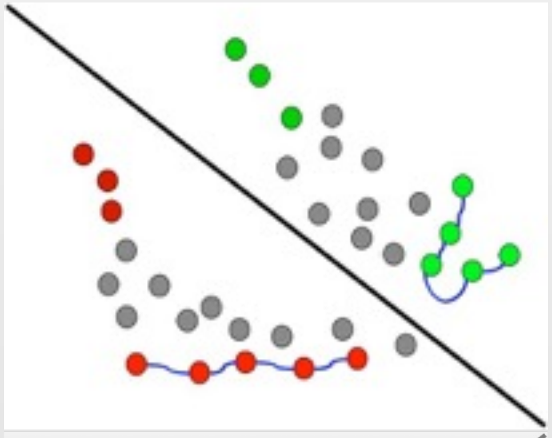
*M*usic Genre Classification:

Semi-supervised Learning with Fuzzy and Hard Clustering

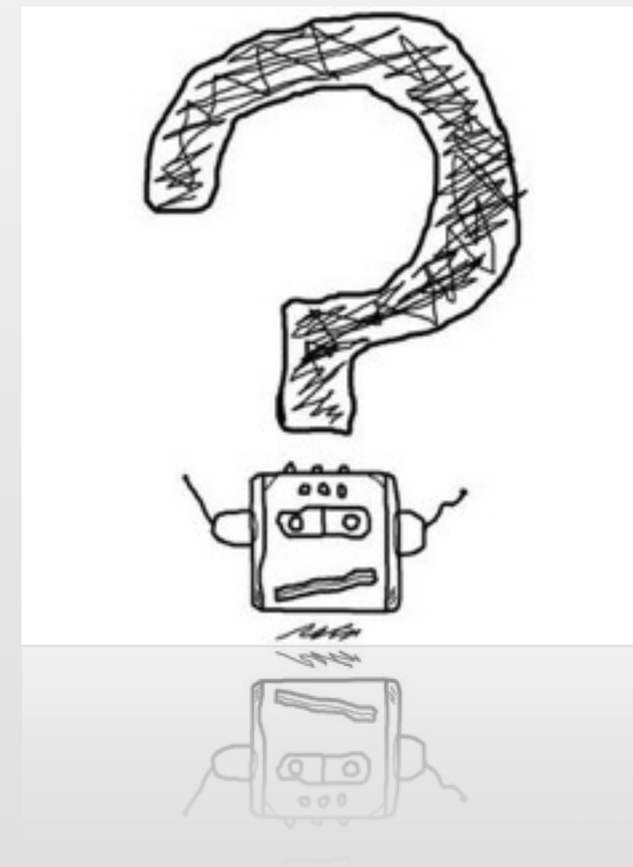


Presentation for Computer Music Seminar SS2016
Supervised by Prof. Paolo Bientinesi
RWTH Aachen University

Tina Raissi, 29 June 2016

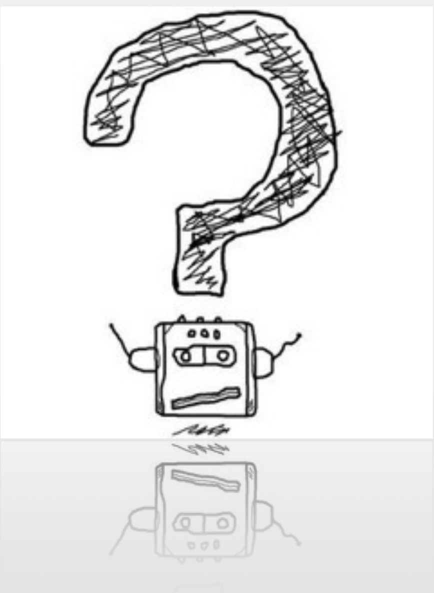
Introduction to the problem: Why semi-supervised?

