BIOSTAT 527

4/26/23

Lecture 10

Hierarchical clustering
(NOT distances)

SVM - linear separable primal

- Hw2 - pb2 removed

  Project t.b.p.

  LV SVM

  LV.1 RKHS

  Bootstrap
- Boosting

  GP, RFF, DD
- ManiL; DP mix

Lecture IV – Hierarchical clustering. Comparing clusterings

Marina Meilă
mmp@stat.washington.edu

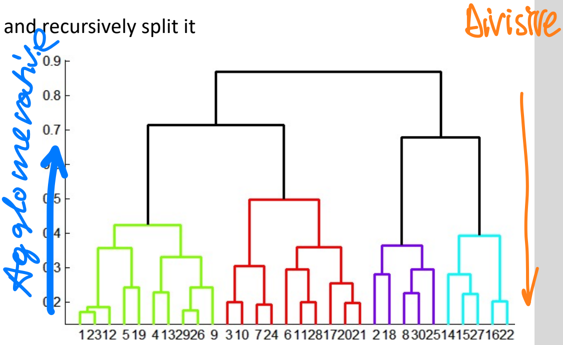
Department of Statistics
University of Washington

STAT/BIOST 527

Hierarchical Methods of Clustering

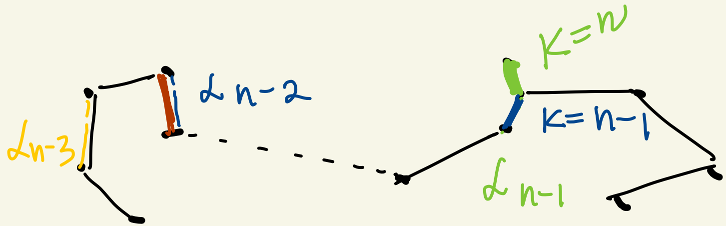
- **Agglomerative** (bottom up):
  - Initially, each point is a cluster
  - Repeatedly combine the two "nearest" clusters into one
- **Divisive** (top down):
  - Start with one cluster and recursively split it

greedy



Loss $\mathcal{L}(\Delta_K)$ = at level K $\qquad K = 1:n$

1. Single Linkage $\mathcal{L}_k \equiv \mathcal{L}(\Delta_k) = -\min\limits_{k \neq k'} \min\limits_{\substack{i \in C_k \\ j \in C_{k'}}} \|x^i - x^j\|$



Min Spanning Tree

Min → log sum edge length

Spanning → contains all nodes $\equiv$ connected $\equiv n-1$ edges

Tree → no cycles

Agglomeration

$K = n$

$K = n-1$

$K = n-2$

...

$\Rightarrow$ tree (no cycles)

(MST)

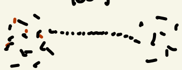
Min Spanning Tree

$\Downarrow$

can be run as Divisive too !!

Pb: NOT ROBUST

100 points

1M  1M $\qquad K=2$

Loss $\mathcal{L}(\Delta_K)$ = at level K $\quad K = 1:n$

1. Single Linkage $\mathcal{L}_k \equiv \mathcal{L}(\Delta_k) = -\min\limits_{k \neq k'} \min\limits_{\substack{i \in C_k \\ j \in C_{k'}}} \|x^i - x^j\|$

2. Least squares

$$\mathcal{L}(\Delta_K) = \sum_{k=1}^{K} \sum_{i \in C_k} \|x^i - \mu_k\|^2$$

3. Mixture log-likelihood

$$\mathcal{L}(\Delta_k) = \sum_{k=1}^{K} \sum_{i \in C_k} \ln\Big(\underbrace{\pi_k f_k(x^i)}_{\ln \pi_k + \ln f_k(x^i)}\Big)$$

CONDITIONAL log-Likelihood

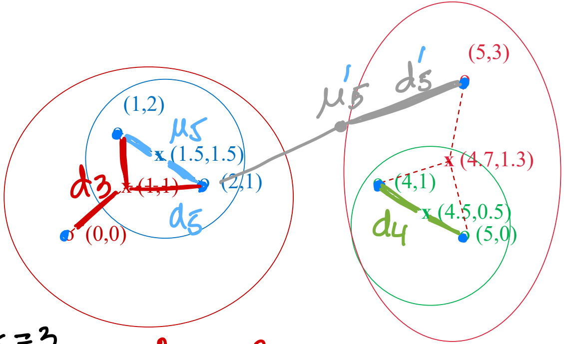
Hierarchical clustering – Overview

(Dendrograms)

