Core concepts in data science

Week 1

What is core data analysis

Data mining is looking for patterns in data stored in databases as part of knowledge discover

Core data analysis: two main goals for knowledge enhancing

Core data analysis does not care of databases, but specific patterns in data patterns (which is about concepts and relationship between concepts) and knowledge enhancement (how: summarization by developing concepts and correlation by deriving statements of relation between concepts)

Two major formats: quantitative, and categorical

Summarization methods:

Quantitative principal component analysis

Categorical cluster analysis

Correlation methods

Quantitative regression

Categorical classifier

Visualization: highlighting and

Sample 1:

Kepler 1st and 2nd Law: Planets circle sun

3rd Law: is any relation between speed/period and distance?

Log (period) ~ log (distance)

Sample #2:

Pattern found success if compatible with existing knowledge

Advice for data analysis: find a pattern, interpret using existing knowledge, care not whether interpretation is compatible

Cluster analysis

Classical mathematical statistics: data is just a vehicle to fit and test mathematical models in the applied domain

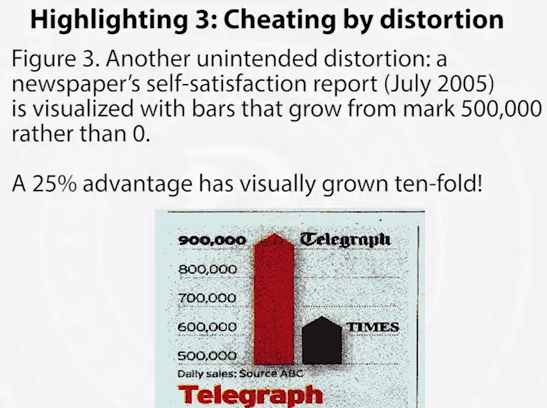
Machine learning:

Prediction rules to be built incrementally

Data mining

Adding new knowledge by finding interesting patterns in databases, which is initial stage of knowledge discover

Visualization of data is an important activity assisting data analysis by a human in many ways including A. highlighting; B. Integrating different aspects. C. Manipulating



Illustrative data

Nominal number (or catogrical number) is only meaningful when comparing them to see whether they are identical or not.

Binary scale: yes or no

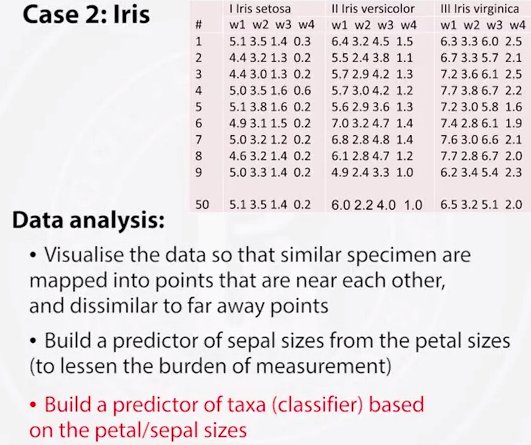
Data analysis:

How to map companies to the screen with their similarity reflected in distances between points? (summarization/visulization)

Would clustering of companies reflect the product? What features would be involved then? (summarization)

Can rules be derived to predict the product for another company, coming outside of the table (correlation)

Is there any relation between the structural features and market related features? (correlation)



Week 2

1D summarization

Histogram: Gaussian and power law

Central values and dispersion values

Data recovery view and probabilistic view

Validation with Bootstrap

Histogram:

First of all, I keep data and metadata in different files so that text do not interfere with numbers.

Data, in a file iris.dat representing iris data just as a flat file of 150 rows (entities) and 4 columns (features). Feature names are kept in a different file, iris.var.

Note:

All entries must be filled in a data table.

No missed data please

If your data do have missing entries, you should cure your data of missing entries before hand. (There are some methods for working with missing entries)

With this arrangement, I point MATLAB running memory to the folder in which my data files are stored and load the dataset:

Iris=load(‘iris.dat’)

Then I get the feature I mentioned (sepal width is in 2nd column)

Sepal\_Width=iris(:,2);

At last, generate the histogram using this:

hist(Sepal\_Width)

So what does the command hist() does?

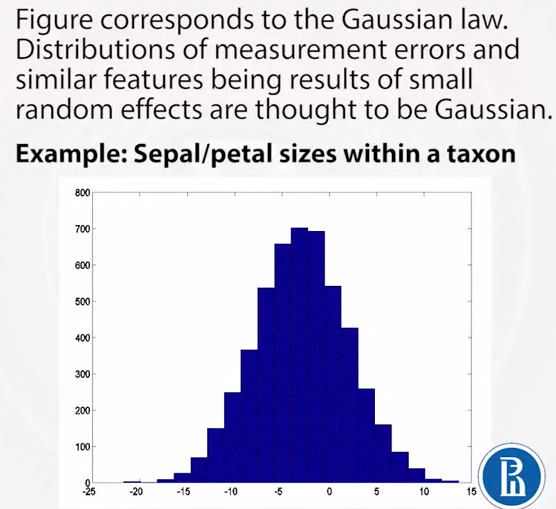
Take interval between min and max and divides it into 10 intervals of equal length (bins)

For each bin, counts the number of entities whose values fall in the bin sequency.