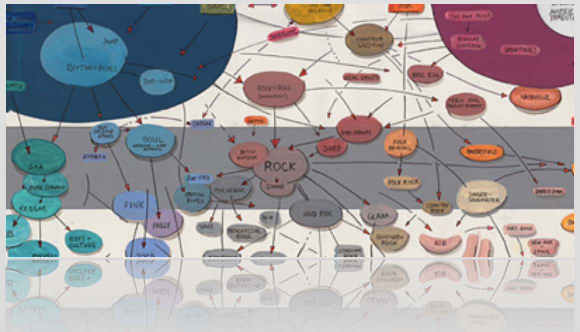
- Growth of musical collections in internet and the necessity of automatic processing
- Huge amount of unlabeled data time consuming process of manual classification
- No reliable boundaries between clusters



Introduction to the problem: Why semi-supervised?

- Growth of musical collections in internet and the necessity of automatic processing
- Huge amount of unlabeled data time consuming process of manual classification
- No reliable boundaries between clusters





What if we learn from both labeled and unlabeled data?

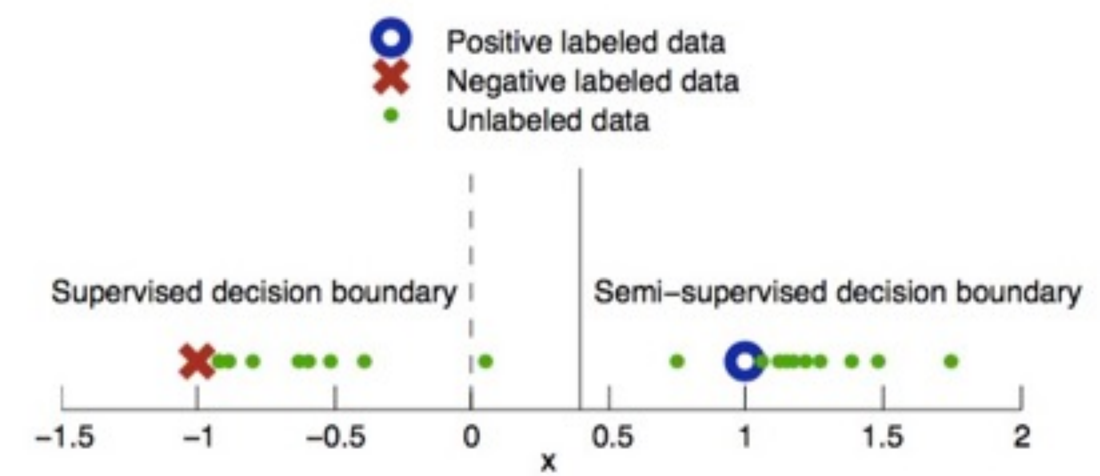
Semi-supervised Learning

Given: Labeled training data $L = \{x_i, y_i\}^{l}_{i=1}$ and unlabeled data $U = \{x_i\}^{u}_{i=l+1}$

Goal: Learning a classifier $f : \mathrm{X} \rightarrow \mathrm{Y}$

Key Concept

- Linking between the distribution of unlabeled data $P(x)$ and the target label.
- Cluster assumption : local and global consistency
- Transductive or Inductive?



A simple example to demonstrate how semi-supervised learning is possible.

Classification process

- Split in two stages:

1. **Feature extraction**: content-based extraction of Musical Surface, Rhythmic content and Pitch content features

2. **Multi-class classification**: binary classifiers extended to multi-class and collection of binary problems

