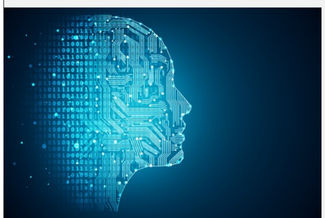


Data Science for Researchers and Scholars

Vasant G. Honavar
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

A deeper dive into data

- What do we mean by data?
  - Digital representation of objects, entities, persons, events, processes, etc. in the real world
    - Employees
    - Genomic sequences
    - Social relationships
    - Images
    - Documents
    - Medical histories
    - .....

Tabular data

- Objects or entities are represented by rows in a table.
- Columns of the table encode properties or characteristics, features, of the objects
- Each object is represented by specifying the values of each attribute
- We call the set of all possible values of an attributes its domain
  - Domain of Refund is {Yes, No}
  - Domain of Taxable Income is $\Re^+$ (positive real numbers)

Attributes

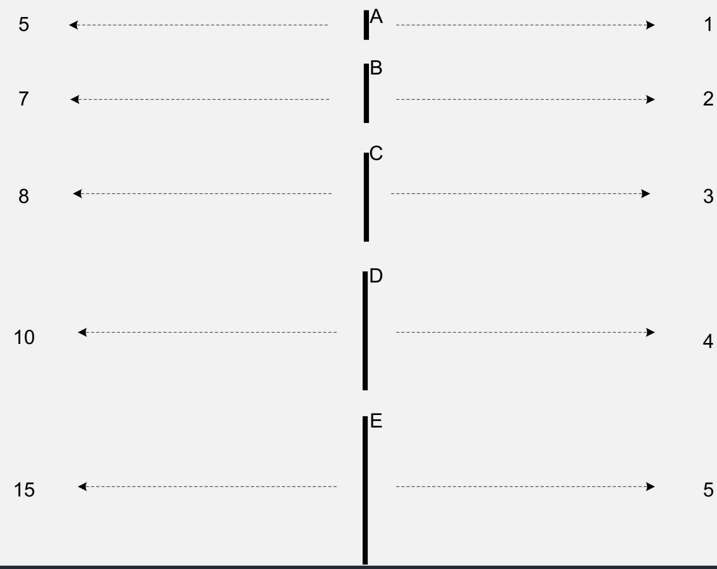
Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

The way you encode an attribute has consequences

- Two different encodings of lengths of objects

This scale preserves only the ordering property of length.



This scale preserves the ordering and additivity property of length.

Attributes come in many flavors

- There are different types of attributes
- Nominal
  - Examples: ID numbers, eye color, zip codes
- Ordinal
  - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
- Interval
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- Ratio
  - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

Properties of Attribute Values

- Different types of attributes possess different properties:
  - Distinctness: = ≠
  - Order: < >
  - Meaningfulness of differences + -
  - Meaningfulness of ratios * /
  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & meaningfulness of differences
  - Ratio attribute: All 4 properties

Measurement is a tricky subject

- Temperature is measured in Kelvin, degrees Celsius, and degrees Fahrenheit
  - Temp in Kelvin = Temp in degrees Celsius + 273.15
  - Temp in Fahrenheit = (Temp in degrees Celsius)(9/5)+32
- Is it physically meaningful to say that a temperature of 10 ° Celsius is twice as high as 5° Celsius?
- Depends
- On what?
  - the measurement scale!!!
- Consider measuring height
  - If Bill's height is three inches above average and Bob's height is six inches above average, then would we say that Bob is twice as tall as Bill?

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*hard, medium, soft*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

	Attribute Type	Transformation	Comments
Categorical Qualitative	Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
	Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Numeric Quantitative	Interval	*new_value = a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Takes real numbers as values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point numbers.

Asymmetric Attributes

- Only presence (a non-zero attribute value) matters
  - Words present in documents
  - Items present in customer transactions

- If you run into a friend at the grocery store would you ever say the following?

