December, 2016

AA ML Course – Theoretical Session #1

Avrahami Israeli

Advanced Analytics
We mine your business

Agenda

1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. Outliers
8. Normalization & Transformation
9. Discretization
10. Unbalanced data

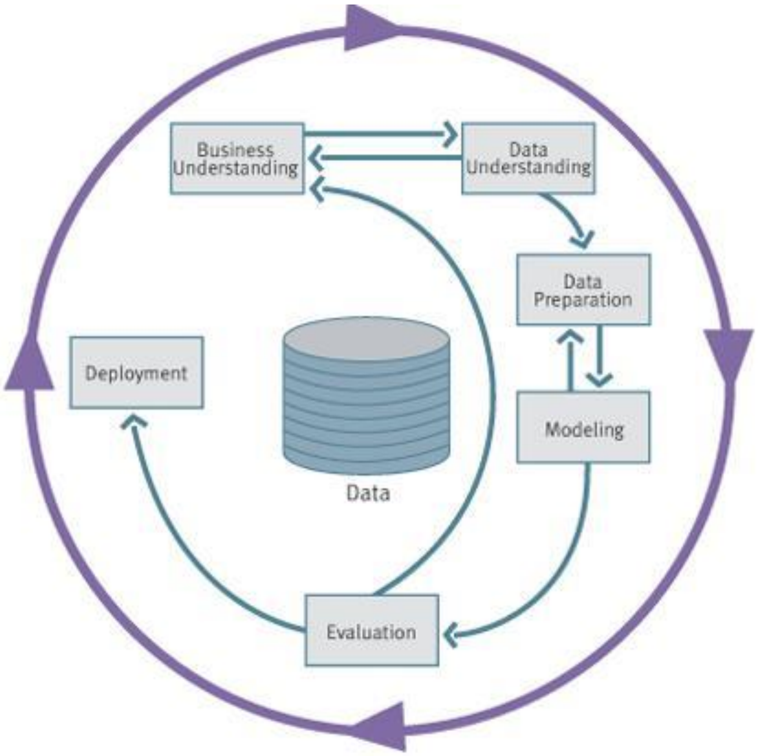
Introduction (1)

- Today - the first part of the CRISP-DM (and most important one!)
- What is NOT going to be covered here
- Statistical session VS ML ‘hard core’ session
- Not all topics involve heavy theoretical material

Introduction (2) - CRISP-DM

CRISP-DM breaks the process of data mining into six major phases

1. Business Understanding
2. **Data Understanding**
3. **Data Preparation**
4. Modeling
5. Evaluation
6. Deployment



The sequence of the phases is not strict and moving

back and forth between different phases may be required

Introduction (3)

Why is data preprocessing important?

- "Garbage in → Garbage out"
  - No quality data – no quality mining results!
  - Irrelevant data
  - Redundant data
  - Too much data (e.g. outliers, curse of dimensionality)
  - Data representation (e.g. - zip-code)

Introduction (4)

Major tasks in data preprocessing

- Data cleaning (e.g. missing values handling)
- Data transformation (e.g. normalization)
- Data discretization
- Data reduction

Agenda

1. Introduction
2. **Data types**
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. Outliers
8. Normalization & Transformation
9. Discretization
10. Unbalanced data

Data types (1)

	Type	Example
I.	Numerical data (double)	Income (e.g. 650.34)
II.	Numerical data (int)	# of children (e.g. 4)
III.	Boolean	Gender (e.g. male)
IV.	Categorical data	Colors (e.g. green)
V.	Ordinal data	Satisfaction (e.g. 2/5)
VI.	Others	Comments

Data types (2)

Why is it so important ??

- A-normal input for modeling
- Distance measures
- Models results are based on this input

Ideal - Subset of HSW22 MIDAS Operation unit level data.txt

Console ~/

```
> head(car.test.frame)
                 Price   Country Reliability Mileage  Type Weight Disp.  HP
Eagle Summit 4    8895       USA           4      33 Small   2560    97 113
Ford Escort   4   7402       USA           2      33 Small   2345   114  90
Ford Festiva 4    6319     Korea           4      37 Small   1845    81  63
Honda Civic 4     6635 Japan/USA           5      32 Small   2260    91  92
Mazda Protege 4   6599     Japan           5      32 Small   2440   113 103
Mercury Tracer 4  8672    Mexico           4      26 Small   2285    97  82
> sapply(car.test.frame, class)
      Price     Country Reliability     Mileage        Type      Weight       Disp.          HP
  "integer"    "factor"   "integer"   "integer"    "factor"   "integer"   "integer"   "integer"
> |
```

For Help, press F1 | Table: Subset of HSW22 MIDAS C | Variables: 7 | Samples: 101085

Agenda

1. Introduction
2. Data types
3. **Distance measures**
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. Outliers
8. Normalization & Transformation
9. Discretization
10. Unbalanced data

Distance Measures

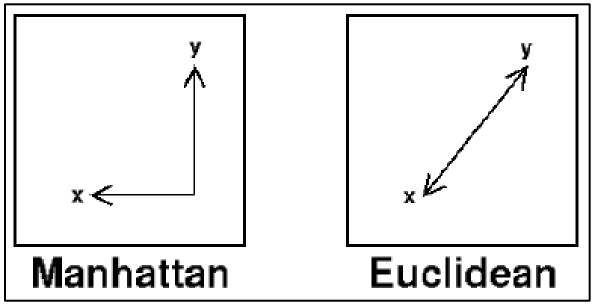
- Distance measure most satisfy some basic rules (e.g. $d(x, y) \geq 0$)

- Distance measure examples:

  - ***Euclidean** (l_2) **distance:*** $d(x,y) = \sqrt{(x-y)^T(x-y)} = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$

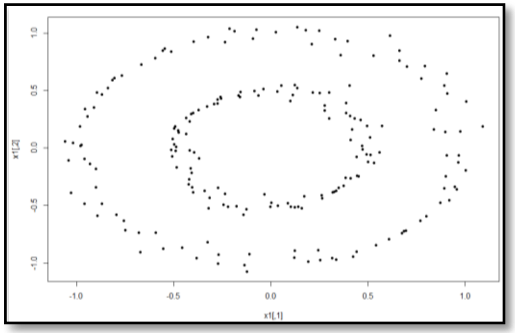
  - ***Manhattan** (l_1) **distance:*** $d(x,y) = \sum_{j=1}^{p} |x_j - y_j|$

  - *l_d–**distance:*** $d(x,y) = \{\sum_{j=1}^{p} |x_j - y_j|^d\}^{1/d}$



- Related examples: K-means, K-nn, Recommendation System algorithms

- Let's have a look in R to see why it is critical



Agenda

1. Introduction
2. Data types
3. Distance measures
4. **Correlation and Mutual information**
5. Data distribution
6. Missing values
7. Outliers
8. Normalization & Transformation
9. Discretization
10. Unbalanced data

Correlation

- Definition: Correlation refers to **any** of a broad class of statistical relationships involving dependence
- How is this related to our discussion ?
- Where else will we use these measures?
- Most common correlations:

