Introduction to Probability Theory 4

Clayton Greenberg

CoLi, CS, MMCI, LSV, CRC 1102 (IDeaL) B4

October 29, 2014

Schedule

22.10.2014	Calculate the probability of a given parse
23.10.2014	Solve the medical test Bayes' Rule problem
27.10.2014	Create a code for simplified Polynesian
29.10.2014	Identify types of machine learning problems
31.10.2014	Find a regression line for 2D data

Huffman coding review

File :

b	p	`	m	j	o	d	a	i	r	u	l	s	e	
1	1	2	2	3	3	3	4	4	5	5	6	6	8	12

Green statement review

- probability = what you want / what is possible
- “and” = * (times) [if independent]
- “or” = + (plus) [if mutually exclusive]
- surprisal = the negative logarithm of probability
- conditional = joint / normalizer
- chain rule: joint = conditional of last * joint of rest
- probability of a tree (PCFG) = product of its rules
- probability of a string (PCFG) = sum of its trees
- Bayes’ rule: posterior = likelihood * prior / normalizer
- expectation = weighted average of random variable
- entropy = expected surprisal
- KL-divergence = how different two distributions are

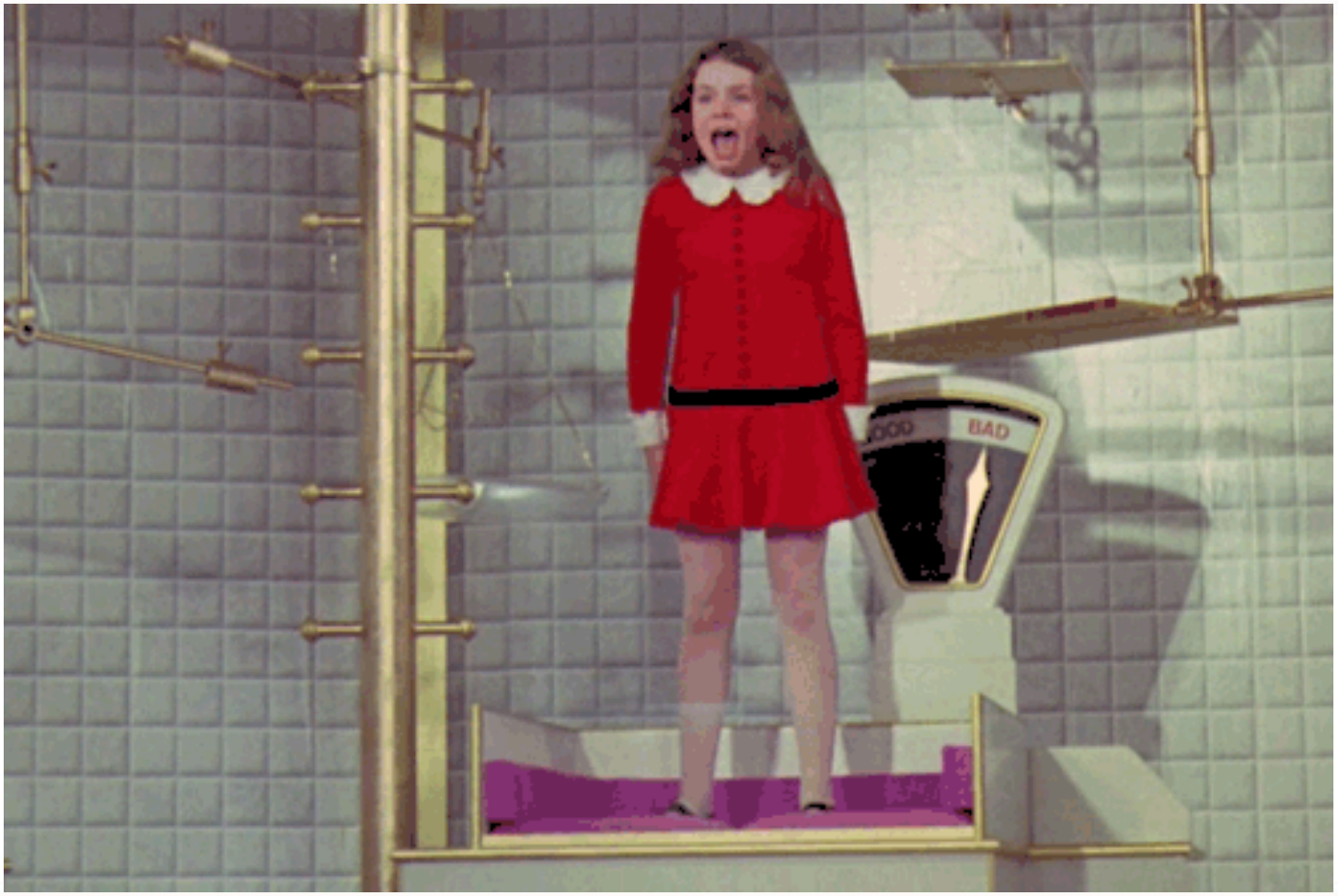
Machine learning

- How to teach a computer to complete a task.
- Humans use natural language (NL) for many tasks.
- NL tasks seem to require “understanding.”
- Computers must be provided with something that approximates “understanding.”
- *Machine learning* creates this mechanism.

