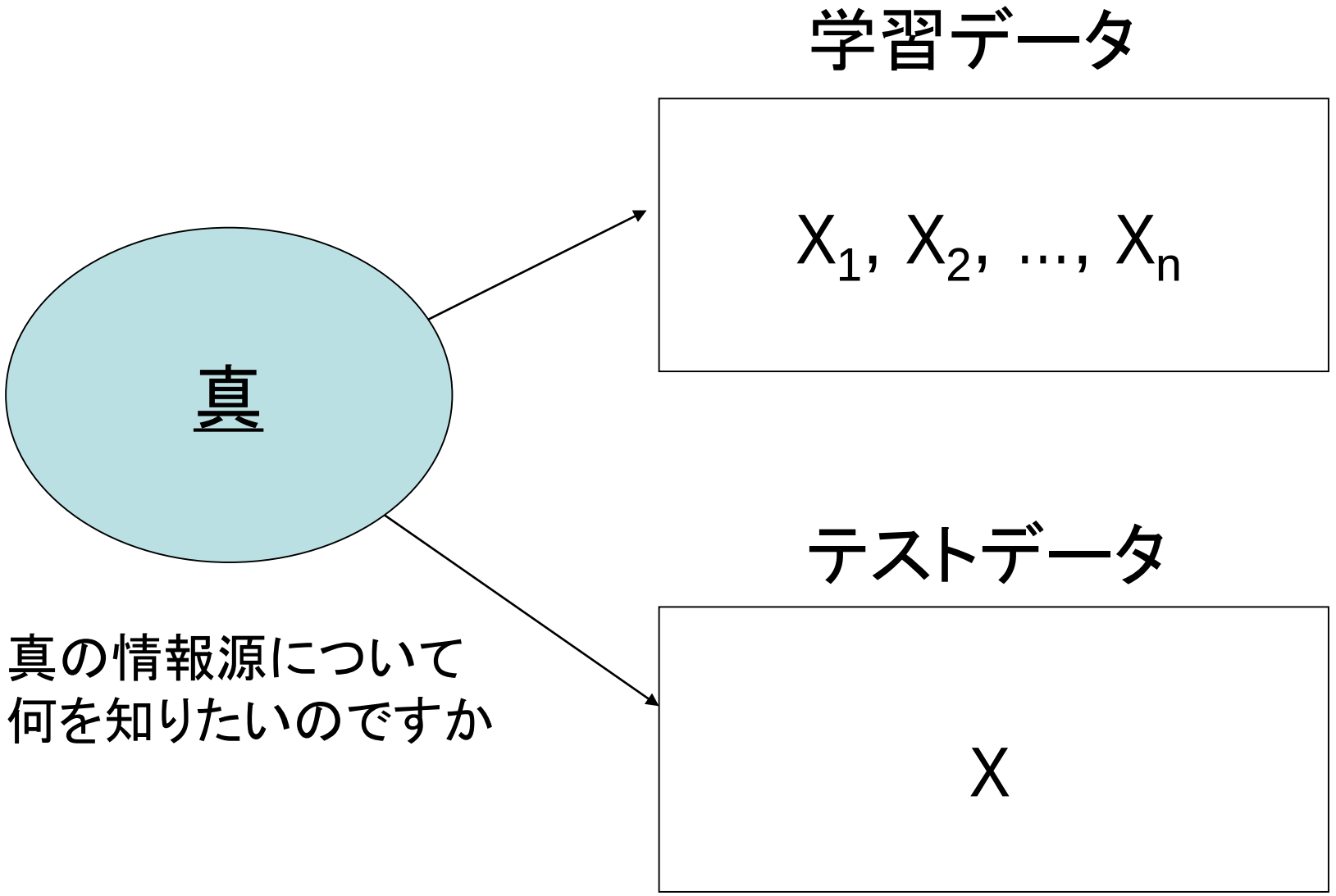
情報学習理論

渡辺澄夫
東京工業大学

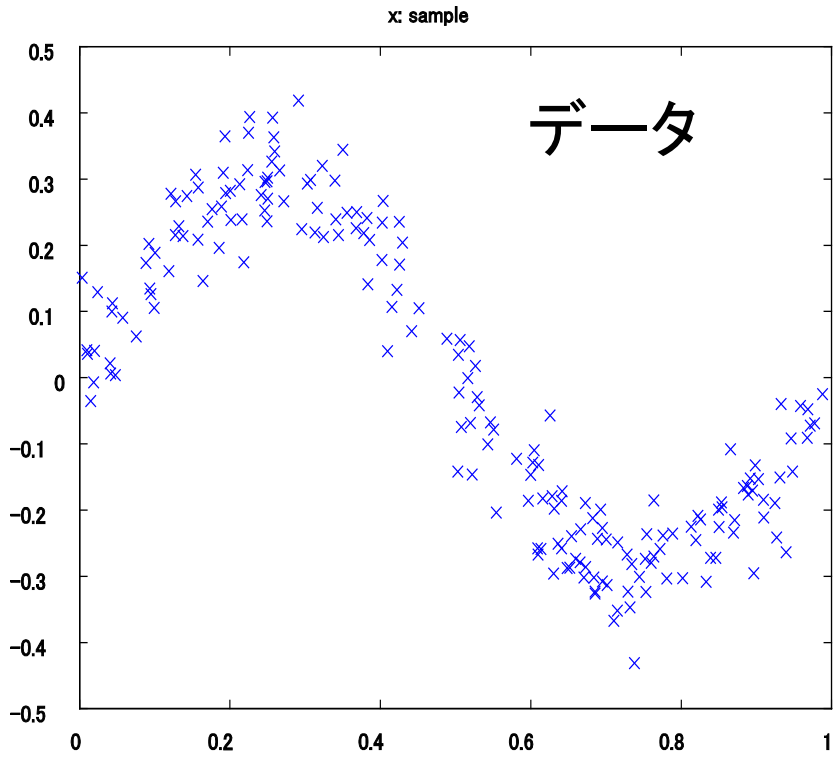
教師なし学習

