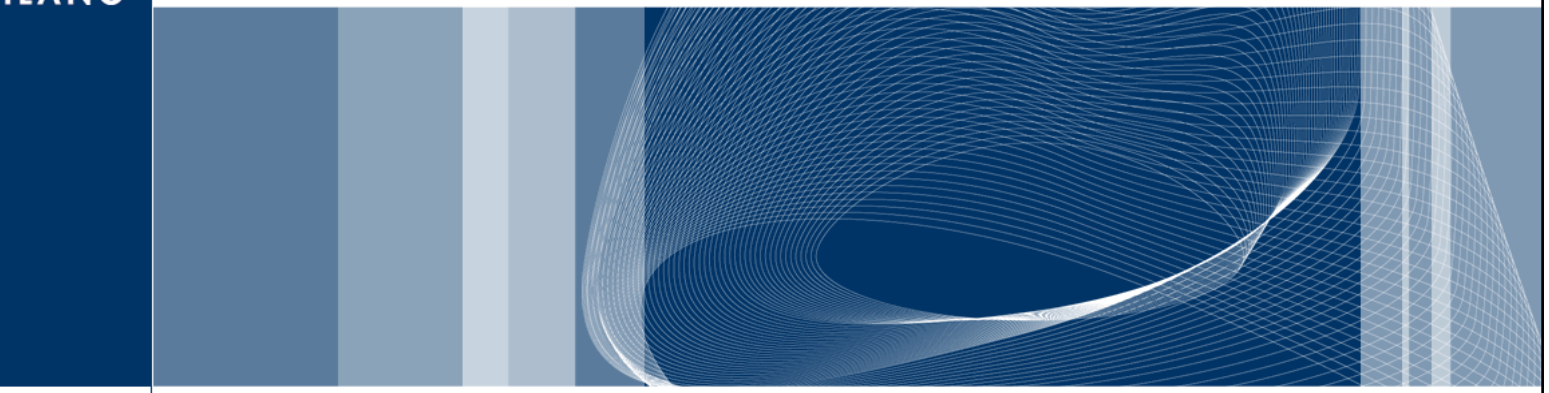
POLITECNICO DI MILANO



Data Representation

Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

Prof. Pier Luca Lanzi

Readings

- "Data Mining and Analysis" – Chapter 1
- "Mining of Massive Datasets" – Chapter 1

describing data

Contact Lenses Data

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

Data Representation

- Data are typically abstracted as a matrix, with n rows and d columns, given as

$$\mathbf{D} = \begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}$$

- Rows are called instances, examples, records, transactions, objects, points, feature-vectors, etc.

- Columns are called attributes, properties, features, dimensions, variables, fields, etc.

	Sepal length X_1	**Sepal width** X_2	**Petal length** X_3	**Petal width** X_4	**Class** X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

Instances, Attributes, Concepts

- Instances (observations, case)
  - The atomic elements of information from a dataset
  - Also known as records, prototypes, or examples
- Attributes (variable)
  - Measures aspects of an instance
  - Also known as features or variables
  - Each instance is composed of a certain number of attributes
- Concepts
  - Special content inside the data
  - Kind of things that can be learned
  - Intelligible and operational concept description

Two Versions of the Weather Data

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
…	…	…	…	…

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
…	…	…	…	…

attribute types

Attributes

- Numeric Attributes
  - Real-valued or integer-valued domain
  - Interval-scaled when only differences are meaningful (e.g., temperature)
  - Ratio-scaled when differences and ratios are meaningful (e.g., Age)
- Categorical Attributes
  - Set-valued domain composed of a set of symbols
  - Nominal when only equality is meaningful (e.g., domain(Sex) = { M, F})
  - Ordinal when both equality (are two values the same?) and inequality (is one value less than another?) are meaningful (e.g., domain(Education) = { High School, BS, MS, PhD})

Numerical Attributes

- Not only ordered but measured in fixed and equal units
- Examples
  - Attribute "temperature" expressed in degrees
  - Attribute "year"
- Characteristics
  - Difference of two values makes sense
  - Sum or product doesn't make sense
  - Zero point is not defined
- Sometimes they are divided into "discrete" and "continuous"

Nominal Attributes (or Categorical)

- Values are distinct symbols
- Values themselves serve only as labels or names

- Example
  - Attribute "outlook" from weather data
  - Values: "sunny", "overcast", and "rainy"

- Characteristics
  - No relation is implied among nominal values
  - No ordering
  - No distance measure
  - Only equality tests can be performed

Other Types of Attributes

- Ratio Attributes
  - Numerical attributes for which the measurement scheme defines a zero point (e.g., an attribute representing distance)
- Ordinal Attributes
  - Categorical attributes with an imposed order on values
  - No distance between values defined
  - For instance, the attribute "temperature" in weather data "hot" > "mild" > "cool"

Nominal or Ordinal?