Draws a bar over each bin; its height reflects the frequency

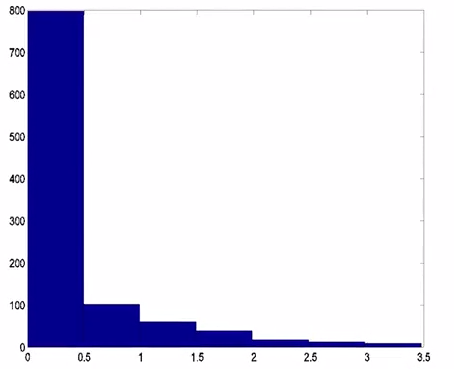
Mode

Gaussian and power law histograms and associated density functions



Power law:

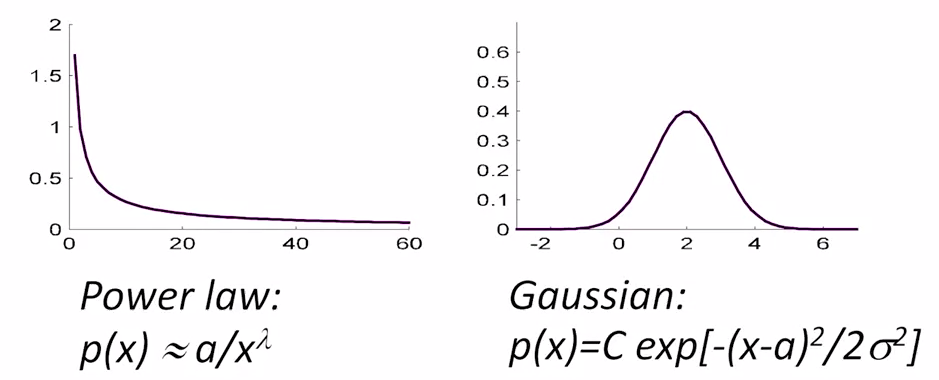
Distributions of human efforts, income distribution, word distribution in a long text, citations in science, web-site popularity



Matthew effect

A win makes it more likely a next win, whereas a loss makes it more likely a next loss

Density function is a model of histogram in the mathematical probability theory



In the Gaussian: the sums of small random variables tend to converge to a Gaussian

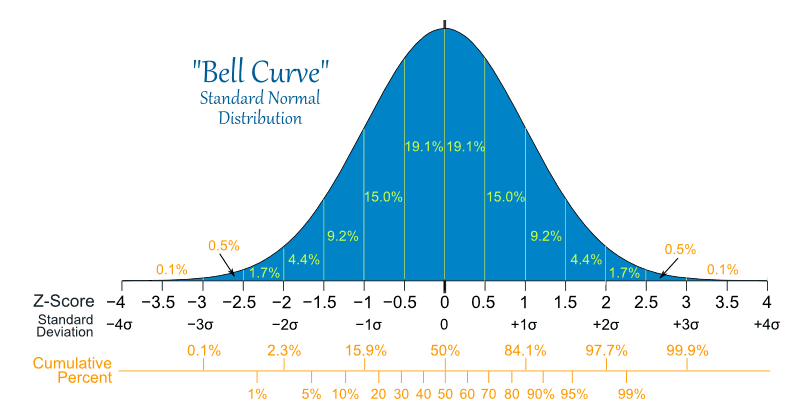
Has two parameters, a – center of the Gaussian bell, σ^2 bell’s width (variance)

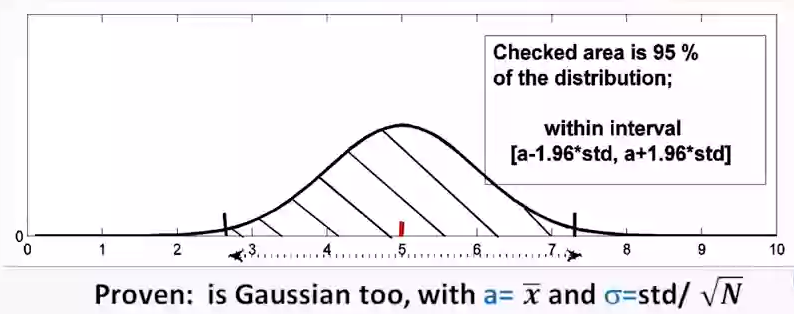
Histograms and associated density functions: clash of civilizations

|  |  |  |
| --- | --- | --- |
|  | Probabilistic mathematical statistics | Data analysis |
| Feature | a random variable (density function) | A column in data table (multidimensional vector) at which a density function may be associated if useful; or not, if not |
| Data | random sample | A table (matrix) |
| Goal of analysis | to estimate parameters of features and functions over them | To enhance knowledge via summarization or correlation |