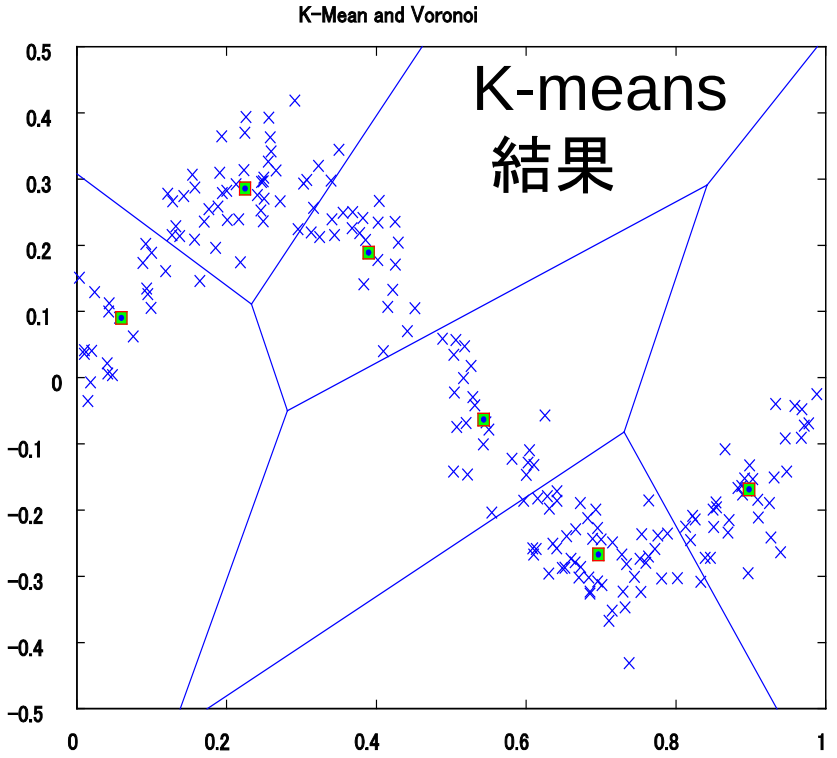


復習： K-means 法

ラベルのないデータがたくさんある → どうしよう

→ クラスタに分けて代表点を見つけた