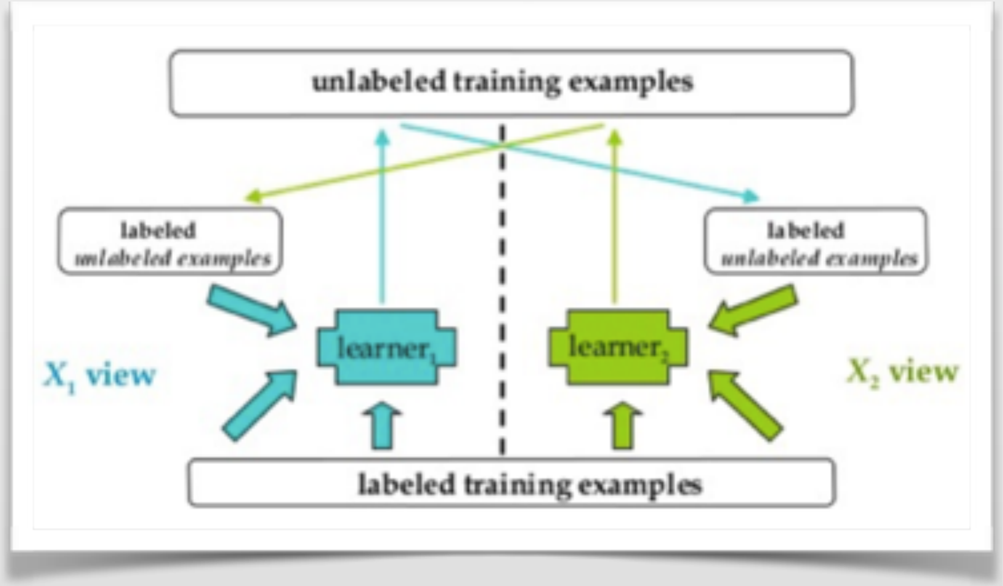
Dealing with Features

- Increasing the accuracy of classification phase by an adequate organization of features.

- Common method: Vector of features

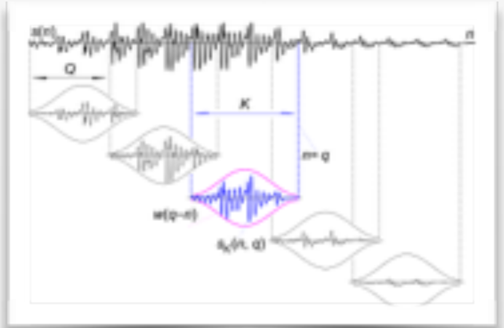
  ✧ *Problem: loss of original physical meaning*

  ✦ *Solution: multi-view features*



Multi-view: According to extraction method, physical definition and classification method the features are divided in subsets.

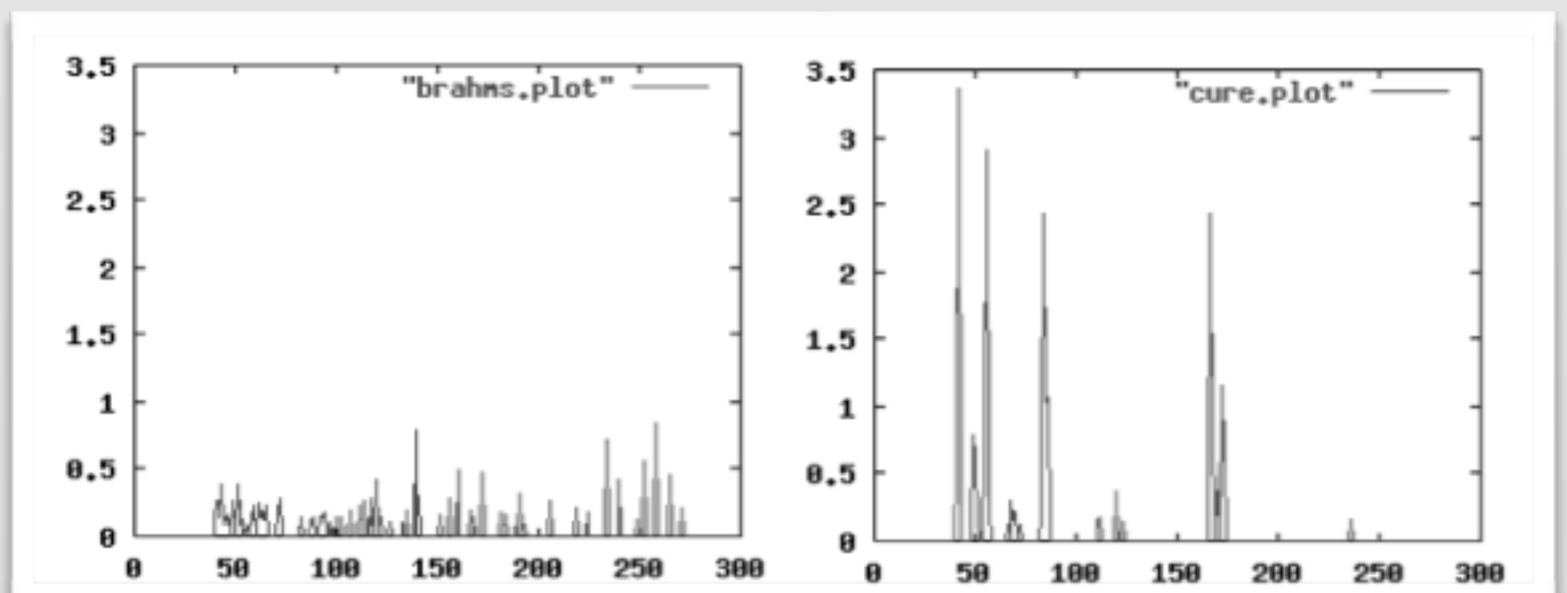
1. **Short time:** Frames, with relative signal assumed to be statically stationary and independent from others.