Further summarization: center and spread

|  |  |  |
| --- | --- | --- |
|  | Spread | Center |
|  | Range | Midrange |
|  | Standard deviation | Mean |
|  |  | Median |



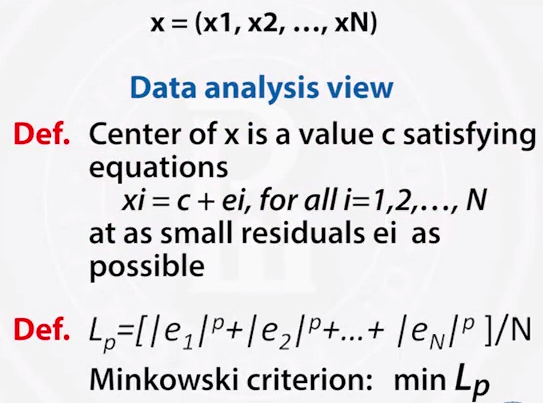


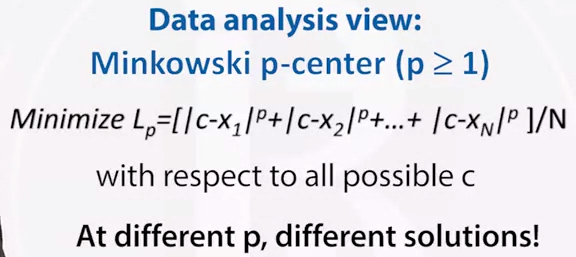
In matlab,

Sort(x)

Median(x)

Substantiate center/spreads





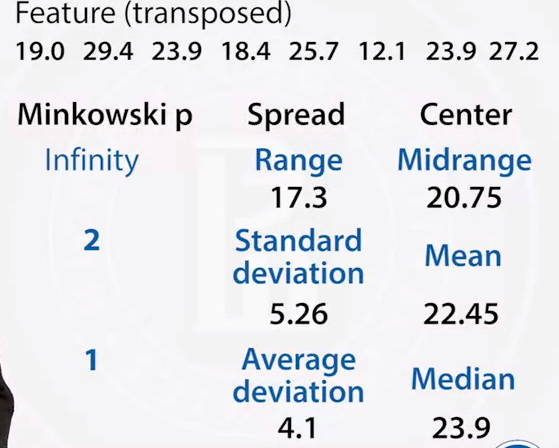
when p = 2, then Lp is quadratic. First order minimum condition can be applied, it leads to optimal c = Mean(x)

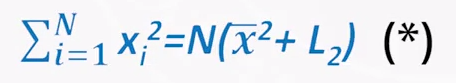
At this c, L2 is the square of the standard deviation

The minimum L2 is referred to as the variance, and its square root, as the standard deviation.

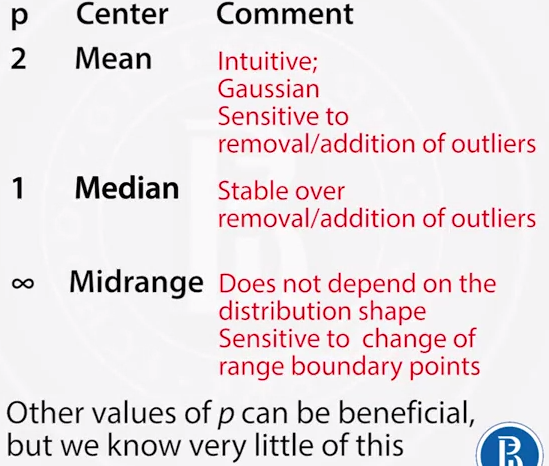
When p = 1, then the minimum of L1 is reached at c being the median, which should be used as the corresponding spread.