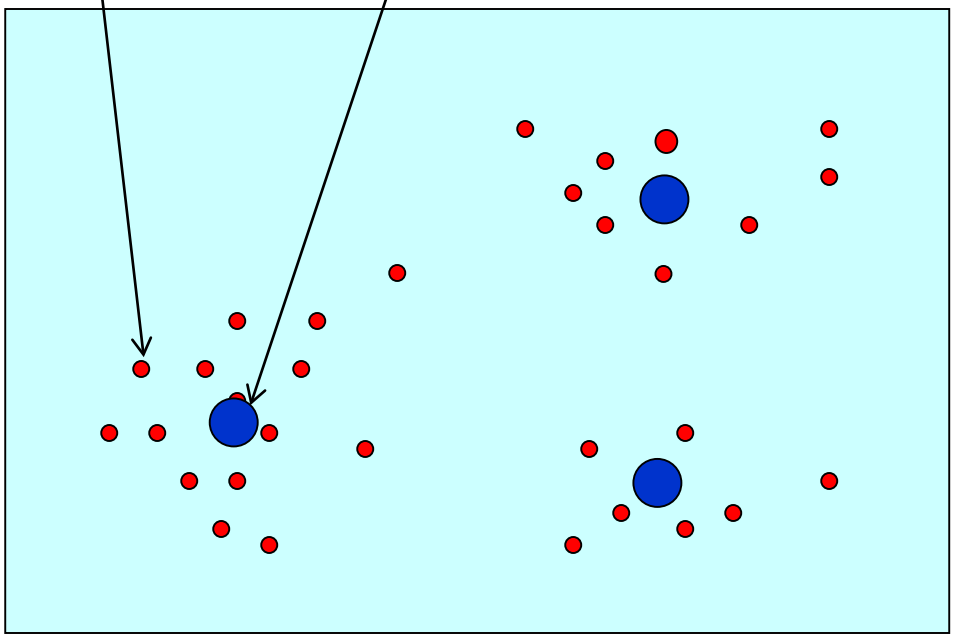




復習： クラスタリングと代表点

データ X_i （とても多い）

代表点 y_k （少ない）



代表点を選び出した!