✧ *Problem: referred to speech recognition*

2. **Long-time**: Integration of several frames in a pre-fixed time window using statistic measures.

3. **Beat**: Transforming audio signal to human-recognizable terms such as mood and emotion.



Example of a Beat Histogram

Different Features for different classification results

	KNN	SVM	LDA	Hastie LDA	Co-Training
STFT+MFCC+DWT	81.38	81.38	83.75	83.84	
STFT+MFCC	83.04	83.04	83.74	83.50	
STFT	82.80	82.80	82.21	82.90	**88.05**
MFCC	65.37	65.37	63.66	66.89	
DWT	71.88	71.88	71.16	70.02	

Multi-view features in Xu et al.

Feature Combination	Fuzzy	SVM	Fuzzy + SVM
Long-time features	59.12%	61.20%	63.25%
Short-time features	42.54%	44.15%	48.92%
Long-time + short-time features	68.21%	71.24%	75.34%
Beat features	39.15%	39.46%	41.27%
Long-time + short-time + semantic	76.33%	87.45%	96.23%
Long-time + beat features	68.67%	72.35%	76.25%
Long + short + semantic + fuzzy vector	79.21%	—	**97.10%**

P-dimensional numerical vectors in Poria et al.

Short-time and Long-time Features compared

Features	Methods					
	SVM1	SVM2	MPSVM	GMM	LDA	KNN
DWCHs	74.9(4.97)	78.5(4.07)	68.3(4.34)	63.5(4.72)	71.3(6.10)	62.1(4.54)
Beat+FFT+MFCC+Pitch	70.8(5.39)	71.9(5.09)	66.2(5.23)	61.4(3.87)	69.4(6.93)	61.3(4.85)
Beat+FFT+MFCC	71.2(4.98)	72.1(4.68)	64.6(4.16)	60.8(3.25)	70.2(6.61)	62.3(4.03)
Beat+FFT+Pitch	65.1(4.27)	67.2(3.79)	56.0(4.67)	53.3(3.82)	61.1(6.53)	51.8(2.94)
Beat+MFCC+Pitch	64.3(4.24)	63.7(4.27)	57.8(3.82)	50.4(2.22)	61.7(5.23)	54.0(3.30)
FFT+MFCC+Pitch	70.9(6.22)	72.2(3.90)	64.9(5.06)	59.6(3.22)	69.9(6.76)	61.0(5.40)
Beat+FFT	61.7(5.12)	62.6(4.83)	50.8(5.16)	48.3(3.82)	56.0(6.73)	48.8(5.07)
Beat+MFCC	60.4(3.19)	60.2(4.84)	53.5(4.45)	47.7(2.24)	59.6(4.03)	50.5(4.53)
Beat+Pitch	42.7(5.37)	41.1(4.68)	35.6(4.27)	34.0(2.69)	36.9(4.38)	35.7(3.59)
FFT+MFCC	70.5(5.98)	71.8(4.83)	63.6(4.71)	59.1(3.20)	66.8(6.77)	61.2(7.12)
FFT+Pitch	64.0(5.16)	68.2(3.79)	55.1(5.82)	53.7(3.15)	60.0(6.68)	53.8(4.73)
MFCC+Pitch	60.6(4.54)	64.4(4.37)	53.3(2.95)	48.2(2.71)	59.4(4.50)	54.7(3.50)
Beat	26.5(3.30)	21.5(2.71)	22.1(3.04)	22.1(1.91)	24.9(2.99)	22.8(5.12)
FFT	61.2(6.74)	61.8(3.39)	50.6(5.76)	47.9(4.91)	56.5(6.90)	52.6(3.81)
MFCC	58.4(3.31)	58.1(4.72)	49.4(2.27)	46.4(3.09)	55.5(3.57)	53.7(4.11)
Pitch	36.6(2.95)	33.6(3.23)	29.9(3.76)	25.8(3.02)	30.7(2.79)	33.3(3.20)

