Data Mining Notes by Emel Kayacı & Mehmet Anıl Taysi

Contents

1. Data vs. Knowledge

Data (Operation) → Information (Analytic) → Knowledge

Veri ambarları üretimi Data mining kullanımı

Data: En temelde yer alan hiç işlem görmemiş (raw) yapı, sadece operasyonel ihtiyaçlara yöneliktir. Objektiftir.

Information: Data'nın işlenmiş, yeniden organize edilmiş, gruplandırılmış halidir. Objektiftir.

Knowledge: Information içinde yer alan çeşitli örüntü, ilişki veya modelleri (pattern, model, finding) keşfetmek ve bu sayede geleceğe dair tahminlerde bulunmak. Subjektiftir.

- Bu keşifler sayesinde daha önce bilinmeyen, üstü kapalı (implicit) bir şekilde duran önemli bilgileri (non-trivial) elde ederiz.
- Hem information hem de knowledge reusable'dır.

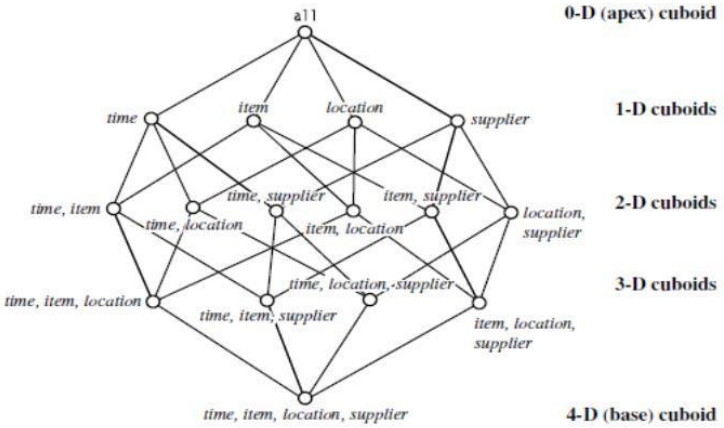
Wisdom: Knowledge'ın bir araya mükemmel bir uyumla bir araya gelmesi. Şimdilik bu seviyeye ulaşamıyoruz. Yukarıdakilerden hepsi aktarılabilir ve tekrar tekrar kullanılabilir burada ise böyle bir durum yoktur.

2. Data Warehouse (Veri Ambarı)

- Verimli olmaları için merkezi (integrated) olmalıdırlar. Heterojen kaynaklardan verinin birleştirilmesi sırasında yapılan işlemler:
  1. Data **cleaning**
  2. Data **integration** (consolidation, combining and storing of varied data in a single place)
- Non-volatile (kalıcı, uçucu olmayan) yapıları olmalıdır çünkü iyi çıkarım yapabilmek için uzun süredir saklanmış bilginin elimizde olması gerekmektedir.
- Subject-oriented (Konulara uygun düzenlenmiş olmalıdır, örneğin müşteri, ürün, tedarikçi vs.)
- Time-variant Çoğu veriden iyi çıkarım yapabilmek için zamana göre nasıl değiştiğini de görebilmemiz gerekmektedir. Veri ambarındaki her anahtar yapı, örtük veya açık olarak bir zaman unsuru içerir.

3. OLAP (Online Analytical Processing) Cube

- OLAP is a powerful technology for **data discovery**, including capabilities for **limitless report** viewing, complex analytical calculations, and predictive "what if" scenario (budget, forecast) planning.



- Yukarıda bir OLAP küpündeki boyutlar (dimension) görülmektedir. Bu boyutlar node olarak gösterilmiştir. Her bir node hangi boyutları içeriyorsa o boyutlarda çeşitli raporlar elde edilebileceğini göstermektedir.

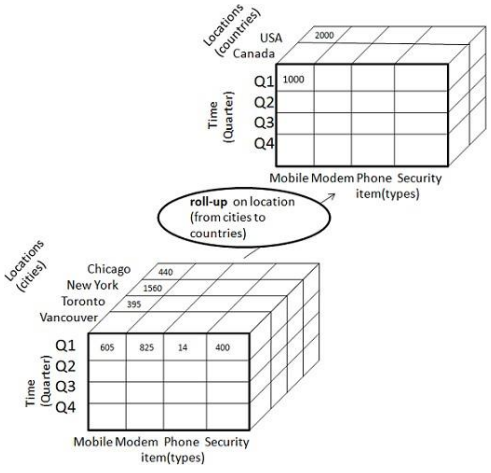
- Bu bir data mining çalışması değildir, analiz aşamasıdır.

- Object oriented yaklaşımdan yararlanılır.

Örnek: 10 boyutlu bir obje raporlama hizmetine sunulursa (OLAP küpü yaratılırsa) kaç farklı rapor elde edilebilir?

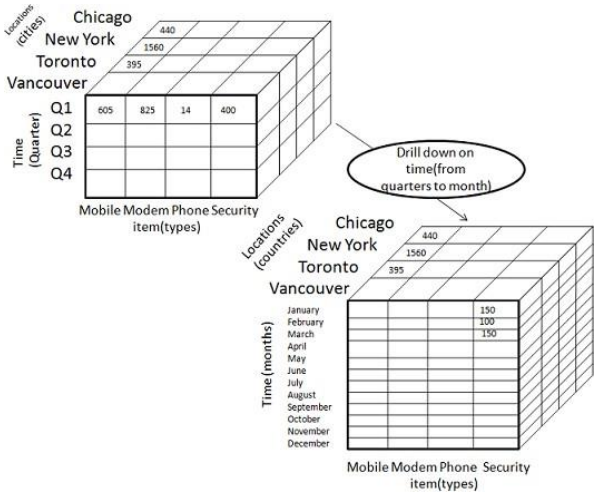
2^{10}-1 adet (Alt küme sayısından 1 çıkarıyoruz, buradaki 1 boş kümeyi ifade eder.)