→ で，それで？

→ 真実が見つかった？

→ 何に使うのですか

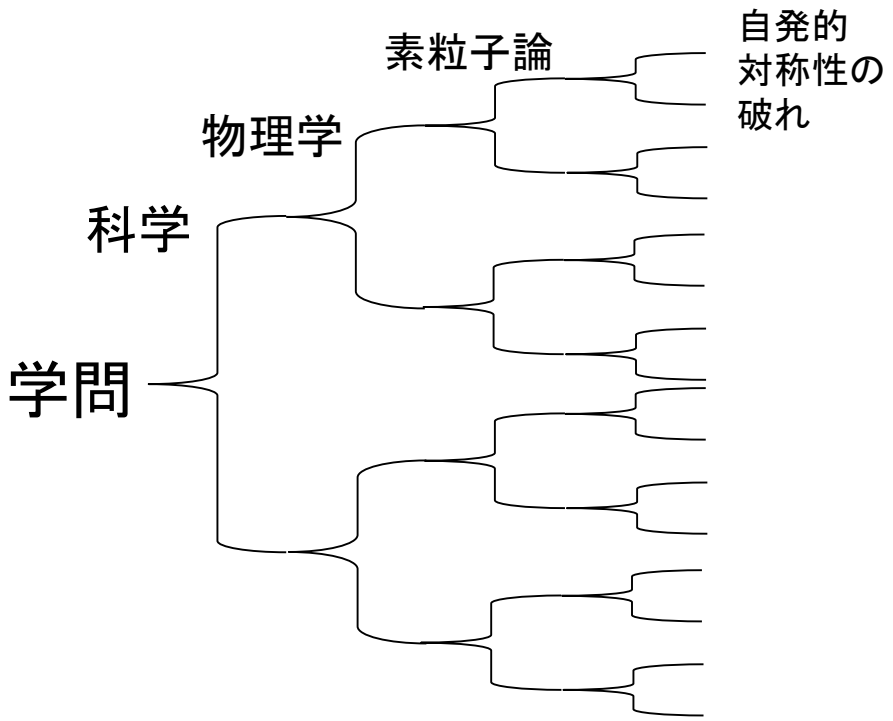

森の中から宝物を見つける

教師なし学習の応用(1)

○　階層的クラスタリング

多くの情報から構造を抽出して
理解しやすい形で提示する

経済学
　◎◎経済学
　□□経済学
科学
　宇宙論
　　ダークマター
　　ビッグバン
　物性基礎論
　　・・・

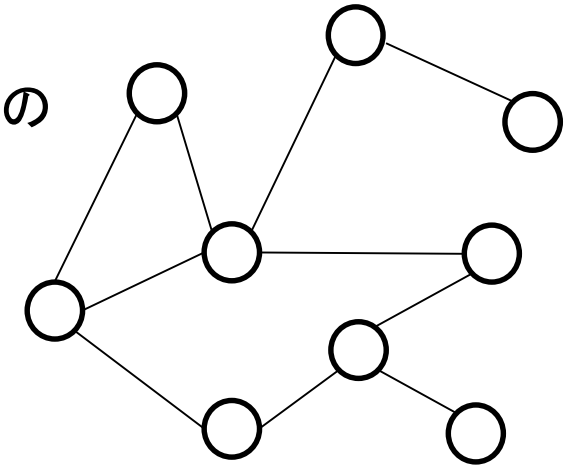


教師なし学習の応用(2)

○ 推薦システム

「＿＿＿＿が好きな人は
＿＿＿＿も好きなのでは」

関連度の
グラフ



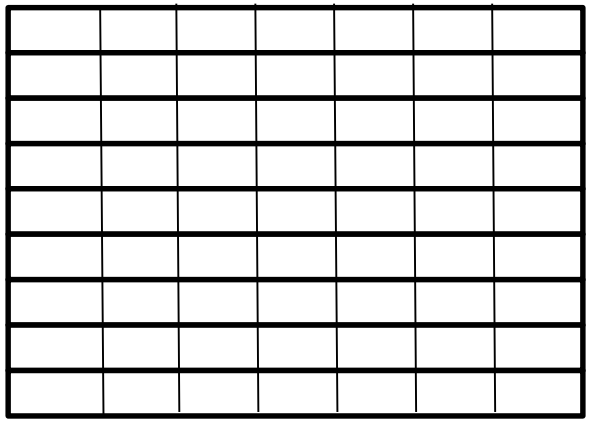
○ 共起関係の手がかり

(A) は (B) を (C) する

同じ動詞の目的語になる名詞は
似ているはず

目的語

動詞



教師なし学習の応用(3)

