Automated text categorization using machine learning techniques (**Distance Metric Learning**)

Presented by Phuong (Jinny) Nguyen

[Borrows some slides from Christopher Manning and Prabhakar Raghavan]

Outline

- Text classification (k Nearest Neighbor)
- Motivations for distance metric learning
- Learning global distance metric
- Future work

Classification definition

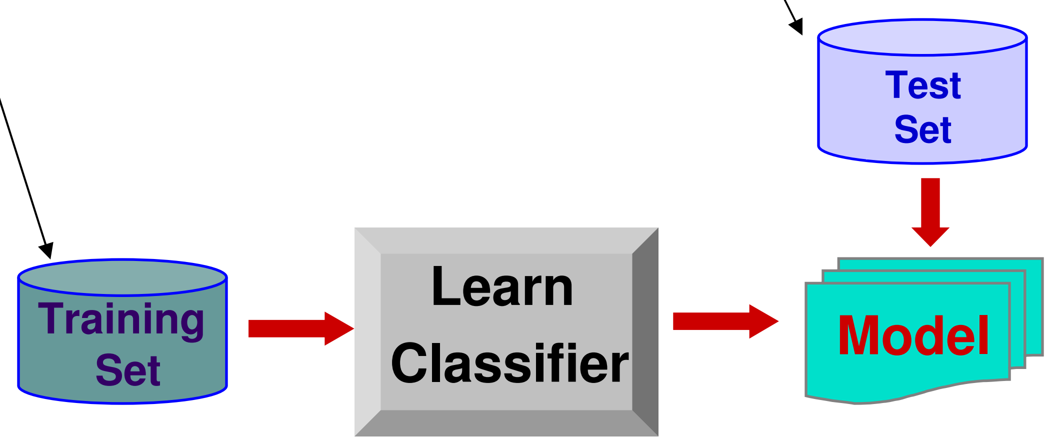
- **Given:** a collection of labeled records *(training set)*
  - Each record contains a set of features *(attributes),* and the true class *(label)*
- **Find:** a model for the class as a function of the values of the features
- **Goal:** previously unseen records should be assigned a class as accurately as possible
  - A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it

Text Classification: An Example

text class

Ex#		Hooligan
1	An English football fan …	Yes
2	During a game in Italy …	Yes
3	England has been beating France …	Yes
4	Italian football fans were cheering …	No
5	An average USA salesman earns 75K	No
6	The game in London was horrific	Yes
7	Manchester city is likely to win the championship	Yes
8	Rome is taking the lead in the football league	Yes

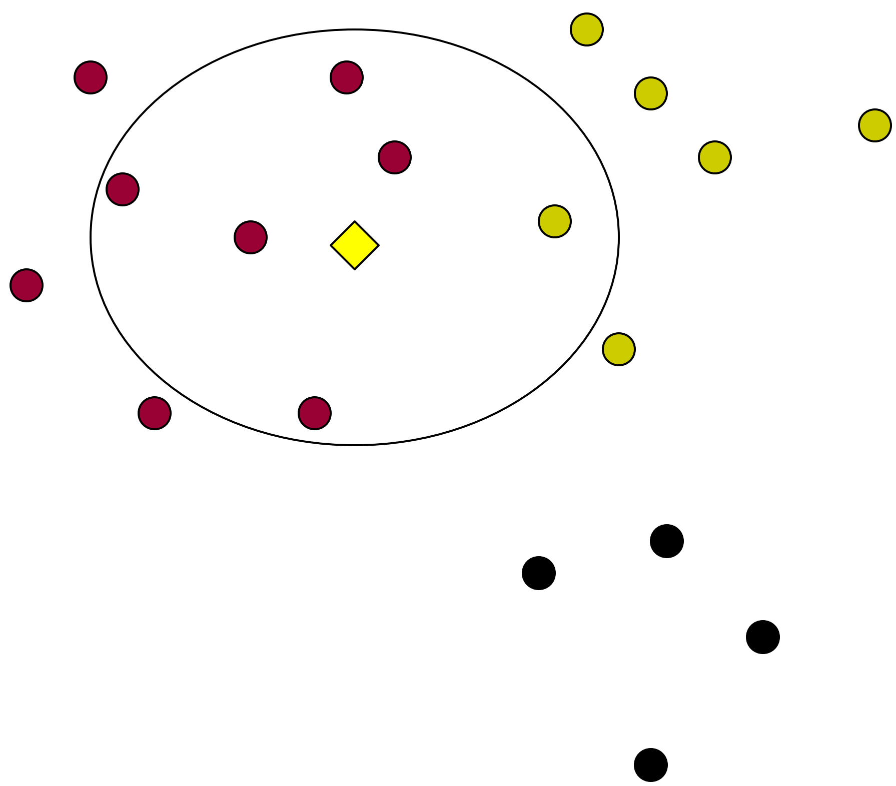
	Hooligan
A Danish football fan	?
Turkey is playing vs. France. The Turkish fans …	?



k Nearest Neighbor Classification

- To classify document *d* into class c
- Define *k*-neighborhood N as *k* nearest neighbors of *d*
- Count number of documents i in N that belong to c
- Estimate P(c|*d*) as i/k
- Choose as class argmax_c P(c|*d*) [= majority class]

Example: k=6 (6NN)



P(science|◇)?