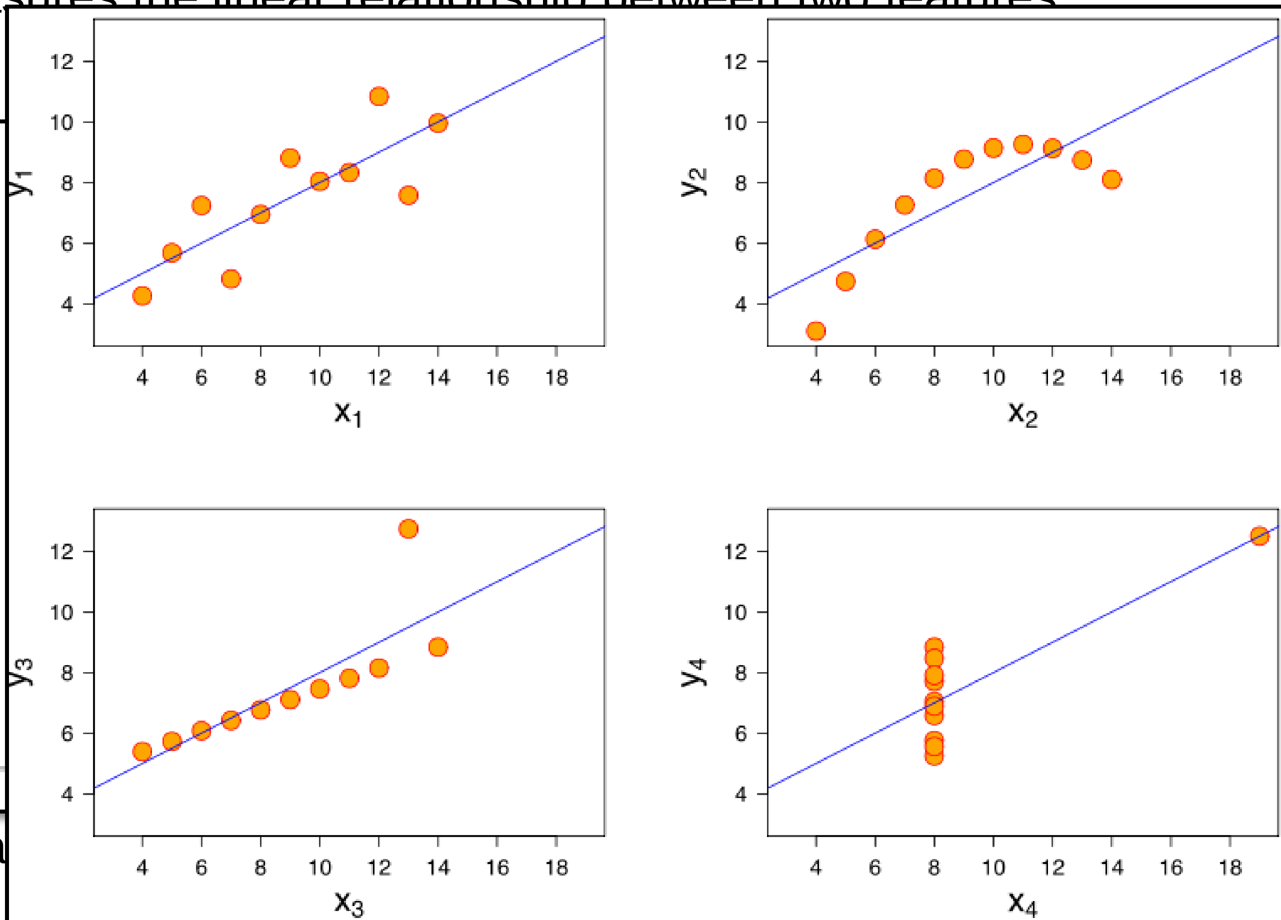
1. **Pearson Correlation** – measures the degree of **linear dependence** between two variables

2. **Spearman correlation** – measures how well the relationship between two variables can be described using a **monotonic function**

3. **Kendall's tau correlation** - measures the **“ordering” dependency** between two variables

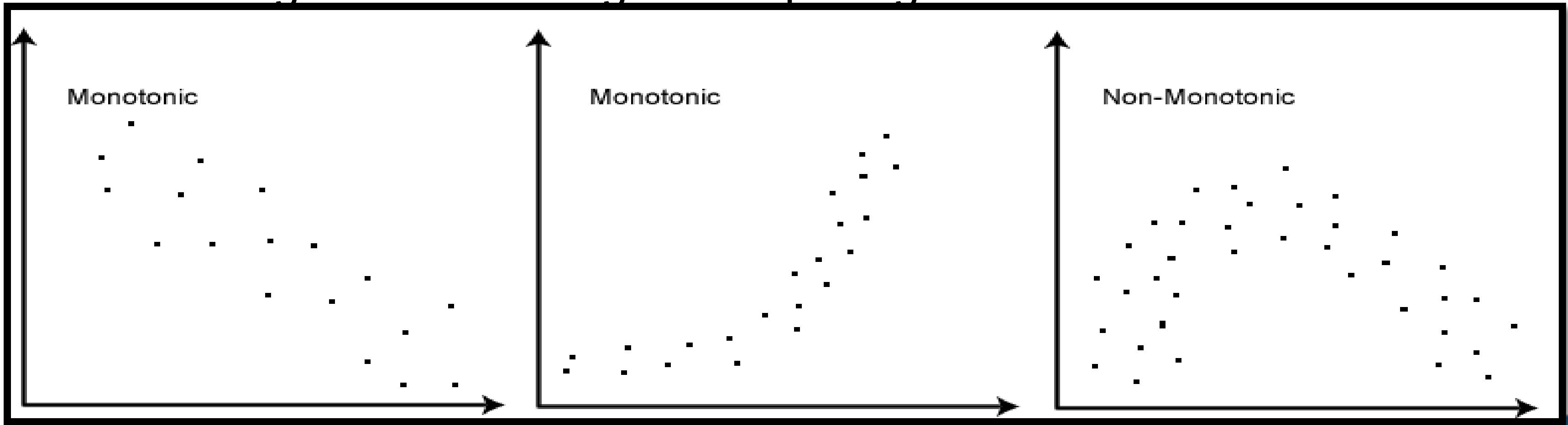
Pearson Correlation

- Measures the linear relationship between two features
- Def
- Sar
- $r_{(X,Y}$
- Rar
- $r_{(X,Y}$
- Wha



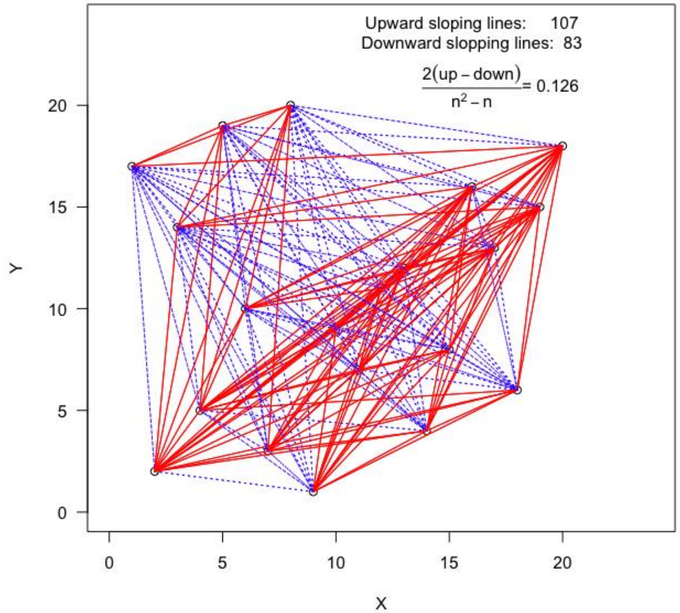
Spearman Correlation

- Measures the **monotonic** behavior relationship between two features
- In some way, 'connects' between Pearson and Kendall's tau
- Definition : $r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} * \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$ (BUT – the x_i & y_i are the **ranked** features!)
- What about ties?
- Range: [-1,1] (what do the -1,0,1 values mean?)
- Advantages/disadvantages comparing to Pearson correlation



Kendall's Tau Correlation

- Measures the **rank** correlation
- Definition : $\tau_{X,Y} = \dfrac{(\#\ of\ concordant\ pairs) - (\#\ of\ discordant\ pairs)}{\frac{1}{2}n(n-1)}$
- What about ties?
- Range: [-1,1] (what does the -1,0,1 values mean?)
- Advantages/disadvantages comparing to Spearman correlation



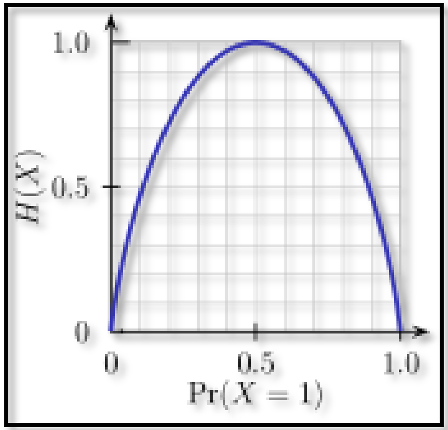
- Yalla, let's use R!

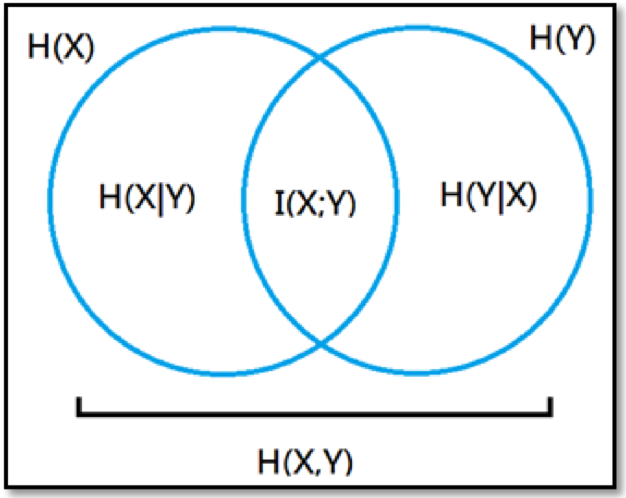
Mutual Information

- **Mutual information** is one of many quantities that measures how much one random variables tells us about another
- Can catch **non-linear** relationship between features
- Definition:
  a. Discrete random variables X and Y:
  $$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$
  b. Continuous random variables X and Y
  $$I(X,Y) = \int_X \int_Y f(x,y) \log\left(\frac{f(x,y)}{f(x)f(y)}\right) dydx$$
- Can also be expressed using the entropy measure:
  $$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X,Y) - H(X|Y) - H(Y|X)$$
  (<-> the amount of uncertainty in X which is removed by knowing Y)

Shannon Entropy

- One out of many information theory measures
- Def. (In the context of information theory) : a measure of the **uncertainty** in a random variable
- $H(X) = \sum_i p(x_i) I(x_i) = -\sum_i p(x_i) log_b(p(x_i))$
- $H(X|Y) = -\sum_{i,j} p(x_i, y_i) log_b(\frac{p(y_i)}{p(x_i,y_i)})$
- Range: [0, ?]. When do we get maximum value?