元画像

8種類の色で作成

クラスタ中心を

(r_k,g_k,b_k) : k=1,2,…,8

出やすい情報を中心に
圧縮したい

代表ベクトルを K-means で
選んで符号化する

領域の分割にも利用できる

実際に教師なし学習を応用してみると・・・

１．まず「当たり前の構造」が見つかる（注１）。

２．データのランダムネスに起因する偶然により
「みかけの構造」が見つかる場合がある。

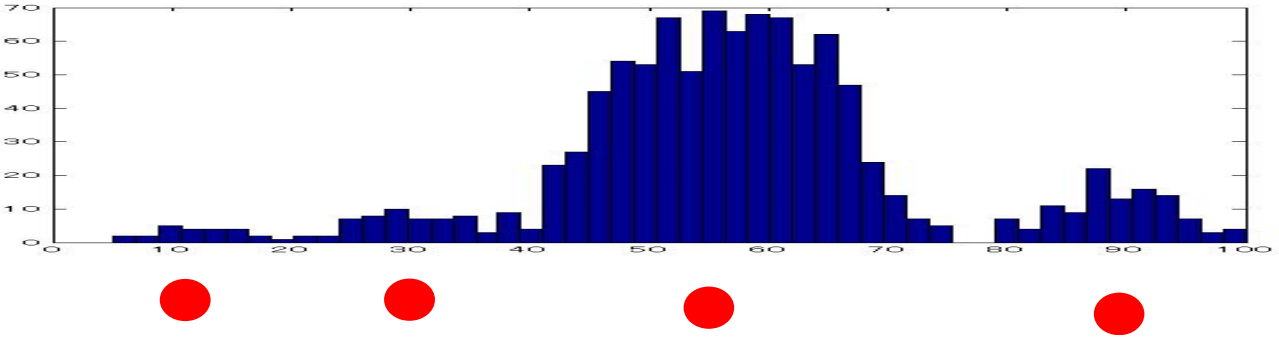
３．「発見」のように見える情報があったとき
偶然が原因の錯覚なのか、真の構造なのかを
見分けるにはどうしたらよいか　が問題（注２）

（注１）「当たり前の構造」が見つかった場合でもデータに基づいた情報は単なる想像とは異なる意義があります。

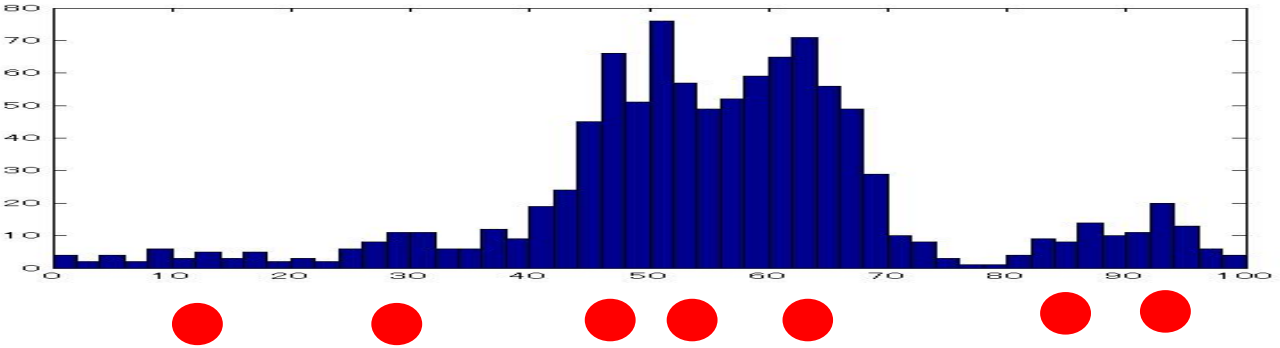
（注２）新しい構造が発見できたとすれば、そのデータ解析は大成功ですがそのように見える時ほど慎重な判断が必要です。

真の構造　VS　見かけの構造

データ例
と
学習の
結果１



データ例
と
学習の
結果２



真の情報源

（注）真の構造か偶然の錯覚かを人間の感覚で判定することはできません。
分布を目で見て決めるのは無理です。数理的な基盤が必要になります。

問1

画像のデータ圧縮に K-means を使ったときの二乗誤差と圧縮率を計算してみよう

K	4	8	12	16	20
二乗誤差					
圧縮率					

$$\text{圧縮率} = \frac{\text{縦} \times \text{横} \times \log_2 K + K \log_2 3 \text{（ビット）}}{\text{縦} \times \text{横} \times 3 \times 8 \text{（ビット）}}$$