  "We have similar taste because I did not buy almost every item that you also did not buy"

Points to remember about attribute types

- The types of operations you choose should be "meaningful" for the type of data you have
  - Distinctness, order, meaningful intervals, and meaningful ratios are only four (among many possible) properties of data
  - The data type you see – often numbers or strings – may not capture all the properties or may suggest properties that are not present
  - Analysis may depend on these other properties of the data
  - In the end, what is meaningful may be domain-specific

Important Characteristics of Data

- Dimensionality (number of attributes)
- Sparsity
- Resolution
- Size

Types of data

- Tabular data
- Document Data
- Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
  - Social networks
- Ordered
  - Clinical histories
  - System call sequences
  - Genome Sequences Sequence Data

Tabular data

- Data that consists of a collection of records, each of which encoded by a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Tabular data

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such a data set can be represented by an *m* by *n* matrix, where there are *m* rows, one for each object, and *n* columns, one for each attribute

Document Data

- Each document is encoded using a vector of word frequencies
  - Each term is a component (attribute) of the vector
  - The value of each component is the number of times the corresponding word occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

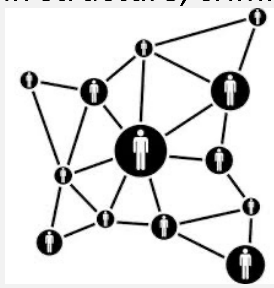
Transaction Data

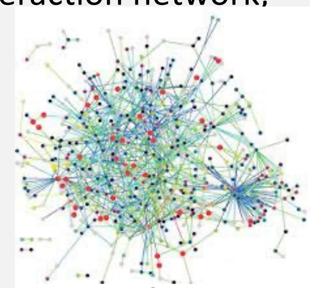
- A special type of data, where
  - Each transaction involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.
  - Can represent transaction data as record data

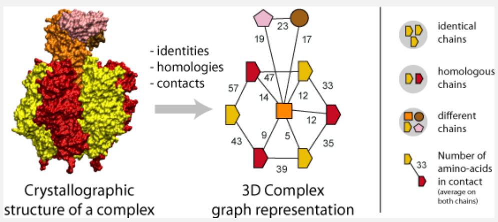
TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

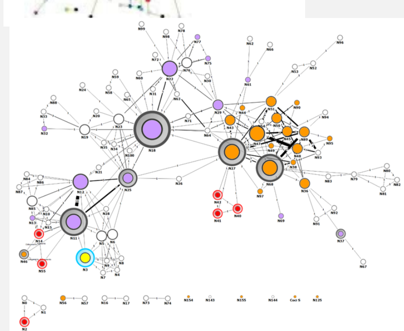
Graph Data

- Examples: Social network, protein interaction network, protein structure, criminal network









Ordered Data

- Genomic sequence data

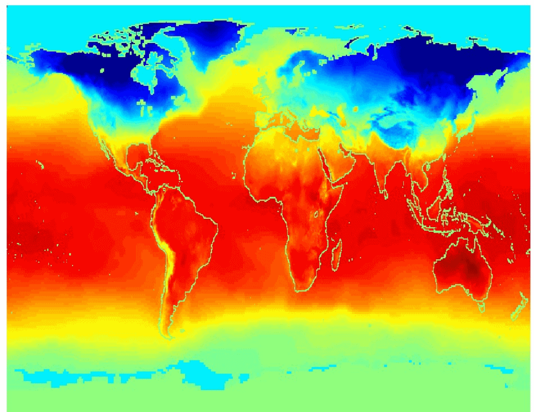
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Ordered Data

- Spatially indexed temporal data

Average Monthly Temperature of land and ocean

Jan



Function approximation (Regression)

- Function approximation is like classification except the labels are real valued

Example applications:

Predicting

- Stock value
- Income
- Power consumption

K nearest neighbor Function Approximator

Learning Phase

For each training example $(X_i, f(X_i))$, store the example in memory

Approximation phase

Given a query instance X_q, identify the k nearest neighbors $X_1 \ldots X_k$ of X_q

$$g(X_q) \leftarrow \frac{\sum_{l=1}^{K} f(X_i)}{K}$$

Value of a function (e.g., price of a product) at a query point is simply the average or inverse distance weighted average of the value of the function at the k nearest neighbors of the query point