3.1 OLAP Operasyonları



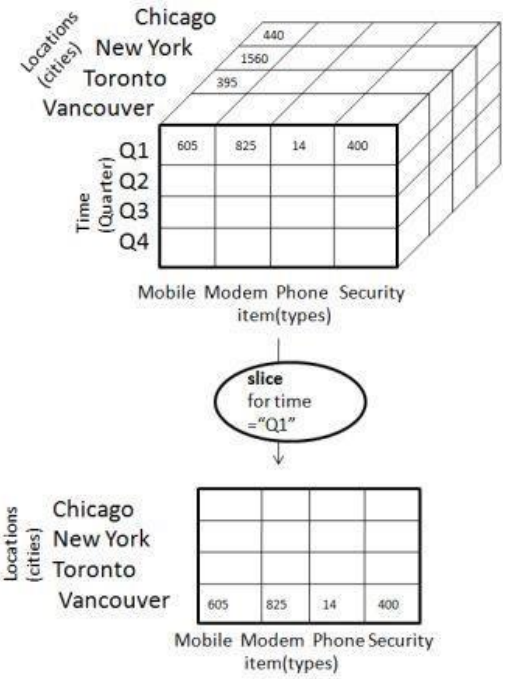
1. Roll-up (Aggregation)

Main goal is reducing dimensions and climbing up concept hierarchy. Concept hierarchy is a system of grouping things based on their order or level.



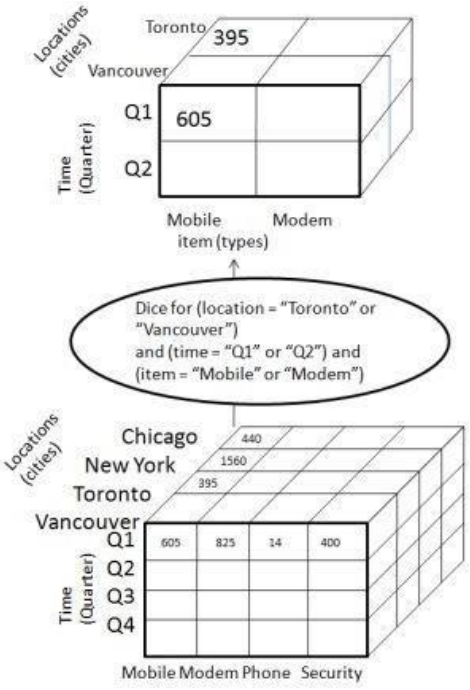
2. Drill-down

Opposite of rollup process where data is fragmented into smaller parts.



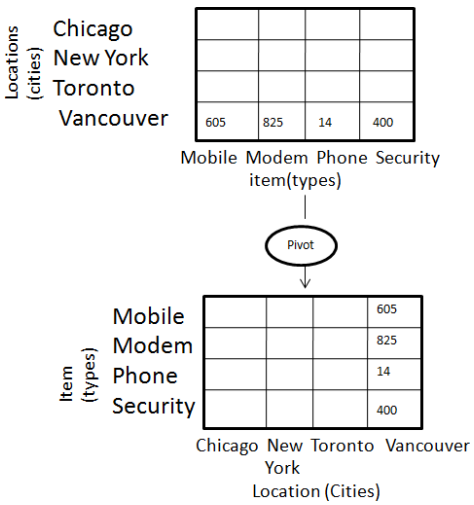
3. Slice

One dimension is selected, and a new sub-cube is created.



4. Dice

This operation is similar to a slice. The difference in dice is you select 2 or more dimensions that result in the creation of a sub-cube.



5. Pivot

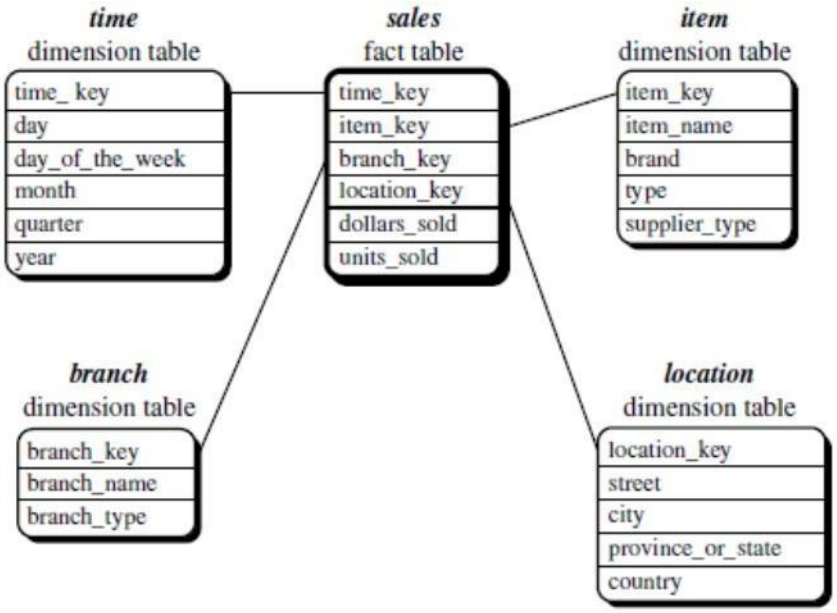
Rotate the data axes to provide a substitute presentation of data.