When p = Inf, then c tends to midrange



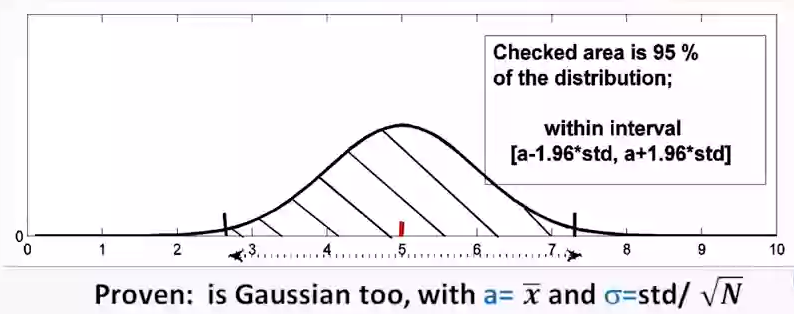


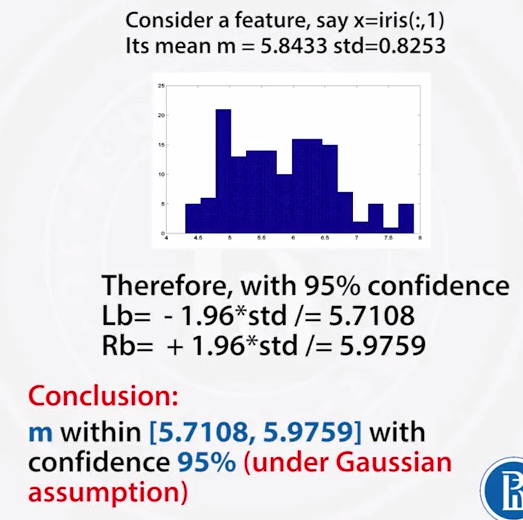
Probabilistic perspective center/spread



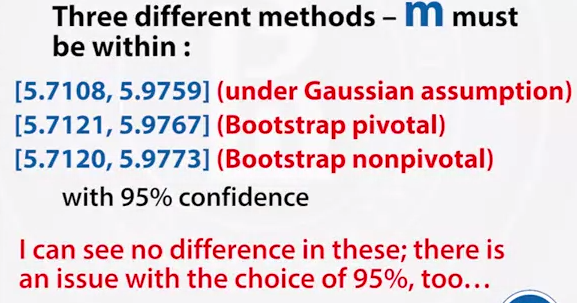
Computational validation of mean using bootstrap

Plausible boundaries for mean





In MATLAB, 



Week 3

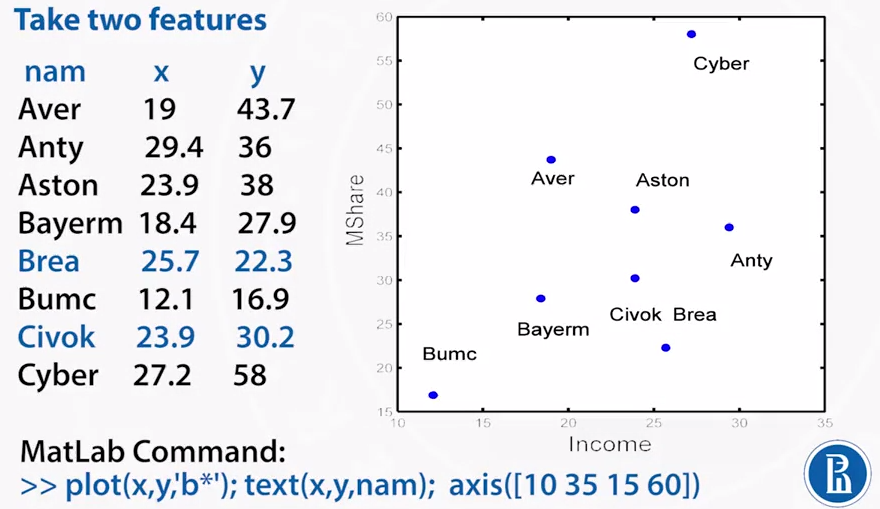
2D two quantitative features

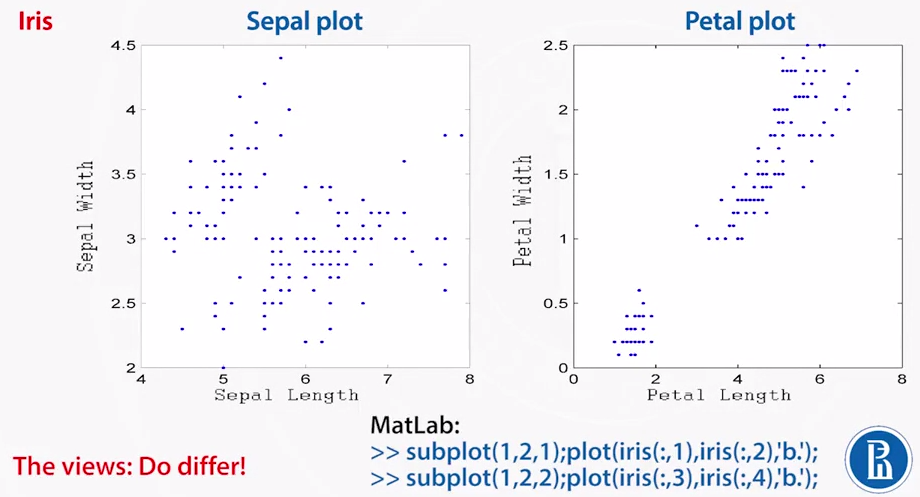
Scatterplot

Linear regression

Correlation and determinacy coefficients: properties and meaning

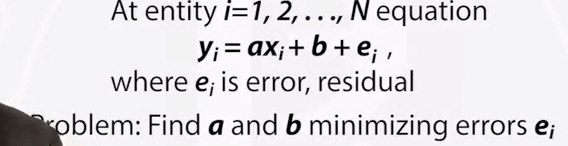
Correlation and regression: case studies