- Attribute "age" nominal
  - If age = young and astigmatic = no
    and tear production rate = normal
    then recommendation = soft
- Attribute "age" ordinal
  (e.g. "young" < "pre-presbyopic" < "presbyopic")
  - If age≤pre-presbyopic and astigmatic = no
    and tear production rate = normal
    then recommendation = soft

Why Specifying Attribute Types?

- Some algorithms fit some specific data types best
- Express the best possible patterns into data
- Make the most adequate comparisons
- Example
  - Outlook > "sunny" does not make sense, while
  - Temperature > "cool" or
  - Humidity > 70 does
- Additional uses of attribute type
  - Check for valid values
  - Deal with missing values, etc.

missing values

Why Missing Values Exist?

- Faulty equipment, incorrect measurements, missing cells in manual data entry, censored/anonymous data
- Review scores for movies, books, etc.
- Very frequent in questionnaires for medical scenarios
- Censored/anonymous data
- In practice, a low rate of missing values may be suspicious
- Interview data ("Did you ever …")

Missing Values

- Frequently indicated by out-of-range entries (e.g. max/min float), Nan or special values (e.g., zero)
- Missing value may have significance in itself
  - E.g. missing test in a medical examination
- Most schemes assume that is not the case
  - "missing" may need to be coded as additional value
- Does absence of value have some significance?
  - If it does, "missing" is a separate value
  - If it does not, "missing" must be treated in a special way

What Types of Missing Values?

- Missing completely at random (MCAR)
  - The distribution of an example having a missing value for an attribute does not depend on either the observed data or the missing data
  - Example: some survey questions contain a random sample of the whole questionnaire
- Missing at random (MAR)
  - The distribution of an example having a missing value for an attribute depends on the observed data, but does not depend on the missing data
  - Missing at Random means the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data
  - Whether or not someone answered #13 on your survey has nothing to do with the missing values, but it does have to do with the values of some other variable
  - For example, people who don't declare their salary not because of the amount of it but just because they don't want to.
- Not missing at random (NMAR)
  - The distribution of an example having a missing depends on the missing values.
  - For example, respondents with high income less likely to report income
- Note that NMAR and MAR might be difficult to identify and often require domain knowledge

Dealing with Missing Values

- Use what you know
  - Why data is missing
  - Distribution of missing data
- Decide on the best strategy to yield the least biased estimates
  - Deletion Methods (listwise deletion, pairwise deletion)
  - Single Imputation Methods (mean/mode substitution, dummy variable method, single regression)
  - Model-Based Methods (maximum Likelihood, multiple imputation

Strategies for missing values handling

- The handling of missing data depends on the type
- Discarding all the examples with a missing values
  - Simplest approach
  - Allows the use of unmodified data mining methods
  - Only practical if there are few examples with missing values. Otherwise, it can introduce bias
- Fill in the missing value manually ☺
- Convert the missing values into a new value
  - Use a special value for it
  - Add an attribute that indicates if value is missing or not
  - Greatly increases the difficulty of the data mining process
- Imputation methods
  - Assign a value to the missing one, based on the rest of the dataset. Use the unmodified data mining methods.

Listwise Deletion (Complete Case Analysis)

- Only analyze cases with available data on each variable
- Simple, but reduces the data
- Comparability across analyses
- Does not use all the information
- Estimates may be biased if data not MCAR

Gender	8th grade math test score	12th grade math score
~~F~~	~~45~~	.
~~M~~	.	~~99~~
F	55	86
F	85	88
F	80	75
.	~~81~~	~~82~~
F	75	80
~~M~~	~~95~~	.
M	86	90
F	70	75
~~F~~	~~85~~	.

Pairwise deletion (Available Case Analysis)

- Analysis with all cases in which the variables of interest are present
- Advantage
  - Keeps as many cases as possible for each analysis
  - Uses all information possible with each analysis
- Disadvantage
  - Can't compare analyses because sample different each time

Gender	8th grade math test score	12th grade math score
F	45	.
M	.	99
F	55	86
F	85	88
F	80	75
.	81	82
F	75	80
M	95	.
M	86	90
F	70	75
F	85	.