- **Agglomerative** (bottom up)
  - **Single linkage**
    - based on Minimum Spanning Tree
    - $\mathcal{O}(n^2 \log n)$
    - sensitive to outliers
  - Heuristics – average linkage
  - **Agglomerative K-means**
    - Loss $\mathcal{L}(\Delta_K) = 0$ for $K = n$
    - When $K \leftarrow K - 1$ (two clusters merged), $\mathcal{L}$ increases
    - For $K = n, n-1, \ldots 2$, iteratively merge the 2 clusters that minimize increase of $\mathcal{L}$
    - $\mathcal{O}(n^3)$ – too expensive for big data
- **Divisive** (bottom down)
  - Recursively split $\mathcal{D}$ into $K = 2$ clusters
  - almost any clustering algorithm (e.g. K-means, min diameter)
  - notable example Spectral clustering (later)
  - Advantages
    - most important splits are first
    - can stop after only a few splits

Example: Hierarchical clustering

$n=6$
$K=6 \quad \mu_\ell = x^\ell$
$\mathcal{L}_6 = 0$



$K=5 \quad \mathcal{L}_5 = 0\times 4 + 2d_5'^2$ BAD

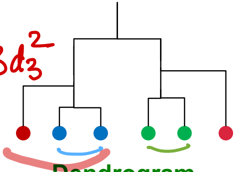
$\mathcal{L}_5 = 0\times 4 + 2d_5^2$

$\mathcal{L}_4 = 0\times 2 + 2d_5^2 + 2d_4^2$

$K=3$

$\mathcal{L}_3 = 0 + 2d_4^2 + 3d_3^2 = \mathcal{L}_4 - 0 - 2d_5^2 + 3d_3^2$

Data:
o … data point
x … centroid



Dendrogram

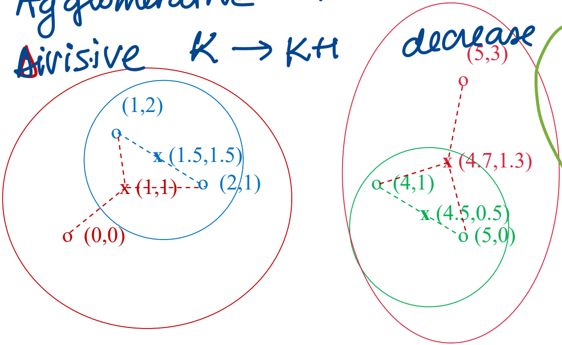
Example: Hierarchical clustering

Agglomerative $K \to K-1$ increase $\mathcal{L}$ least

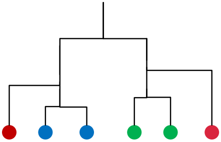
Divisive $K \to K+1$ decrease $\mathcal{L}$ most

$K=1$: 2^n options

$K=n$: $\frac{n(n-1)}{2}$ opti



Data:
o … data point
x … centroid



Dendrogram

Lecture V: Support Vector Machines and Kernel Machines

Isabelle Guyon
Corinna Cortes
Vladimir Vapnik

1. max margin
2. convex opt
3. kernel trick

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

April, 2023

Linear SVM's

why

The margin and the expected classification error
Maximum Margin Linear classifiers 1,2
Linear classifiers for non-linearly separable data

Non linear SVM

The "kernel trick"
Kernels
Prediction with SVM

Extensions

L_1 SVM
Multi-class and One class SVM
SV Regression

Reading AoNPS Ch.: Ch. 12.1–3, HTF Ch.: Ch 14 (14.1,14.2–14.2.4 kernels, 14.4 and equations (14.28,14.29) kernel trick, 14.5.1.–3 Support Vector Machines)7.1–7.4, 7.7
Additional Reading: C. Burges - "A tutorial on SVM for pattern recognition"
These notes: Appendices (convex optimization) are optional.