(注意）RGBの各値は0から 255の整数（2^8)

雪の結晶は7種類に分類できる

教師なし学習の実際

実際にデータを解析するときには

(1) 目的もなくデータを解析してもうまくいきません。目的を明確にして、どんなデータを集めるべきかを考えましょう。

(2) 新しい発見をするためには、考察しているデータについての詳しい知識も必要になります。

(3) データについての専門家と協力する必要があるときにはデータ解析者は、データ解析の結果を分かりやすく説明する能力も必要になります。

教師なし学習の例

選手	パス成功	ゴール	・・・	走行距離
1	3	2		6
2	1	5		0
3	5	1		3
4	2	2		2
5	0	0		1
6	10	0		0
・				
・	3	3		2

（例）

サッカー選手には
大別して4種類あり
FW, MF, DF, GK
がある.

当たり前（？）

教師なし学習の例

消費者	ジュース	ビスケット	・・・	ケーキ
1	3	2		6
2	1	5		0
3	5	1		3
4	2	2		2
5	0	0		1
6	10	0		0
・				
・	3	3		2

（例）
顧客には大別して
7タイプがあり
セールを気にする
品質を気にする
などがある.

商品には大別して
20種類があり
固定客のあるもの
季節によるもの
などがある.

教師なし学習の例

歌	1	2	・・・	n
1	C	F		G
2	C	G		Am
3	Am	Em		Dm
4	C	Am		Em
5	C	Fm		G
6	C	G		C
・				
・				

（例）
音楽のコード進行には
大別して10種類あり
カノン、王道、
ツーファイブ、・・・
などがある.

音楽のジャンルには
大別して5種類あり、
頻繁に用いられる
コード進行は・・・である

当たり前（？）

教師なし学習の応用

仕事名	技術割合	折衝割合	・・・	管理割合
研究	80	10		10
生産	60	30		10
コンサル	20	70		10
社長	10	30		60
販売	30	60		10
サービス	50	40		10
・				
・				

（例）
仕事には大別して
30種類あり、
研究開発、
サービス、
販売、・・・
などがある.

当たり前のことですが
名称と実務が
対応していないことも
あるので注意必要.

教師なし学習の例

俳優名	主人公	ライバル	・・・	仲間
1	3	2		6
2	1	5		0
3	5	1		3
4	2	2		2
5	0	0		1
6	10	0		0
・				
・	3	3		2

（例）
俳優女優には大別して
１０種類あり、演じる
役柄の傾向は
勇ましい主人公、
知的な主人公、
きざな敵、
悪意に満ちた敵、
特技で助けてくれる人、
主人公を好きになる人、
・・・
などがある.

「気づく」ためにはどうしたら良いですか

多くのデータから背後にある構造や法則を見つけるには

(0) 大発見に王道なし。
(1) 対象の分野を心から好きになる。
(2) いろいろなアルゴリズムを適用して違いを見る。
(3) データをスパース化してみる。
(4) 近づいたり遠ざかったりしてみる。
(5) WWW との関わりを断つ。
(6) とことん考えて煮詰まってから散歩する。

問2

あなたはあるゲームの新しいキャラクターを作る仕事を
することになりました。そのための準備として小説・映画・ゲームに
現れる登場人物について典型的なタイプを抽出するための
データベースを作ることになりました。

どのようなデータベースを作成したらよいと思いますか。あなたの
好きなジャンルをひとつ決めて、その評価ベクトルの例を示し
登場キャラクター５名の例を示してください。

（例） ジャンル：三国志

登場人物	決断	戦略	戦闘	調整	温かさ
曹操	10	9	7	7	1
劉備	7	4	3	7	10
孫権	6	6	5	10	4
呂布	8	2	10	1	2
諸葛亮	4	10	1	5	6