Generative models for classification

Basic Probability Theory

- A random **experiment** has a set of potential outcomes, e.g., throw a die, "look at" another data sample
- The **sample space** of an experiment is the set of all possible outcomes, e.g., {1, 2, 3, 4, 5, 6}
- For machine learning the sample spaces can be ***very*** large

Basic Probability Theory

An **event** is a subset of the sample space

Dice rolls

- {2}
- {3, 6}
- even = {2, 4, 6}
- odd = {1, 3, 5}

Machine learning

- A chosen feature has particular values
- A data sample is described by the values of features

Events

We're interested in probabilities of events

- p({2})
- p(label=cancer)
- p(tumorpresent = 1)
- p(smoker = 1)

ariables

HT	HTH	HTT	THH	THT	TTH	TTT
2	2	1	2	1	1	0

ble is a mapping from the sample space to a
events)

vhose values we want to measure in an experiment

ndom variable, *X*, could be the number of heads for

Random variables

- We're interested in the probability of the different values of a random variable
- The definition of probabilities over *all* of the possible values of a random variable defines a **probability distribution**

	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
X	3	2	2	1	2	1	1	0

Probability distribution

To be explicit

- A probability distribution assigns probability values to *all possible values* of a random variable
- These values must be >= 0 and <= 1
- These values must sum to 1 for all possible values of the random variable

Unconditional/prior probability

Simplest form of probability is

- P(X)

Prior probability: In the absence of any additional information, what is the probability of an outcome of interest

- What is the probability of heads?
- What is the probability of surviving cancer?
- What is the probability of a wine review containing the word “pinot”?
- What is the probability of a passenger on the titanic being under 21 years old?
- …

Joint distribution

We can also talk about probability distributions over multiple variables

P(X,Y)

- Joint probability of X *and* Y
- A distribution over the cross product of possible values

DSPass AND HCIPass	P(DsPass, HCIPass)
true, true	.80
true, false	.01
false, true	.04
false, false	.15

Joint distribution

Still a probability distribution

- all values between 0 and 1, inclusive
- all values sum to 1

All questions/probabilities of the two variables can be calculate from the joint distribution

What is P(HCIPass)?

DSPass AND HCIPass	P(DsPass, HCIPass)
true, true	.80
true, false	.01
false, true	.04
false, false	.15

Joint distribution

Still a probability distribution

- all values between 0 and 1, inclusive
- all values sum to 1

All questions/probabilities of the two variables can be calculate from the joint distribution

DSPass AND HCIPass	P(DsPass, HCIPass)
true, true	.80
true, false	.01
false, true	.04
false, false	.15

P(HCIPass)=0.84

How did you figure that out?

Joint distribution

DSPass AND HCIPass	P(DsPass, HCIPass)
true, true	.80
true, false	.01
false, true	.04
false, false	.15

$$P(x) = \sum_{y \in Y} p(x,y)$$

DSPass	P(DSPass)
true	0.81
false	0.19

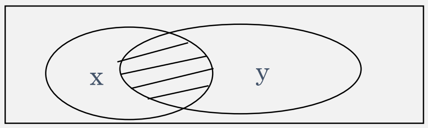
HCIPass	P(HCIPass)
true	0.84
false	0.16

Conditional probability

- As we acquire more information, we can update our probability distribution
- $P(X|Y)$ models this (read "probability of X *given* Y")
  - What is the probability of a heads *given* that both sides of the coin are heads?
  - What is the probability the document is about Chardonnay, given that it contains the word "Pinot"?
  - What is the probability of the word "noir" given that the sentence also contains the word "pinot"?
- Notice that $P(X|Y)$ is still a distribution over the values of X

Conditional probability

$$p(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$



Given that y has happened, in what proportion of those events does x also happen?

Conditional probability

$$p(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

Given that y has happened, what proportion of those events does x also happen

DSPass AND HCIPass	P(DsPass, HCIPass)
true, true	.80
true, false	.01
false, true	.04
false, false	.15

What is:
p(DSPass=true | HCIPass=false)?