Government

Science

Arts

Nearest-Neighbor Learning Algorithm

- Learning is just storing the representations of the training examples in *D*.
- Testing instance *x*:
  - Compute similarity between *x* and all examples in *D*.
  - Assign *x* the category of the most similar example in *D*.
- Does not explicitly compute a generalization or category prototypes.

Similarity Metrics

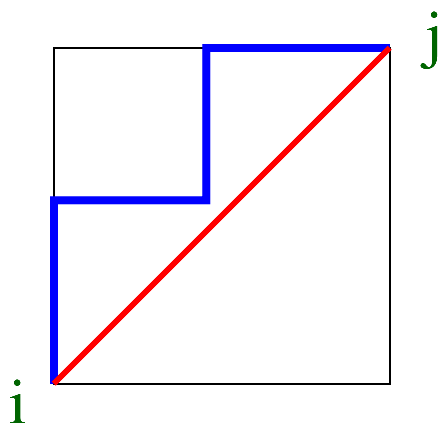
- Nearest neighbor method depends on a similarity (or distance) metric.
- Simplest for continuous *m*-dimensional instance space is *Euclidean distance*.
- Simplest for *m*-dimensional binary instance space is *Hamming distance* (number of feature values that differ).
- For text, cosine similarity of tf.idf weighted vectors is typically effective.

Similarity Metrics

- Manhattan
  - (city-block)
- Euclidean

$$EuclideanD_{ij} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$Manhat\tan D_{ij} = | x_2 - x_1 | + | y_2 - y_1 |$$



Motivations for distance metric learning

- Definition
  - Distance Metric learning is to learn a distance metric for the input space of data from a given collection of pair of similar/dissimilar points that preserves the distance relation among the training data pairs.
- Importance
  - Many machine learning algorithms, heavily rely on the distance metric for the input data patterns. e.g. kNN
  - A learned metric can significantly improve the performance in classification, clustering and retrieval tasks:

    e.g. KNN classifier, spectral clustering, content-based image retrieval (CBIR).

Supervised Global Distance Metric Learning (Xing et al. 2003)

Equivalence constraints: $S = \{(x_i, x_j) \mid x_i \text{ and } x_j \text{ belong to the same class}\}$

Inequivalence constraints: $D = \{(x_i, x_j) \mid x_i \text{ and } x_j \text{ belong to different classes}\}$,

$\mathrm{d}_{\mathrm{A}}^2(x, y) = \|x - y\|_A^2 = (x - y)^T A(x - y),\ A \in S_+^{m\times m}$ is the distance metric

- Goal : keep ***all*** the data points within the same classes close, while separating ***all*** the data points from different classes.

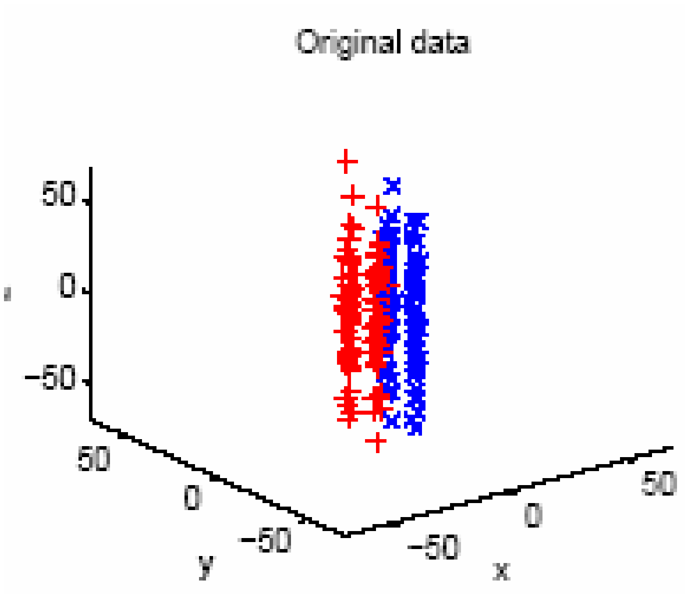
- Formulate as a constrained convex programming problem
  - minimize the distance between the data pairs in S
  - Subject to data pairs in D are well separated

Global Distance Metric Learning (Cont'd)