Table 1: Classification accuracy of the learning methods tested on Dataset A using various combinations of features. The accuracy values are calculated via ten-fold cross validation. The numbers within parentheses are standard deviations. SVM1 and SVM2 respectively denote the pairwise SVM and the one-versus-the-rest SVM.

1 to 1 mapping problem with music

- **2 step reduction:**

1. Reducing number of features to those that better represents the dataset

2. Defining a small number of classes (classifier design)

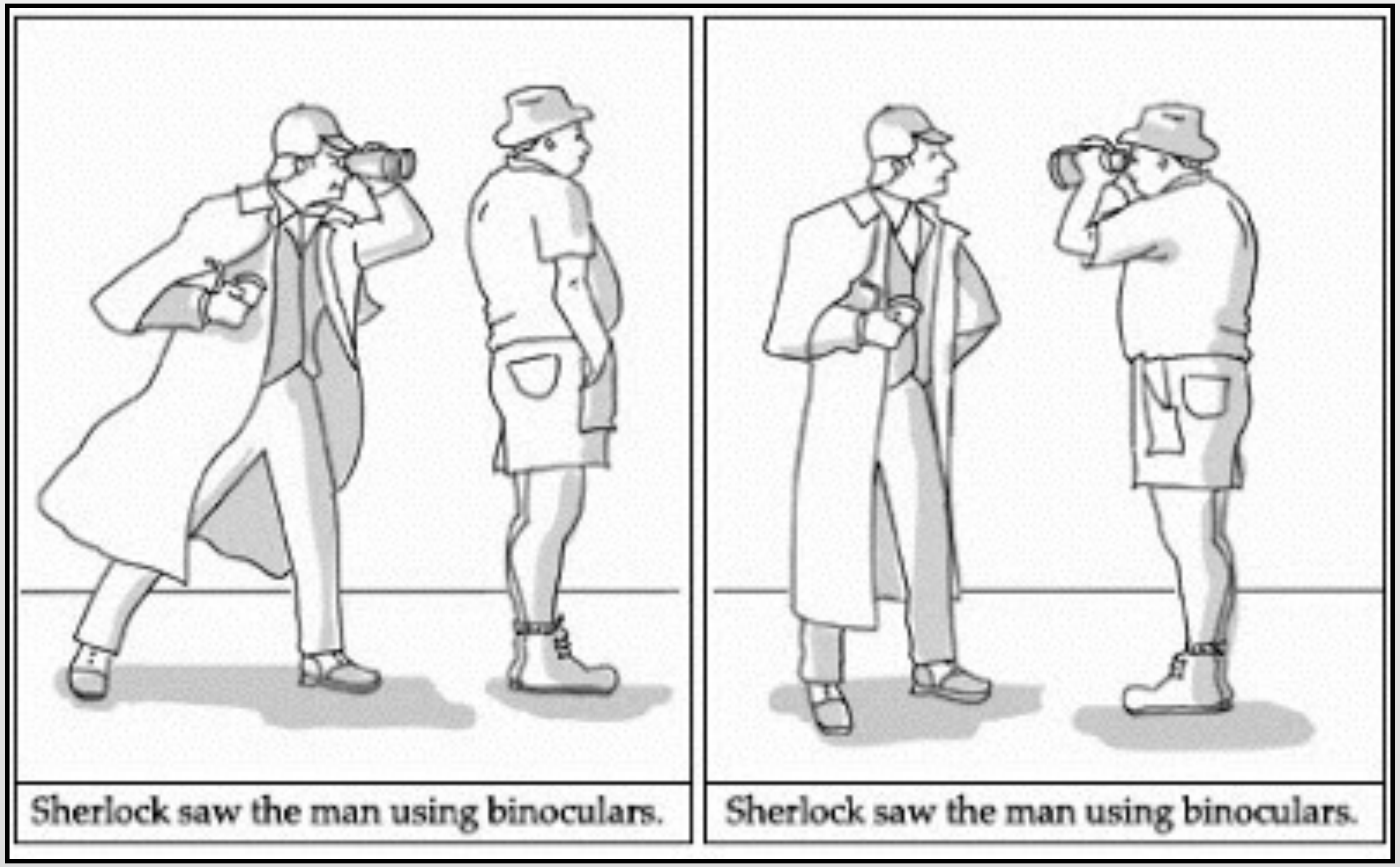
- **Crispness/Uncertainty**

Crisp ➪ deterministic yes/no: you must know the structure and parameters and have a precise description of overall system and process

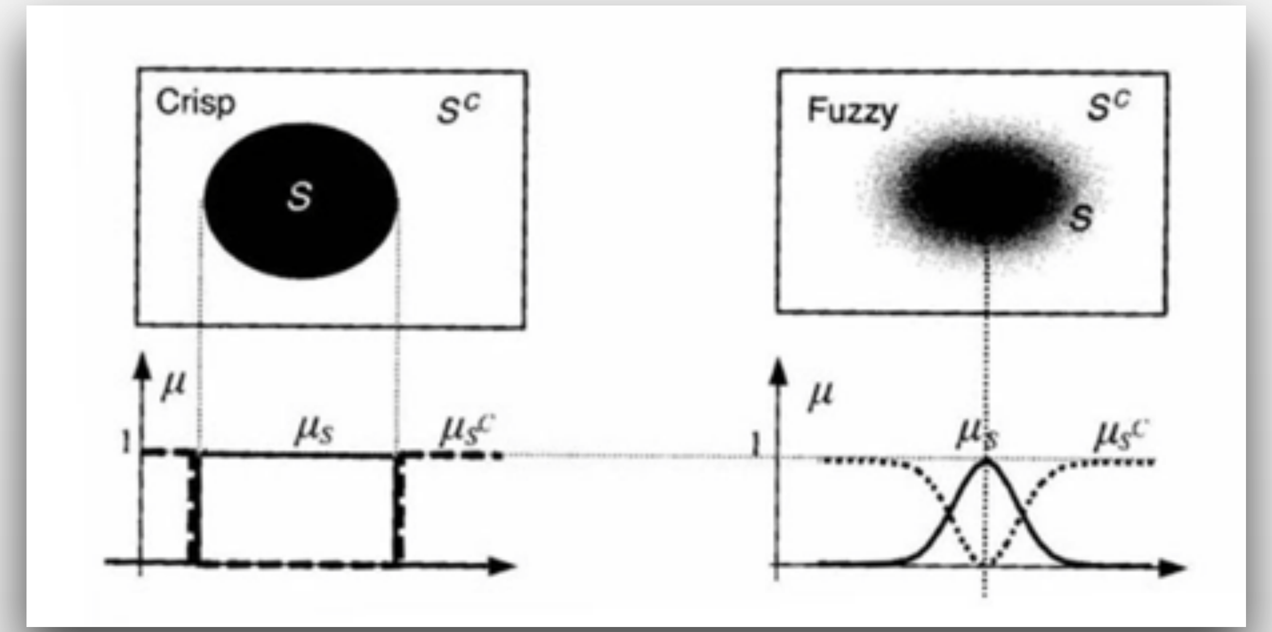
Uncertain ➪ probabilistic uncertainty (stochastic process) or fuzziness.

