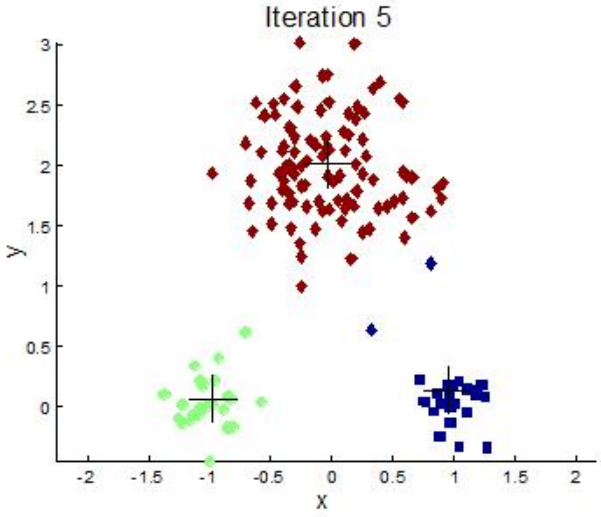
Clustering (k-means)

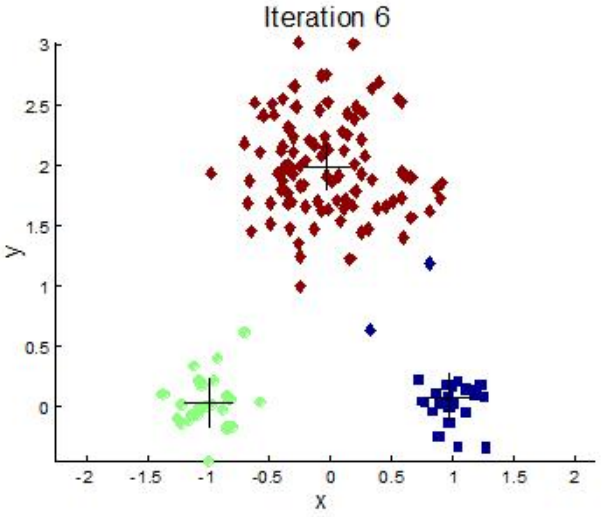












Clustering (k-means)

Subtopics

- Types of clustering methods
- Measures of cluster goodness (SSE, silhouette score)
- The k-means algorithm
- Ways of choosing K (elbow method)

Classification (decision trees)

Subtopics

- Characterizing purity (Gini/Info)
- Splitting based on features to improve purity (trees)
- Improving the generalizability of training trees (pruning)

Classification (decision trees)

Subtopics

- Feed forward neural networks
- Matrix multiplication of layers
- Backpropagating error (conceptual)
- Activation functions

Combining classifiers (ensembles)

Subtopics

- Simple combiners
- Bagging (e.g., random forests)
- Boosting (Adaboost)
- Blending/Stacking

Evaluating model generalizability (cross-validation)

Subtopics

- Error metrics (confusion matrix based & continuous)
- Training, validation, and testing sets
- Cross-validation
- Model selection