Classification

Educated Eggdicator clip removed for file size

Feature evaluation



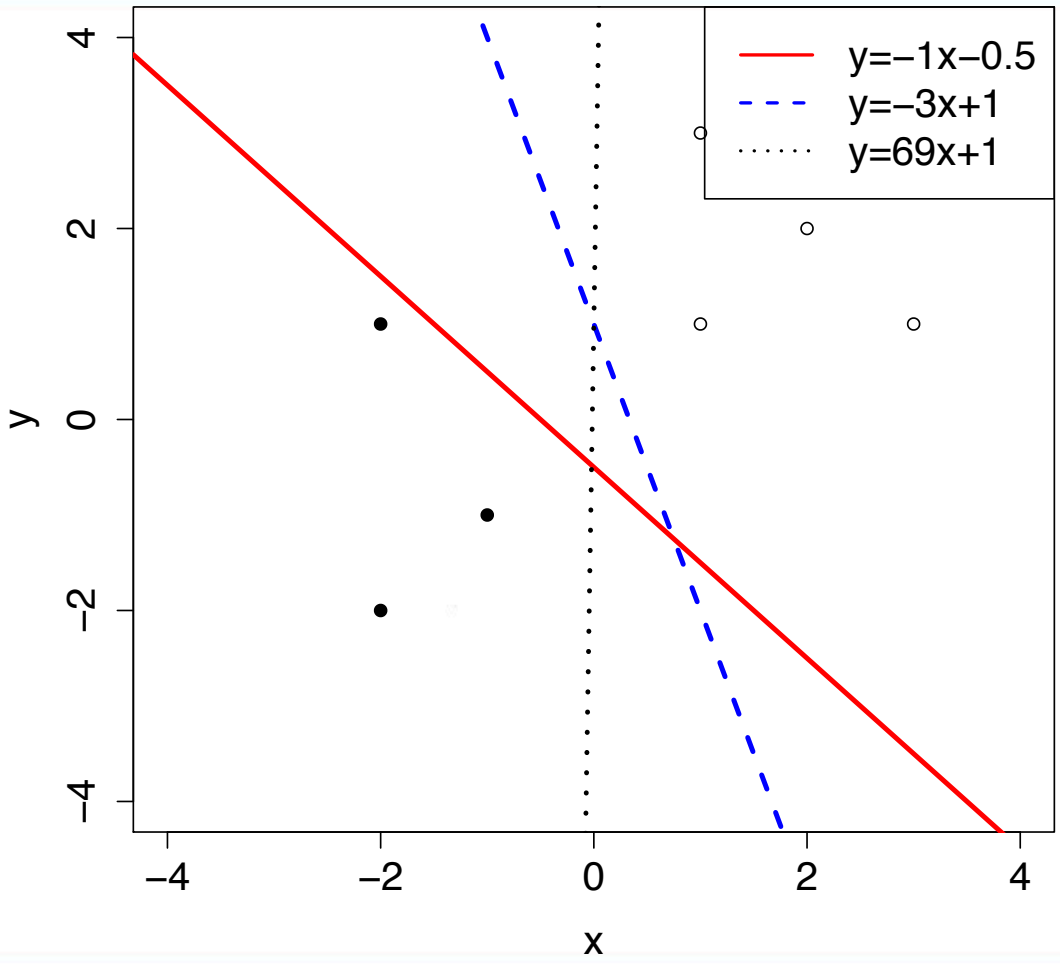
Examples of food features

- Sight: size, color, shape
- Hearing: acoustic signal, chemical processes
- Touch: shape, texture, temperature, weight
- Smell: ingredients, chemical processes
- Taste: nutrients, ingredients

Examples of NLP features

- Topic
- Sentence
- Referent
- Discourse relation
- Parse tree
- POS tag
- Animacy
- Word
- Morpheme
- Phoneme
- Formant
- Frequency