3.3 OLAP vs. OLTP

	Data Warehouse (OLAP)	Operational Database (OLTP)
1	Involves historical processing of information.	Involves day-to-day processing.
2	OLAP systems are used by knowledge workers such as executives, managers and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
3	Useful in analyzing the business.	Useful in running the business.
4	It focuses on Information out.	It focuses on Data in.
5	Based on Star Schema, Snowflake, Schema and Fact Constellation Schema.	Based on Entity Relationship Model.
6	Contains historical data.	Contains current data.
7	Provides summarized and consolidated data.	Provides primitive and highly detailed data.
8	Provides summarized and multidimensional view of data.	Provides detailed and flat relational view of data.
9	Number or users is in hundreds.	Number of users is in thousands.
10	Number of records accessed is in millions.	Number of records accessed is in tens.
11	Database size is from 100 GB to 1 TB	Database size is from 100 MB to 1 GB.
12	Highly flexible.	Provides high performance.

3.4 Schemas

3.4.1 Star Schema



Bu sayede many to many ilişkiler üçüncü bir tablo olan fact table ile one to many haline dönüştürülür. Örneğin müşteri ile ürün arasında satış tablosunun eklenmesi. Birden fazla müşteri birden fazla ürün alabilirken şimdi bir müşterinin birden fazla satış bilgisi olabilir, bir satış da yalnızca bir müşteriye ait olabilir.

Fact table: Yalnızca key ve unique identifier barındırır. Olay bu tablo üzerinden gerçekleşir. Örnekte bu olay satıştır. Unique identifier doğrudan sales işlemi ile alakalı özelliklerdir.

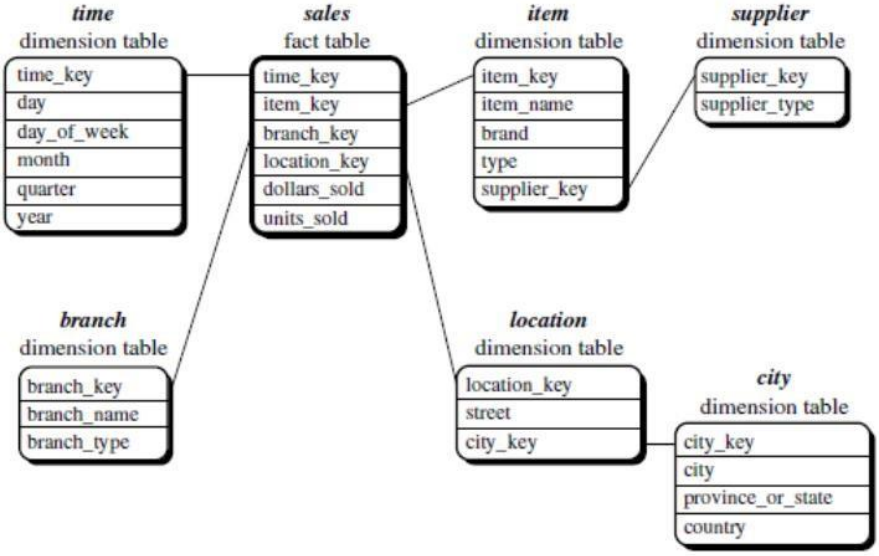
Dimension: Her bir key için açıklamalar bulunur.

Raporlar oluşturulurken genellikle bu sıra izlenir:

Join → Filter → Group → Aggregate

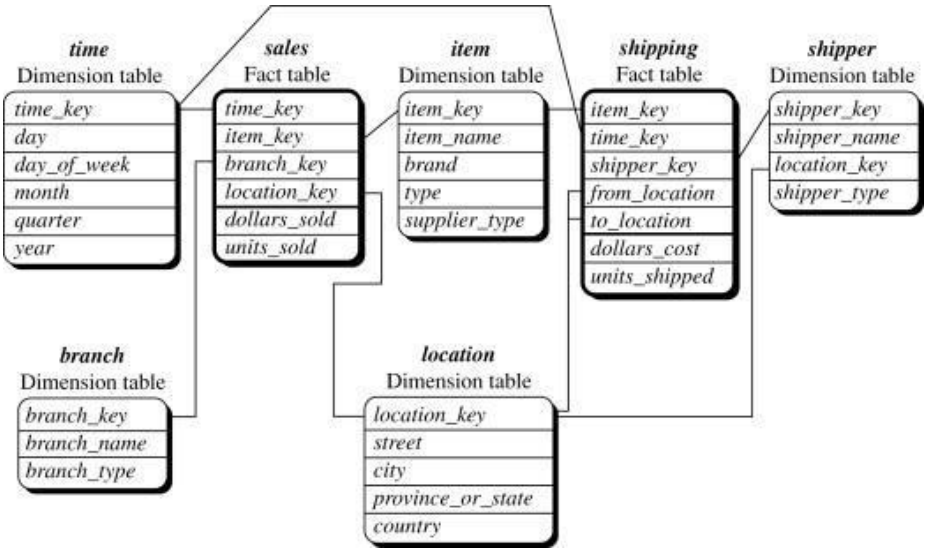
1. Gereken bilgilerin olduğu tablolar birleştirilir. (Join)
2. Sadece 2015 veya sadece İstanbul'da olan satışlar gibi filtreleme işlemleri uygulanır.
3. İstanbul'daki şubeler gibi gruplama işlemi uygulanır. (Group)
4. Gruplar için ortalama, mod, medyan gibi istatistiksel işlemlerle çeşitli bilgileri elde edilir.

