CS145: INTRODUCTION TO DATA MINING

Midterm Review

Instructor: Yizhou Sun

yzsun@cs.ucla.edu

November 8, 2017

Learnt Algorithms

	Vector Data	Set Data	Sequence Data	Text Data
Classification	**Logistic Regression; Decision Tree**; **KNN; SVM**; **NN**			Naïve Bayes for Text
Clustering	**K-means; hierarchical clustering; DBSCAN; Mixture Models**			PLSA
Prediction	**Linear Regression** GLM*			
Frequent Pattern Mining		Apriori; FP growth	GSP; PrefixSpan	
Similarity Search			DTW	

Midterm Exam

- Time
  - 11/13, 10-11:50am
- Location
  - In Class
- Policy
  - Closed book exam
  - You can take a "reference sheet" of A4 size
  - You can bring a simple calculator

Type of Questions

- True or false
- Conceptual questions
- Computation questions

True or False Questions

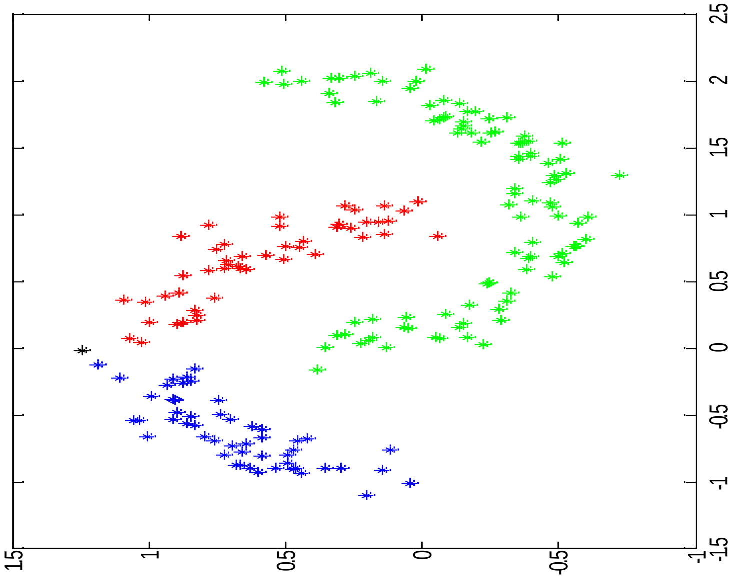
- ___Logistic regression is a linear classifier.
- ___K-means can detect arbitrary shapes of clusters.
- ___GMM does not need to specify the number of clusters.

Conceptual Questions

- What are the stopping conditions when constructing a decision tree?

- What are the main difference between KNN classifier and other learned classifiers, such as decision tree, logistic regression, and SVM?

- Suppose under a parameter setting (e.g., Eps = 0.2; minpts = 4) for DBSCAN, we get the following clustering results. How shall we change the two parameters (eps and minpts) if we want to get two clusters?



Computation Questions

- Information gain computation in decision tree

- SVM problem in HW2

- Classification evaluation