2D linear classification



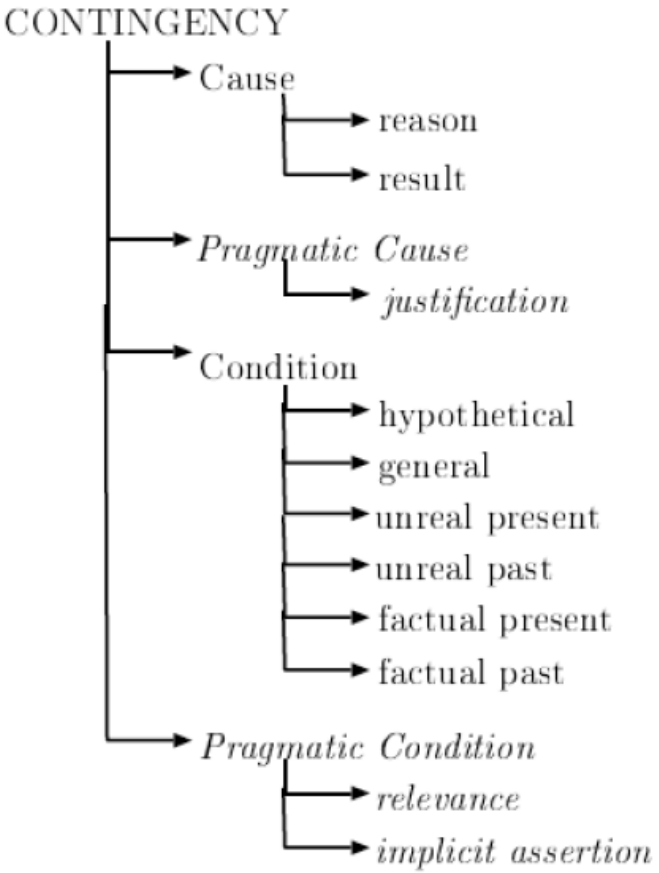
Probabilistic classifier

- Evaluates how well each item fits each category/ class (CAT).
- The classifier chooses the highest scoring CAT for each item.
- p(CAT | item) = p(item | CAT) * p(CAT) / p(item)
- Classification = anything goes in (features), discrete CATs come out.

What if there are no CATs?

- Option 1: make them (annotation scheme)
- Option 2: induce them (clustering)
- Option 3: use scores directly (regression)

Annotation schemes



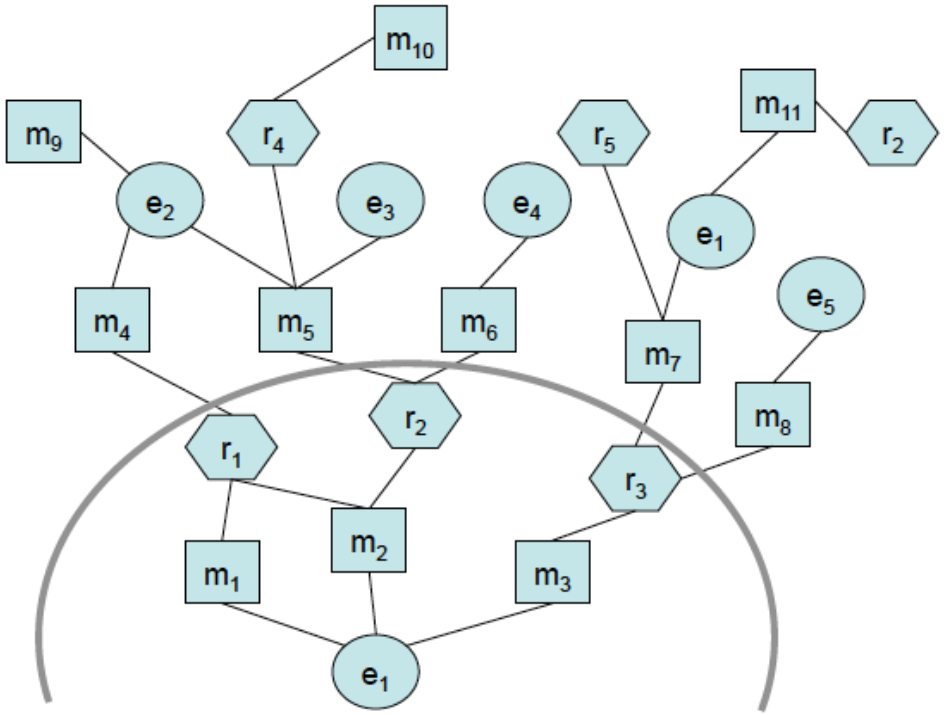
Scheme generation cycle:

1. Exploration
2. Execution
3. Application

Supervision

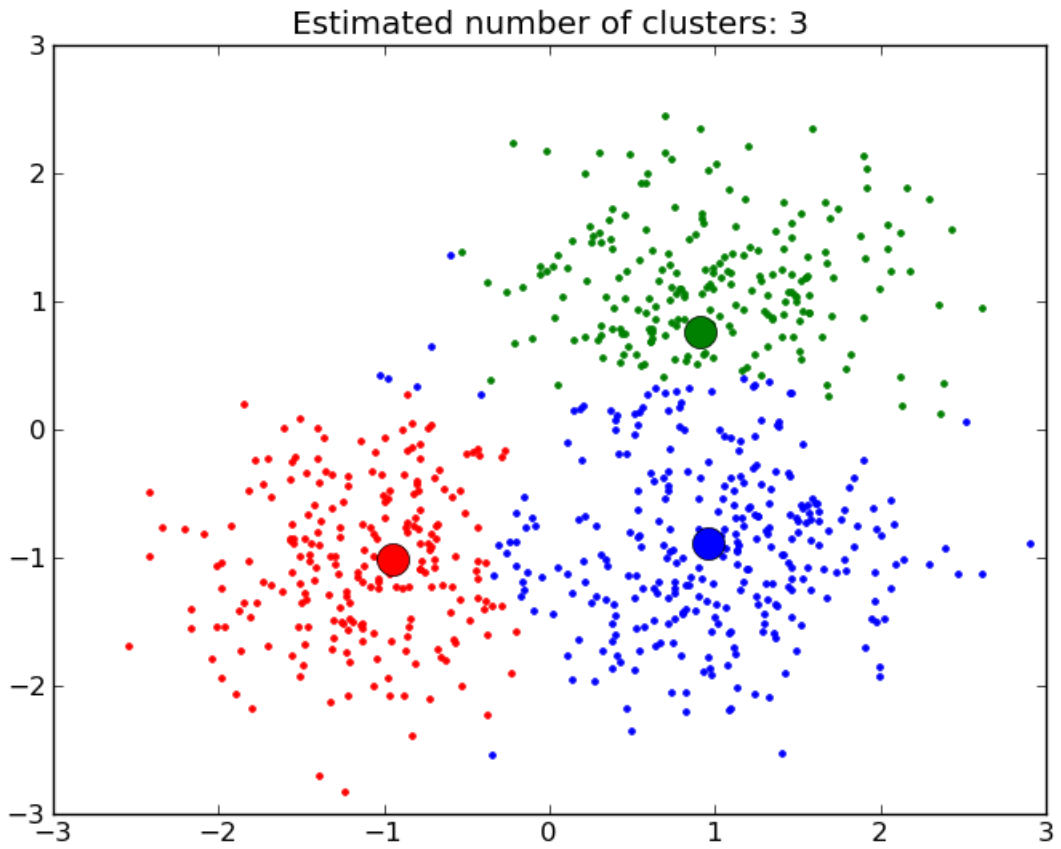
Supervised = computer has access to items that have been classified already (assumed accurate)

Semi-supervised learning



- Use a few labeled points and many unlabeled points to train.
- Common method: *bootstrapping*. Find unlabeled instances similar to the labeled ones, then use all of these to train further.
- Example bootstrapping system: Domain Adaptive Relation Extraction (DARE).

Clustering (unsupervised)



Steps:

1. Choose number of clusters (k)
2. Initialize randomly
3. Minimize distortion

Clustering = classification into machine-made groups

- Hard clustering: points are assigned to clusters during training.
- Soft clustering: each point has a posterior for each cluster.

