**Data Mining**

**Regression**

# Looking for Interest Patterns in Data

* Classification rules:

IF buy\_time in December

AND cost > 500

AND type\_of\_item = electronics AND location = overseas

AND ...etc...

THEN possibly\_fraudulent = yes

# Data Mining Tasks

* Classification (predictive)
* Regression (predictive)
* Clustering (descriptive)
* Association rule discovery (descriptive)

# Description of OneR

For each attribute A:

For each value V of that attribute, create a rule:

* 1. count how often each class appears
  2. find the most frequent class, c
  3. make a rule "if A=V then C=c"

Calculate the error rate of this rule

Pick the attribute whose rules produce the lowest

error rate

## Summary of ID3 Algorithm

1. For each attribute, compute its entropy with

respect to the conclusion (the target attribute)

1. Select the attribute (say A) with lowest entropy
2. Divide the data into separate sets so that within a set, A has a fixed value (eg *Colour=green* in one set, *Colour=brown* in another, etc)
3. Build a tree with branches:

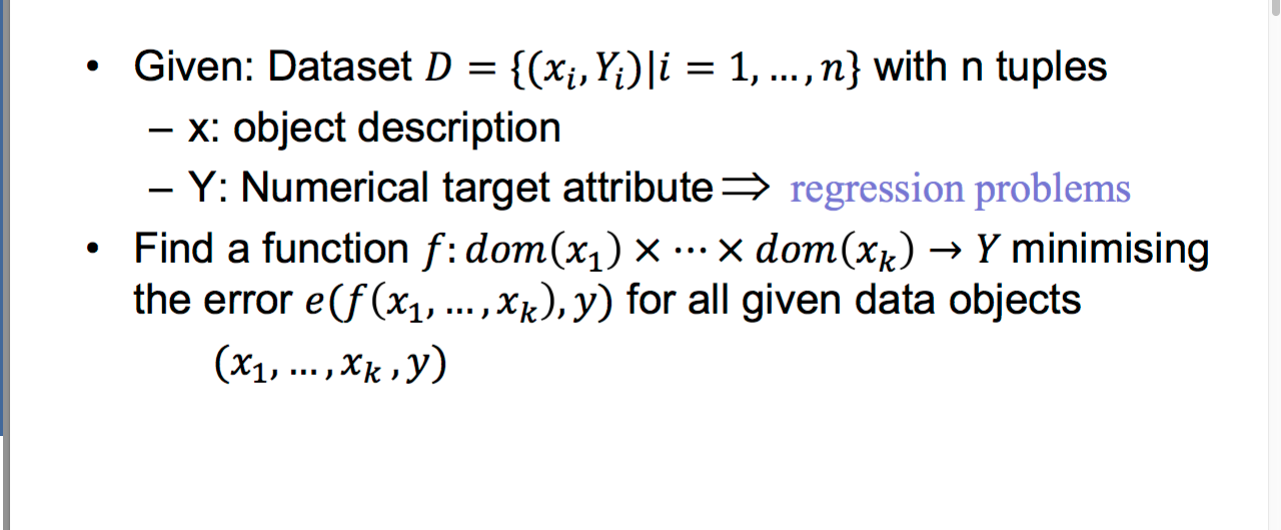
if *A=a1* then ... (subtree1) if *A=a2* then ... (subtree2)

...etc...

Summary of ID3 Algorithm – Cont’d

1. For each subtree, repeat this process from step 1
2. At each iteration, one attribute gets removed from consideration. The process stops when there are no attributes left to consider, or when all the data being considered in a subtree have the same value for the conclusion (e.g. they all say *Conclusion=safe*).

# Regression

****

Remember

* Instead of finding structure in a data set, we are now focusing on methods that find explanations for an unknown dependency within the data.
* Supervised (because we know the desired outcome)
* Descriptive (because we care about explanation)

# Regression line

* Given: A data set for two continuous attributes x

and y.

* It is assumed that there is an approximate linear

dependency between x and y:

*y* = *ax*  *b*

* Find a regression line (i.e. determine the parameters a and b) such that the line fits the data as good as possible.

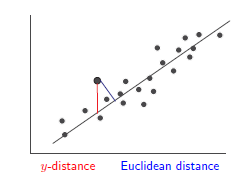
# Regression line -- Example

* Trend estimation (e.g. oil price over time)
* Epidemiology (e.g. cigarette smoking vs. lifespan )
* Finance (e.g. return on investment vs. return on all

risky assets)

* Economics (e.g. consumption vs. available income)

# Regression line



What is a good fit?