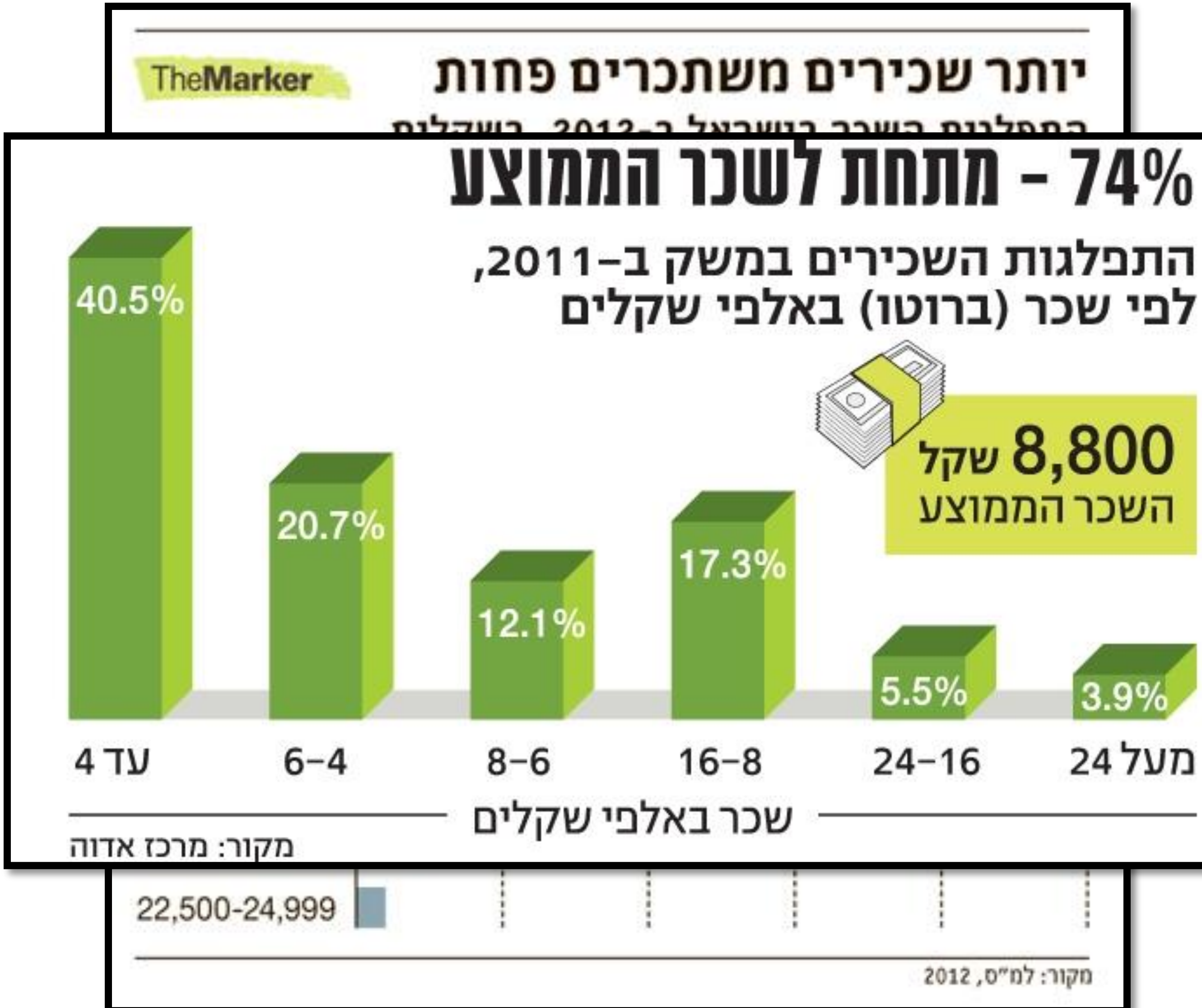


Correlation and MI

- So – when should we use each measure?
  - Discrete features – MI
  - Continuous features – Start with correlation
  - Continuous features – always check MI (what is the most critical decision now?)
- In the 'correlation world' – which measure to use?
  - Care about the actual values? If so – Pearson
  - Care only about the **rank** of value? If so – Spearman
  - Care about the **order** of the value? If so – Kendell's tau
  - Don't care? If so – so do I

Agenda

1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. **Data distribution**
6. Missing values
7. Outliers
8. Normalization & Transformation
9. Discretization
10. Unbalanced data


TheMarker
יותר שכירים משתכרים פחות
74% - מתחת לשכר הממוצע
התפלגות השכירים במשק ב-2011,
לפי שכר (ברוטו) באלפי שקלים
8,800 שקל
השכר הממוצע
40.5%
20.7%
12.1%
17.3%
5.5%
3.9%
עד 4
6-4
8-6
16-8
24-16
מעל 24
שכר באלפי שקלים
מקור: מרכז אדווה
22,500-24,999
מקור: למ"ס, 2012

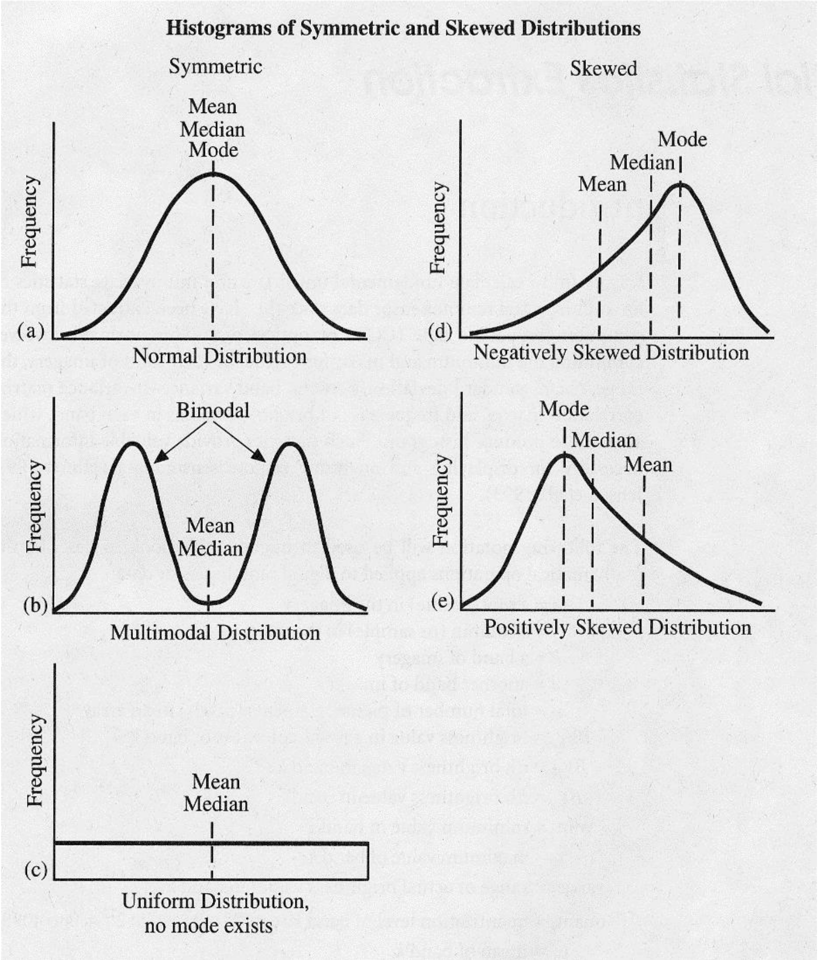
Basic measures (1)

Many statistical tests assume values are normally distributed, but this is not always the case

- Examine data prior to processing

Comparing Mean, Median & Mode

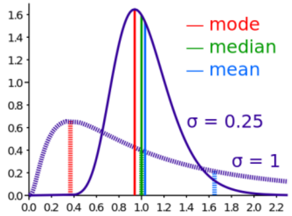
- Mode (שכיח)
  - Good for nominal variables
  - Quick and easy
- Median (חציון)
  - Robust central tendency statistics
    - Less sensitive to outliers and extreme values
  - Good for "bad" distributions
- Mean (ממוצע)
  - Most commonly used statistic for central tendency
    - Generally preferred except for "bad" distribution
  - Based on all data in the distribution
  - Used for inference as well as description
    - best estimator of the parameter