3.4.2 Snowflake Schema



Star şemasından farkı, dimension tablolarının sub dimension tablolarının bulunabilmesidir.

3.4.3 Fact Constellation Schema



Snowflake şemasından farkı, birden fazla fact tablosu barındırabilmeleridir.

3.4.4 Şemalar nerelerde kullanılır?

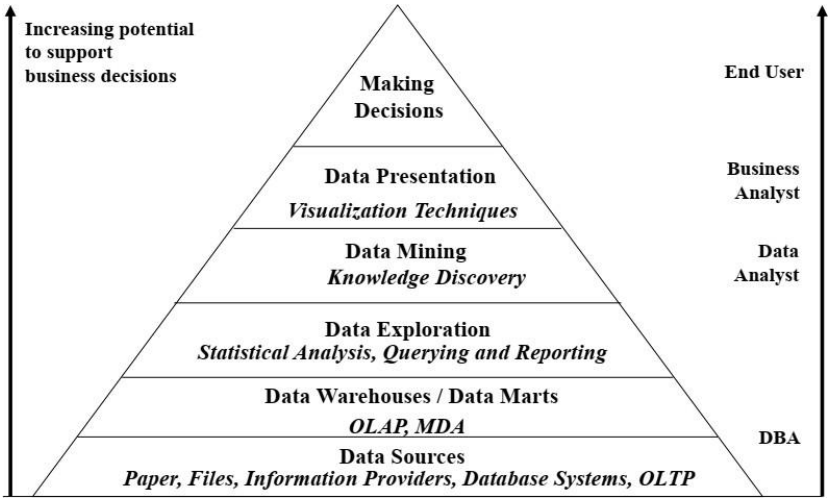
Veri ambarlarında birden fazla birbirleriyle ilişkileri bulunan objeler bulunduğundan fact constellation şemaları kullanılırken data martlarda tek obje modellendiğinden star veya snowflake şemaları kullanılır.

Şemalarda bahsedilen fact data OLAP küplerinde tutulan bilgileri göstermektedir.

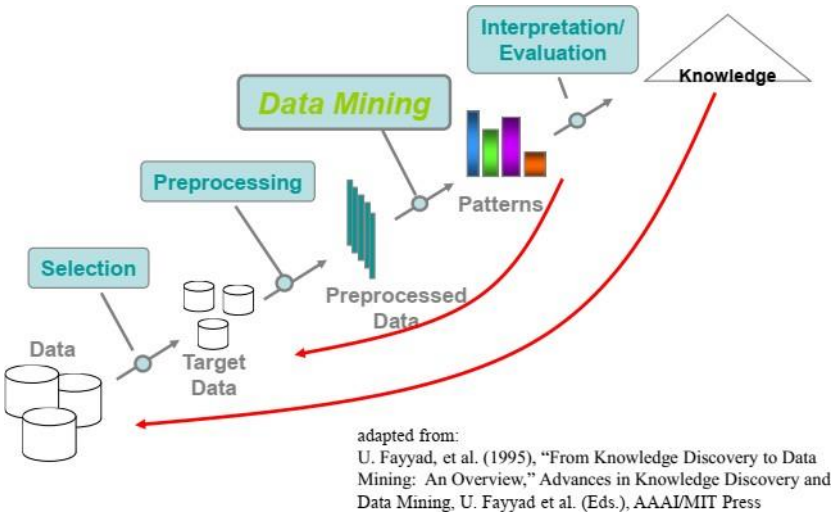
3.5 OLAP vs Data Mining

1. Advantages relative to data mining
    - Can obtain a wider variety of results
    - Generally faster to obtain results
2. Disadvantages relative to data mining
    - User must “ask the right question”
    - Generally used to determine high-level statistical summaries, rather than specific relationships among instances

4. Data to Knowledge Representations



Data to Knowledge Pyramid



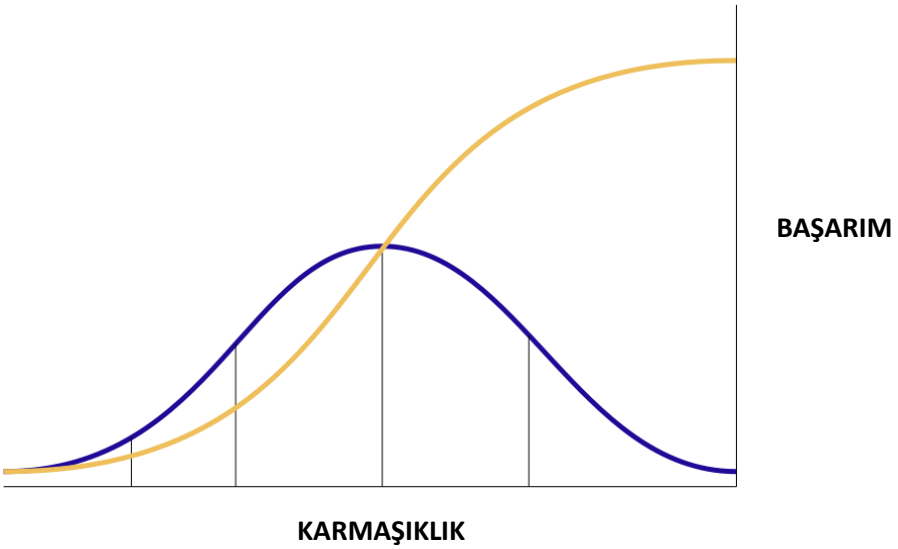
Data ile gösterilen kısım salt data değildir. Veri ambarında bulunan veridir. Target data adımında datamart oluşturuluyor. Datamart: Veri ambarından çeşitli özelleşmiş problemler bazında kesitlerin alınmasıdır.