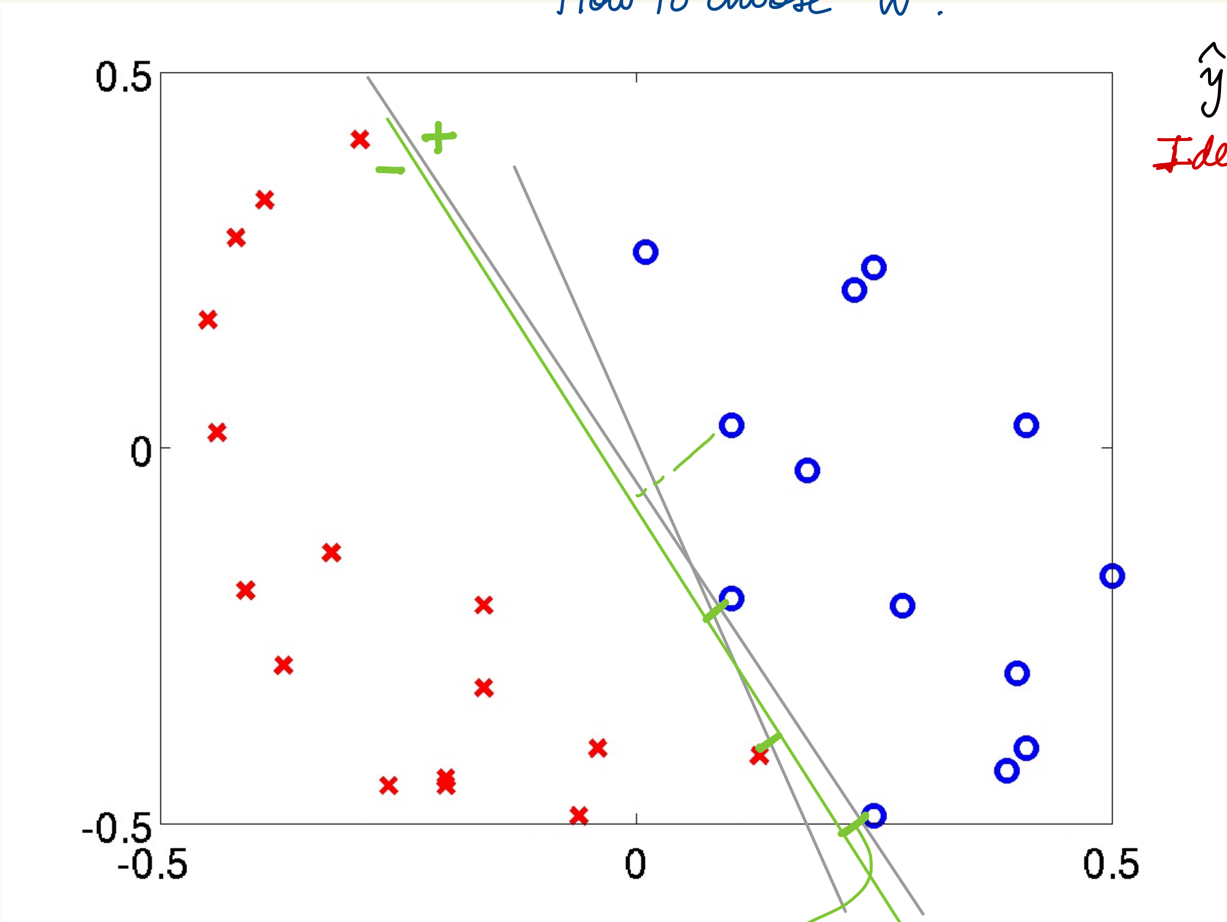
Data Linearly separable $\Rightarrow$ ∞ linear classifiers

How to choose $\hat{w}$?

$$f(x) = w^T x + b$$
$$x, w \in \mathbb{R}^d$$

$$\hat{y} = \text{sign}\, f(x)$$

Idea 1. Max Margin



Parametric
LDA
LR
Perceptron

ρ = margin of $f_{w,b}$

In $\mathbb{R}^d$: $\underbrace{d+1 \text{ points}}_{\text{support vectors}}$ to determine (w^*, b^*) = max margin hyperplane

support vectors
" points in $\mathbb{R}^d$

SVM = max margin classifier
" classifier

Robustness :

- any x^i can be perturbed by $\leq \rho$ and not change label
- all x^i's NOT SV - can be perturbed more
  - have no influence on w^*, b^*

$\Rightarrow$ fewer params ?

