Data Mining: Data

Lecture Notes for Chapter 2

Introduction to Data Mining

by

Tan, Steinbach, Kumar

Orijilal slaytların Türkçe çevirisidir.

What is Data?

- ***Veri nesneleri*** ve onların ***özniteliklerinin*** koleksiyonu

- Öznitelik (***attribute***), bir nesnenin karakteristiği veya özelliğidir. Örnek: kişinin göz rengi, sıcaklık, vb.
  - Attribute is also known as variable, field, characteristic, or feature
- Bir öznitelik koleksiyonu bir nesneyi (***object***) tanımlar
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values (Öznitelik değerleri**)**

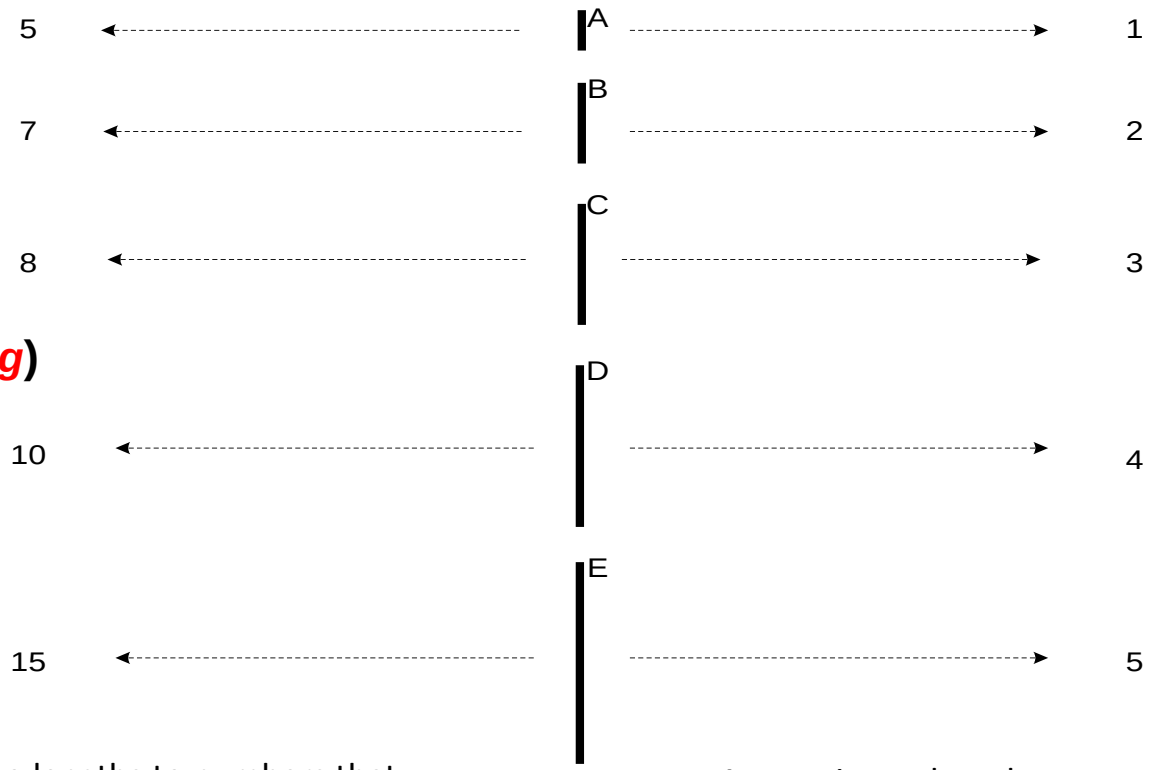
- Öznitelik değerleri, bir özniteliğe atanan sayılar veya sembollerdir

- Öznitelikler ve öznitelik değerleri arasındaki ayrım
  - Aynı öznitelik farklı öznitelik değerlerine izdüşürülebilir
    - Örnek: yükseklik metre veya feet olarak ölçülebilir
  - Farklı öznitelikler aynı değer kümesine eşlenebilir
    - Örnek: Kimlik numarası (*ID*) ve yaş (*age*) için öznitelik değerleri tamsayıdır
    - Fakat öznitelik değerlerinin özellikleri farklı olabilir
      - Kimlik numarasında sınırlama yoktur ancak yaşın maksimum ve minimum değeri vardır

Measurement of Length

- Bir özniteliği ölçme şekliniz, öznitelik özellikleriyle eşleşmeyebilir.

Bu ölçek yalnızca uzunluğun sıra (*ordering*) özelliğini korur.



Bu ölçek uzunluğun sıra (*ordering*) ve toplanırlık (*additvity*) özelliğini korur.

A mapping to lengths to numbers that captures only the order properties of length

A mapping to lengths to numbers that captures both order and additivity properties of length

Thus, an attribute can be measured in a way that does not capture all the properties of the attribute.

Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

Properties of Attribute Values
(Öznitelik değerlerinin özellikleri)

- Bir (öz)niteliğin türü, aşağıdaki özelliklerden hangisine sahip olduğuna bağlıdır :
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

Attribute Type	Description	Examples	Operations
Nominal	Nominal bir niteliğin değerleri sadece farklı isimlerdir, yani nominal nitelikler sadece bir nesneyi diğerinden ayırt etmek için yeterli bilgi sağlar. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male*, *female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	Bir ordinal niteliğin değerleri, nesneleri sıralamak için yeterli bilgi sağlar. (<, >)	hardness of minerals, {*good*, *better*, *best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	Aralık (Interval) nitelikleri için, değerler arasındaki farklar anlamlıdır, yani bir ölçü birimi mevcuttur. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	Oran (Ratio) değişkenleri için, hem farklar hem de oranlar anlamlıdır. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

This categorization of attributes is due to S. S. Stevens

Categorical (or qualitative) ***attribute***

Numeric (Quantitative) ***attributes***

Attribute Level	Transformation	Comments
Nominal	Her türlü permütasyon (Any permutation of values)	Tüm çalışan kimlik numaraları (ID) yeniden atansa, herhangi bir fark yaratır mı?
Ordinal	Değerlerin sırasını muhafaza eden bir değişiklik, yani *new_value = f(old_value)* burada f monotonik bir fonksiyondur.	İyi, daha iyi en iyi kavramını kapsayan bir öznitelik, başka değerlerle de aynı şekilde temsil edilebilir{1, 2, 3} veya { 0.5, 1, 10} ile
Interval	*new_value =a * old_value + b* burada *a* ve *b* sabitdir	Buradan hareketle, Fahrenheit ve Santigrat sıcaklık ölçekleri sıfır değerlerinin nerede olduğu ve bir birimin (derece) büyüklüğü açısından farklılık gösterir.
Ratio	*new_value = a * old_value*	Uzunluk metre veya feet olarak ölçülebilir.

Nitelik türleri, bir niteliğin anlamını değiştirmeyen dönüşümler (transformations) olarak da tanımlanabilir.

Discrete and Continuous Attributes

- Ayrık Nitelik (Discrete Attribute)
  - Sonlu bir değer kümesine sahiptir
  - Örnekler: posta kodu, sayılar veya bir belge koleksiyonundaki kelime kümesi
  - Genellikle tamsayı değişkenleri olarak gösterilir.
  - Not: ikili öznitelikler (***binary attributes***), ayrık özniteliklerin özel bir durumudur
    - **Sadece iki değer alır, e.g., true/false, yes/no, male/female, or 0/1.**
- Sürekli Nitelik (Continuous Attribute)
  - Öznitelik değerleri olarak gerçek sayılar vardır
  - Örnek: sıcaklık, yükseklik veya ağrılık.
  - Pratikte, gerçek değerler sadece sınırlı sayıda basamak kullanılarak ölçülebilir ve temsil edilebilir.
  - Sürekli öznitelikler genellikle kayan nokta değişkenleri olarak temsil edilir.

Asymmetric Attributes

- Yalnızca varoluş/mevcudiyet (sıfır olmayan bir öznitelik değeri) önemli olarak kabul edilir
  - Dokumanlarda geçen kelimeler
  - Müşteri işlemlerinde mevcut olan kalemler

- Markette bir arkadaşla karşılaşsak şunu söyler miydik?

  "Aynı şeylerin çoğunu almadığımız için alımlarımızın çok benzer olduğunu görüyorum."

- Sıfır olmayan değerlere odaklanmak daha anlamlı ve daha verimlidir.
- Sadece sıfır olmayan değerlerin önemli olduğu ikili niteliklere asimetrik ikili öznitelikler (**asymmetric binary attributes**) denir.
  - Birliktelik analizinde asimetrik öznitelikler kullanılır.

Types of data sets

- **Record**
  - **Data Matrix**
  - **Document Data**
  - **Transaction Data**
- **Graph-based**
  - **World Wide Web**
  - **Molecular Structures**
- **Ordered**
  - **Spatial Data**
  - **Temporal Data**
  - **Sequential Data**
  - **Genetic Sequence Data**

Important Characteristics of Data

- **Dimensionality** (number of attributes)
  - Curse of Dimensionality
- **Sparsity**
  - Only presence counts
- **Resolution**
  - Patterns depend on the scale
- **Size**
  - Type of analysis may depend on size of data

Record Data

- Her biri sabit bir öznitelik kümesinden (fixed set of attributes) oluşan kayıt koleksiyonu

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

- Veri nesneleri aynı sabit sayısal öznitelik kümesine sahipse, veri nesneleri (data objects) çok boyutlu bir uzayda noktalar (points in a multi-dimensional space) olarak düşünülebilir; burada her boyut farklı bir özniteliği temsil eder
- Bu veri seti, her nesne için bir tane olmak üzere m satır ve her bir öznitelik için bir tane olmak üzere n sütun ile, yani bir mxn matrisi ile temsil edilebilir.

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

- Her belge bir 'terim' vektörü olur
  - her terim, vektörün bir bileşenidir (özniteliğidir)
  - her bileşenin değeri, karşılık gelen terimin belgede kaç kez geçtiğini gösterir.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

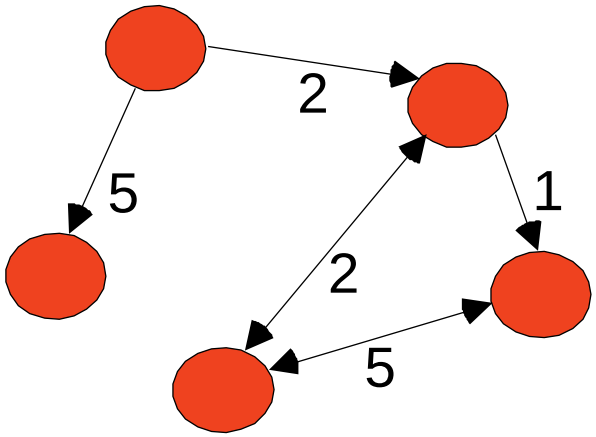
Transaction Data

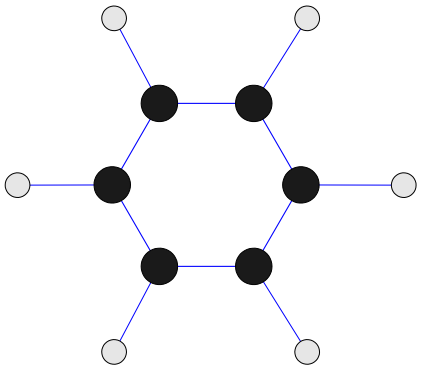
- Özel bir kayıt verisi türü, burada
  - Her kayıt (işlem/ transaction) bir dizi maddeyi içerir.
  - Örneğin, bir marketi düşünün. Bir müşterinin bir alışveriş gezisi sırasında satın aldığı ürün grubu bir işlem (***transaction***) oluştururken, satın alınan tekil ürünler öğelerdir (***items***).