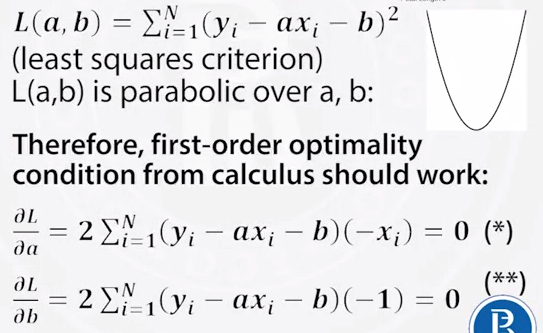




2D linear regression

Regression: the coefficients of regression against or on another variable





Correlation and determinacy coefficients: properties and meaning

Properties:

The determinacy coefficient rho^2 is within interval [0,1]; the correlation coefficient rho, within [-1,+1]

Coefficient rho is 1 or -1 iff regression equation y=ax+b is true for every i = 1,2,…,N with no errors.

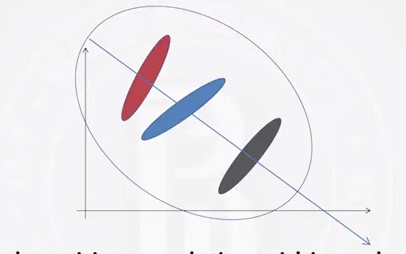
Correlation cefficient rho is 0 iff the slope a = 0

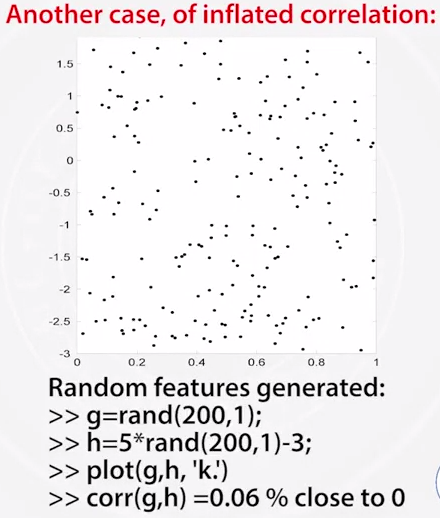
The sign of rho is the sign of the slope a; therefore, x and y are related positively if rho > 0, and negatively, if rho < 0.

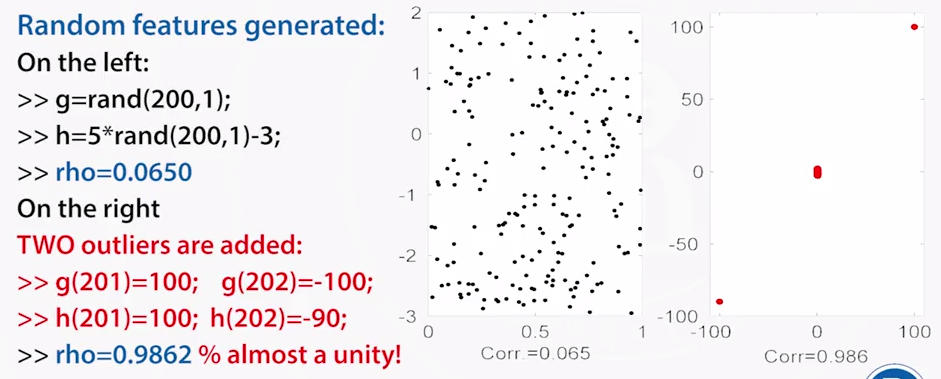
Trick:

A high positive correlation within each group, red, blue, grey; yet a negative correlation overall

Instances of data manipulation, sometimes unintentionally, should make great politicians to say of three levels of lie: a lie, damned lie and statistics.







Summary: scatter plot:

just a Cartesian representation 2D

linear regression: a convenient format to summarize two features

correlation and determinacy coefficients: these are due the linearity and least-squares criterion; rho^2 scoring the extent of y-variance taken into account; rho expressing, vaguely, extent of linear relation between x and y

correlation and regression: usefrul, but be aware of “lies”

week 4

2D features, both nominal

Contents:

Contingency table

Conditional probability and Quetelet index for capturing correlation

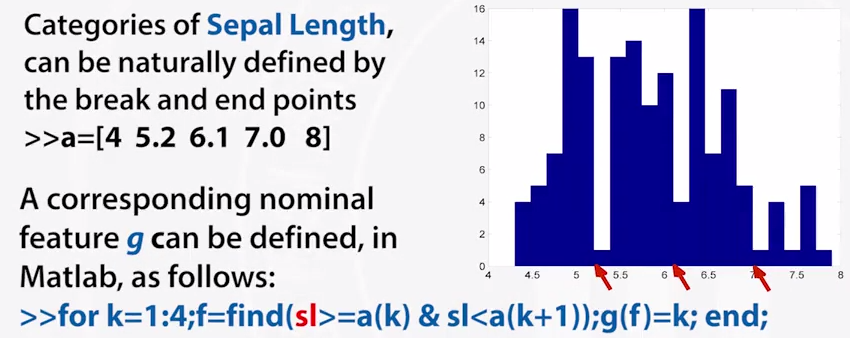
Statistical independence and pearson chi-squared coefficient