5. Memorization, adaptation ve learning

1. Memorization: Salt bilgidir.

2. Adaptation: Salt bilgiden elde edilen pattern'lardır. Central limit theorem ile açıklanır.

3. Learning: Bulunan adaptation bilgisi birleşerek learning kavramına ulaşmamızı sağlar. Birbirleriyle uyumlu adaptation'lar diyebiliriz. Grafiği çan eğrisi şeklindedir. Amaç bu grafikte de görüldüğü üzere overfit veya underfit durumlarında bulunmadan en uygun noktada adaptasyonun yönetilmesidir.



Sarı ile gösterilen eğri adaptasyonu ifade eder. Veri (row) sayısı arttıkça verileri arasında desen (pattern) bulunma olasılığı da artmaktadır. Yani karmaşıklık arttıkça başarı da artmaktadır.

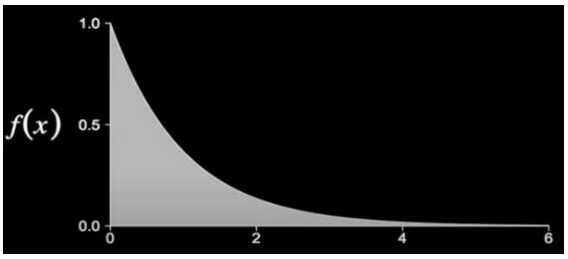
Mavi ile gösterilen eğri ise öğrenmeyi ifade eder. Tepe noktası ideal olup model daha önce hiç karşılaşmadığı veri hakkında yorum yapabilmektedir. Tepenin solunda kalan kısım underfit durumunu sağında kalan kısım ise overfit durumunu ifade eder.

İki grafik karşılaştırıldığında adaptasyonun artmasının overfit'e neden olduğu gözükmektedir. Çünkü model yalnızca veriyi iyi bilmekle kalmayıp bulundurduğu çelişkilerden de haberdardır bu nedenle yorum yapabilmesi de oldukça zordur.

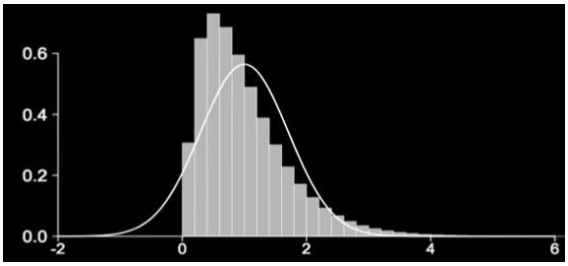
6. Central Limit Theorem

Örneklem ortalamasının (kendi iç dağılımından bağımsız) dağılımının örneklem büyüklüğü arttıkça normal dağılıma yakınsamasıdır. Ana çıkarım, veri yeterince sonsuz olduğunda birim hatanın öneminin olmamasıdır.

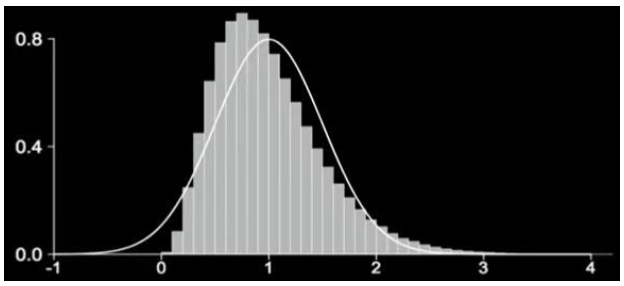
Elimizde bir milyon kayıt bulunan veri (iç dağılımı exponential) olsun.



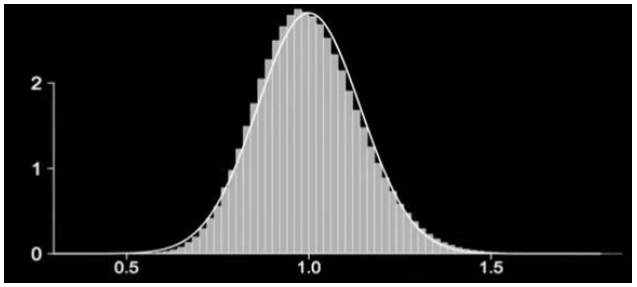
Bu bir milyon kayıttan 2 büyüklüğünde örneklem (sample) alıp ortalamalarını bulup bir grafiğe yerleştirelim ve bu işlemi n kez tekrar edelim.



Şimdi de 4 büyüklüğünde örneklem alıp işlemi n kez tekrar edelim.



Ve 50 büyüklüğünde örneklem alıp işlemi n kez tekrar edelim.



6.1 Teorem hangi problemlerde işe yarar?

Büyük bir şirketteki maaşların ortalaması $62.000 ve standard sapması $32.000 olsun. Rastgele seçilen bir çalışanın maaşının $66.000'den fazla olma olasılığı nedir?

Bu soruya CLT ile cevap veremeyiz ve cevap verebilmek için gerçek dağılımın nasıl olduğunu bilmemiz lazım.

Büyük bir şirketteki maaşların ortalaması $62.000 ve standard sapması $32.000 olsun. Rastgele seçilen 100 çalışanın maaşının $66.000'den fazla olma olasılığı nedir?