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data

- Examples: Generic graph, a molecule, and webpages





Benzene Molecule: C6H6

Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

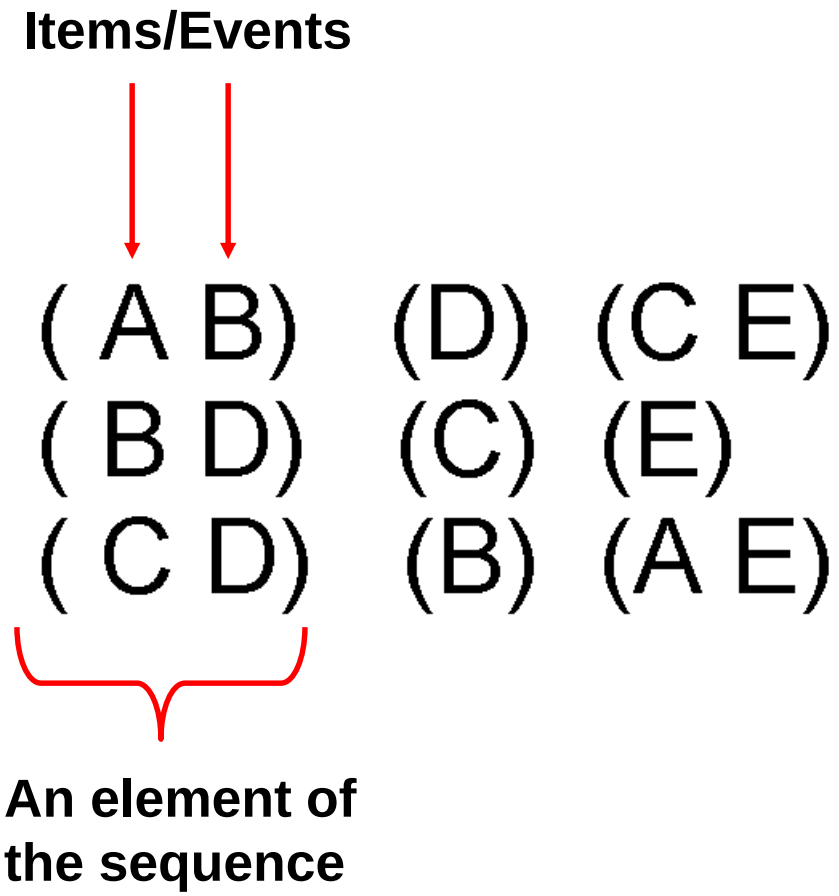
General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Ordered Data

- Sequences of transactions



Ordered Data

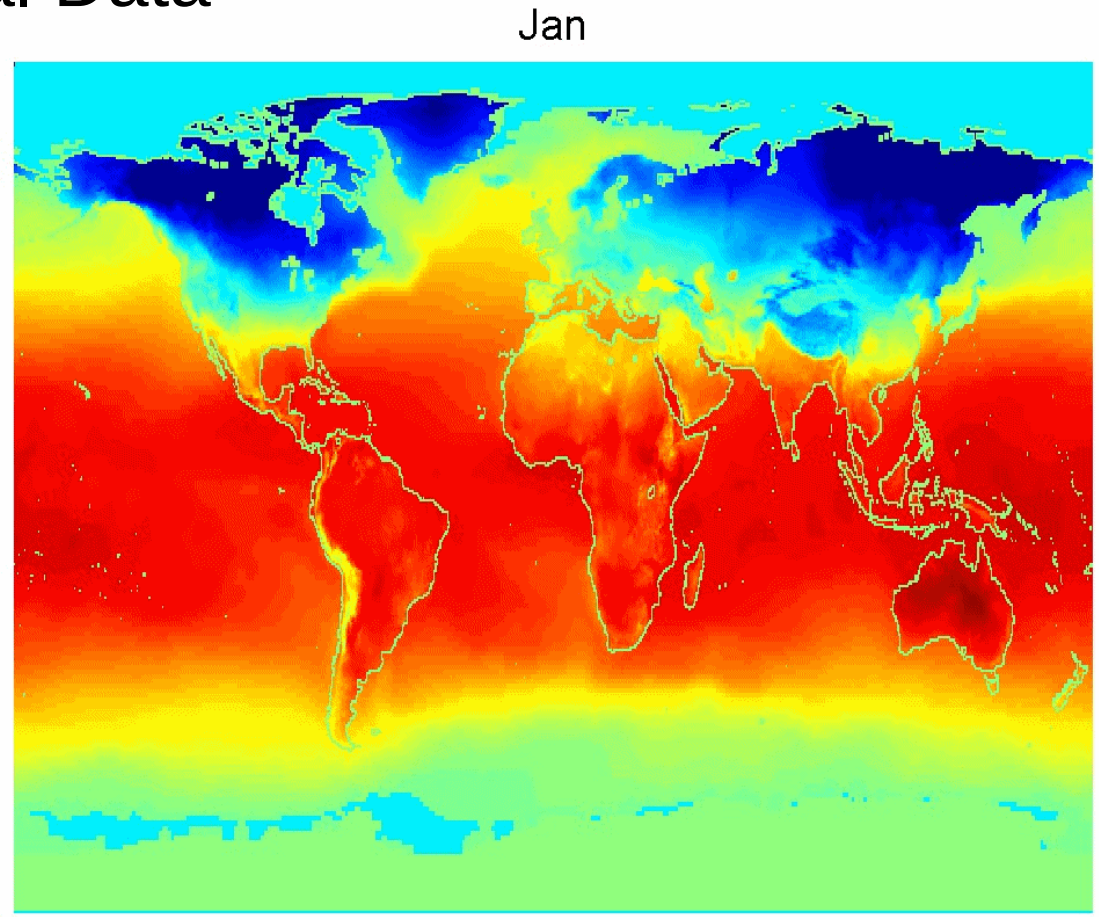
- Genomic sequence data (Gen dizilim verisi)

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Ordered Data

- Spatio-Temporal Data

Kara ve okyanusların Aylık Ortalama Sıcaklık verisi



Data Quality

- Yetersiz veri kalitesi, birçok veri işleme çabasını olumsuz etkiler

"En önemli nokta, düşük veri kalitesinin gelişen bir felaket olmasıdır.

Düşük veri kalitesi, tipik bir şirketin gelirinin en az yüzde onuna (%10) mal olur; Yüzde yirmi (%20) muhtemelen daha iyi bir tahmin."

Thomas C. Redman, DM Review, August 2004

- Veri madenciliği örneği: kredi riski olan kişileri tespit etmek için bir sınıflandırma modeli yetersiz/eksik veriler kullanılarak oluşturulmuştur
  - Bazı krediye değer adayların kredileri reddedildi
  - Temerrüde düşen kişilere daha fazla kredi verildi

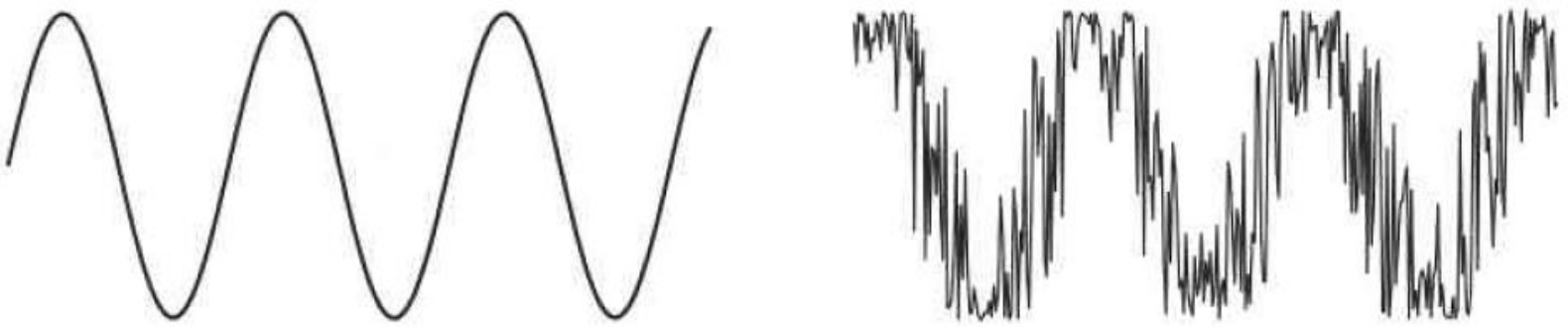
Data Quality

- Ne tür veri kalitesi sorunları?
- Verilerle ilgili sorunları nasıl tespit edebiliriz?
- Bu sorunlar hakkında neler yapabiliriz?

- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data

Noise

- Gürültü (***noise***), orijinal değerlerin değiştirilmesi anlamına gelir
  - Örnekler: kaklitesiz bir telefonda konuşurken kişinin sesinde bozulma ve televizyon ekranında "karlanma"



(a) Time series.

(b) Time series with noise.

Figure 2.5. Noise in a time series context.

Outliers

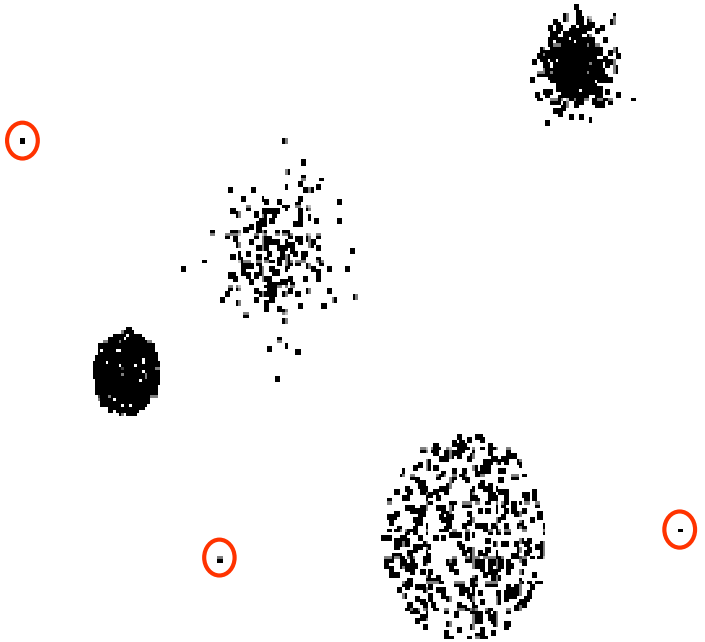
- ***Outliers*** (uç/aykırı değerler) veri kümesindeki diğer veri nesnelerinin çoğundan önemli ölçüde farklı özelliklere sahip veri nesneleridir
  - **Case 1:** *Outliers*,

    veri analizine müdahale

    eden gürültüdür

  - **Case 2:** *Outliers* analizimizin

    hedefidir
    - Credit card fraud
    - Intrusion detection



Missing Values

- Eksik değerlerin nedenleri
  - Bilginin toplanamadığı durumlar
    (ör. insanlar yaşlarını ve kilolarını vermeyi reddederler)
  - Nitelikler tüm durumlar için geçerli olmayabilir
    (ör. yıllık gelir çocuklar için geçerli değildir)

Age	Income	Team	Gender
23	24,200	Red Sox	M
39	?	Yankees	F
45	45,390	?	F

?: missing value

- Eksik verilerle başa çıkma
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

Duplicate Data

- Veri kümesi, yinelenen (*duplicate*) veya neredeyse birbirinin kopyası olan veri nesnelerini içerebilir
  - Heterojen kaynaklardan gelen verileri birleştirirken önemli sorun
- Örnekler :
  - Birden çok e-posta adresine sahip aynı kişi:
- Data cleaning (Veri temizleme)
  - Tekrarlı veri sorunlarıyla ilgilenme süreci

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature Subset Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformation

Aggregation

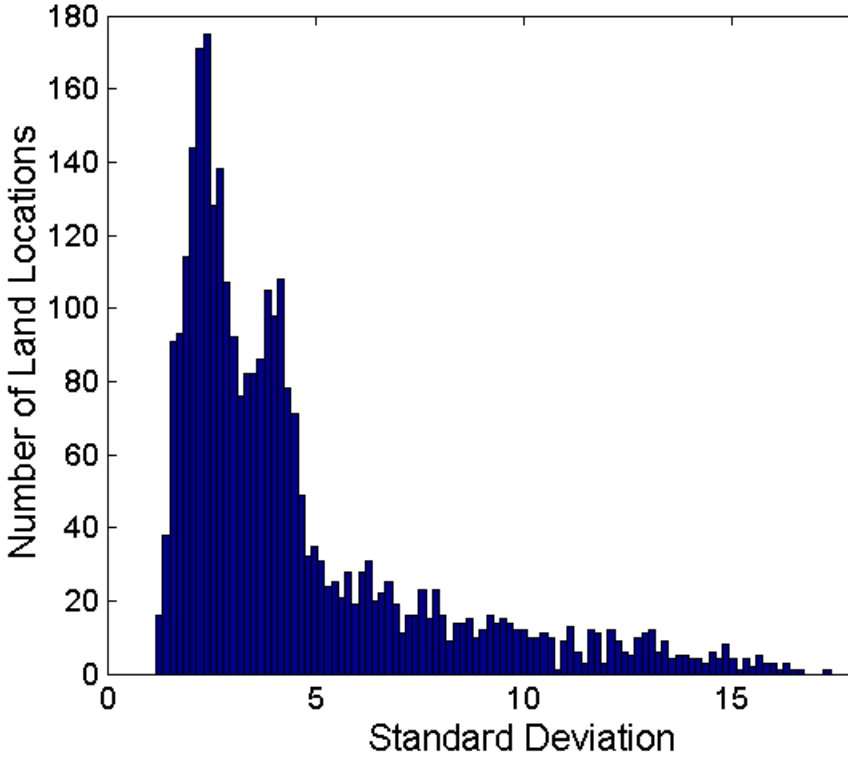
- İki veya daha fazla özniteliği (veya nesneyi) tek bir öznitelikte (veya nesnede) birleştirmek

- Amaç
  - Veri azaltma (*Data reduction*)
    - Özniteliklerin (attributes) veya nesnelerin (objects) sayısını azaltma
  - Ölçek değişikliği (*Change of scale*)
    - Bölgeler, eyaletler, ülkeler vb. şeklinde birleştirilmiş şehirler
  - Daha "kararlı" (*stable*) veriler
    - Birleştirilmiş veriler daha az değişkenliğe/oynaklığa sahip olma eğilimindedir

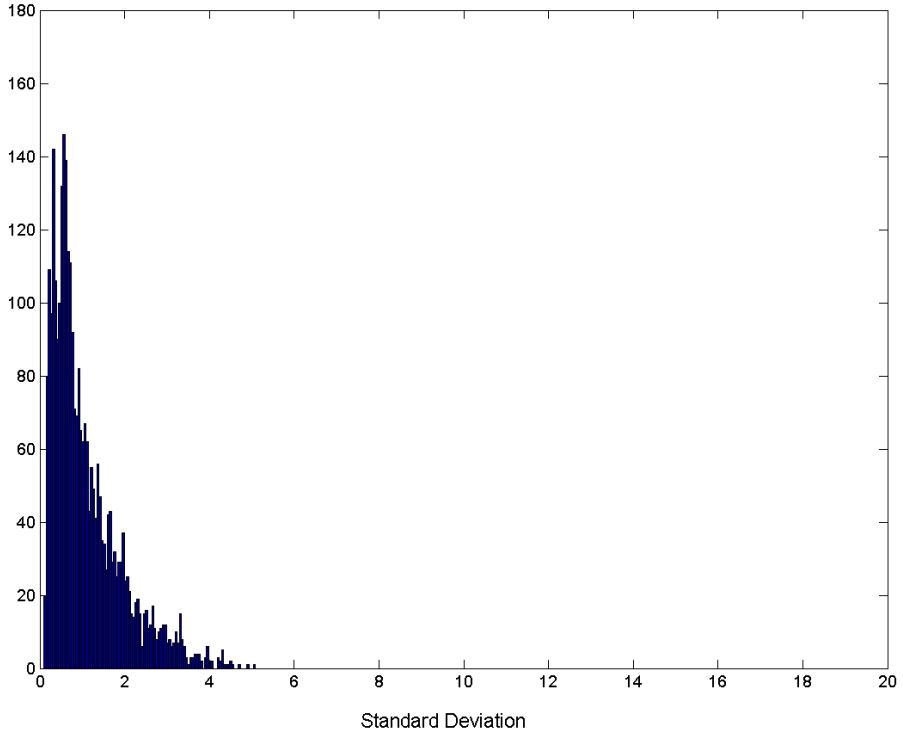
Aggregation

Avustralya'daki Yağış (Precipitation) Değişimi

Aggregation sayesinde std. dev. miktarında belirgin azalma



Ortalama Aylık Yağışların Standart Sapması



Ortalama Yıllık Yağışların Standart Sapması

Sampling

- **Veri seçimi (data selection) için kullanılan ana teknik örneklemedir.**
  - **Genellikle hem verilerin ön araştırması, hem de nihai veri analizi için kullanılır.**

- **İstatistikçiler örnekleme yapar çünkü ilgilenilen tüm veri setini elde etmek çok pahalı veya zaman alıcıdır.**

- **Örnekleme, veri madenciliğinde kullanılır çünkü ilgilenilen tüm veri kümesinin işlenmesi çok pahalı (*expensive*) veya zaman alıcıdır (*time consuming*).**

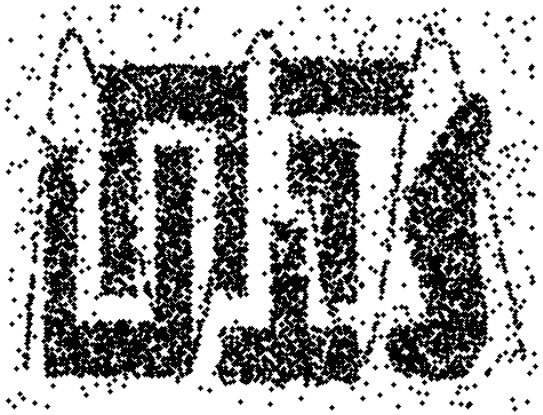
Sampling …

- Etkili örnekleme için temel ilke şudur:
  - Eğer seçilen örneklemin temsil gücü yüksek ise, bir örneklem kullanmak **neredeyse** tüm veri setini **kullanmak kadar** işe yarayacaktır.
  - Bir örneklem, orijinal veri kümesiyle yaklaşık olarak (ilgili) aynı özelliğe sahipse temsilcidir (representative).

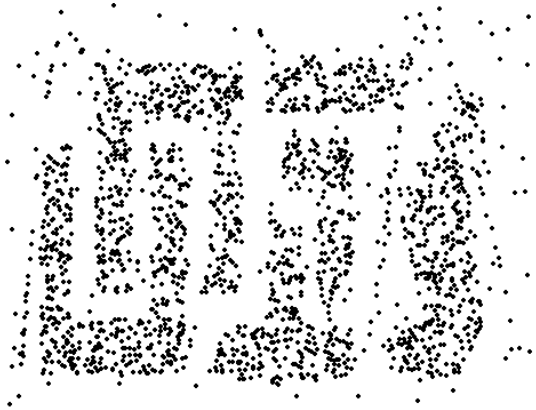
Types of Sampling

- Simple Random Sampling
  - Herhangi bir belirli öğeyi seçme konusunda eşit bir olasılık vardır
- Sampling without replacement
  - Her öğe seçildikçe popülasyondan çıkarılır.
- Sampling with replacement
  - Nesneler, örneklem için seçildikçe popülasyondan çıkarılmaz.
    - Aynı nesne birden fazla kez alınabilir.
- Stratified sampling
  - Verileri birkaç bölüme (*partition*) ayırın; sonra her bölümden rastgele örnekler alın

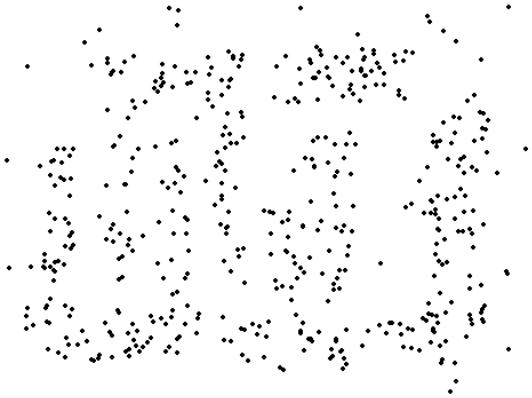
Sample Size



8000 points



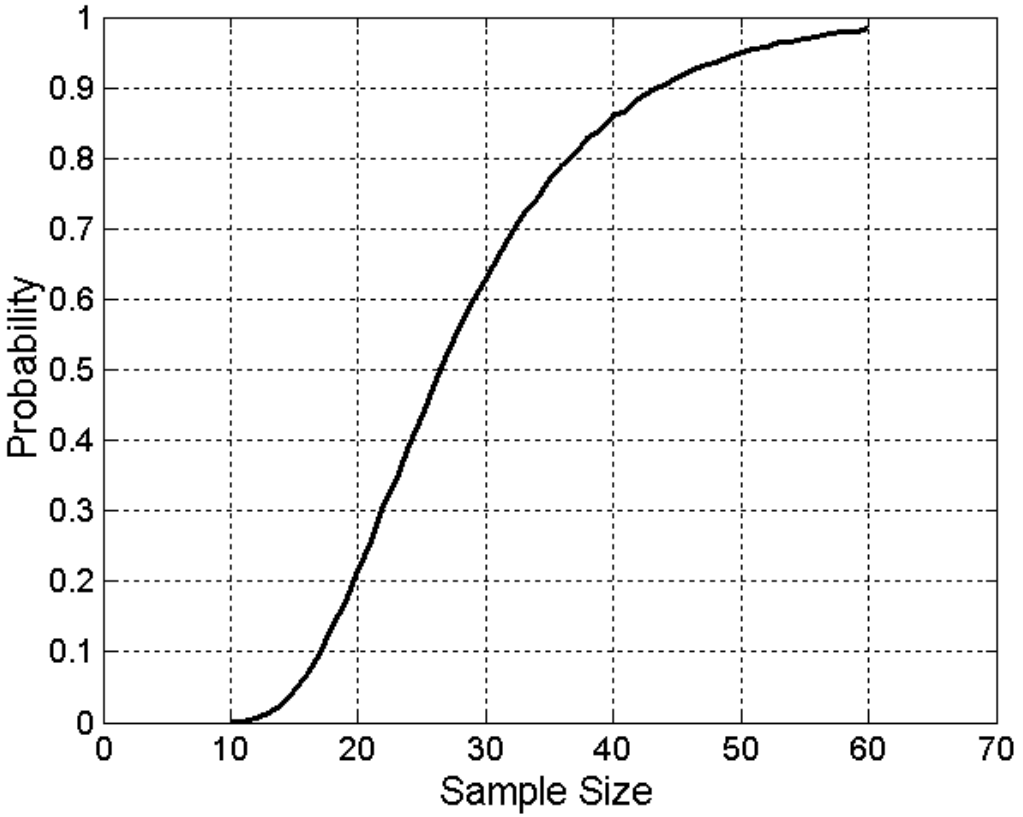
2000 Points



500 Points

Sample Size

- 10 eşit büyüklükteki grubun her birinden en az bir nesne elde etmek için hangi örneklem boyutu gereklidir?