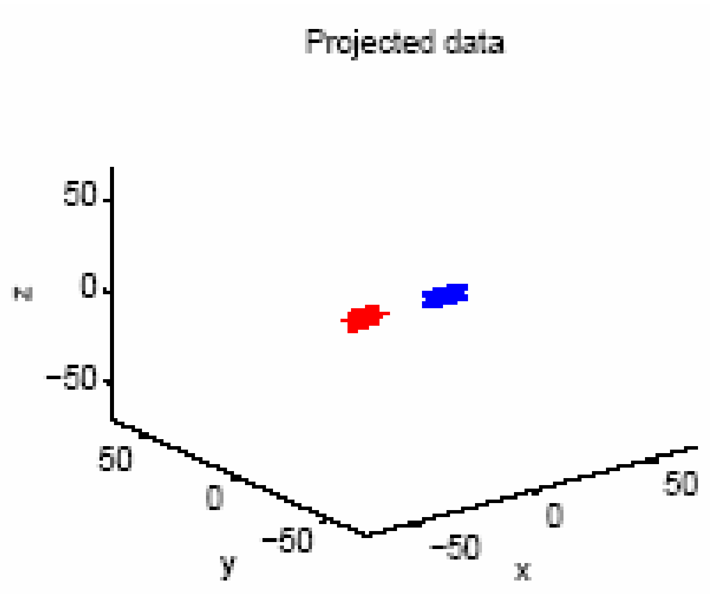
$$\min_{A \in R^{m \times m}} \sum_{(x_i, x_j) \in S} \left\| x_i - x_j \right\|_A^2 \qquad s.t. \qquad A \succ= 0, \sum_{(x_i, x_j) \in D} \left\| x_i - x_j \right\|_A^2 \geq 1$$

- *A* is positive semi-definite
  - Ensure the negativity and the triangle inequality of the metric
- The number of parameters is quadratic in the number of features
  - Difficult to scale to a large number of features
- Simplify the computation

Global Distance Metric Learning: Example I



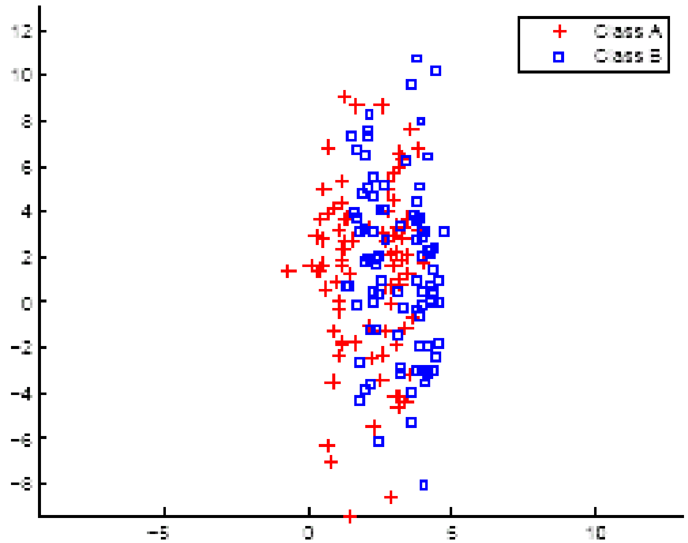
(a) Data Dist. of the original dataset



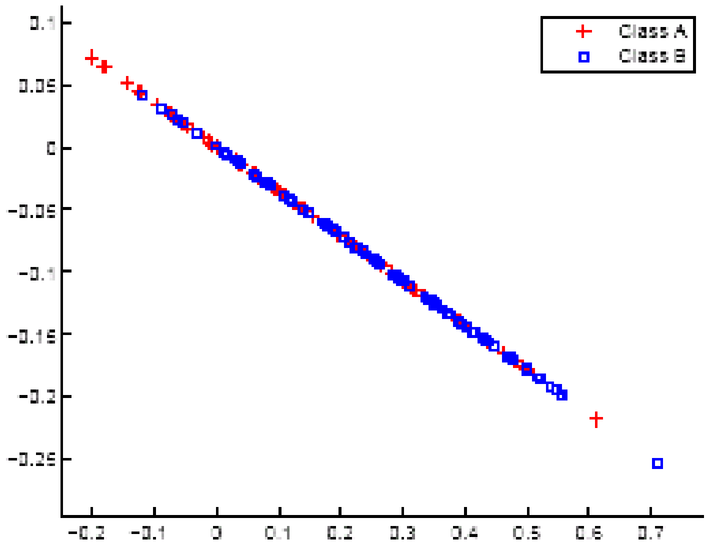
(b) Data scaled by the global metric

- Keep ***all*** the data points within the same classes close
- Separate ***all*** the data points from different classes

Problems with Global Distance Metric Learning



(a) Data Dist. of the original dataset



(b) Data scaled by the global metric

Multimodal data distributions prevent global distance metrics from simultaneously satisfying constraints on within-class compactness and between-class separability.

Local distance metric learning (Liu Yang et al 2006)

- Presents a novel probabilistic framework that learn local distance metric
  - Bring pairs from the same mode of a class closer
  - Separate nearby pairs from different classes
- How
  - Initialize the algorithm using Euclidean metric to identify the initial set of local constraints.
  - Alternately iterate between the step of local distance metric learning and the step of refining the subset of local constraints until convergence is reached

Issues

- Learn an explicit nonlinear distance metric in the local sense.
- Efficiency issue.

Thank you

Question ?