Bu soruya CLT ile cevap verebiliriz çünkü rastgele seçilen 100 adet çalışan bir normal dağılım eğrisi oluşturmuş olabilir.

7. Bias nedir?

Bias bir tercihtir, mevcut adaptasyonlara bakarak aralarında tercih yapılmasıdır. Adaptasyonların baskınlığı (görülme sıklıkları) birbirinden farklıdır. Bazıları tesadüfi bile olabilir. **Bias yoksa öğrenme de yoktur!**

3 türü bulunur.

1. Inductive (Restrictive, Language) Bias: Bias daha en başta alternatiflerini kısıtlıyorsa yani baştan elimine ediyorsa bazı çözümleri inductive bias olarak isimlendirilir. Yalnızca en başta eleme işlemi olduğundan data driven (veri odaklı) yaklaşımlarla uyumlu değildir.
    - Örneğin lineer regresyonda model veriler arasındaki ilişkinin lineer olduğuna dair önyargılıdır.
    - Öneri sistemlerinde yüksek güven değerine sahip 2 ürünün müşteriler tarafından çoğu zaman birlikte satın alınacağına dair önyargılıdır.
2. Search (Preference) Bias: Arama kümesinde dolaştıkça (satırları gezdikçe) bazı çözümleri elimine eder. Akıl yürütme (reasoning) de içerir. Örneğin karar ağaçları bu şekilde çalışmaktadır.
3. No bias: Öğrenme değildir, ezberdir.

8. Major tasks in data pre-processing

1. Data cleaning

- Eksik verileri çeşitli istatistiksel yöntemlerle doldur.
- Outlier değerlerini keşfet.

- Uyumsuz, birbirleriyle çelişen verileri düzelt.
- Veriler entegre olurken tekrarlar oluşabileceğinden bunları yok et.

2. Data integration (Farklı kaynaklardan gelen verileri tek bir merkezde birleştirmek)

3. Data transformation

1. Smoothing: Remove noise from data

2. Aggregation: Summarization, data cube construction

3. Generalization: Concept hierarchy climbing

4. Normalization: Scaled to fall within a small, specified range min-max normalization. Min-max normalization yalnızca bir örnektir ayrıca z-skoru normalizasyonu gibi çeşitli yöntemleri bulunur.

$$y = \frac{x - min(x)}{old\_range} * new\_range + new\_min$$

Örnek: Maaş verisi $12.000 ile $98.000 arasında değişiyor olsun ve [0.0, 1.0] aralığında normalize edilmek istensin. Bu halde $73.600'ün değeri ne olur?

(73.600 – 12.000) / (98.000 – 12.000) * (1.0 – 0.0) + 0 = 0.716

Z skoru normalizasyonu: **(x – μ) / σ**

x: Original value

μ: Mean of data

σ: Standard deviation of data

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

5. Attribute/feature construction: New attributes constructed from the given ones.

Not: Çoğu kaynakta discretization bu işlemin alt başlığı olarak yer almaktadır.

4. Data reduction (Sütun sayısını azaltmak veya çeşitli satırları dahil etmemek)

5. Data discretization

Gerçek dünya verilerinde sayısal değerleri kategorik değerlere en az veri kaybıyla dönüştürmek işlem hızını oldukça arttırır. Çeşitli yöntemleri bulunur.

Örneğin cluster analysis kullanarak veri unsupervised learning kullanılarak kümelere ayrılabilir. Karar ağaçları benzer verileri aynı gruplar içerisine toplamak için kullanılabilir.

Attribute	Age	Age	Age	Age
	1,5,4,9,7	11,14,17,13,18,19	31,33,36,42,44,46	70,74,77,78
After Discretization	Child	Young	Mature	Old

Binning de bu yöntemlerden biridir. İki türü bulunur. Frekans bazlı olanı gerçek hayat problemlerinde bolca kullanılırken uzaklığa bağlı olan çok kullanılmaz çünkü her bir aralıkta eşit sayıda değer bulundurmayı garanti etmez.

1. Equi-depth (frequency) partitioning

Data: 15, 8, 21, 4, 24, 21, 34, 25, 28

Step 1: Sort the data in ascending order

Data: 4, 8, 15, 21, 21, 24, 25, 28, 34

Step 2: Find Width (W = (Max-Min) / Number of data)

W = (34 – 4) / 9 = 3.33 = 3

Data should be in 3 for each bin

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

Amaç değer sayısını azaltmaktı. Şimdilik bir azaltma bulunmamaktadır. Azaltmak için çeşitli yöntemler bulunmaktadır. Bunlardan biri her bir bin değerlerini ortalama ile değiştirmektir.

Step 3: Smoothing by bin means

Bin 1: 9, 9, 9 (replace the data)

Bin 2: 22, 22, 22

Bin 3: 29, 29, 29

Step 3 (Alternative): Smoothing by bin boundaries

1. Pick the minimum and maximum value, put the minimum on the left side and maximum on the right side.
2. Middle values in bin boundaries move to its closest neighbor value with less distance.

Bin 1: 4, 4, 15 (8-4=4, 15-8=9 hence 8 is closer to 4)

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

2. Equi-width (distance) partitioning

Data: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Step 1: Find the width (max-min / number of bins)

Number of bins is optional and given by user. Bu bizim belirlediğimiz aralık sayısına **quantile** denir. In this example consider the number of bins is 3.