- Theorem (s) $\xrightarrow{\text{YES}}$

The margin and the expected classification error

Theorem Let $\mathcal{F} = \{\mathrm{sgn}\,(w^T x),\ ||w|| \leq \Lambda,\ ||x|| \leq R\}$ and let $\rho > 0$ be any "margin". Then for any $f \in \mathcal{F}$, w.p $1-\delta$ over training sets

generalization err

$$L_{01}(f) \;\leq\; \hat{L}_\rho + \sqrt{\frac{c}{n}\left(\frac{R^2\Lambda^2}{\rho^2}\ln n^2 + \ln\frac{1}{\delta}\right)} \qquad \frac{1}{\sqrt{n}}\sqrt{\frac{1}{\delta^2}+\cdots} \tag{5}$$

classif err on $\mathcal{D}$

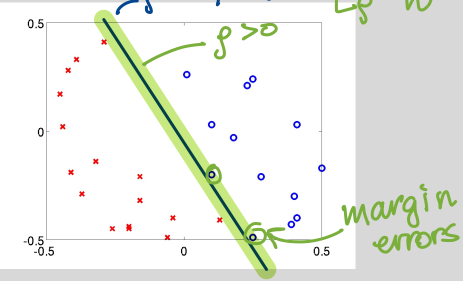
where c is a universal constant and $\hat{L}_\rho$ is the fraction of the training examples for which

including margin errors

$$y^i w^T x_i < \rho \tag{6}$$

- a data point i that satisfies (6) for some ρ is called a **margin error**
- For $\rho = 0$ the margin error rate $\hat{L}_\rho$ is equal to $\hat{L}_{01}$

$\rho=0,\ \hat{L}_\rho=0$ $\qquad$ $\hat{L}_\rho=\frac{2}{n}$



The margin and the expected classification error

Theorem Let $\mathcal{F} = \{\mathrm{sgn}\,(w^T x),\ ||w|| \leq \Lambda,\ ||x|| \leq R\}$ and let $\rho > 0$ be any "margin". Then for any $f \in \mathcal{F}$, w.p $1-\delta$ over training sets

$$L_{01}(f) \ \leq\ \hat{L}_\rho + \sqrt{\frac{c}{n}\left(\frac{R^2\Lambda^2}{\rho^2}\ln n^2 + \ln\frac{1}{\delta}\right)} \qquad (5)$$

where c is a universal constant and $\hat{L}_\rho$ is the fraction of the training examples for which

$$y^i w^T x_i < \rho \qquad (6)$$

- a data point i that satisfies (6) for some ρ is called a **margin error**
- For $\rho = 0$ the margin error rate $\hat{L}_\rho$ is equal to $\hat{L}_{01}$

Another theorem

$$\mathcal{F}_\rho = \{ f_{w,b},\ w, x \in \mathbb{R},\ \text{margin} \geq \rho \}$$

$$\text{VCdim} = \min\{ d+1,\ \frac{1}{\rho^2} \}$$

Maximum Margin Linear classifiers

How to estimate?

all x^i away from bdary

Support Vector Machines appeared from the convergence of Three Good Ideas

Assume (for the moment) that the data are linearly separable.

- Then, there are an infinity of linear classifiers that have $\hat{L}_{01} = 0$. Which one to choose?

First idea Select the classifier that has **maximum margin** ρ on the training set.

- For any parameters (w, b) that perfectly classify the data $\hat{L}(w, b) = 0$.
- Among these, the best (w, b) is the one that minimizes ρ in 5
- Hence, we should choose

① wanted

$$\underset{\rho, w, b : \hat{L}(w,b)=0}{\operatorname{argmax}} \rho, \quad \text{s.t. } d(x, H_{w,b}) \geq \rho \text{ for } i = 1 : n, \tag{7}$$

all (x^i, y^i) correct

where $d()$ denotes the Euclidean distance and $H_{w,b} = \{ x \mid w^T x + b = 0 \}$ is the decision boundary of the linear classifier.

hyperplane

- Because $d(x, H_{w,b}) = \frac{|w^T x + b|}{||w||}$ (proof in a few slides) (7) becomes

②

$$\underset{\rho, w, b : \hat{L}(w,b)=0}{\operatorname{argmax}} \rho, \quad \text{s.t. } \frac{|w^T x^i + b|}{||w||} \geq \rho \text{ for } i = 1 : n, \tag{8}$$

replace $d(\,)$

Maximum Margin Linear classifiers

sign $w^T x + b = y$ $\quad y \in \{\pm 1\}$

$y(w^T x + b) > 0$

We continue to transform (8)