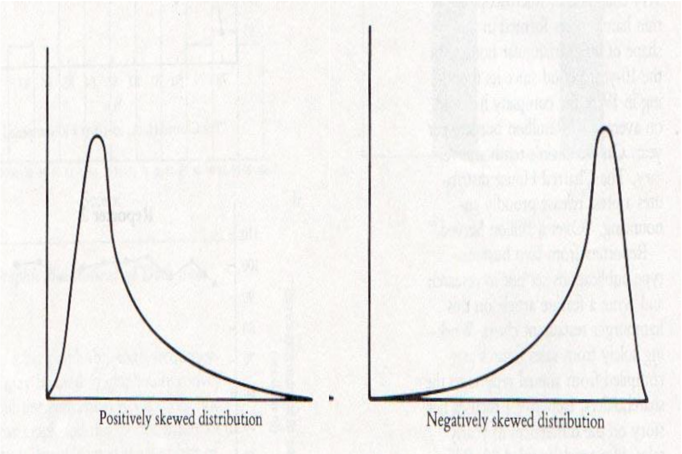


Basic measures (2)

- ## Skewness *(tails)*

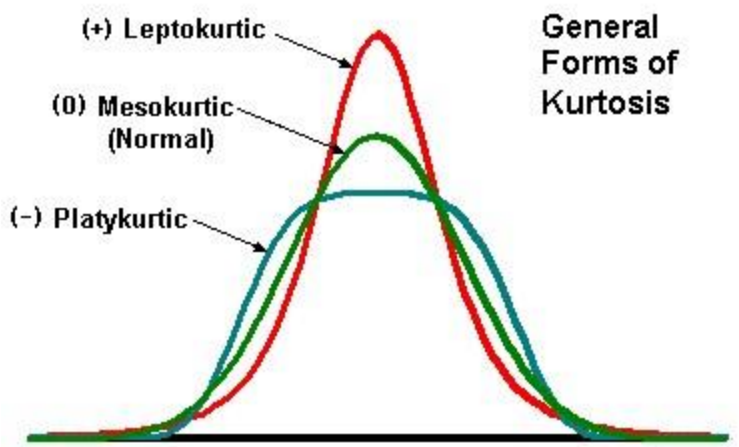
- Skewness is a measure of the asymmetry of the probability distribution
  - $\alpha_3 = \frac{E[(X-\mu)^3]}{\sigma^3} = \frac{\mu_3}{\sigma_3}$
    - Right skew - $\alpha_3 > 0$
    - Left skew - $\alpha_3 < 0$
    - Symmetric - $\alpha_3 = 0$





- ## Kurtosis *(shoulders, heavy tail)*

- Kurtosis is the degree of peakedness of a distribution relative to a normal distribution
  - $\alpha_4 = \frac{E[(X-\mu)^4]}{\sigma^4} - 3 = \frac{\mu_4}{\sigma_4} - 3$
    - A normal distribution is a *mesokurtic* distribution
    - A pure *leptokurtic* distribution has a higher peak than the normal distribution and has heavier tails.
    - A pure *platykurtic* distribution has a lower peak than a normal distribution and lighter tails.



Data distribution (1)

Normal (Gaussian) Distribution

- $X \sim N(\mu, \sigma^2)$
  - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
- Z-score
  - $z = \frac{x-\mu}{\sigma}$
  - The distance of a value from the mean measured in standard deviations

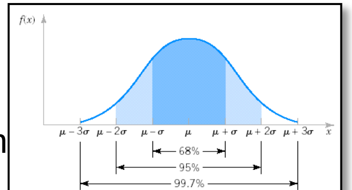
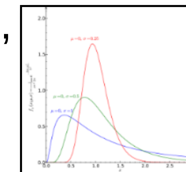


Figure 3-13 Probabilities associated with a normal distribution.

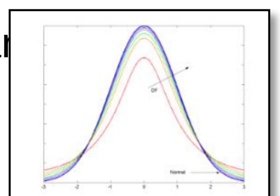
Log-normal Distribution

- $X \sim \ln N(\mu, \sigma^2)$, $x = e^z$, $z \sim N(\mu, \sigma^2)$
  - $f(x; \mu, \sigma) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}$
- Used to model a variable which is a product of positive i.i.d vars,
  - A compound return from a sequence of many trades
  - Measures of size of living tissue



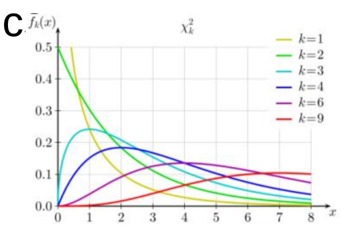
Student's t-Distribution (Gosset 1908)

- Sampling distrib. (i.i.d measures) of
  - $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$
- Approaches the Gaussian distrib. when
  - $n > 30$ or $s = \sigma$
- Used for
  - Test the diff. between two sample means
  - Inference when (μ, σ^2) are unknown



The χ^2 Distribution with *k* D.F

- $X \sim \chi_k^2$, $\chi_k^2 = \sum_{i=1}^{k} z_i^2$, $Z \sim N(0,1)$
  - $f(x; k) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^k \Gamma\left(\frac{k}{2}\right)}$
- Heavily used in statistic
  - Estimating variance
  - Goodness-of-fit test



Data distribution (2)

- ## Bernoulli Distribution
  - Bernoulli trial
    - A trial with only two possible outcomes
  - Bernoulli Distribution
    - Represents success/failure (e.g. accuracy of prediction)
    - $X \in [0,1] \sim Bernoulli(p)$
  - $f(x;p) = p^x(1-p)^x$

  $(\ \Pr[X=1] = p\ )$

- ## Binomial distribution
  - Number of success in *n* independent trials
  - $K \sim B(p,n), \quad K = \sum_{i=1}^{n} z_i\ , \ Z \sim Bernoulli(p)$
    - $f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$

  If *n* is large, then:
  $Z \sim N(np, np(1-p))$
  is a good approximation
  for $K \sim B(p,n)$

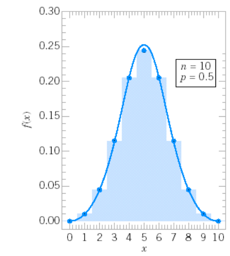
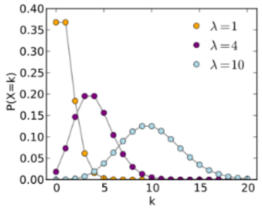


Figure 3-36 Normal approximation to the binomial distribution.

- ## Multinomial Distribution
  - Categorical Distribution
    - A trial with *k* possible outcomes
    - $f(x_1, \dots, x_k; p_1, \dots, p_k) = \prod_{i=1}^{k} p_i^{x_i}$
      where $x_i \in \{0,1\}$ and $\sum_{i=1}^{k} p_i = 1,\ p_i \in [0,1]$
  - Multinomial Distribution
    - Number of occurrences of *k* categories in *n* independent trials
    - $f(n_1, \dots, n_k; n, p_1, \dots, p_k)$
      $= \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}$
      where $n_i \in \mathbb{N}, \sum_{i=1}^{k} n_i = n$

- ## Poisson Distribution
  - Number of events occurring within a fixed time interval (or space)
    - λ , the shape param., indicates the average number of events in the given time interval
  - $K \sim Pois(\lambda),\ K \in \mathbb{N},\ \lambda > 0$
    - $f(k;\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$



  - If λ is large, then $Z \sim N(\lambda, \lambda)$ is a good approximation for $K \sim Pois(\lambda)$

Testing the data distribution

Parametric Hypothesis and general test

- Statistical tests to check the mean/variance
- Q-Q plot

Testing a general distributions

- Shapiro's test for normality
- Kolmogorov–Smirnov test
- Cramér–von Mises criterion
- Anderson–Darling test

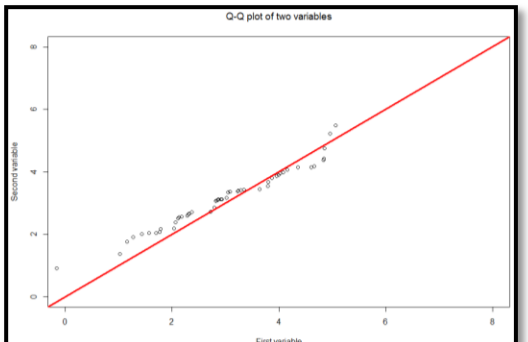
Testing the data distribution

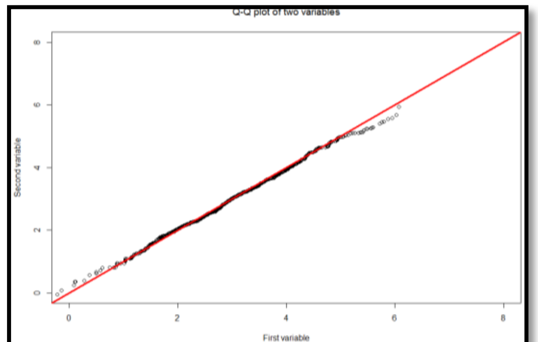
Data comparisons you are making	Data are normally distributed	Data are not normally-distributed, or are ranks or scores	Data are Binomial (Possess 2 possible values)
Compare one set of data to a hypothetical value	One-sample t-test	Wilcoxon test	χ^2 test
Compare two sets of independently-collected (unpaired) data	Unpaired t-test	Mann-Whitney test	χ^2 test or Fisher test
Compare two sets of data from the same subjects under different circumstances (paired)	Paired t-test	Wilcoxon test	McNemar's test
Compare three or more sets of data	One-way ANOVA	Kruskal-Wallis test	χ^2 test
Look for a relationship between two variables	Pearson Correlation coefficient	Spearman correlation coefficient	Contingency Correlation coefficients
Look for a linear relationship between two variables	Linear regression	Nonparametric linear regression	Simple logistic regression
Look for a non-linear relationship between two variables	Non-linear regression	Nonparametric non-linear regression	