(215-5) / 3 = 70

Step 2: Partition the bins

Bin 1 (5 + 70 = 75 (from 5 (minimum) to 75): 5, 10, 11, 13, 15, 35, 50, 55, 72

Bin 2 (70 + 75 = 145): 92

Bin 3 (70 + 145 = 215): 204, 215

9. Bazı İstatistik Hatırlatmaları

- Sıralı olmayan bir listede maksimum veya minimum elemanı bulma: O(n)
- Sıralı listede maksimum veya minimum elemanı bulma: O(1)
- Listede ortalama bulmakta sıralı veya sırasız olması fark etmiyor her türlü O(n)
- Sıralı olmayan bir listede range bulunması (en büyük eleman ile en küçük eleman arasındaki fark): O(n)
- $mean - mode = 3 \times (mean - median)$

- Tüm listeyi sıralamadan median bulmanın kısa bir yolu vardır. Bu algoritmaya median of median denir. Klasik yöntem olan liste sıralamada zaman karmaşıklığı nlogn iken bu yöntemle n'de düşürülmüştür. Yaklaşık değer bulmaz, tam değere ulaşırız.

9.1 Median of Medians

Median of Medians

Partitioning on Array

Partitioning bir dizide pivot seçimine dayanır.
Seçilen pivota göre, pivottan küçük olanlar pivotun solunda,
pivottan büyük olanlar ise pivotun sağında yer alır.

Pivot seçiminin kontrol edilmediği partitioning işlemlerinde,
eğer pivot zaten yukarıdaki koşulları sağlıyor ise $O(n)$ best case
eğer karşılanmıyor ve dizinin maksimum veya minimumu ise $O(n^2)$ worst case

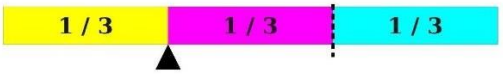
Pivot seçimi için uygun yöntem bulmak gerekmektedir.
İyi bir pivot dizisi, her parça dizinin boyutunun sabit bir kısmı olacak şekilde bölmelidir.

bu bölmeler eşit uzunlukta olmayabilir.

Pivot Bulma Yöntemi

- Dizinin ilk 2/3'sinin medyan değerinin recursive olarak bulunması
- Bu medyan değerinin partition yaparken pivot olarak kullanılması

Bu işlem dizinin daha önceden bölünmüş olan ilk 2/3'lük kısmının da kendi içinde parçalanmasını sağlar.



- Our algorithm
  - Recursively calls itself on the first 2/3 of the array.
  - Runs a partition step.
  - Then, either immediately terminates, or recurses in a piece of size $n / 3$ or a piece of size $2n / 3$.

$T(1) = \Theta(1)$

$T(n) \leq 2T(2n/3) + \Theta(n)$

Şimdi bu pivot seçme yöntemimizin zaman karmaşıklığı denklemine "Master Theorem" uygulayalım.

$T(1) = \Theta(1)$ iken $T(n) = 2T(2n/3) + \Theta(n)$ (Big theta)

a $1/b$ d

Master theorem: $\log_b a > d$

$$\log_b a > d \rightarrow O(n^{\log_b a}) \quad 1.$$

$$\log_b a = d \rightarrow O(n^d \log n) \quad 2.$$

$$\log_b a < d \rightarrow O(n^d) \quad 3.$$

Örneğe geri dönersek

$\log_{3/2} 2 > 1$ yani ilk seçenek oluyor.

$O(n^{\log_{3/2} 2}) \cong O(n^{1.26})$

Şimdi pivot seçme konusunda daha iyi bir yöntem inceleyeceğiz.

Median of Medians Algoritması

9.2 İki sütun birbirleriyle ne kadar ilişkilidir?

Bunun için öncelikle her iki sütunun temizlenmiş olması gerekmektedir. Böylece bulundurdukları verilerin karakteristliği daha iyi ortaya çıkabilmektedir. Sayısal veriler için covariance kullanılırken kategorik verilerde chi-square kullanılabilir.

Chi-Square Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

Parantez içindeki değerler şu şekilde hesaplanır: 1500 kayıttan 300'ü satranç oynuyormuş, eğer bu trend aynı olsaydı her yerde o zaman 450'de de 5'te 1'i yani 90 kayıt olurdu.

Beklentiler ile gerçek değerler arasında ne kadar sapma varsa değişkenler de o kadar bağımlıdır (correlated).

9.3 Conditional Probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of A given B — Probability of A and B — Probability of B

B koşulu altında A'nın gerçekleşme olasılığı nedir?

1. 10 kez para atıldığında 5'inde tura gelme olasılığı nedir?

   Cevap: TTTTTYYYYY
   Bunlar değişik şekillerde sıralanabilir. 10! / 5!*5! = 10*9*7*6/5*4*3*2 = 252
   On kez para atıldığında 2^10 = 1024 şekilde sıralanabilir.
   252/1024 = 63 / 256

2. 10 kez para atıldığında ard arda 5 kez tura gelme olasılığı nedir?
   Turalar yan yana geleceklerinden bir grup olarak düşünürüz.
   Şu şekilde sıralanabilirler:
   TTTTT????? 1 X 2 X 2 X 2 X 2 X 2 = 32
   YTTTTT???? 1X1X2X2X2X2 = 16
   ?YTTTTT??? 2X1X1X2X2X2 = 16
   ??YTTTTT?? 2X2X1X1X2X2 = 16

???YTTTTT? 2X2X2X1X1X2 = 16
????YTTTTT 2X2X2X2X1X1 = 16
32 + 16*5 = 112
On kez para atıldığında 2^{10} = 1024 şekilde sıralanabilir.
112/1024 = 7/64

3. 10 kez atılan parada 5'inin tura geldiği bilindiğine göre bunların ard arda olma olasılığı nedir?
Cevap: 2.soru / 1.soru
Koşullu olasılık sorularında payda artık tüm durumlar değil soruda bilindiğine göre kalıbından önce gelen ifade olmaktadır.