- If all data correctly classified, then $y^i(w^T x^i + b) = |w^T x^i + b|$. Therefore (8) has the same solution as

③
$$\operatorname*{argmax}_{\rho, w, b} \rho, \quad \text{s.t. } \frac{y^i(w^T x^i + b)}{||w||} \geq \rho \text{ for } i = 1:n, \tag{9}$$

- Note now that the problem (9) is underdetermined. Setting $w \leftarrow Cw, b \leftarrow Cb$ with $C > 0$ does not change anything.
- We add a cleverly chosen constraint to remove the indeterminacy; this is $||w|| = 1/\rho$, which allows us to eliminate the variable ρ. We get

④
$$\operatorname*{argmax}_{w,b} \frac{1}{||w||}, \quad \text{s.t. } y^i(w^T x^i + b) \geq 1 \text{ for } i = 1:n, \tag{10}$$

Note: the successive problems (7),(8),(9),... are equivalent in the sense that their optimal solution is the same.

$$\operatorname*{argmax}_{w,b} \underbrace{\frac{1}{||w||}}_{\rho} \quad \text{s.t. } \frac{y^i(w^T x^i + b)}{\cancel{||w||}} \geq \frac{1}{\underbrace{\cancel{||w||}}_{\rho}}$$

Alternative derivation of (10)

st idea Select the classifier that has maximum margin on the training set, by the alternative definition of margin.
Formally, define $\min_{i=1:n} y^i f(x^i)$ be the **margin of classifier f on $\mathcal{D}$**. Let $f(x) = w^T x + b$, and choose w, b that

$$\text{maximize}_{w \in \mathbb{R}^n, b \in \mathbb{R}} \min_{i=1:n} y^i (w^T x^i + b) \; s.t. \; \hat{L}(w, b) = 0$$

- Remarks
  - (if data is linearly separable), there exist classifiers with margins > 0
  - one can arbitrarily increase the margin of such a classifier by multiplying w and b by a positive constant.
  - Hence, we need to "normalize" the set of candidate classifiers by requiring instead

$$\text{maximize} \min_{i=1:n} d(x, H_{w,b}), \;\text{s.t.}\; y^i (w^T x^i + b) \geq 1 \text{ for } i = 1:n, \tag{11}$$

    where $d()$ denotes the Euclidean distance and $H_{w,b} = \{ x \,|\, w^T x + b = 0 \}$ is the decision boundary of the linear classifier.
  - Under the conditions of (11), because there are points for which $|w^T x + b| = 1$, maximizing $d(x, H_{w,b})$ over w, b for such a point is the same as

$$\max_{w,b} \frac{1}{||w||}, \;\text{s.t.}\; \min_i y_i (w^T x + b) = 1 \tag{12}$$

Second idea

5 $\min_{w,b} \|W\|$ st $y^i(W^Tx^i+b) \geq 1$ for $i=1:n$

The **Second idea** is to formulate (10) as a **quadratic** optimization problem.

$$6 \quad \min_{w,b} \boxed{\frac{1}{2}||w||^2} \text{ s.t } \boxed{y^i(w^Tx^i+b) \geq 1} \text{ for all } i = 1:n \tag{13}$$

This is the **Linear SVM (primal) optimization problem**

- This problem has a strongly convex **objective** $||w||^2$, and **constraints** $y^i(w^Tx^i + b)$ linear in (w, b).
- Hence this is a convex problem, and can be studied with the tools of convex optimization.

objective
quadratic
convex

constraints
linear
$\Rightarrow \{$feasible $w,b\}$ convex

The distance of a point x to a hyperplane $H_{w,b}$

$$d(x, H_{w,b}) \quad = \quad \frac{|w^T x + b|}{||w||} \tag{14}$$

Intuition: denote

$$\tilde{w} \ = \ \frac{w}{||w||}, \ \tilde{b} \ = \ \frac{b}{||w||}, \ x' \ = \ \tilde{w}^T x. \tag{15}$$

Obviously $H_{w,b} = H_{\tilde{w},\tilde{b}}$, and x' is the length of the projection of point x on the direction of w. The distance is measured along the normal through x to H; note that if $x' = -\tilde{b}$ then $x \in H_{w,b}$ and $d(x, H_{w,b}) = 0$; in general, the distance along this line will be $|x' - (-\tilde{b})|$.