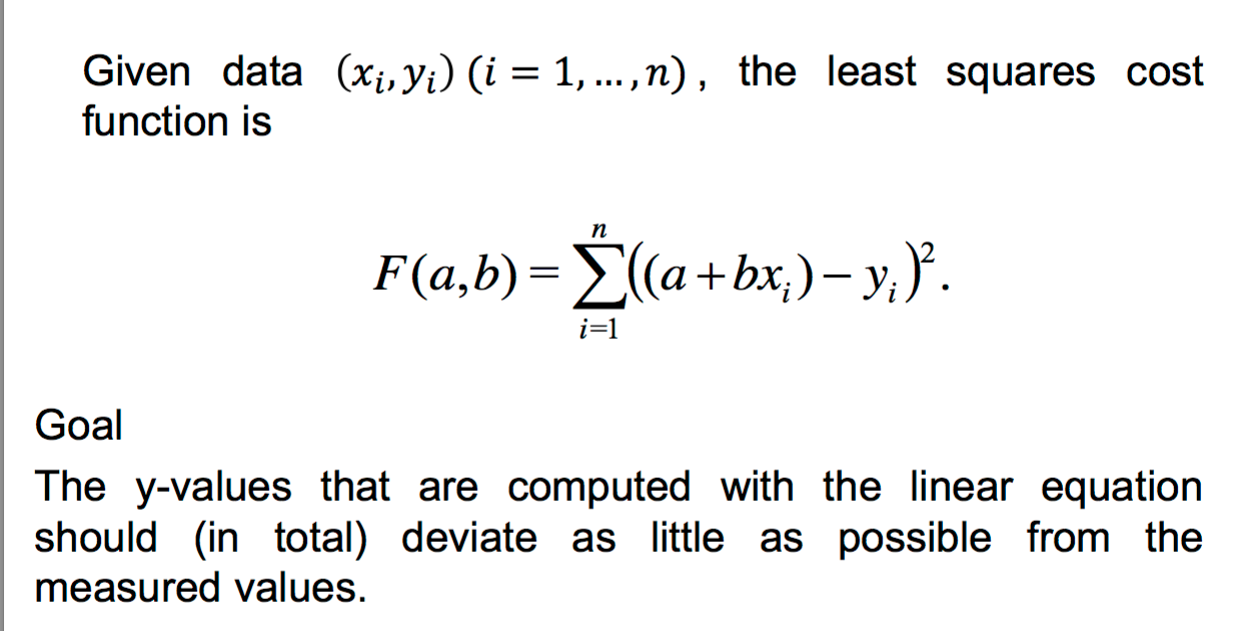
# Cost functions

sually, the sum of square errors in y-direction is chosen as cost function (to be minimized).

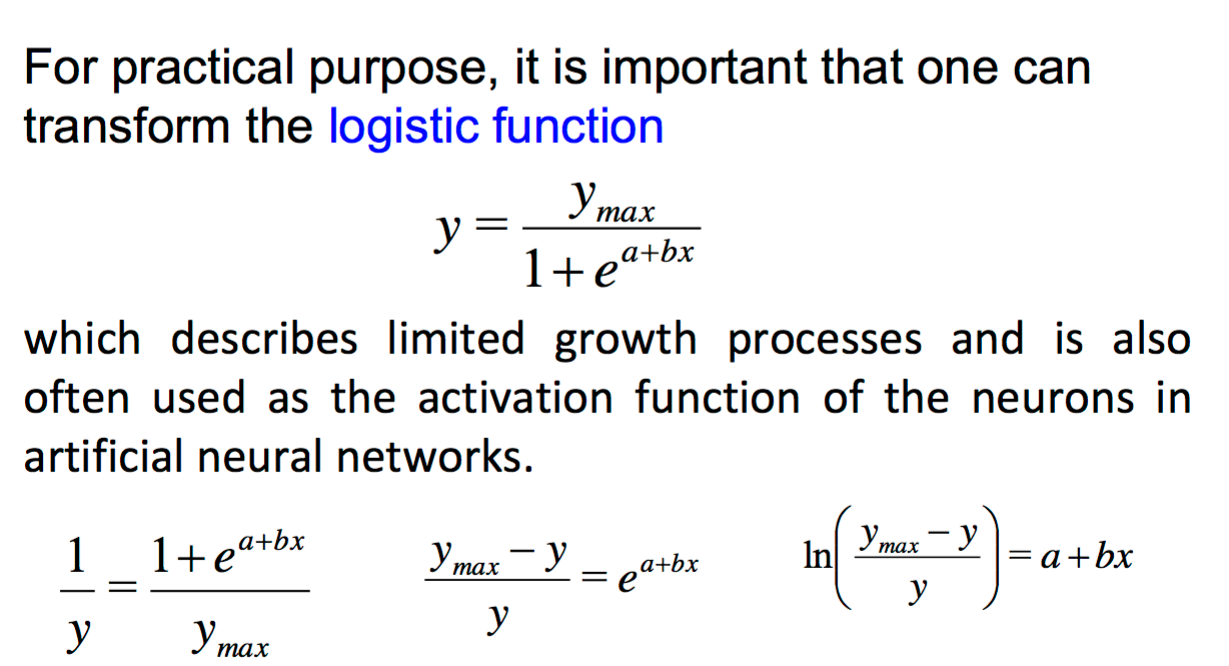
* Other reasonable cost functions:
* mean absolute distance in y-direction
* mean Euclidean distance
* maximum absolute distance in y-direction (or equivalently: the maximum squared distance in y- direction)
* maximum Euclidean distance