Conditional probability

DSPass AND HCIPass	P(DsPass, HCIPass)
true, true	.80
true, false	.01
false, true	.04
false, false	.15

$$p(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

What is: $P(DSPass = True \mid HCIPass = False)$?

$$\frac{P(\ DSPass = True, HCIPass = False)}{P(HCIPass = False)} = \frac{0.01}{0.15 + 0.01} = \frac{1}{16}$$

Notice this is very different than $P(DSPass = true)\ =\ 0.81$

Both are distributions over X

Unconditional/prior probability

$$p(X)$$

MLPass	P(MLPass)
true	0.81
false	0.19

Conditional probability

$$p(X \mid Y)$$

MLPass	P(MLPass|EngPass=false)
true	0.0625
false	0.9375

A note about notation

- When talking about a particular random variable value, you should technically write $P(X = x)$, etc.
- We may write $P(x)$ to mean probability that X takes any particular value, i.e. $P(X = x)$

Chain rule (aka product rule)

$$p(X \mid Y) = \frac{P(X,Y)}{P(Y)} \Rightarrow p(X,Y) = P(X \mid Y)P(Y)$$

We can view calculating the probability of X *AND* Y occurring as two steps:

- Y occurs with some probability $P(Y)$
- Then, X occurs, given that Y has occurred

Chain rule

$$p(X,Y,Z) = P(X \mid Y,Z)P(Y,Z)$$

$$p(X,Y,Z) = P(X,Y \mid Z)P(Z)$$

$$p(X,Y,Z) = P(X \mid Y,Z)P(Y \mid Z)P(Z)$$

$$p(X,Y,Z) = P(Y,Z \mid X)P(X)$$

$$p(X_1, X_2, \cdots X_n) = \prod_{i=1}^{n} P(X_i | X_{1,} \cdots X_{i-1})$$

Applications of the chain rule

We saw that we could calculate the individual prior probabilities using the joint distribution

$$p(x) = \sum_{y \in Y} p(x, y)$$

What if we don't have the joint distribution, but do have conditional probability information:

- P(Y)
- P(X|Y)

$$p(x) = \sum_{y \in Y} p(y) p(x \mid y)$$

This is called "summing over" or "marginalizing out" a variable

Bayes' rule (theorem)

$$p(X \mid Y) = \frac{P(X,Y)}{P(Y)} \Rightarrow p(X,Y) = P(X \mid Y)P(Y)$$

$$p(Y \mid X) = \frac{P(X,Y)}{P(X)} \Rightarrow p(X,Y) = P(Y \mid X)P(X)$$

$$p(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

Bayes' rule

- p(disease | symptoms)
  - For everyone who had the symptoms, what fraction had the disease?
- p(symptoms|disease)
  - For everyone that had the disease, what fraction had this symptom?

- p(label| features)
  - For all examples that had those features, what fraction had that label?
- p(features | label)
  - For all the examples with that label, what fraction had this feature

Bayes Rule

Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.

$$P(cancer) = \qquad P(\neg cancer) =$$

$$P(+ \mid cancer) = \qquad P(- \mid cancer) =$$

$$P(+ \mid \neg cancer) = \qquad P(- \mid \neg cancer) =$$

Bayes Rule

Does patient have cancer or not?

$$P(cancer) = 0.008 \qquad P(\neg cancer) = 0.992$$
$$P(+ \mid cancer) = 0.98 \qquad P(- \mid cancer) = 0.02$$
$$P(+ \mid \neg cancer) = 0.03 \qquad P(- \mid \neg cancer) = 0.97$$

$$P(cancer|+) = \frac{P(+|cancer)P(cancer)}{P(+)};$$

$$P(\neg cancer|+) = \frac{P(+|\neg cancer)P(\neg cancer)}{P(+)}$$

$$P(cancer|+)P(+) = 0.98 \times 0.008 = 0.0078;$$

$$P(\neg cancer|+)P(+) = 0.03 \times 0.992 = 0.0298$$

$$P(+) = 0.0078 + 0.0298$$

$$P(cancer \mid +) = 0.21; \qquad P(\neg cancer \mid +) = 0.79$$

The patient, more likely than not, does not have cancer