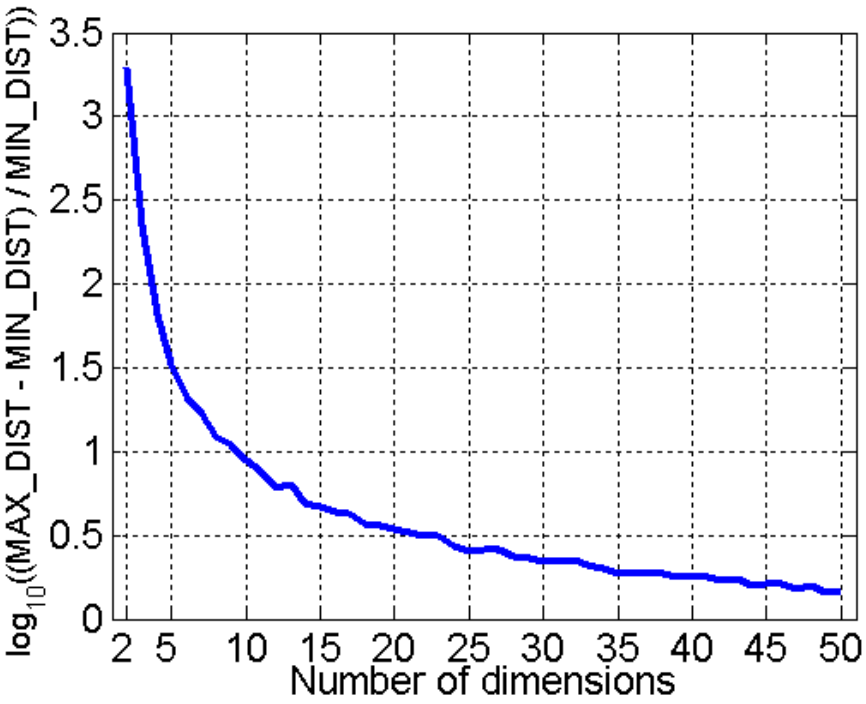


The figure showing an idealized set of clusters (groups) from which these points might be drawn

Curse of Dimensionality

- Boyut arttığında (dimensionality increases), veri kapladığı alanda giderek daha seyrek (sparse) hale gelir

- **Kümeleme (*clustering*) ve aykırı değer tespiti (*outlier detection*) için kritik olan** yoğunluk (*density*) ve noktalar arasındaki mesafe tanımları daha az anlamlı hale gelir



- **Rastgele 500 nokta oluşturun**
- **Herhangi bir nokta çifti arasındaki maksimum ve minimum mesafe arasındaki farkı hesaplayın**

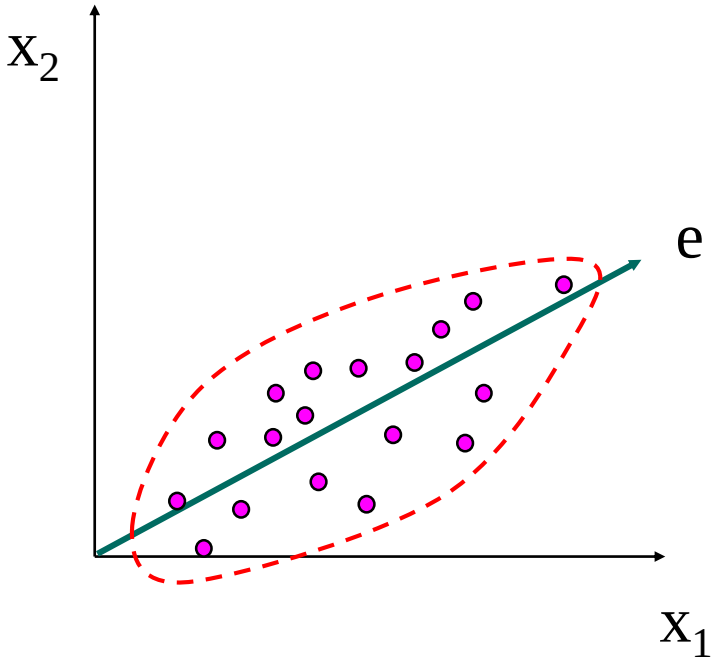
Dimensionality Reduction

- Amaç:
  - Çok boyutlululuğun getirdiği sıkıntıdan kurtulmak
  - Veri madenciliği algoritmalarının gerektirdiği süre (time) ve bellek (memory) miktarını azaltmak
  - Verilerin daha kolay görselleştirilmesine olanak tanır
  - Alakasız özellikleri (irrelevant features) ortadan kaldırmaya veya gürültüyü (noise) azaltmaya yardımcı olabilir

- Teknikler
  - Principle Component Analysis (PCA)
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

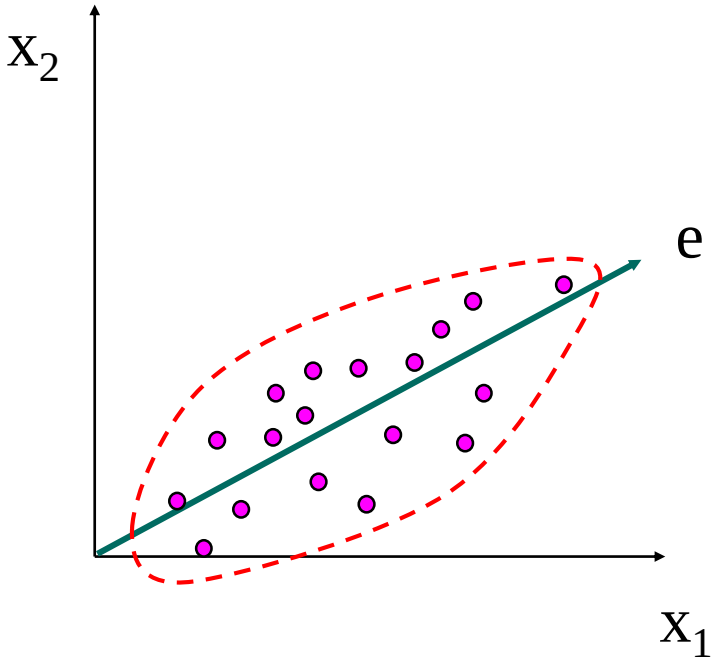
Dimensionality Reduction: PCA

- Amaç, verilerdeki en büyük miktarda varyasyonu yakalayan bir projeksiyon bulmaktır.



Dimensionality Reduction: PCA

- Kovaryans matrisinin özvektörlerini bulunur
- Özvektörler yeni uzayı tanımlar



Dimensionality Reduction: PCA

- Temel Bileşenler Analizi (PCA) sürekli öznitelikler için yeni öznitelikler (**temel bileşenleri**) bulan bir lineer cebir tekniğidir ve bu bileşenler
  - (1) orijinal özelliklerin lineer kombinasyonlarıdır,
  - (2) birbirlerine diktir (orthogonal)
  - (3) verilerdeki maksimum varyasyon miktarını yakalar

Örneğin, ilk iki temel bileşen, orijinal niteliklerin doğrusal kombinasyonları olan iki ortogonal nitelik ile mümkün olduğu kadar verideki varyasyonun çoğunu yakalar.

Dimensionality Reduction: PCA

Feature Subset Selection

- Verilerin boyutunu azaltmanın başka bir yolu

- **Redundant features** (yedekli özellikler)
  - bir veya daha fazla başka öznitelikte bulunan bilgilerin çoğunu veya tamamını tekrarlama (***duplicate***)
  - Örnek: bir ürünün satın alma fiyatı ve ödenen satış vergisi tutarı

- **Irrelevant features** (alakasız özellikler)
  - eldeki veri madenciliği görevi için yararlı hiçbir bilgi içermez
  - Örnek: öğrencilerin kimliği (ID) genellikle öğrencilerin not ortalamasını (GPA) tahmin etme görevi ile ilgisizdir

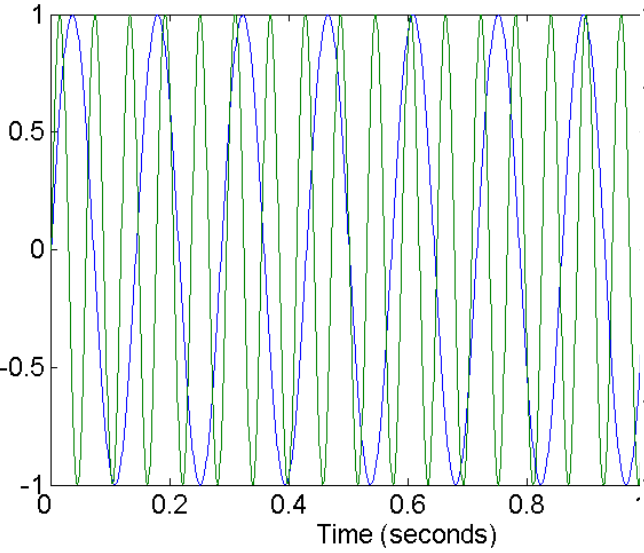
Feature Creation

- Bir veri kümesindeki önemli bilgileri orijinal özniteliklerden çok daha verimli bir şekilde yakalayabilen yeni öznitelikler oluşturma

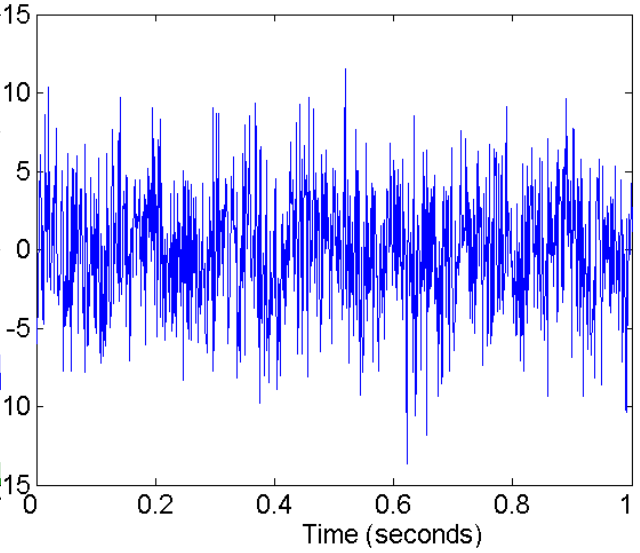
- Üç genel metodoloji :
  - Feature extraction (*öznitelik çıkarımı*)
    - Örnek: görüntülerden kenarları çıkarma
  - Feature construction (*öznitelik oluşturma*)
    - Örnek: yoğunluğu elde etmek için kütleyi hacme bölme
  - Mapping data to new space (*Verileri yeni uzaya izdüşürme*)
    - Örnek: Fourier and wavelet analizi