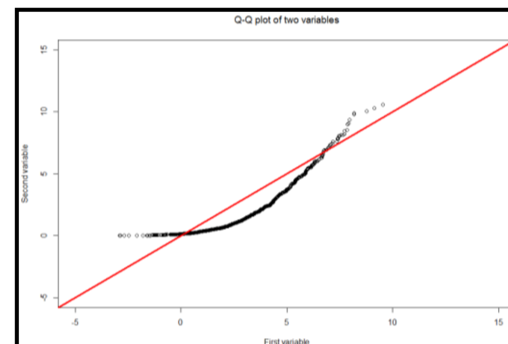
Let's see some examples how to run these tests

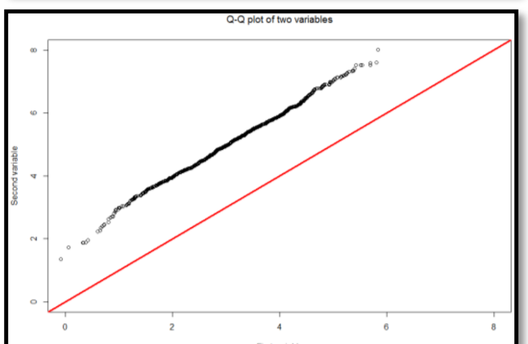
Q-Q plot

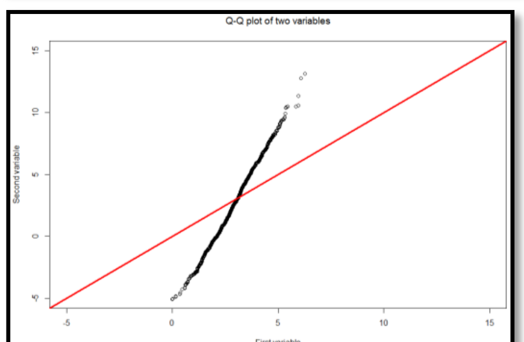
- A plot of the quantiles of the first data set against the quantiles of the second data set
- Data sets sizes don’t have to be equal
- The **greater** the departure from the 45 deg. reference line, the **greater** the evidence for the conclusion that the two data sets have come from populations with **different** distributions











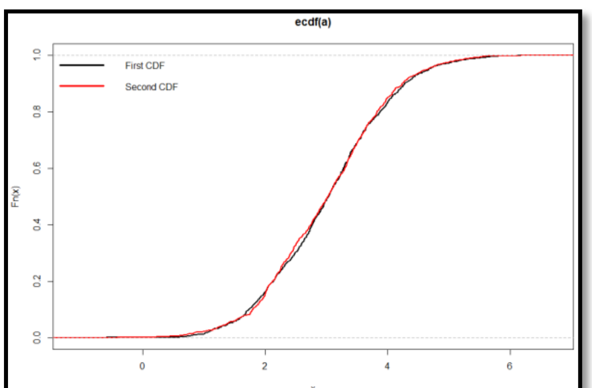
Kolmogorov–Smirnov test

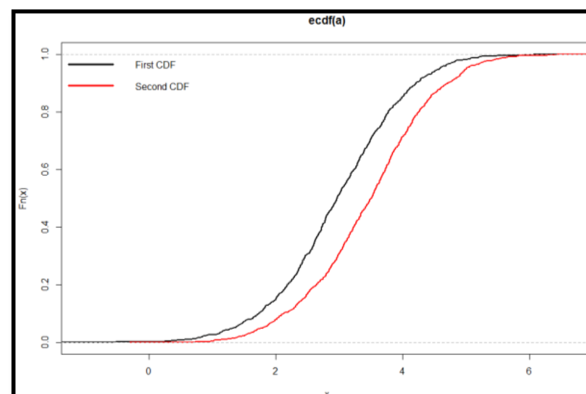
- A non-parametric test for the **equality** of continuous, one-dimensional probability distribution
- Can be applied to test a dataset distribution against a **known distribution** OR against **another dataset distribution**

H_0: The data follow a specified distribution

H_1: The data does not follow a specified distribution

- The K-S statistics is defined as: $\boxed{D_n = \sup_x |F_n(x) - F(x)|}$
- Let's have an example in R





Agenda

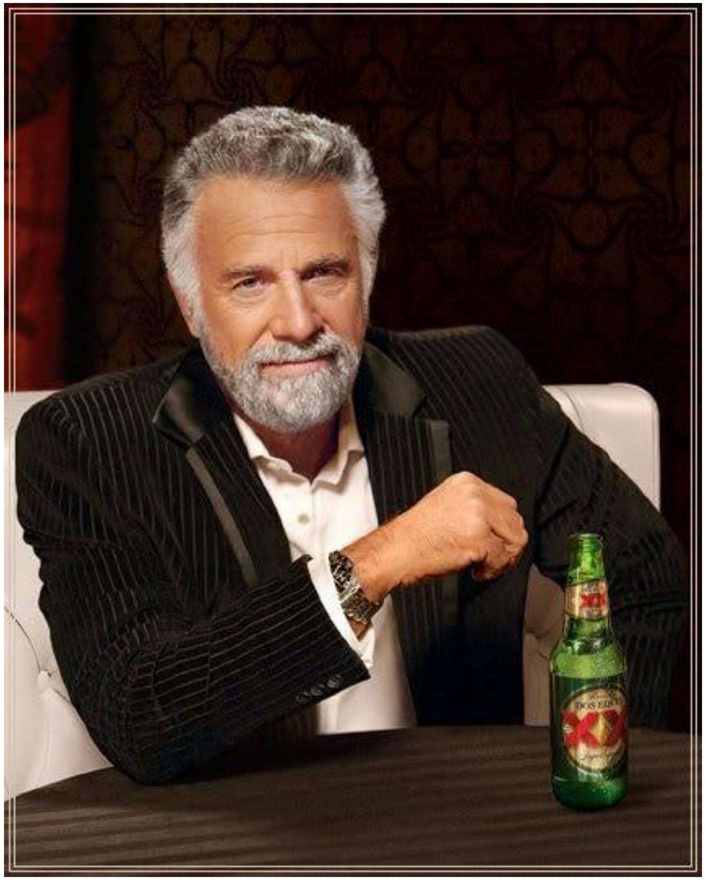
1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. **Missing values**
7. Outliers
8. Normalization & Transformation
9. Discretization
10. Unbalanced data

Missing values handling (1)

- We don't always need to handle missing value

- But when we do…

- Any ideas?



Missing values handling (2)

- Ignore the entire tuple/feature

	Price	Country	Reliability	Mileage	Type	Weight	Disp.	HP
~~Hyundai Sonata 4~~	~~9999~~	~~Korea~~	~~NA~~	~~23~~	~~Medium~~	~~2885~~	~~143~~	~~110~~
Mazda 929 V6	23300	Japan	5	21	Medium	3480	180	158
~~Nissan Maxima V6~~	~~17899~~	~~Japan~~	~~5~~	~~22~~	~~NA~~	~~3200~~	~~180~~	~~160~~
Oldsmobile Cutlass Ciera 4	13150	USA	2	21	Medium	2765	151	110
~~Oldsmobile Cutlass Supreme V6~~	~~14495~~	~~NA~~	~~1~~	~~21~~	~~Medium~~	~~3220~~	~~189~~	~~135~~
Toyota Cressida 6	21498	Japan	3	23	Medium	3480	180	190
Buick Le Sabre V6	16145	USA	3	23	Large	3325	231	165
Chevrolet Caprice V8	14525	USA	1	18	Large	3855	305	170
Ford LTD Crown Victoria V8	17257	USA	3	20	Large	3850	302	150
~~Chevrolet Lumina APV V6~~	~~13995~~	~~USA~~	~~NA~~	~~18~~	~~Van~~	~~3195~~	~~151~~	~~110~~
Dodge Grand Caravan V6	15395	USA	3	18	Van	3735	202	150

  - Simple
  - Reduces statistical power, estimation might be biased if data is missing on purpose.