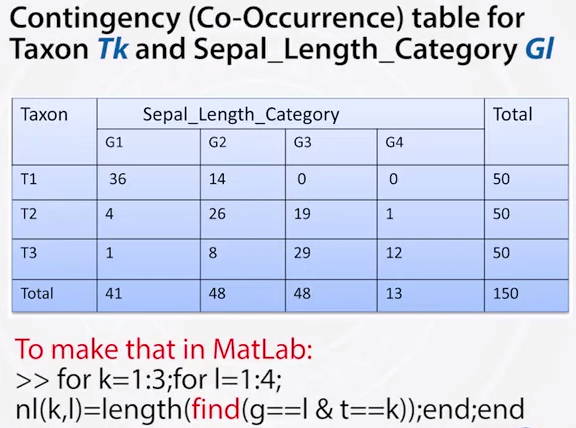
Pearson chi-squared as the summary quetelet index

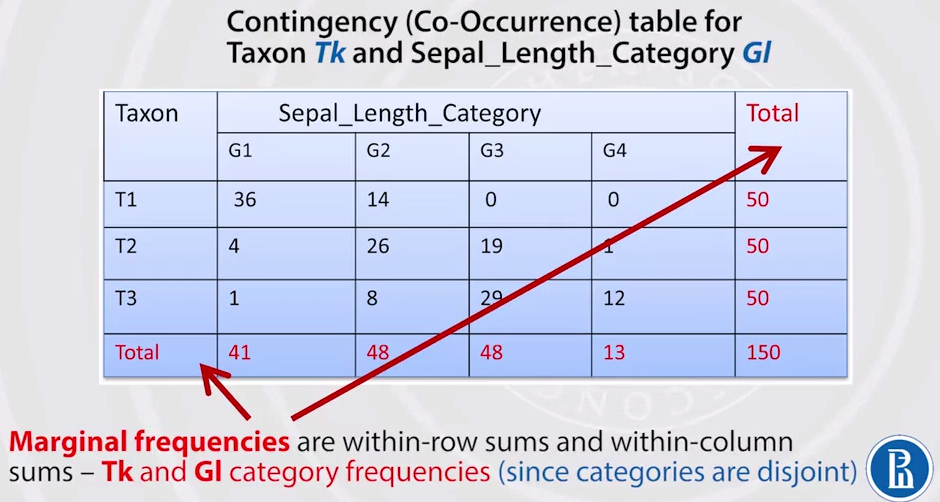
Visualization of association patterns

Contingency table

The minima define natural boundaries between categories to be in the histogram.

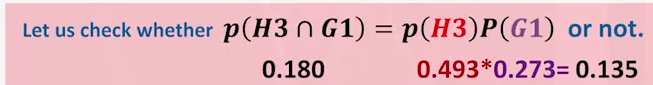


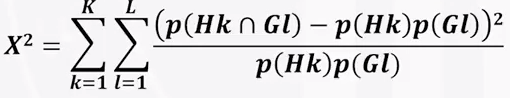




Chi-squared coefficient: a different development

Two features are independent if and only if





Why chi^2?

Pearson: the density function of random variable NX^2 tends to distribution chi^2 with (K-1)(L-1) degrees of freedom, under the hypothesis that the features are independent in the population, and entity sampling has been done randomly and independently.

Note:

In many textbooks, it is the value of NX^2 which is denoted by X^2, while denoting our X^2 by phi^2.

This theorem allows for probabilistic testing of the hypothesis that the features are independent.

The chi-squared is not recommended as a measure of association between features.

(How to calculate chi-squared ? Because I got a wrong answer)

Pearson’s chi-squared is Q.

2D part 5: Visualization of Q=X^2, useful at dull tables.

Summary of the lecture

Contingency table: bivariate distribution, co-occurrence frequencies

Conditional probability and Quetelet index for capturing correlation: cond. Probability is good at strong correlations; if those are weak, the Quetelet index can help – it scoures the relative change of probability of a category, when a crossing category is given. The summary quetelet measures the average relative change.

Statistica lindependence and pearson chi-squared coefficient: these are rather standard, pearson’s is for testing the hypothesis of statistical independence.

Pearson chi-squared as the summary Quetelet index: rather unconventionally, the chi-squared equals the summary Quetelet index and so appears to be a measure of association, in contrast to conventional wisdom. Moreover, it has something to do with conceptual relation “if … then …”

Visualization of association patterns; conventional is due to pearson indices. Yet the summary Quetelet suggests another, more natural way of visualization of the pattern of decomposition of the chi-squared over the contingency table.