Mapping Data to a New Space

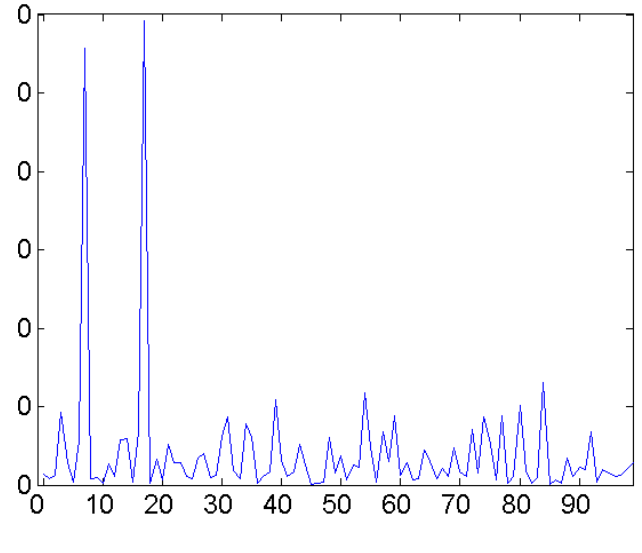
- Fourier transform
- Wavelet transform



Two Sine Waves



Two Sine Waves + Noise



Frequency

Discretization

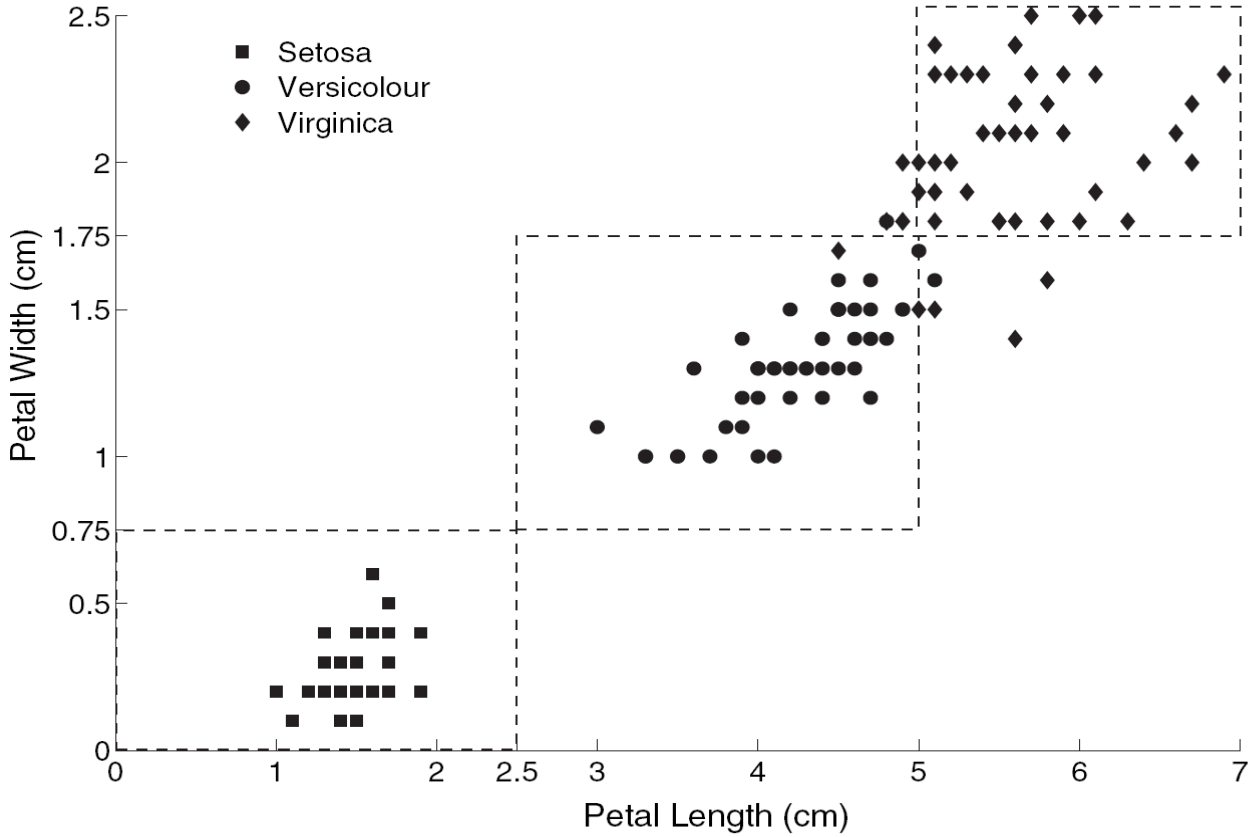
- Ayrıklaştırma (Discretization), sürekli (***continuous***) bir özniteliği sırasal (***ordinal***) özniteliğe dönüştürme sürecidir.
  - Potansiyel olarak sonsuz sayıda değer, az sayıda kategoriye eşlenir
  - Ayrıklaştırma genellikle sınıflandırmada kullanılır
  - Birçok sınıflandırma algoritması, hem **bağımsız** hem de **bağımlı değişkenleri** yalnızca birkaç değere sahipse en iyi şekilde çalışır
  - Iris veri setini kullanarak ayrıklaştırmanın yararlılığına dair bir örnek…

Iris Sample Data Set

- Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Versicolour
    - Virginica
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

Discretization: Iris Example



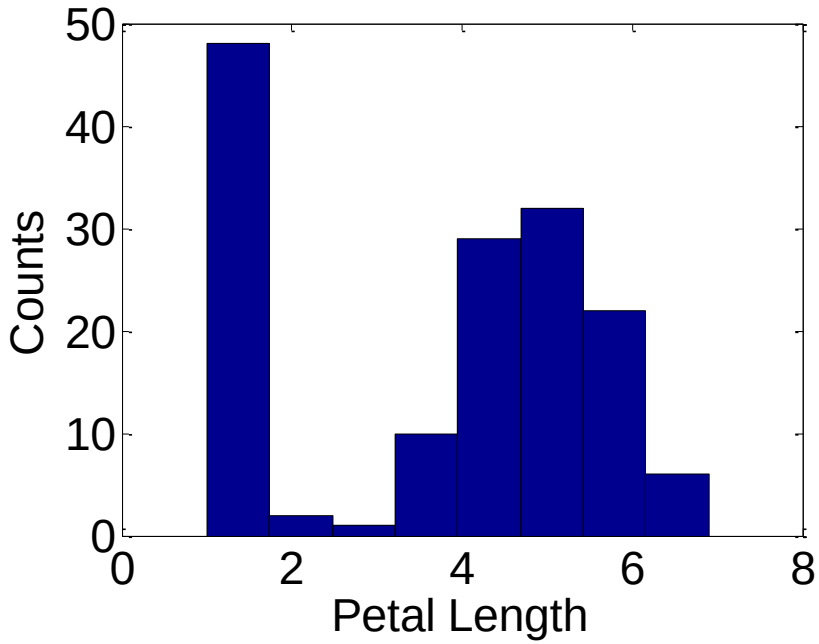
Petal ->taç yaprak

Petal genişliği düşük veya petal uzunluğu düşük, **Setosa** anlamına gelir.
Petal genişliği orta veya petal uzunluğu orta, **Versicolour** anlamına gelir.
Petal genişliği yüksek veya petal uzunluğu yüksek, **Virginica** anlamına gelir.

Discretization: Iris Example ...

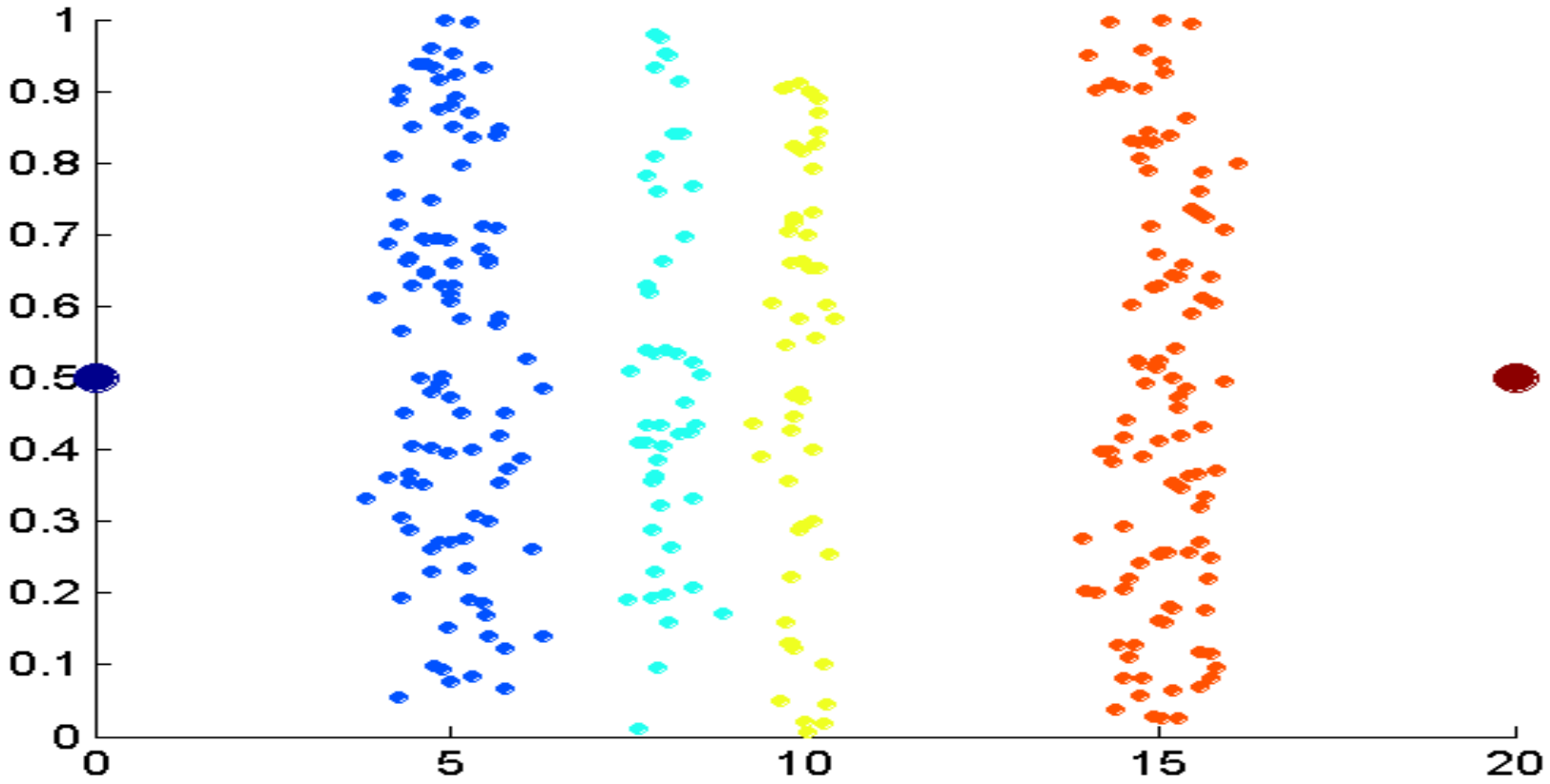
En iyi ayrıklaştırmanın (best discretization) ne olduğunu nasıl anlayabiliriz?

- Unsupervised discretization: veri değerindeki kırılmaları (breaks) bulmak
  - Example: Petal Length



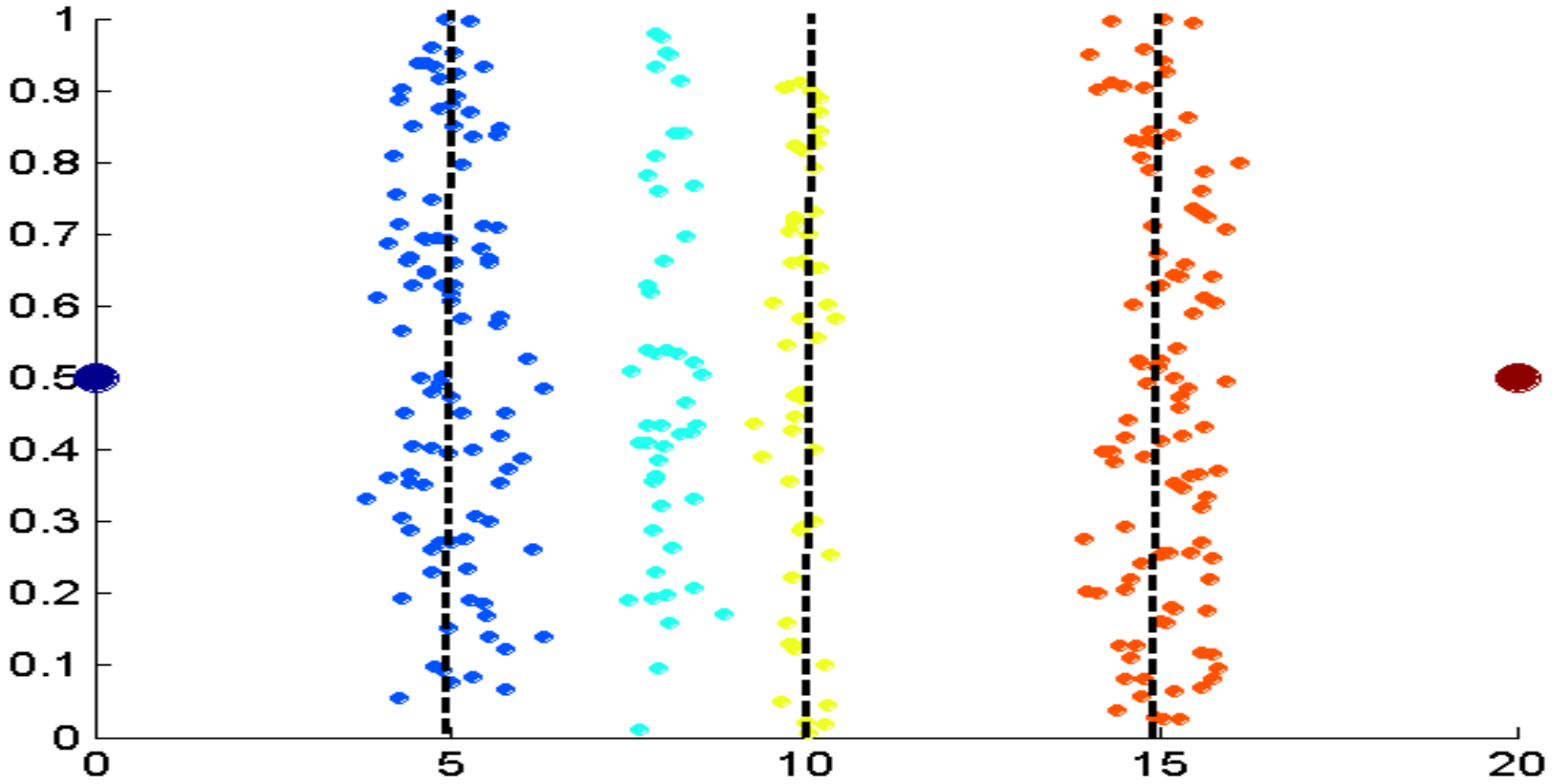
- Supervised discretization: Kırılmaları bulmak için sınıf etiketleri kullanmak