Missing values handling (3)

- Analyze only cases in which the relevant variables are present (Pairwise deletion)

	Price	Country	Reliability	Mileage	Type	Weight	Disp.	HP
Hyundai Sonata 4	9999	Korea	~~NA~~	23	Medium	2885	143	110
Mazda 929 V6	23300	Japan	5	21	Medium	3480	180	158
Nissan Maxima V6	17899	Japan	5	22	~~NA~~	3200	180	160
Oldsmobile Cutlass Ciera 4	13150	USA	2	21	Medium	2765	151	110
Oldsmobile Cutlass Supreme V6	14495	~~NA~~	1	21	Medium	3220	189	135
Toyota Cressida 6	21498	Japan	3	23	Medium	3480	180	190
Buick Le Sabre V6	16145	USA	3	23	Large	3325	231	165
Chevrolet Caprice V8	14525	USA	1	18	Large	3855	305	170
Ford LTD Crown Victoria V8	17257	USA	3	20	Large	3850	302	150
Chevrolet Lumina APV V6	13995	USA	~~NA~~	18	Van	3195	151	110
Dodge Grand Caravan V6	15395	USA	3	18	Van	3735	202	150

  - Uses all possible information with each analysis

Missing values handling (4)

- Use attribute **mean**, **median** or **mode** to complete the missing data

	Price	Country	Reliability	Mileage	Type	Weight	Disp.	HP
Hyundai Sonata 4	9999	Korea	NA	23	Medium	2885	143	110
Mazda 929 V6	23300	Japan	5	21	Medium	3480	180	158
Nissan Maxima V6	17899	Japan	5	22	NA	3200	180	160
Oldsmobile Cutlass Ciera 4	13150	USA	2	21	Medium	2765	151	110
Oldsmobile Cutlass Supreme V6	14495	NA	1	21	Medium	3220	189	135
Toyota Cressida 6	21498	Japan	3	23	Medium	3480	180	190
Buick Le Sabre V6	16145	USA	3	23	Large	3325	231	165
Chevrolet Caprice V8	14525	USA	1	18	Large	3855	305	170
Ford LTD Crown Victoria V8	17257	USA	3	20	Large	3850	302	150
Chevrolet Lumina APV V6	13995	USA	NA	18	Van	3195	151	110
Dodge Grand Caravan V6	15395	USA	3	18	Van	3735	202	150

Mean (Reliability): (5+5+2+1+3+3+1+3+3)/9 = **2.88**

Median (Reliability): 1 1 2 3 **3** 3 3 5 5

Mode (Country): **USA = 6**, Japan = 3, Korea = 1.

Missing values handling(5)

- Use attribute mean, median or mode to complete the missing data – **restricted to a cla[ss]**

	Price	Country	Reliability	Mileage	Type	Weight	Disp.	HP	Class
Hyundai Sonata 4	9999	Korea	NA	23	Medium	2885	143	110	A
Mazda 929 V6	23300	Japan	5	21	Medium	3480	180	158	A
Nissan Maxima V6	17899	Japan	5	22	NA	3200	180	160	A
Oldsmobile Cutlass Ciera 4	13150	USA	2	21	Medium	2765	151	110	A
Oldsmobile Cutlass Supreme V6	14495	NA	1	21	Medium	3220	189	135	B
Toyota Cressida 6	21498	Japan	3	23	Medium	3480	180	190	B
Buick Le Sabre V6	16145	USA	3	23	Large	3325	231	165	B
Chevrolet Caprice V8	14525	USA	1	18	Large	3855	305	170	B
Ford LTD Crown Victoria V8	17257	USA	3	20	Large	3850	302	150	C
Chevrolet Lumina APV V6	13995	USA	NA	18	Van	3195	151	110	C
Dodge Grand Caravan V6	15395	USA	3	18	Van	3735	202	150	C

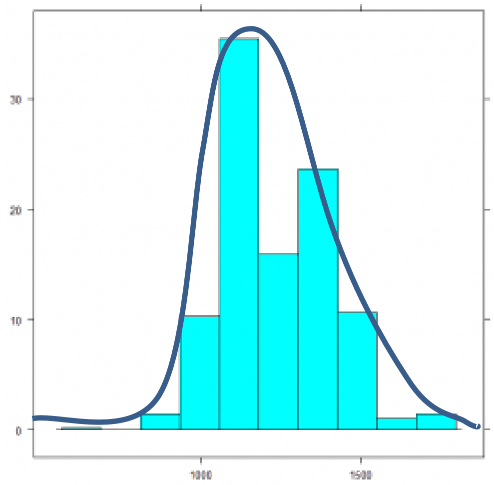
Hyundai.**Mean** (Reliability): (5+5+2)/3 = **4**

Hyundai.**Median** (Reliability): 2 **5** 5

Oldsmobile cutlass supreme.**Mode** (Country): **USA = 2**, Japan = 1

Missing values handling (6)

- Sampling
  - If distribution is known, sample from it
  - Else, sample from all possible values

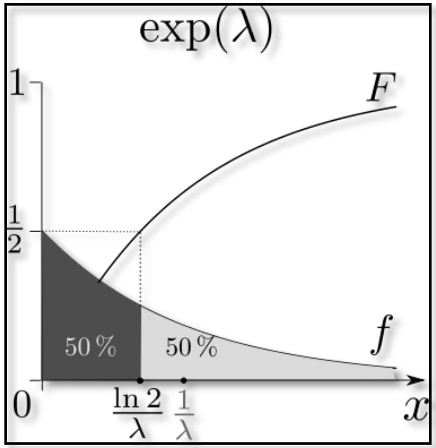


- Sampling from related class (as seen in previous slide)

Missing values handling (7)

- Sampling (cont.)
  - So – how does the sampling “algorithm” works?
    I. Generate random number c=rand() (uniform [0,1])
    II. Find the cumulative distribution function (F(b)) (remember, $0 = F(-\infty) \leq F(b) \leq F(\infty) = 1$)
    III. Calc $b=F^{-1}(c)$

    E.g. exponential distribution $y=F(x)=1-e^{-\lambda x}$
    $x=\log(1-y)/(-\lambda)$



  - Same with discrete values (staircase function)

Missing values handling (8)

- Use global closest fit to K nearest neighbors (take the value from the closest tuple.

	Price	Country	Reliability	Mileage	Type	Weight	Disp.	HP
Hyundai Sonata 4	9999	Korea	NA	23	Medium	2885	143	110
Mazda 929 V6	23300	Japan	5	21	Medium	3480	180	158
Nissan Maxima V6	17899	Japan	5	22	NA	3200	180	160
Oldsmobile Cutlass Ciera 4	13150	USA	2	21	Medium	2765	151	110
Oldsmobile Cutlass Supreme V6	14495	NA	1	21	Medium	3220	189	135
Toyota Cressida 6	21498	Japan	3	23	Medium	3480	180	190
Buick Le Sabre V6	16145	USA	3	23	Large	3325	231	165
Chevrolet Caprice V8	14525	USA	1	18	Large	3855	305	170
Ford LTD Crown Victoria V8	17257	USA	3	20	Large	3850	302	150
Chevrolet Lumina APV V6	13995	USA	NA	18	Van	3195	151	110
Dodge Grand Caravan V6	15395	USA	3	18	Van	3735	202	150

- If K > 1, you can use either mean, median, mode or sampling to select the best fit.

Missing values handling (9)

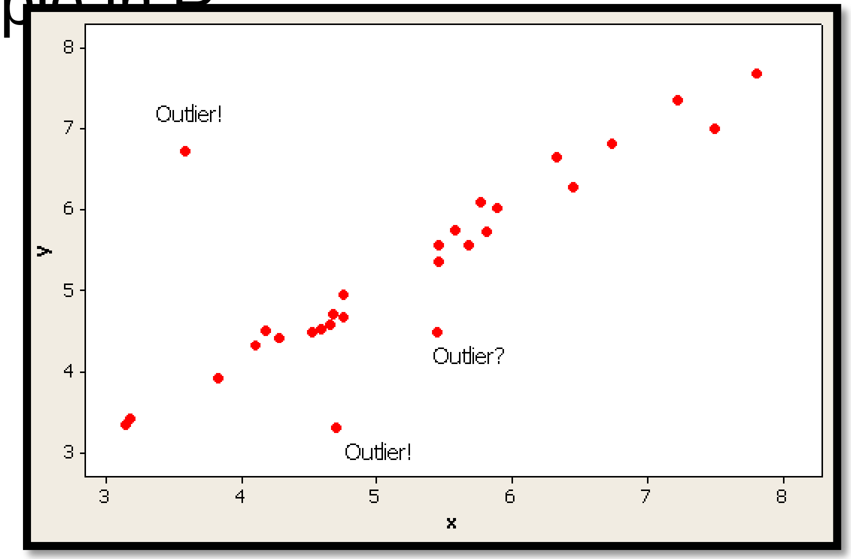
- EM (Expectation-Maximization) algorithm
  - Replace each missing value by an estimate (conditional expectation)
  - Then estimate the parameters (data distribution parameters) using the new “complete data”
  - Continue until converged…

Agenda

1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. **Outliers**
8. Normalization & Transformation
9. Discretization
10. Unbalanced data

Outliers (1)

- Definition (Wikipedia): "An observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error"
- Lets see an introduction example in R



Outliers (2)

- Identifying observation as an outlier:

  a. Distance based Methods (e.g. +- 3*SD)

  b. Statistical Methods

- Formal outlier tests
  - Differ in their distributional model
    - Usually assume approximately normal
    - Univariate VS Multivariate
  - A single outlier VS multiple outliers tests

- OK – what should we do with these outliers??