WEEK 5

Correlation rules

Popular correlation structures

Bayes approach to prediction and Naïve Bayes classifier

Assigning articles to categories: Naïve Bayes algorithm.

Bag-of-words text model.

Decision tree and splitting criteria.

Metrics of accuracy

Typically, to analyze relations between different aspects, all the features are divided in two parts:

Input features X

And target features

Then a rule F is sought to establish a relation between the inpur and target features, most usefully such as U=F(X).

Popular types of rule U=F(x)

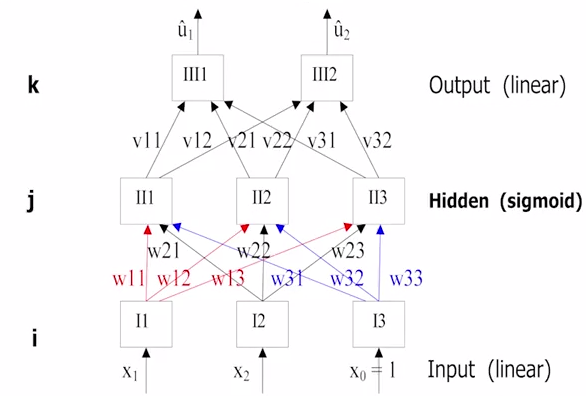
* Econometric structural model
* Hidden Markov chain

Example: use seaweed to predict the weather, using observable states from a simple probabilistic process (markov chain) to predict hidden states.

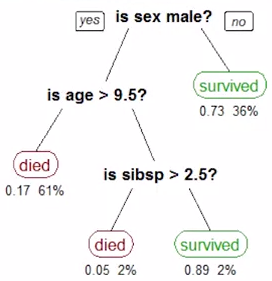
* Bayes network

Example: network of variables describing a class of computer games, with a dependence graph imposed on them. The outcome probabilities depend on those of others according to the graph structure.

* Neutral network



* Decision tree



The bayes approach

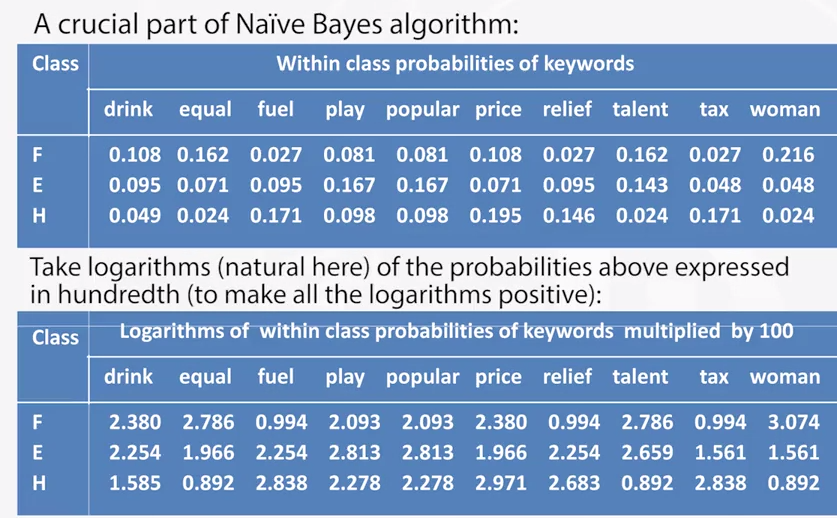
There is a probabilistic distribution over patterns. Observed entities change that from a prior one to a posterior distribution.

Naïve bayes rule:

Naïve bayes decision:

Computing probabilities of keywords within classes: Bag of words model

How: put in a bag all the key words (X1) from the upper line of the table, add all the occurrences of all key words in the category(X2). The probability of a keyword is equal to its total occurrence number plus 1, then divided by the bag size (X1 + X2).



Correlation rules. Classification trees

Classification tree building over a training set, given a target partition H.

Goal: to build a partition G maximally similar to H by sequentially splitting a cluster over a feature

Start: G consisting of one cluster, the dataset

A split is chosen as the best of the splits of all clusters available over all feature values available

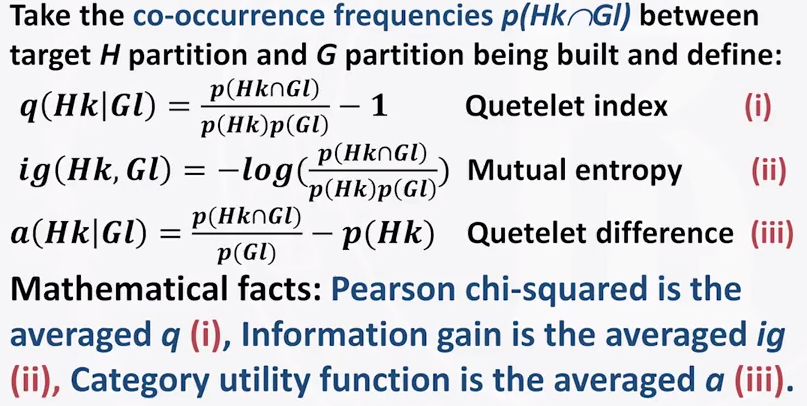
Split scoring function evaluates the similarity between the target partition H and partition G being built. Among most popular scoring functions are Pearson chi-squared (which is used in SPSS), information gain, category utility function

A unifying framework for split scoring functions:

Take contingency table between target H and G being built p(HkGI) (co-occurrence frequency).