Discretization Without Using Class Labels



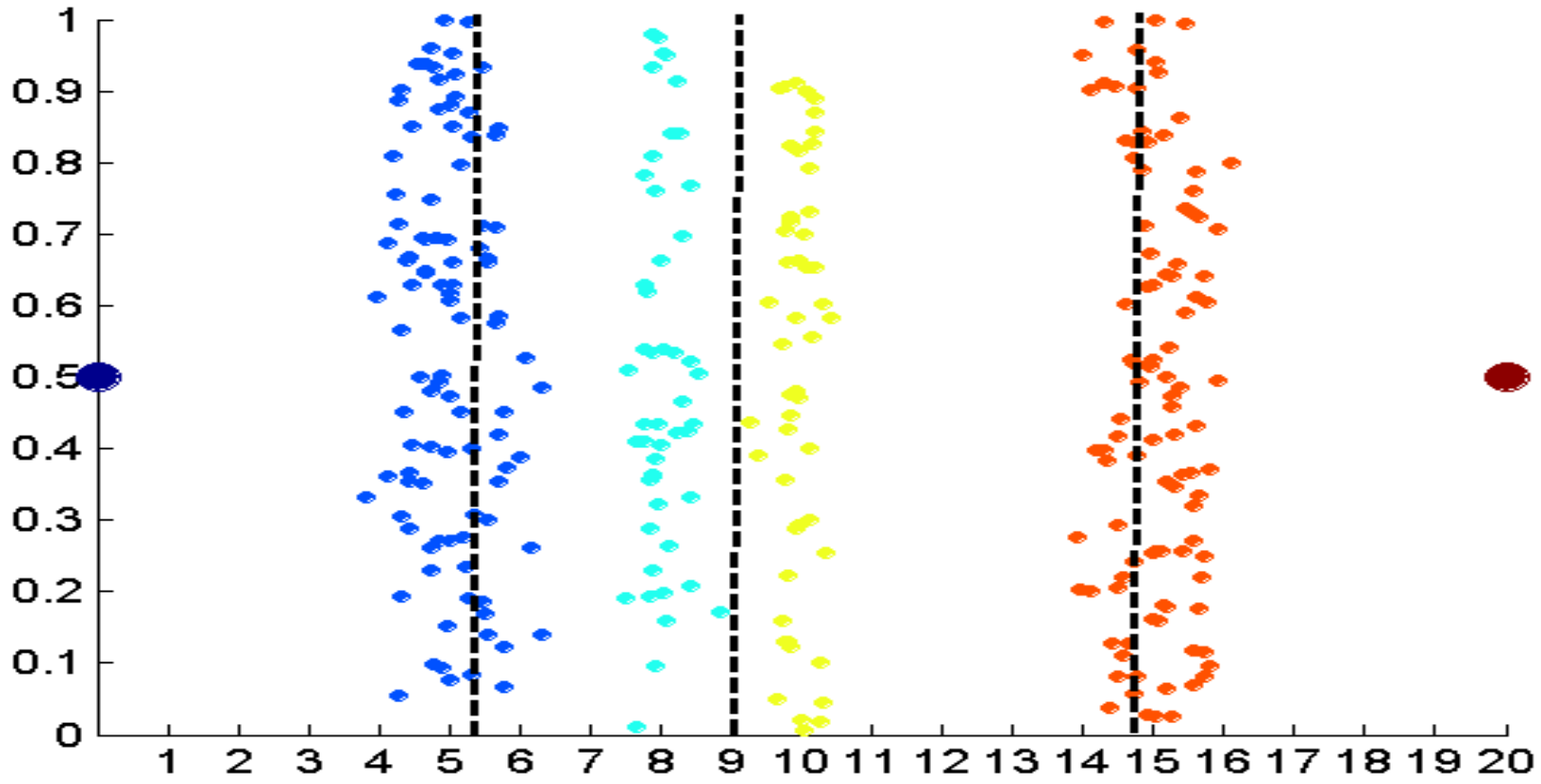
Veriler dört grup noktadan ve iki uç değerden oluşur. Veriler tek boyutludur, ancak çakışmayı azaltmak için rastgele bir y bileşeni eklenir

Discretization Without Using Class Labels



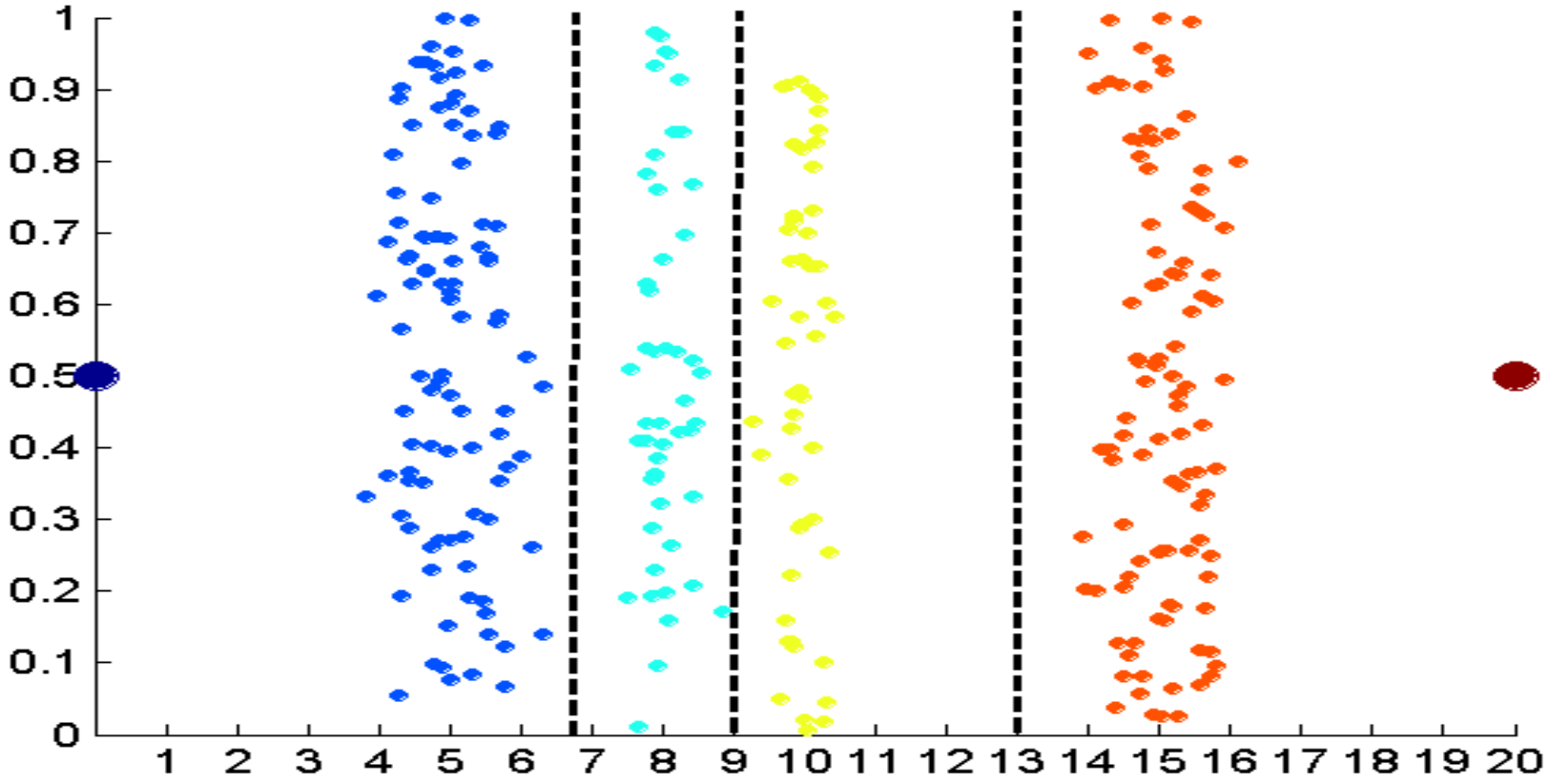
4 değer elde etmek için kullanılan eşit aralık genişliği (Equal interval width) yaklaşımı.

Discretization Without Using Class Labels



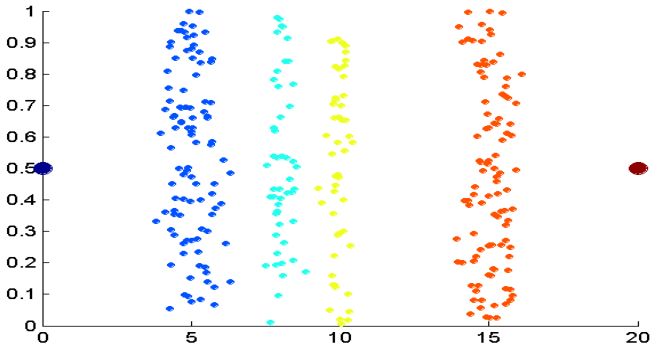
4 değer elde etmek için kullanılan eşit frekans (Equal frequency) yaklaşımı

Discretization Without Using Class Labels

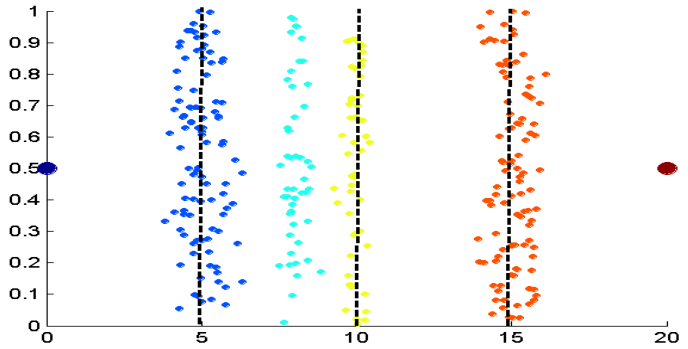


4 değer elde etmek için K-means yaklaşımı

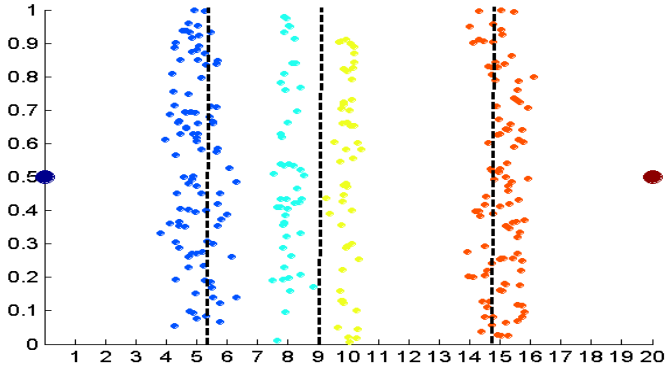
Discretization Without Using Class Labels