Outliers - Univariate (3)

Grubbs' Test (outlier test for normal univariate data)

- Test for a *single* outlier
  - H_0: There is no outlier in data
  - H_A: There is one outlier
- Grubbs' test statistic
  - The largest absolute deviation from the sample mean in units of the sample standard deviation *s*

$$G = \frac{\max_i |X_i - \bar{X}|}{s}$$

- Critical region for significance level α
  - Reject H_0 (the hypothesis of no outliers), if

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/2N, N-2)}}{N - 2 + t^2_{(\alpha/2N, N-2)}}}$$

Outliers - Univariate (4)

Rosner Test (outlier test for normal univariate data)

- Test for *multiple* outliers by sequentially applying Grubbs' Test
  - Detect one outlier at a time, remove the outlier, and repeat
- Critical region for significance level α, at iteration i

$$\lambda_i = \frac{N-i}{\sqrt{N-i+1}} \sqrt{\frac{t^2_{(p,N-i-1)}}{N-i-1+t^2_{(p,N-i-1)}}}$$

Where $p = 1 - \alpha/2(N-i-1)$

  - All (adjusted) test statistics and critical values are being calculated up to a predetermined upper bound
  - The number of outliers is determined by the largest i such that the test statistics is larger than λ_i

Outliers - Multivariate (5)

Nearest Neighbors based approaches

- Compute the distance between every pair of data points
- There are various ways to define outliers

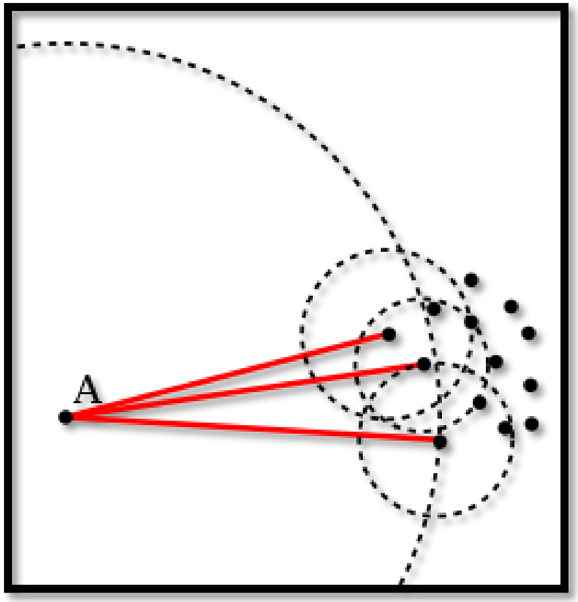
1. Density
   - Data points for which there are fewer than *p* neighbors within a distance *D*

2. Distance
   - The top n data points whose distance to the k^{th} nearest neighbor are the greatest
   - The top n data points whose average distance to the k nearest neighbors are the greatest

3. Local Outlier Factor (LOF)
   - Based on a concept of a local density, where locality is given by k nearest neighbors, whose distance is used to estimate the density
     - Compare the local density of an object to the local densities of its neighbors

4. Class Outlier Factor (COF)
   - A class restricted distance approach

Outliers - Multivariate (6)



Agenda

1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. Outliers
8. **Normalization & Transformation**
9. Discretization
10. Unbalanced data

Normalization (1)

- AKA Feature Scaling
- Why do we need normalize the data?
  - Easy comparison of values
  - In some algorithms, objective functions will not work properly (or quick) without it
- Example:
  - Predict the cost of the house, giving it's size (squared meters) and the # of bedrooms

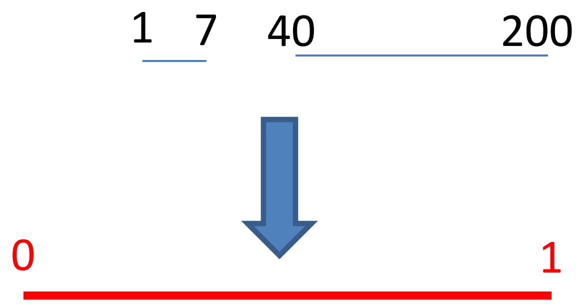
40 200

1 7

Normalization (2)

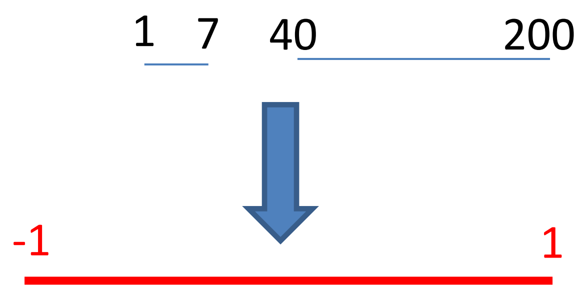
- Min/Max normalization to [0,1]

$$X_{i,\,0\text{ to }1} = \frac{X_i - X_{Min}}{X_{Max} - X_{Min}}$$



- Min/Max normalization to [-1,1] (if we want 0 to be the central point)

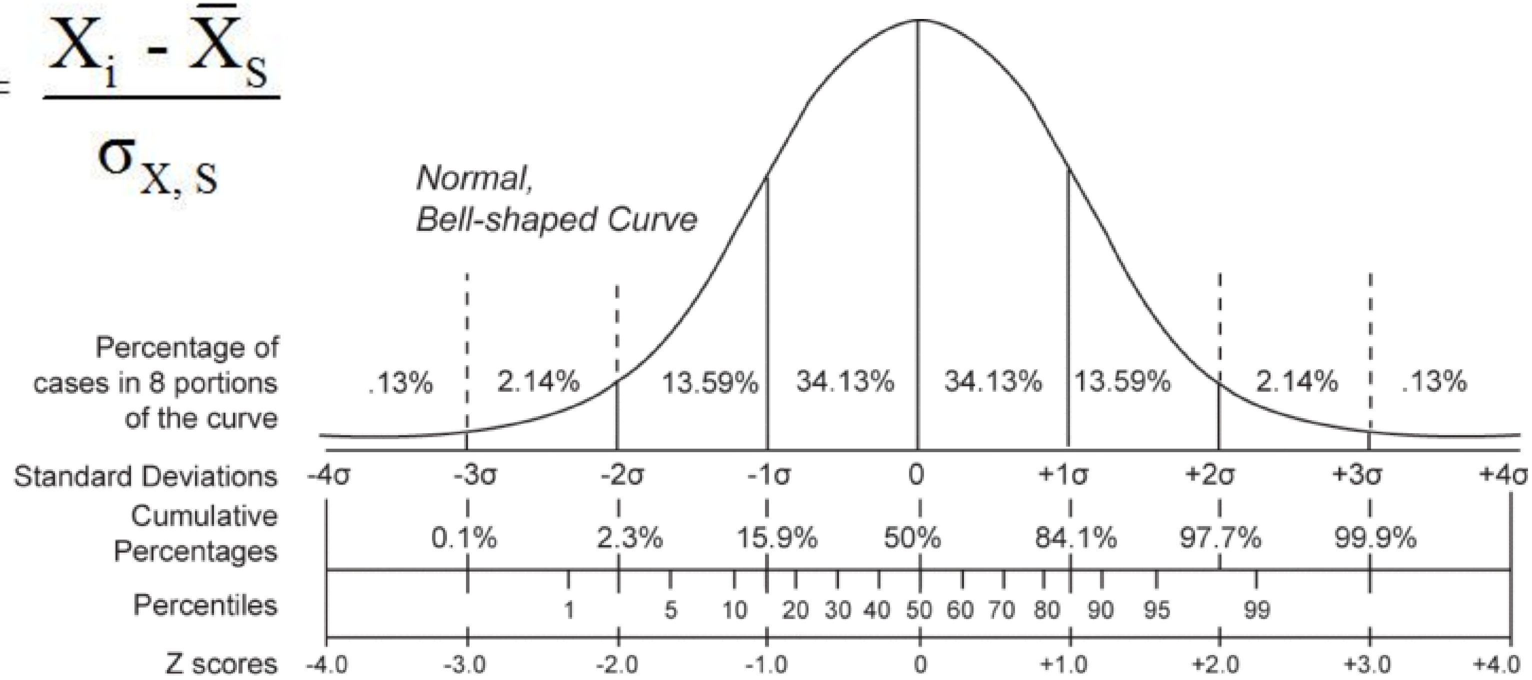
$$X_{i,\,-1\text{ to }1} = \frac{2X_i - X_{Min} - X_{Max}}{X_{Max} - X_{Min}}$$



Normalization (3)

- Standardization (Z – normalization).
  - using mean and standard deviation. Fits normalized-like data.

$$X_{i,\,1\sigma} = \frac{X_i - \bar{X}_S}{\sigma_{X,\,S}}$$



Normalization (4)

- Log normalization
  - Used when values are ranged over several orders of magnitude.
  - $X' = a*\log_b(X)$

Normalization (5)

- But… which normalization method to use?

