UNIVERSITY of DELAWARE UNIVERSITY of DE

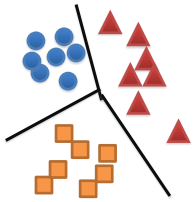
Text Categorization
- Discriminative Methods

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

UNIVERSITY of DELAWARE UNIVERSITY of DE

Discriminative Classifiers

- Rocchio classification
- KNN (K nearest neighbors)
- Support Vector Machines (SVM)



UNIVERSITY of DELAWARE UNIVERSITY of DE

Discriminative Method (1): Rocchio Classification

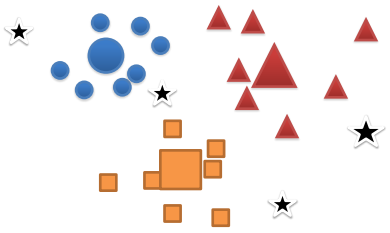
UNIVERSITY of DELAWARE UNIVERSITY of DE

Rocchio classification

- **Main ideas**
  - Representing documents as vectors
  - Assuming we have a distance measure between documents
  - Using centroids to define the boundaries
  - Assigning new documents to the category with the closest centroid

UNIVERSITY of DELAWARE UNIVERSITY of DE

An Example of Rocchio classification



UNIVERSITY of DELAWARE UNIVERSITY of DE

How to compute Centroids?

$$centroid(c) = \frac{\sum_{d \in D_c} v(d)}{|D_c|}$$

where D_c is a set of all documents with label c and $v(d)$ is the vector space representation of d.

UNIVERSITY of DELAWARE

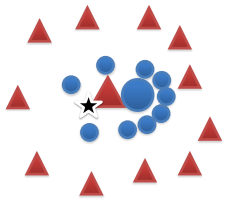
Summary of Rocchio Classification Method

- It represents each class using its centroid.
- Assigning the label based on the distance to the centroids.
- It can not guarantee that classifications are consistent with the given training data
- Only used for text categorization, but quite effective

7

UNIVERSITY of DELAWARE

Limitaiton of Rocchio Classifier



UNIVERSITY of DELAWARE

Discriminative Method (2): KNN

10

UNIVERSITY of DELAWARE

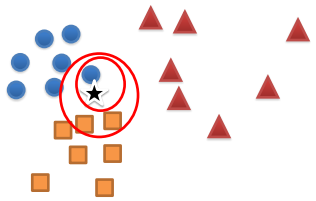
K-Nearest Neighbor Classifier

- It does not learn any model in advance.
- Given a new document, find k training examples (i.e., neighbors) that are most similar to the new document.
- Assign the category that is most common in these neighbor documents

11

UNIVERSITY of DELAWARE

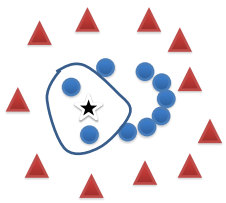
Example of K-NN Classifier



12

UNIVERSITY of DELAWARE

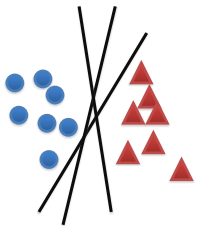
Revisiting the example



K-Nearest Neighbor Classifier

- Advantages
  - No Training needed
  - Intuitive
  - Empirically effective
  - Voting strategy can be improved by weighting the neighbors.
- Disadvantages
  - High time complexity
  - Performance is sensitive to the number of neighbors.

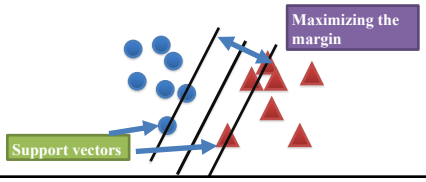
Discriminative Method (3): SVM



Find a,b,c, such that
- *ax+by>c, blue*
- *ax-by < c, red*

Support Vector Machine

- It finds an optimized solution by maximizing the distance between the hyperplane and the "difficult points" close to decision boundary



Discriminative Classifiers

- Rocchio classification
- KNN (K nearest neighbors)
- Support Vector Machines (SVM)