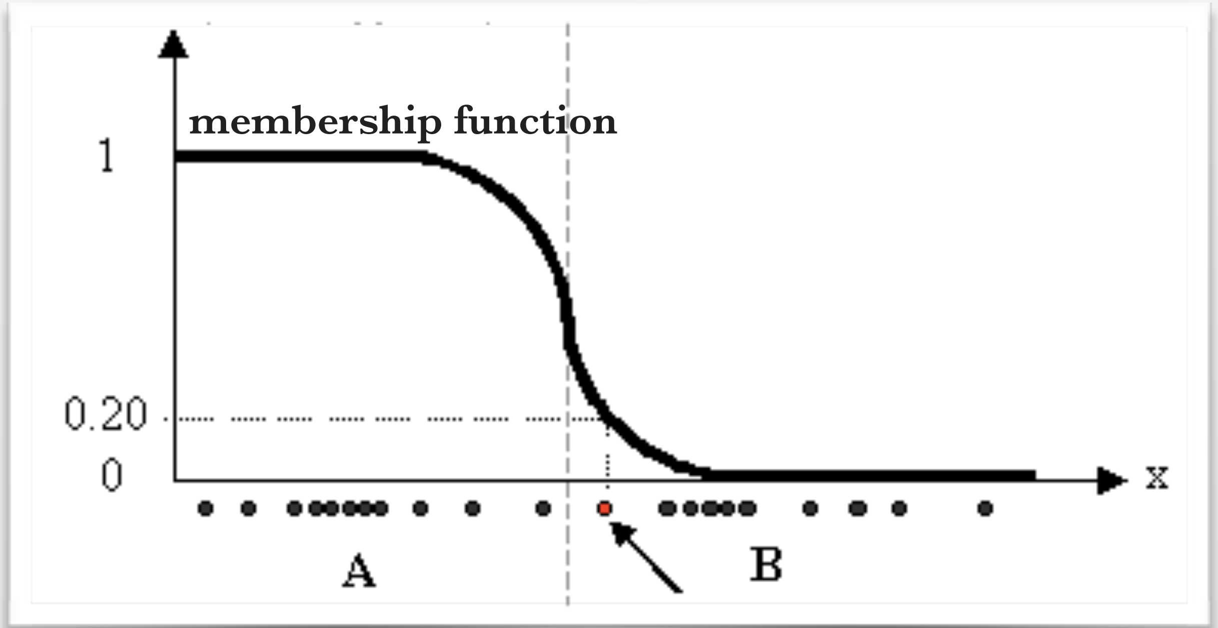
- **Fuzziness**: related to semantic meaning and imprecision due to lack of information

Sherlock saw the man with binoculars



What is a Fuzzy set?





1. Let X be a collection of objects, a fuzzy set $S \in X$ is defined as a set of ordered pair (x, μ_S) with $x \in X$
2. μ_S is called grade of membership of x in S and can range from 0 to 1.

Fuzzy Clustering

$\varepsilon = 0.01$

Input: Dataset GTZAN Genre Collection

$N := 10(Genres) * 100(Tracks)$
$p := |Features|$

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{bmatrix}$$

$$J_p(\mu, v) = \sum_{i=1}^{c} \sum_{k=1}^{N} \mu_{ik}^p \left(\|x_k - v_i\|^2 + \rho \sum_{x \in N_k} \|x - v_i\|^2 \right)$$

Output: 2 sets

1. centroids $C = v_1, v_2, \dots, v_c$ corresponding to c clusters
2. cN membership values $0 < \mu_{ik} < 1, i = 1, \dots, c; k = 1, \dots, N$ and $\sum_{i=1}^{c} \mu_{ik}, k = 1, \dots, N$

$$M = \begin{bmatrix} \mu_{11} & \mu_{12} & \dots & \mu_{1N} \\ \mu_{21} & \mu_{22} & \dots & \mu_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{c1} & \mu_{c2} & \dots & \mu_{cN} \end{bmatrix}$$