Transformations

- Transformation examples – log(X), 1/X, X^2 etc…. Can lead us to non-linear models
- Let's see an example

Agenda

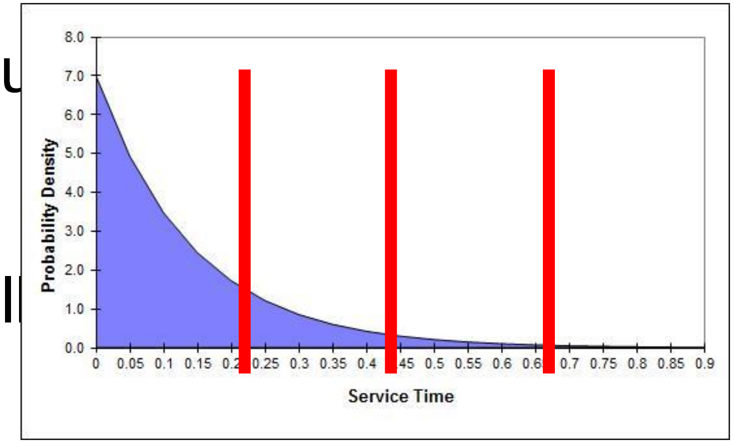
1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. Outliers
8. Normalization & Transformation
9. **Discretization**
10. Unbalanced data

Discretization (1)

- Why do we need to change the data?
  - Some models/measures can’t handle continuous values (i.e. Naïve Bayes, MI)
  - Some numeric values don’t have a meaningful numeric insights (but when taking them as discrete ones – they do have)
  - The business might have useful information to give us.

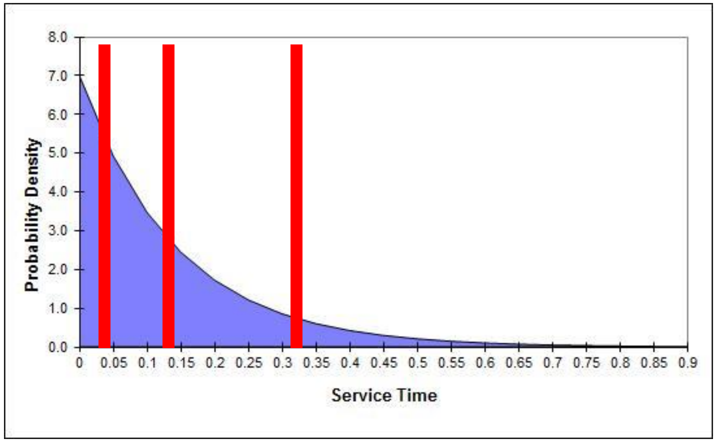
Discretization (2)

- **Equal-width** (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B –A)/N.
  - The most straightforward, but ou
    presentation
  - Skewed data is not handled well



Discretization (3)

- **Equal-depth** (frequency) partitioning
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky



Discretization (4)

- **Entropy based**
  - The entropy (or the information content) is calculated on the basis of the class label.
  - Intuitively, it finds the best split so that the bins are as pure as possible, i.e. the majority of the values in a bin correspond to having the same class label.
  - Formally, it is characterized by finding the split with the maximal information gain.

Agenda

1. Introduction
2. Data types
3. Distance measures
4. Correlation and Mutual information
5. Data distribution
6. Missing values
7. Outliers
8. Normalization & Transformation
9. Discretization
10. **Unbalanced data**

Unbalanced data (1)

- "Unbalanced" is difficult to define precisely
- Generally speaking - unequal numbers of observations in each category (usually related to classification problems)
- Usually talking about the unbalanced regarding the target data – but not always!
- Examples:
  - medical diagnosis: 90% healthy, 10% disease
  - eCommerce: 99% don’t buy, 1% buy
  - Defects in the manufacturing process
- Easily we can build an amazing model

Unbalanced data (2)

- Stratified Sampling – sampling technique in which each subpopulation (stratum) is sampled independently

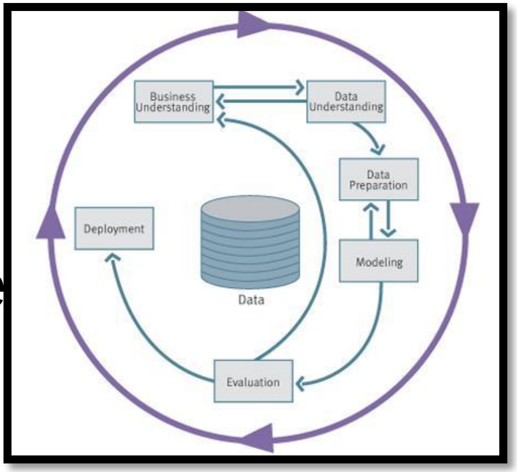
- Ensure that each class is represented with approximately equal proportions in train and test

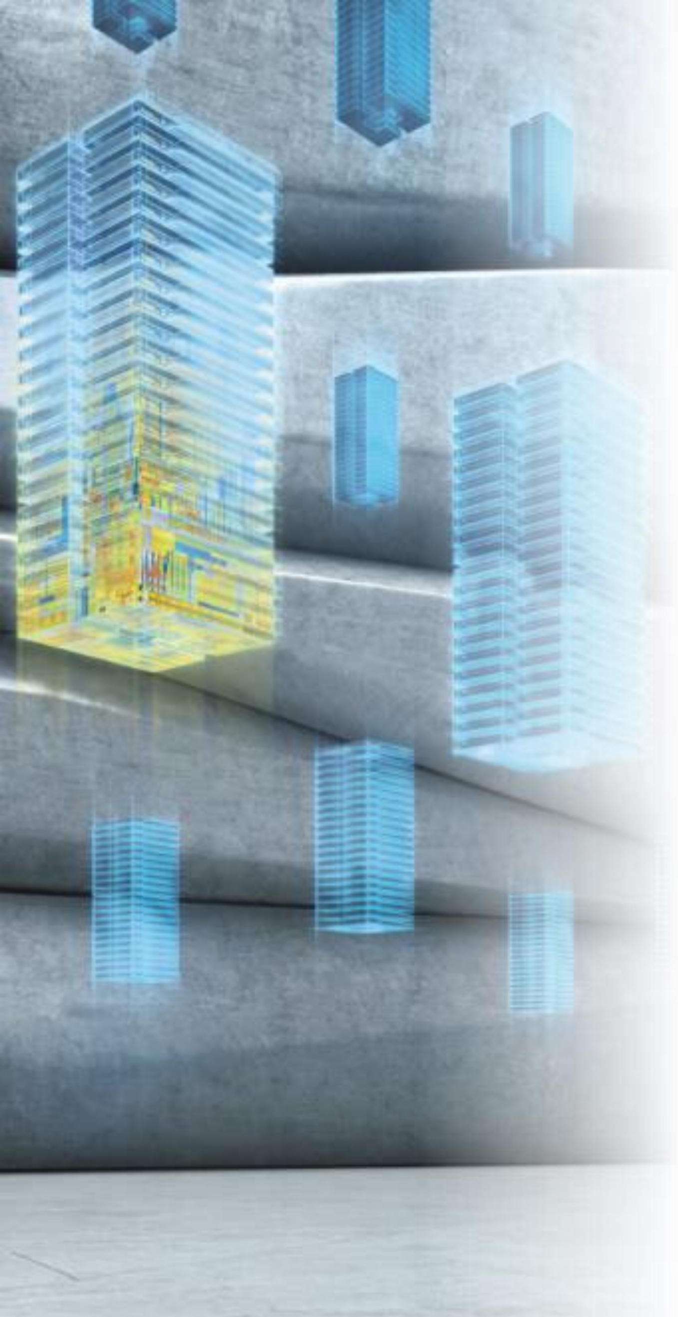
- Estimate the final results using an imbalanced held-out (test) set

- How to create a "balanced" dataset?
  1. Down-sample the large classes
     - Use when majority is very large and minority is extremely small
  2. Bootstrap the smaller classes
     - Use when minority size is large enough to safely resample
  3. Assign weights to the samples
     - A commonly used weighting scheme: is $w_c = \frac{n}{n_c}$
     - Where n_c is the size of the class c and $n = \sum_c n_c$ is the total sample size

Summary

- Topics we have covered
- How CRISP-DM is related to the session
- In practice – what is being done in real life
- Anything else?





Backup