– …

# Construction



# LOGISTIC FUNCTION

****

which describes limited growth processes and is also

## Regression & nominal attributes

If most of the predictor variables are numerical and the few nominal attributes have small domains, a regression function can be constructed for possible combination of the values of the nominal attributes, given that the data set is sufficiently large and covers all combinations.

## Example

|  |  |
| --- | --- |
| **Attribute** | **Type/Domain** |
| Sex | F/M |
| Vegetarian | Yes/No |
| Age | Numerical |
| Height | Numerical |
| Weight | Numerical |

* + Task: Predict the weight based on the other attributes. Possible solution: Construct four separate regression functions for (F,Yes),(F,No),(M,Yes),(M,No) using only age and height as predictor variables.

## Regression & nominal attributes

Alternative approach: Encode the nominal

attributes as numerical attributes.

* + Binary attributes can be encoded as 0/1 or -1/1
  + For nominal attributes with more than two values, a 0/1 or -1/1 numerical attribute should be introduced for each possible value of the nominal attribute.
  + Do not encode nominal attributes with more than two values in one numerical attribute.

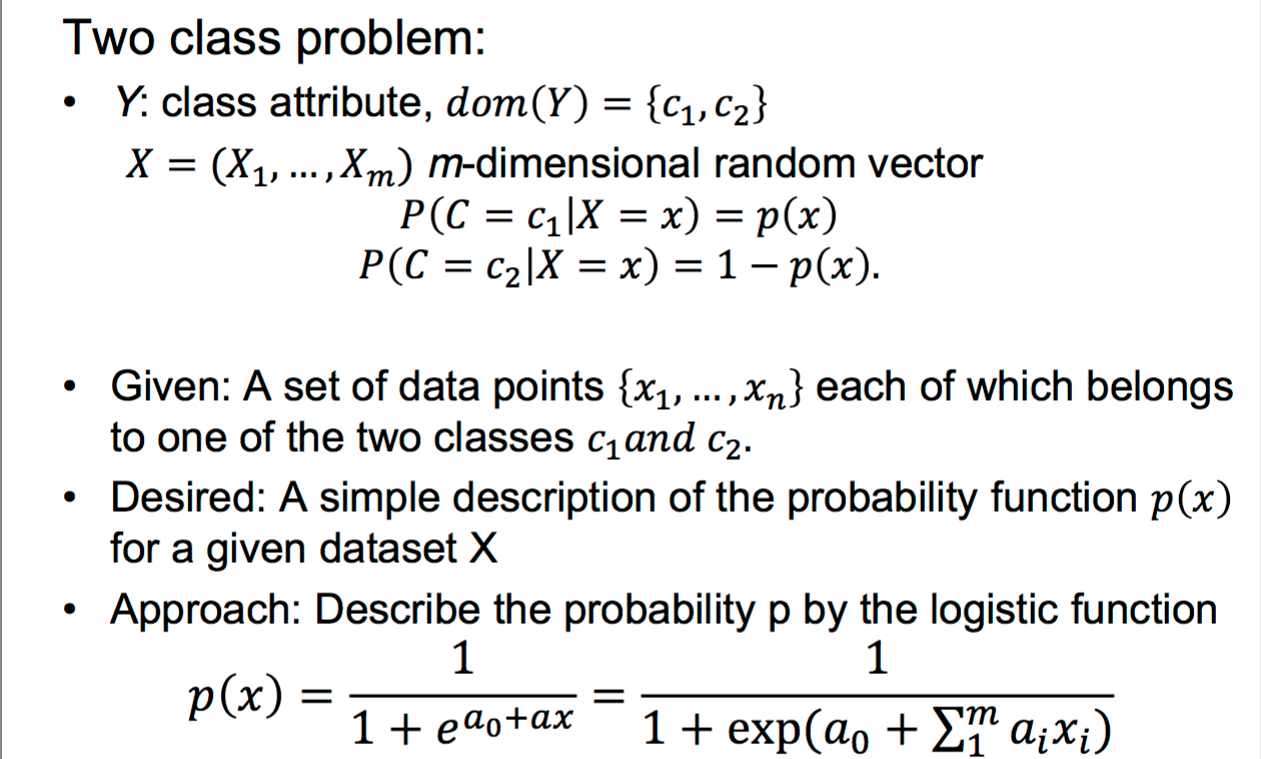
## Classification as regression

* + - A two-class classification problem (with classes 0 and 1) can be viewed as regression problem.
    - The regression function will usually not yield exact outputs 0 and 1, but the classification decision can be made by considering 0.5 as a cut-of value.