4. Assume that a test to detect a disease whose prevalence is (1/1000) has a false positive rate of 5 and a true positive rate of 100%. What is the **probability that a person found to have a positive result** actually has the disease assuming that you know nothing about the person's symptoms?

False positive: Testin sonucu pozitif çıkmış fakat gerçekle uyuşmadığından false denilmiş. Yani gerçekte pozitif (hasta) değil.
True positive: Hem tesin sonucu pozitif hem gerçekten pozitif. Gerçekle uyuştuğundan başına true gelir.

D = Has the disease T = Test result is positive

P(T | Not D) şu şekilde okunur. Sağdaki kısmın hangi durumu belirttiğine bakılır. Bu soruda hasta olmadığını belirtiyor. Daha sonra bilindiğine göre kalıbı gelir ve sol kısım okunur.
Hasta olmadığı bilindiğine göre testin pozitif çıkma olasılığı

Given: P(D) = .001 so P(Not D) = .999
P(T | Not D) = .05
P(T | D) = 1.00
P(D | T) = P(T∩D) / P(T)
P(T) bilinmediğinden koşullu olasılık formüllerinin değişiminden yararlanılarak P(D)'yi kullanarak istediğimiz sonuca ulaşmaya çalışıyoruz.
P(T∩D) = P(D)P(T | D) = .001

Test pozitif ise bu durumu oluşturan iki unsur vardır: Kişi hasta değil ama testi pozitif, kişi hasta ve testi pozitif. Soruda asıl sorgulanan bu test ne kadar güvenilir?
P(T) = P(T∩D) + P(T∩Not D)
= P(T∩D) + P(Not D)P(T | Not D)
= .001 + (.999)(.05)
= .05095
P(D | T) = .001/.05095 = .019627

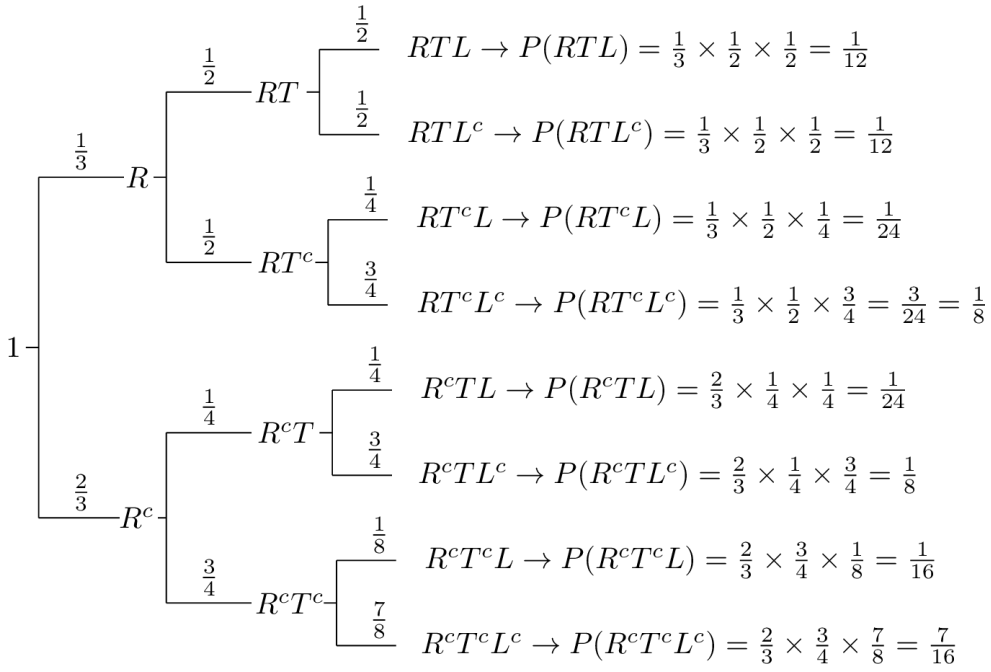
Testin pozitif olduğu bilindiğine göre gerçekten hasta olanları tespit etmesi yüzde 2 olasılıktadır, görüldüğü üzere test çok güvenilir değildir. (Güvenilir gibi gözükse de sınıfların çok imbalance olması burada oranı da etkilemiştir.)

5. In my town, it's rainy one third of the days. Given that it is rainy, there will be heavy traffic with probability 1/2, and given that it is not rainy, there will be heavy traffic with probability 1/4. If it's rainy and there is heavy traffic, I arrive late for work with probability 1/2.

On the other hand, the probability of being late is reduced to 1/8 if it is not rainy and there is no heavy traffic. In other situations (rainy and no traffic, not rainy and traffic) the probability of being late is 0.25. You pick a random day.

a. What is the probability that it's not raining and there is heavy traffic and I am not late?
b. What is the probability that I am late?
c. Given that I arrived late at work, what is the probability that it rained that day?

R = Rainy days T = Heavy traffic L = Late for work



a. 1/8 (Direkt ağaçtan bulunabilir.)
b. 1/12 + 1/24 + 1/24 + 1/16 = 11/48
c. 1/12 + 1/24 = 3/24
3/24*48/11 = 6/11