Imputation methods

- Extract a model from the dataset to perform the imputation
- Suitable for MCAR and, to a lesser extent, for MAR
- Not suitable for NMAR type of missing data
- For NMAR we need to go back to the source of the data to obtain more information
- Survey of imputation methods available at
  http://sci2s.ugr.es/MVDM/index.php
  http://sci2s.ugr.es/MVDM/biblio.php

Single Imputation Methods

- Mean/mode substitution (most common value)
  - Replace missing value with sample mean or mode
  - Run analyses as if all complete cases
  - Advantages: Can use complete case analysis methods
  - Disadvantages: Reduces variability
- Dummy variable control
  - Create an indicator for missing value (1=value is missing for observation; 0=value is observed for observation)
  - Impute missing values to a constant (such as the mean)
  - Include missing indicator in the algorithm
  - Advantage: uses all available information about missing observation
  - Disadvantage: results in biased estimates, not theoretically driven
- Regression Imputation
  - Replaces missing values with predicted score from a regression equation.

Do Not Impute (DNI)

- Simply use the default policy of the data mining method
- Works only if the policy exists

inaccurate values

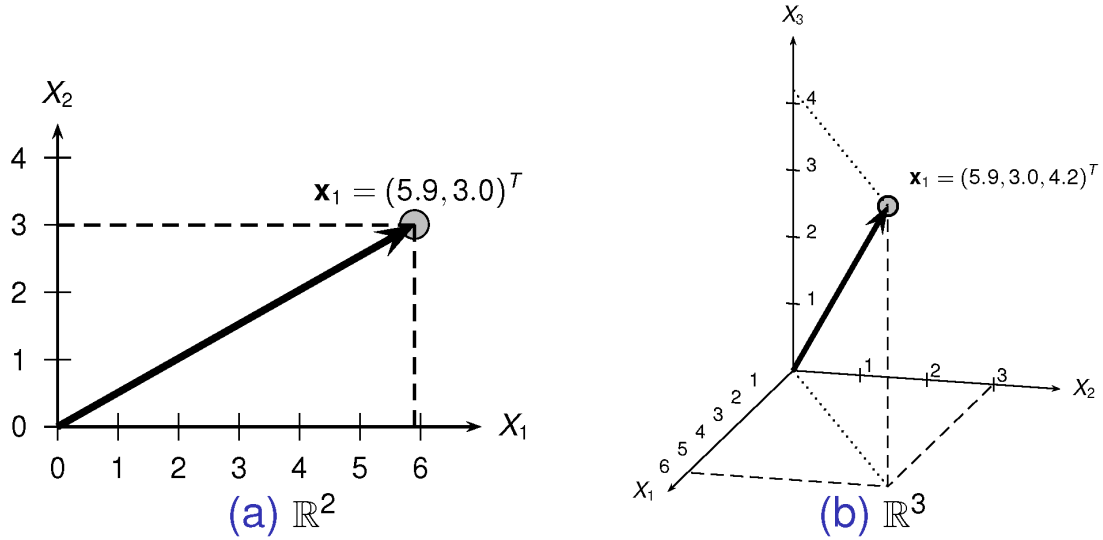
Inaccurate Values

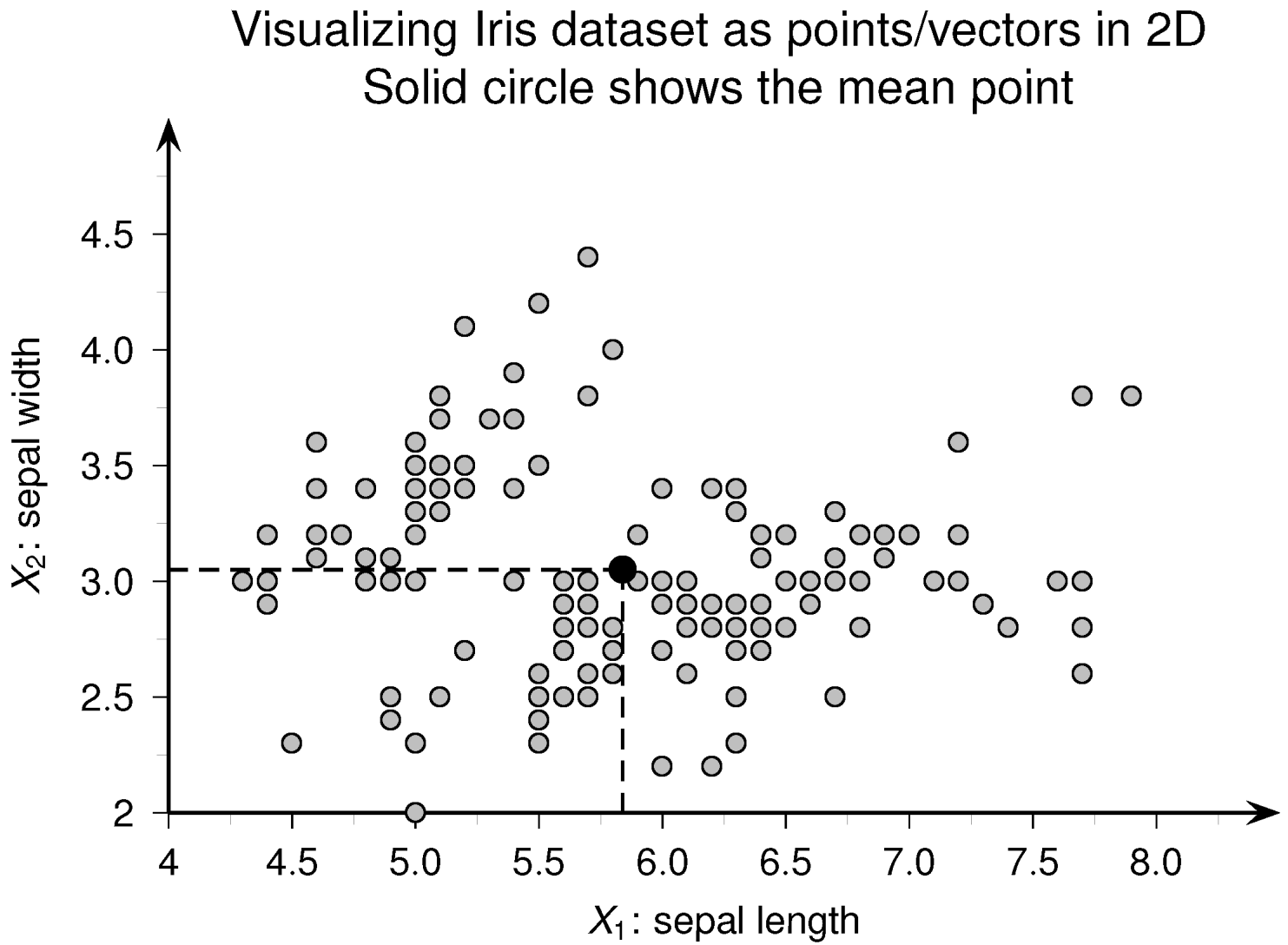
- Data has not been collected for mining it
- Errors and omissions that don't affect original purpose of data (e.g. age of customer)
- Typographical errors in nominal attributes,
  thus values need to be checked for consistency
- Typographical and measurement errors in numeric attributes,
  thus outliers need to be identified
- Errors may be deliberate (e.g. wrong zip codes)

the geometric view

The Geometrical View of the Data

- When the data matrix contains only numerical values
  - Every row can be viewed as a point in a d-dimension space
  - Every column as a point in a n-dimensional space




Visualizing Iris dataset as points/vectors in 2D
Solid circle shows the mean point
X_2: sepal width
4.5
4.0
3.5
3.0
2.5
2
4
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
X_1: sepal length

Given two points $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$, their *dot product* is defined as the scalar

$$\mathbf{a}^T\mathbf{b} = a_1b_1 + a_2b_2 + \cdots + a_mb_m = \sum_{i=1}^{m} a_ib_i$$

The *Euclidean norm* or *length* of a vector $\mathbf{a}$ is defined as

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}^T\mathbf{a}} = \sqrt{\sum_{i=1}^{m} a_i^2}$$