Define a reasonable function f(p(HkGI)) evaluating the extent of correlation between clusters Hk and GI.

Score the similarity between G and H as the sum of all f(p(HkGI)) weighted by their frequencies p(HkGI)



Metrices of accuracy

False positive

False negative

Week 5

Principal component analysis: method and model

Contents

Theoretic introduction: summarization versus correlation

Matrix operations

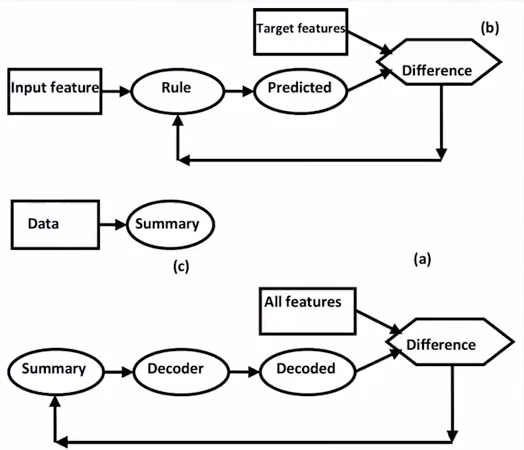
Matrix spectrum, singular value decomposition, approximation

Hidden factor model. Its solution. Principal components, loadings, contributions.

Conventional PCA critetion and method. Relation between the model-based and conventional approaches. Covariance and correlation matrix.

Principal component analysis: summarization versus correlation

1. Data recovery view of summarization (all features are target), where summarization outcome much depends on data preprocessing because the recovery extent is measured by the sum of derivations. Deviations change at changing the scale
2. Correlation problem: data recovery formulation, where target feature scaling predetermines the scale of the output
3. Conventional view of summarization



Company data 8x5 converted to the quantitative format 8x7 (standardization in order to sharpen the data structure)

1. Just centering by subtraction within column means; in order to look at feature values against a normal backdrop
2. Centering and normalization by dividing over column ranges; in order to balance feature weights;
3. Centering and normalization as in 2 plus additionally dividing the three sector columns on the right by square root of 3 to compensate for the multiple columns

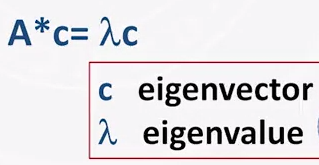
Principal component analysis: matrix algebra

Why matrix algebra is called a linear algebra

Transpose

Multiply A by B = B\*A ?

Spectrum



Geometric explanation

If c is eigenvector for A, then so is alpha\*c. to make c unique, conventionally c is considered normed, abs(c) = 1.

The number of different eigenvalues for an mxm A is not greater than m; the number of nonzero eigenvalues is A’s rank.

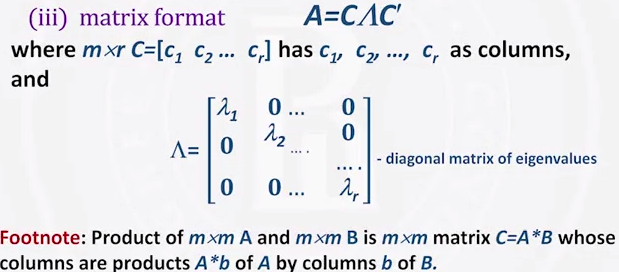
If A is symmetric all its eigenvalue are real; and the eigenvectors, mutually orthogonal.

Spectral decomposition in three equivalent

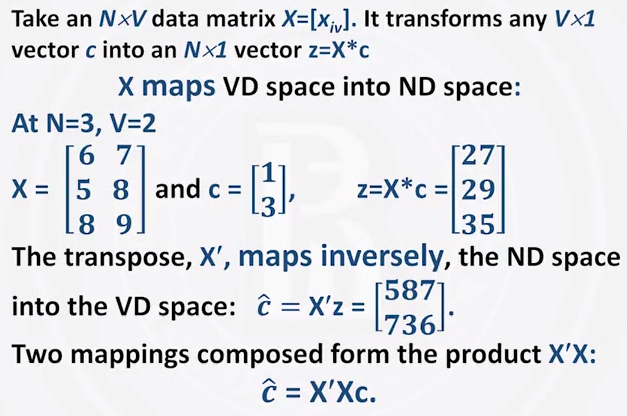








Singular value decomposition, composition



Rank is a mathematical explication of the space dimension

Hidden factor model