Mapping clusters to genre labels

1. For each data point you decide the cluster (centroid) to which it belongs by: $C(x_k) = \arg\max_i \mu_{ik}$

2. For each centroid you decide the target label by majority vote.

✧ *Problem:* **You may have an empty value / tie**

Final crisp decision for target label

$$M^T = \begin{bmatrix} \mu_{11} & \mu_{21} & \cdots & \mu_{c1} \\ \mu_{12} & \mu_{22} & \cdots & \mu_{c2} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{1N} & \mu_{2N} & \cdots & \mu_{cN} \end{bmatrix}$$

Descending Sort

$$\mu_k = \begin{bmatrix} \mu_{i_1 k} & \mu_{i_2 k} & \ldots & \mu_{i_c k} \end{bmatrix}$$

Pick first m

$$\mu_m = \begin{bmatrix} \mu_{i_1 k} & \mu_{i_2 k} & \ldots & \mu_{i_m k} \end{bmatrix}$$

- With ***m* = *1*** no further process
- With ***m* = *10*** you have max uncertainty
- Acceptable values for ***m*** are ***2*** and ***3***

Final crisp decision: A single label for each data point

1. Train ***n*** classifiers using ***10 + p*** features (adding membership values) and relative ***m*** labels.

$$n = \binom{10}{m} = \frac{10!}{m!(10-m)!}$$

2. Use the appropriate classifier for choosing the label for every data point

3. The highest accuracy is obtained by clustering algorithm in Support Vector Machine framework.

Evaluation

- Table I: Fuzzy clustering with different values of m using support vector machine for training the classifier
- Table II: other classifiers not using fuzzy clustering phase

Table 1: Selection of most likely fuzzy cluster

m	Accurarcy
1	76.33%
2	97.10%
3	79.38%
4	77.51%
10	67.45%

Table 2: Comparision with other procedures

Classifier	Accurarcy
KNN	54.21%
Naïve Bayes	65.88%
MLP	74.23%
This procedure	97.10%

Review of available semi-supervised methods (2005-2013)

Table 3: Snapshot of semi-supervised approaches

Researchers	Date	Feature Methods	Classification Methods
S. Poria, A. Gelbukh et al.	2013	CB[1], numerical vectors	Fuzzy & Hard Clustering
Y. Yaslan, Z. Cataltepe	2009	Prob-mRMR [2]	RASCO[3]
Y. Song, C. Zhang et al.	2007	CB, Multi-view vectors	Co-training
Y. Xu, C. Zhang et al.	2005	CB, numerical vectors	Manifold Regulariztion

[1] Content based
[2] Probabilistic minimum Redundancy Maximum Relevance
[3] Random Subspace Method for Co-training

References:

1. S. Poria, A. Gelbukh, D. Das, S. Bandyopadhyay. *Fuzzy Clustering for Semi-Supervised Learning* MICAI 2012.
2. T. Li, M. Ogihara, and Q. Li, *A comparative study on content-based music genre classification,* SIGIR, 2003.
3. G. Tzanetakis, P. Cook. *Musical Genre Classification of Audio Signals*, IEEE Transaction on Speech and Audio Processing, Vol. 10, No. 5, Jul, 2002.
4. Xu, Y., Zhang, C., Yang, J.: *Semi-supervised classification of musical genre using multi-view features,* International Computer Music Conference ICMC, 2005.
5. D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, *Learning with local and global consistency* NIPS-2004.
6. Y. Yaslan and Z. Cataltepe, *Co-training with relevant random subspaces,* Neurocomputing, vol. 73, no. 10-12, 2009.
7. J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.
8. H.-J Zimmermann, *Fuzzy Set Theory and its application*, Third Edition, Kluwer Academic Publishers, Massachusetts, 1996.
9. A. B. Goldberg, *New directions in semi-supervised learning*, phd dissertation, University of Wisconsin-Madison, 2010.
10. http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html.