    - Problem: The objective functions aims at minimizing the function approximation error (for example, the mean squared error), but not misclassifications.

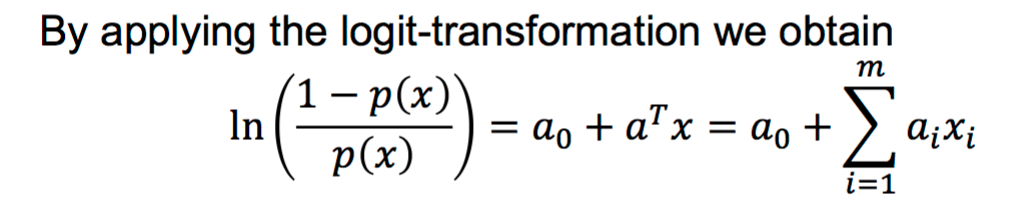
## Logisticregression

## Next page

## 

𝑖=1

Classificatin : Logistic regressions



that is, a multilinear regression problem, which can be

solved with the introduced techniques.

But how do we determine the values p(x) that enter the above equation?

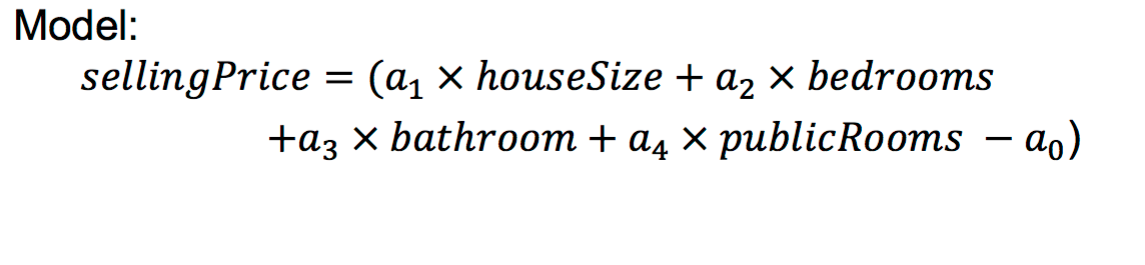
* For a small data space with sufficient many realizations for every possible point the class probability can be estimated by the relative frequencies of the classes.
* If this is not the case, we may rely on an approach known

as kernel estimation.

Next page:

## Linea Regression - Example

* Task: Predict housing price
* Attributes
  + houseSize
  + bedrooms
  + bathroom
  + publicRooms
  + sellingPrice
* Model:



## Pros & Cons

Pros:

Strong mathematical foundation

Simple to calculate and to understand (for a moderate number of dimensions)

High predictive accuracy Cons:

Many dependencies are non-linear

Global model does not adapt to locally different data

distributions

 Locally weighted regression

Slides Edited from "Guide to Intelligent Data Analysis", By M. Berthold et al

# References

* M. Berthold, C. Borgelt, F. Hoppner and F. Klawonn Guide to Intelligent Data Analysis, Spriner-Verlag,

2010

– Chapter 8, Section 8.3

* I. Witten, E. Frank and M. Hall (2011)

*Data Mining: Practical Machine Learning Tools and Techniques*,

Morgan Kaufmann Publishers.

* <http://www.kdnuggets.com/>