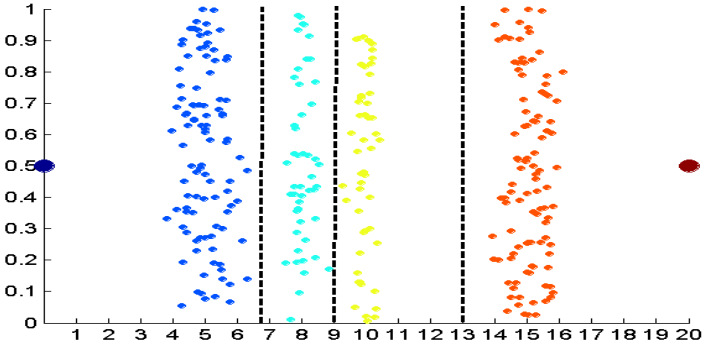
Data



Equal interval width



Equal frequency



K-means

Binarization

- İkilileştirme, sürekli veya kategorik bir özniteliği bir veya daha fazla ikili değişkenle eşler.
- Tipik olarak birliktelik (association) analizi için kullanılır.
- Genellikle sürekli bir öznitelik önce kategorik özniteliğe dönüştürülür ve ardından kategorik öznitelik bir dizi ikili özniteliğe dönüştürülür
  - Birliktelik analizi asimetrik ikili özniteliklere (***asymmetric binary attribute***) ihtiyaç duyar
  - Örnekler: göz rengi ve {low, medium, high} şeklinde ölçülen boy özniteliği

Binarization

Table 2.5. Conversion of a categorical attribute to three binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

Kategorik özniteliğin ikili özniteliğe dönüştürülmesi

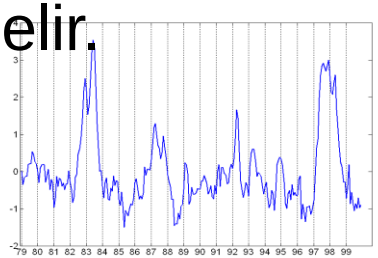
Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

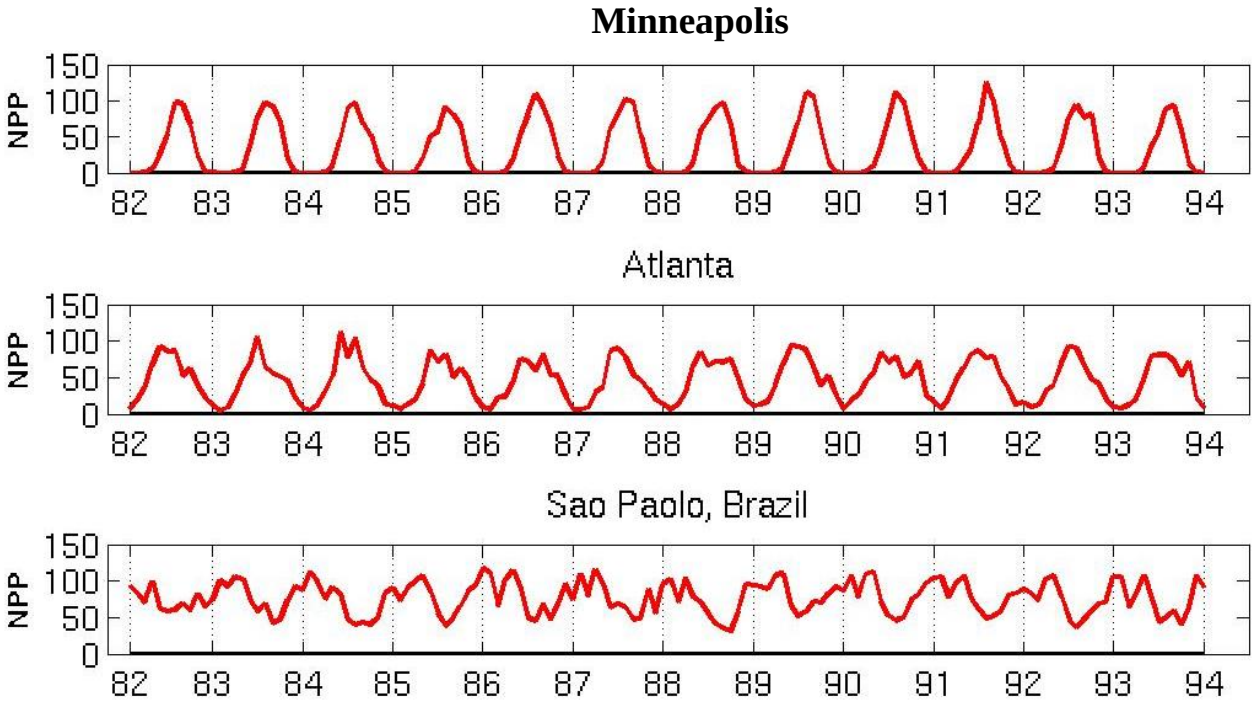
Kategorik özniteliğin 5 tane asimetrik ikili özniteliğe dönüştürülmesi

Attribute Transformation

- Attribute transform: Belirli bir özniteliğin tüm değer kümesini yeni bir ikame değerler kümesiyle eşleştiren bir fonksiyon, böylece her eski değer yeni değerlerden biriyle tanımlanabilir
  - Basit fonksiyonlar: x^k, log(x), e^x, |x|
  - Normalization
    - Ortalama (*mean*), varyans (*variance*), aralık (*range*) açısından özellikler arasındaki farklılıklara uyum sağlamak için çeşitli teknikleri ifade eder.
    - İstenmeyen, ortak sinyali çıkarın, örn. mevsimsellik
  - Standardization, istasistikte ortalamaların çıkarılması ve standart sapmaya bölünmesi anlamına gelir.



Example: Sample Time Series of Plant Growth

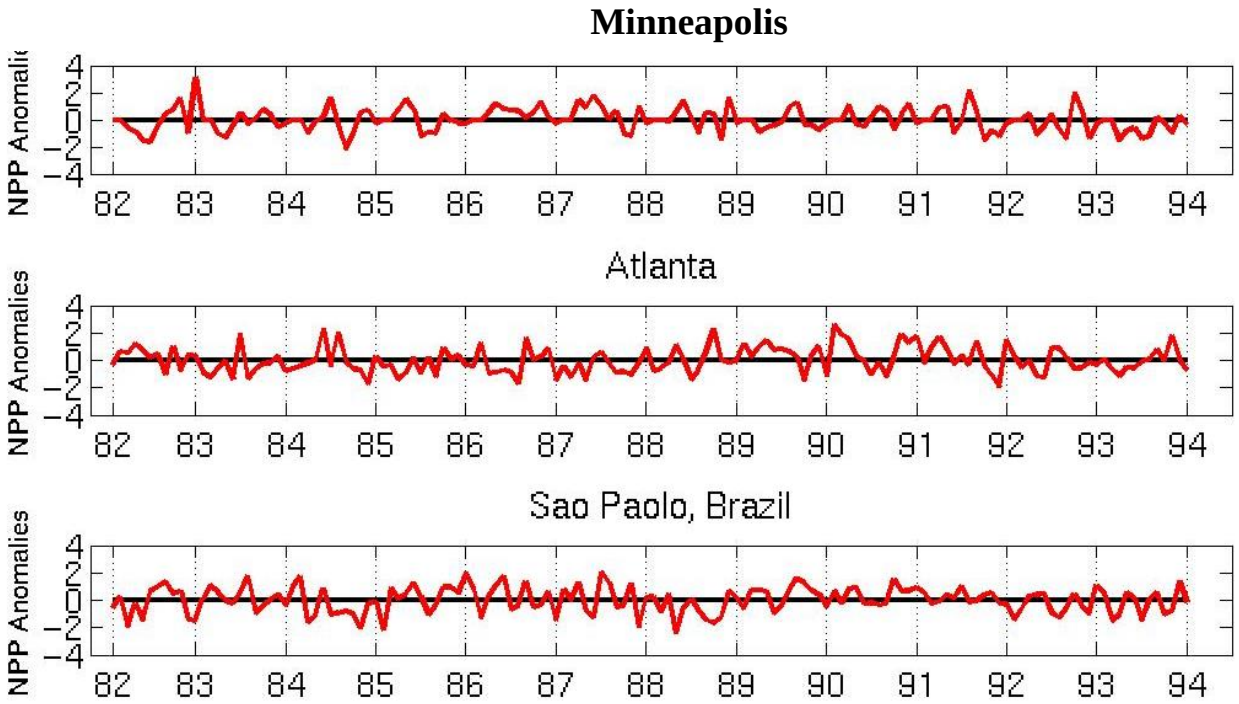


Net Birincil Üretim (Net Primary Production -NPP), ekosistem bilimcileri tarafından kullanılan bitki büyümesinin bir ölçüsüdür.

Zaman serileri arasındaki korelasyon

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paolo	-0.7581	-0.5739	1.0000

Seasonality Accounts for Much Correlation



Korelasyonun büyük bölümü mevsimsellik sebebiyledir

Aylık Z Score kullanılarak normalize edildi

Aylık ortalamayı çıkarın ve aylık standart sapmaya bölün

Correlations between time series

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.0492	0.0906
Atlanta	0.0492	1.0000	-0.0154
Sao Paolo	0.0906	-0.0154	1.0000

Similarity and Dissimilarity

- Similarity (*Benzerlik*)
  - İki veri nesnesinin ne kadar benzer olduğunun sayısal ölçüsü.
  - Nesneler birbirine daha çok benzediğinde daha yüksektir.
  - Genellikle [0,1] aralığına düşer
- Dissimilarity (*Farklılık*)
  - İki veri nesnesinin ne kadar farklı olduğunun sayısal ölçüsü
  - Nesneler birbirine daha çok benzediğinde daha düşük
  - Minimum farklılık genellikle 0'dır
  - Üst limit değişebilir
- Yakınlık (***Proximity***), benzerlik veya farklılığı ifade eder

Similarity/Dissimilarity for Simple Attributes

p and *q* are the attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

Similarity/Dissimilarity transformation examples

For the dissimilaritiy values of 0, 1, 10, 100;

$s = \frac{1}{1+d}$ transformation equation results in similaritiy values of 1, 0.5, 0.09, 0.01, respectively.

$s = 1 - \frac{d - min\_d}{max\_d - min\_d}$ transformation equation results in similaritiy values of 1.00, 0.99, 0.00, 0.00, respectively.

$s = e^{-d}$ transformation equation results in similaritiy values of 1.00, 0.37, 0.00, 0.00, respectively.

Euclidean Distance

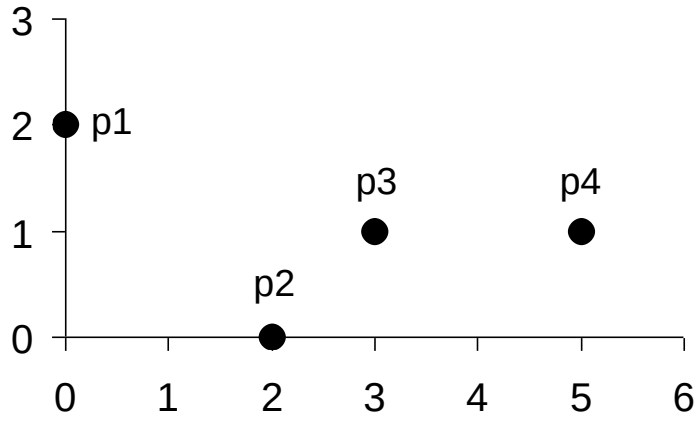
- Euclidean Distance (Öklit Mesfesi)

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Burada n boyut sayısı (öznitelikler) ve p_k ve q_k sırasıyla k'inci öznitelikler (bileşenler) veya p ve q veri nesneleridir.