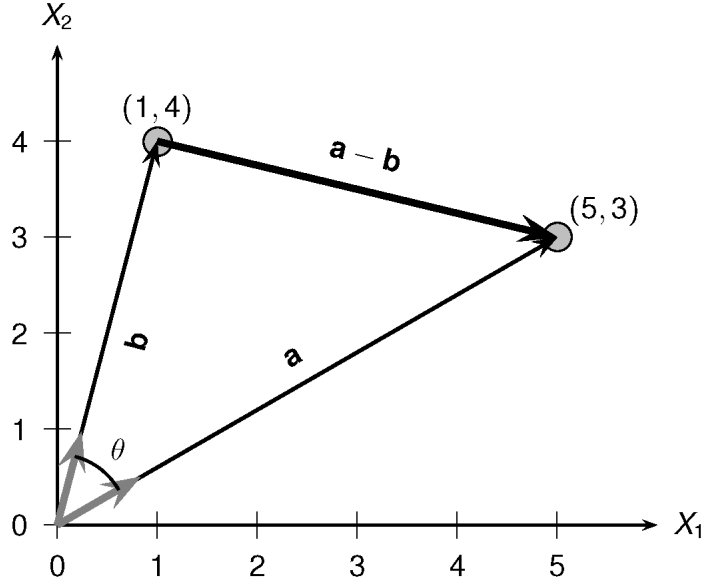
The *unit vector* in the direction of $\mathbf{a}$ is $\mathbf{u} = \frac{\mathbf{a}}{\|\mathbf{a}\|}$ with $\|\mathbf{a}\| = 1$.

Distance between $\mathbf{a}$ and $\mathbf{b}$ is given as

$$\|\mathbf{a} - \mathbf{b}\| = \sqrt{\sum_{i=1}^{m} (a_i - b_i)^2}$$

Angle between $\mathbf{a}$ and $\mathbf{b}$ is given as

$$\cos\theta = \frac{\mathbf{a}^T\mathbf{b}}{\|\mathbf{a}\|\,\|\mathbf{b}\|} = \left(\frac{\mathbf{a}}{\|\mathbf{a}\|}\right)^T \left(\frac{\mathbf{b}}{\|\mathbf{b}\|}\right)$$



the probabilistic view

If X is discrete, the *probability mass function* of X is defined as

$$f(x) = P(X = x) \qquad \text{for all } x \in \mathbb{R}$$

f must obey the basic rules of probability. That is, f must be non-negative:

$$f(x) \geq 0$$

and the sum of all probabilities should add to 1:

$$\sum_x f(x) = 1$$

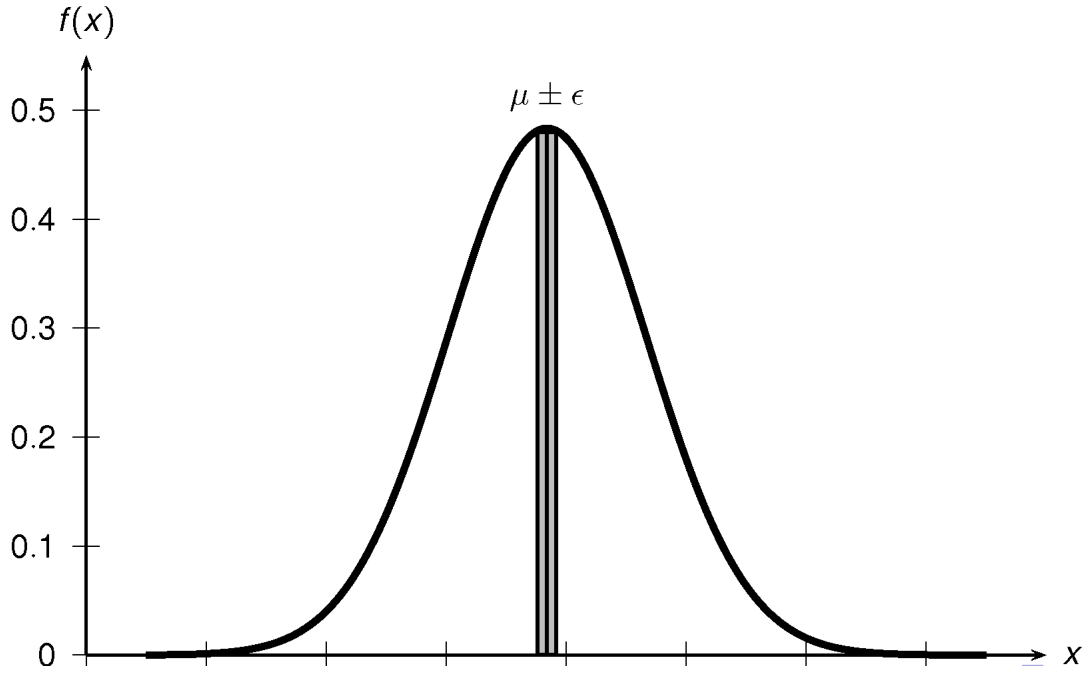
Intuitively, for a discrete variable X, the probability is concentrated or massed at only discrete values in the range of X, and is zero for all other values.

We model sepal length via the *Gaussian* or *normal* density function, given as

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

where $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the mean value, and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ is the variance.

Normal distribution for sepal length: $\mu = 5.84$, $\sigma^2 = 0.681$



data format

Data Format

- Most commercial tools have their own proprietary format
- Most tools import excel files and comma-separated value files

```
Year,Make,Model,Length
1997,Ford,E350,2.34
2000,Mercury,Cougar,2.38
```

```
Year;Make;Model;Length
1997;Ford;E350;2,34
2000;Mercury;Cougar;2,38
```

Attribute-Relation File Format (ARFF)


```
%
% ARFF file for weather data with some numeric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {true, false}
@attribute play? {yes, no}

@data
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
...
```

http://www.cs.waikato.ac.nz/~ml/weka/arff.html

Additional Attribute Types

- ARFF supports string attributes:

```
@attribute description string
```

- Similar to nominal attributes but list of values is not pre-specified

- ARFF also supports date attributes:

```
@attribute today date
```

- Uses the ISO-8601 combined date and time format yyyy-MM-dd-THH:mm:ss

Additional Attribute Types

- ARFF supports sparse data, for instance the following examples,

```
0, 26, 0, 0, 0 ,0, 63, 0, 0, 0, "class A"
0, 0, 0, 42, 0, 0, 0, 0, 0, 0, "class B"
```

- Can also be represented as,

```
{1 26, 6 63, 10 "class A"}
{3 42, 10 "class B"}
```

Missing Values in ARFF


```
@relation labor
@attribute 'duration' real
@attribute 'wage-increase-first-year' real
@attribute 'wage-increase-second-year' real
@attribute 'wage-increase-third-year' real
@attribute 'cost-of-living-adjustment' {'none','tcf','tc'}
@attribute 'working-hours' real
@attribute 'pension' {'none','ret_allw','empl_contr'}
@attribute 'standby-pay' real
@attribute 'shift-differential' real
@attribute 'education-allowance' {'yes','no'}
@attribute 'statutory-holidays' real
@attribute 'vacation' {'below_average','average','generous'}
@attribute 'longterm-disability-assistance' {'yes','no'}
@attribute 'contribution-to-dental-plan' {'none','half','full'}
@attribute 'bereavement-assistance' {'yes','no'}
@attribute 'contribution-to-health-plan' {'none','half','full'}
@attribute 'class' {'bad','good'}
@data
1,5,?, ?, ?,40, ?, ?,2, ?,11,'average', ?, ?,'yes',?,'good'
2,4.5,5.8, ?, ?,35,'ret_allw', ?, ?,'yes',11,'below_average', ?,'full', ?,'full','good'
?, ?, ?, ?, ?,38,'empl_contr', ?,5, ?,11,'generous','yes','half','yes','half','good'
3,3.7,4,5,'tc', ?, ?, ?, ?,'yes', ?, ?, ?, ?,'yes', ?,'good'
```

Attribute Types and Interpretation

- Interpretation of attribute types in ARFF depends on the mining scheme
- Numeric attributes are interpreted as
  - Ordinal scales if less-than and greater-than are used
  - Ratio scales if distance calculations are performed (normalization/standardization may be required)
- Instance-based schemes define distance between nominal values (0 if values are equal, 1 otherwise)
- Integers in some given data file: nominal, ordinal, or ratio scale?

DSPL: Dataset Publishing Language

- Open format by Google available at http://code.google.com/apis/publicdata/
- Use existing data: add an XML metadata file existing CSV
- Read by the Google Public Data Explorer, which includes animated bar chart, motion chart, and map visualization
- Allow linking to concepts in other datasets
- Geo-enabled: allows adding latitude and longitude data to your concept definitions

model representation

Predictive Model Markup Language

- XML-based markup language developed by the Data Mining Group (DMG) to provide a way for applications to define models related to predictive analytics and data mining
- The goal is to share models between applications
- Vendor-independent method of defining models
- Allow to exchange of models between applications.
- PMML Components: data dictionary, data transformations, model, mining schema, targets, output

data repositories

Publicly Available Datasets

- UCI repository
  - http://archive.ics.uci.edu/ml/
  - Probably the most famous collection of datasets
- Kaggle
  - http://www.kaggle.com/
  - It is not a static repository of datasets, but a site that manages Data Mining competitions
  - Example of the modern concept of crowdsourcing
- KDNuggets
  - http://www.kdnuggets.com/datasets/