- **Ölçekler farklıysa standardizasyon gereklidir.**

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
  - Bunun yaygın bir örneği, iki binary vektör arasında farklı olan bitlerin sayısı, Hamming mesafesidir (**Hamming distance**).
- $r = 2$. Euclidean distance (L_2 norm)
- $r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - Bu, vektörlerin herhangi bir bileşeni arasındaki maksimum farktır
- Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

Common Properties of a Similarity

- Similarities, also have some well known properties.

1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$.

2. $s(p, q) = s(q, p)$ for all p and q. (Symmetry)

where $s(p, q)$ is the similarity between points (data objects), p and q.

Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities using the following quantities

  M_{01} = the number of attributes where p was 0 and q was 1

  M_{10} = the number of attributes where p was 1 and q was 0

  M_{00} = the number of attributes where p was 0 and q was 0

  M_{11} = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes

  $= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = number of 11 matches / number of not-both-zero attributes values

  $= (M_{11}) / (M_{01} + M_{10} + M_{11})$

SMC versus Jaccard: Example

p = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

M_{01} = 2 (the number of attributes where p was 0 and q was 1)

M_{10} = 1 (the number of attributes where p was 1 and q was 0)

M_{00} = 7 (the number of attributes where p was 0 and q was 0)

M_{11} = 0 (the number of attributes where p was 1 and q was 1)

$SMC = (M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$

$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$

Cosine Similarity

- If $\mathbf{d}_1$ and $\mathbf{d}_2$ are two document vectors, then

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \|\mathbf{d}_2\| ,$$

where $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ indicates inner product or vector dot product of vectors, $\mathbf{d}_1$ and $\mathbf{d}_2$, and $\| \mathbf{d} \|$ is the length of vector $\mathbf{d}$.

- Example:

$$\mathbf{d}_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$$

$$\mathbf{d}_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$$

$\langle \mathbf{d}_1, \mathbf{d2} \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$\| \mathbf{d}_1 \| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$\| \mathbf{d}_2 \| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.449$

$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0.3150$

Extended Jaccard Coefficient (Tanimoto)

- Variation of Jaccard for continuous or count attributes
  - Reduces to Jaccard for binary attributes

$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

Correlation

- Korelasyon, nesneler arasındaki doğrusal ilişkiyi ölçer
- Korelasyonu hesaplamak için, veri nesnelerini, p ve q'yu standartlaştırıyoruz ve sonra «dot product» alıyoruz

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

Correlation measures the linear relationship between objects

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y}, \quad (2.11$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}) \quad (2.12$$

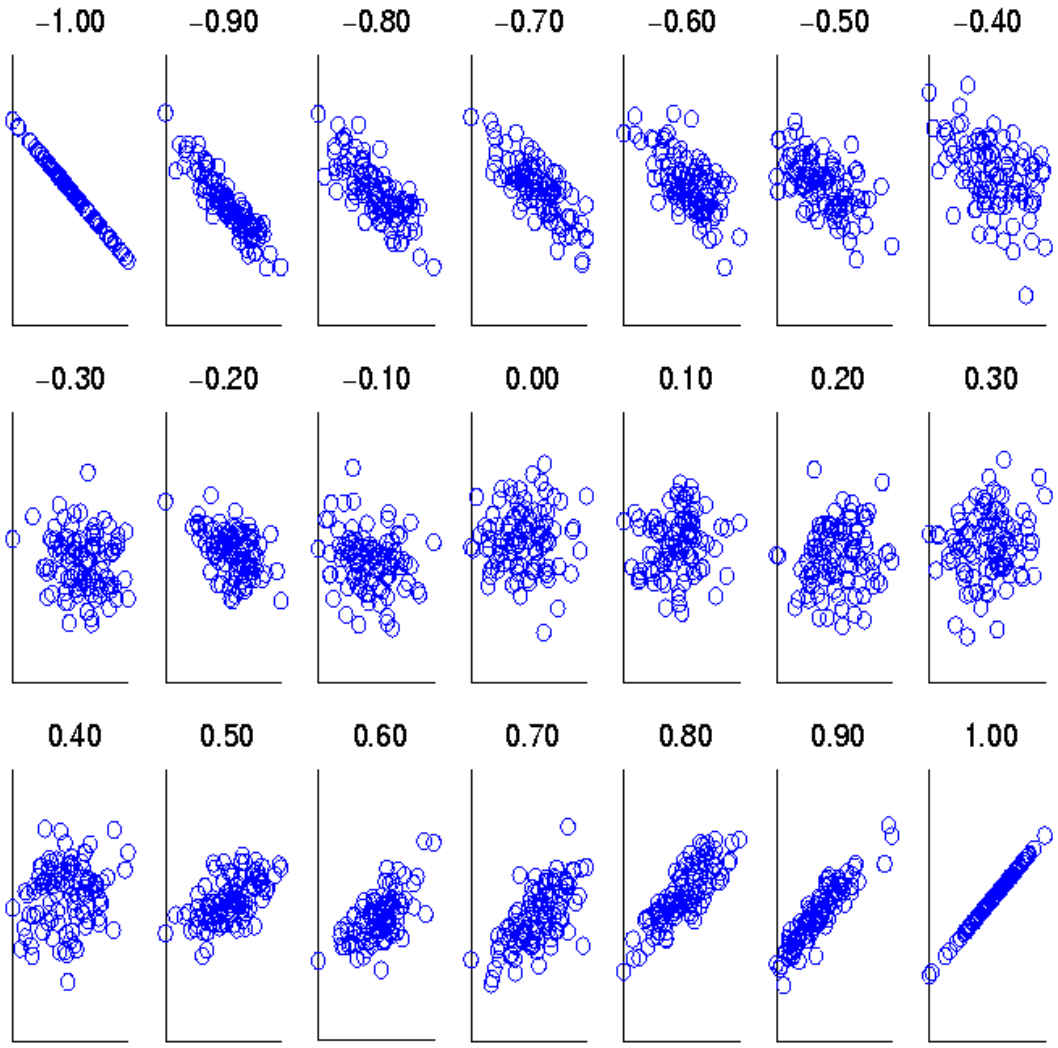
$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2}$$

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Drawback of Correlation

- **x** = (-3, -2, -1, 0, 1, 2, 3)
- **y** = (9, 4, 1, 0, 1, 4, 9)

$y_i = x_i^2$ ←

- mean(**x**) = 0, mean(**y**) = 4
- std(**x**) = 2.16, std(**y**) = 3.74

- corr = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74)

  = 0

If the correlation is 0, then there is no linear relationship between the attributes of the two data objects. However, **non-linear relationships** may still exist as in this example.

General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.

1. For the k^{th} attribute, compute a similarity, s_k, in the range $[0, 1]$.

2. Define an indicator variable, δ_k, for the k_{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have} \\ & \text{a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$similarity(p, q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use weights w_k which are between 0 and 1 and sum to 1.

$$similarity(p,q) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

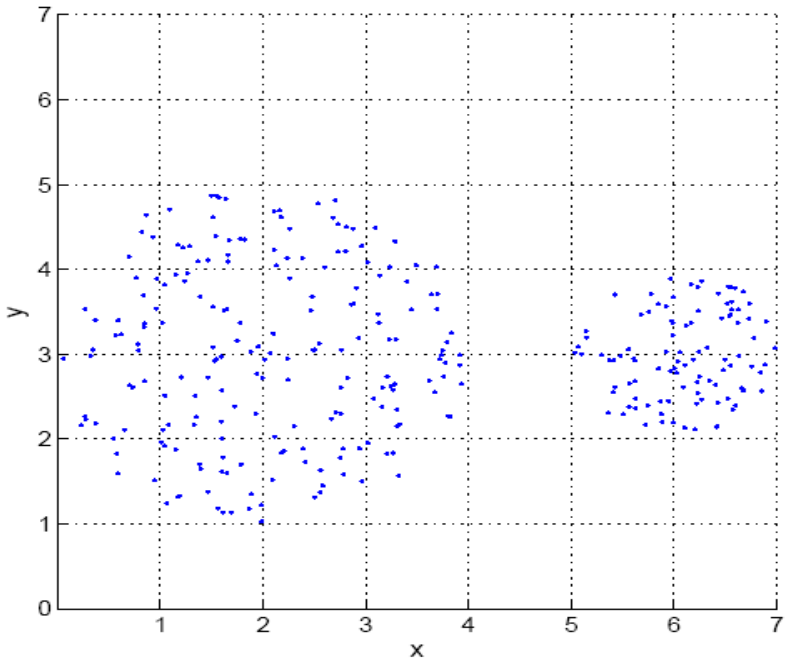
$$distance(p,q) = \left(\sum_{k=1}^{n} w_k |p_k - q_k|^r \right)^{1/r}$$

Density

- Belirli bir alanda veri nesnelerinin birbirine yakın olma derecesini ölçer
- Yoğunluk (***density***) kavramı yakınlık kavramı ile yakından ilgilidir.
- Yoğunluk kavramı tipik olarak kümeleme ve anormallik tespiti için kullanılır
- Examples:
  - Euclidean density
    - Euclidean density = number of points per unit volume
  - Probability density
    - Estimate what the distribution of the data looks like
  - Graph-based density
    - Connectivity

Euclidean Density: Grid-based Approach

- En basit yaklaşım, bölgeyi belirli sayıda eşit hacimli dikdörtgen hücrelere bölmek ve yoğunluğu hücrenin içerdiği nokta sayısı olarak tanımlamaktır.



Grid-based density.

0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Counts for each cell.

Euclidean Density: Center-Based

- Öklid yoğunluğu, belirli bir yarıçapı içindeki noktaların sayısıdır.

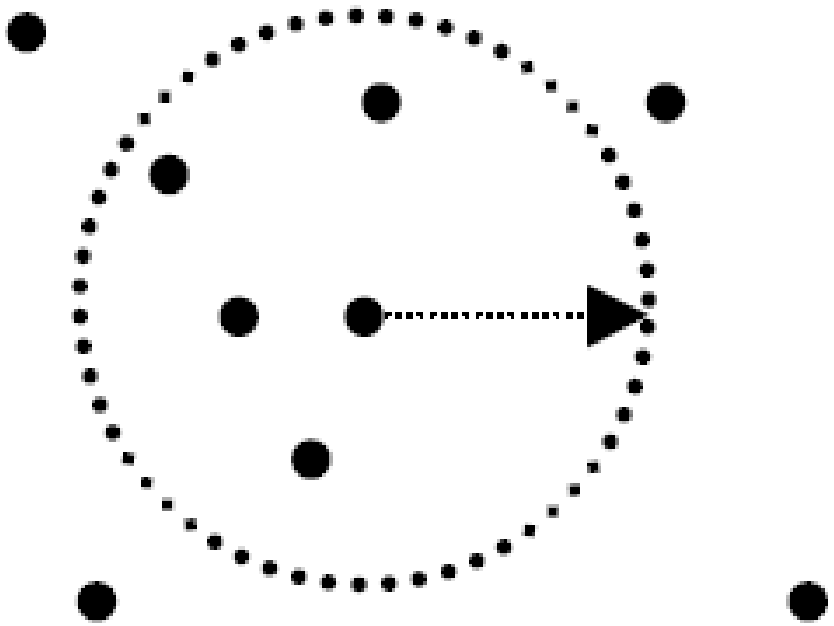


Illustration of center-based density.