Knowledge-based learning

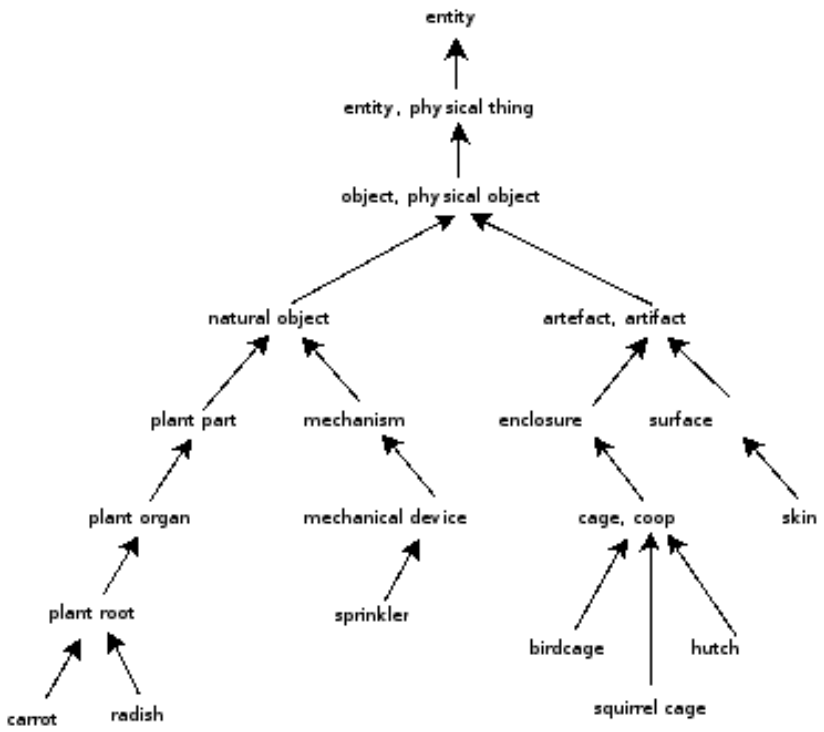
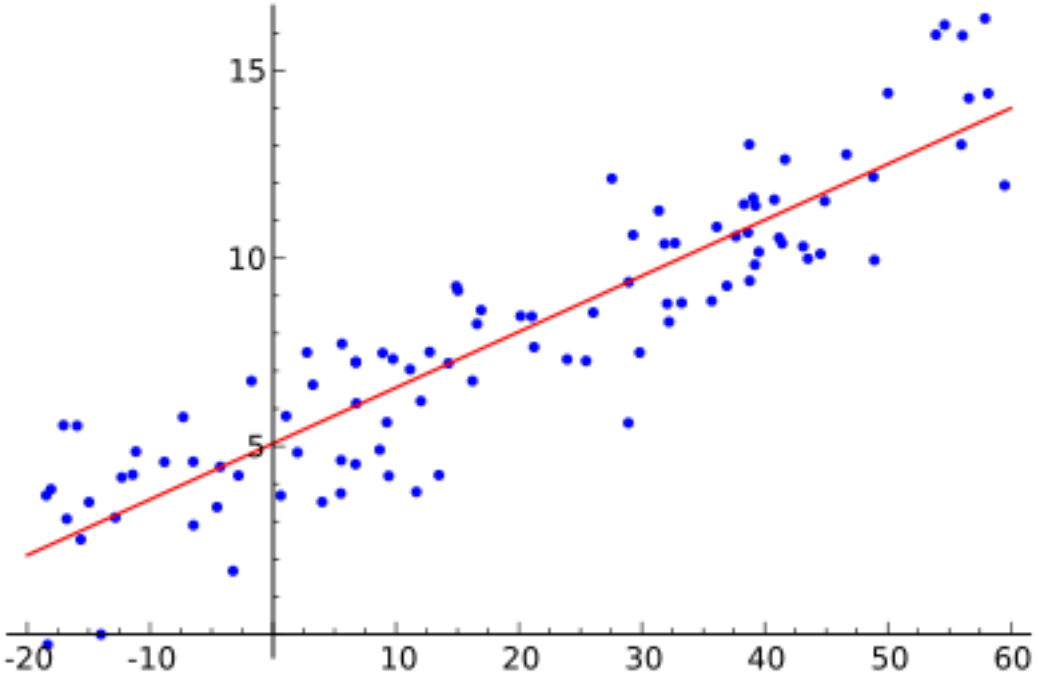


Figure 1. "is a" relation example

- Knowledge-based = unsupervised with a task-general resource
- All training points are still unlabeled.
- For example, allocate one cluster for each SynSet in WordNet.
- Important to acknowledge all task-general resources used when comparing systems.

Regression



Regression = anything goes in (features),

continuous function comes out

Simple linear regression derivation

$$Error(m,b) = \sum_{i=1}^{n} (mx_i + b - y_i)^2$$

$$b = \arg\min_b Error(m,b)$$

$$m = \arg\min_m Error(m,b)$$

$$\frac{\partial Error(m,b)}{\partial b} = \sum_{i=1}^{n} 2(mx_i + b - y_i) = 2n(m\bar{x} + b - \bar{y}) = 0$$

$$\frac{\partial Error(m,b)}{\partial m} = \sum_{i=1}^{n} 2(mx_i + b - y_i)x_i = 2n^2(m\bar{x^2} + b\bar{x} - \bar{xy}) = 0$$

$$b = \bar{y} - m\bar{x}$$

$$m = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2}$$

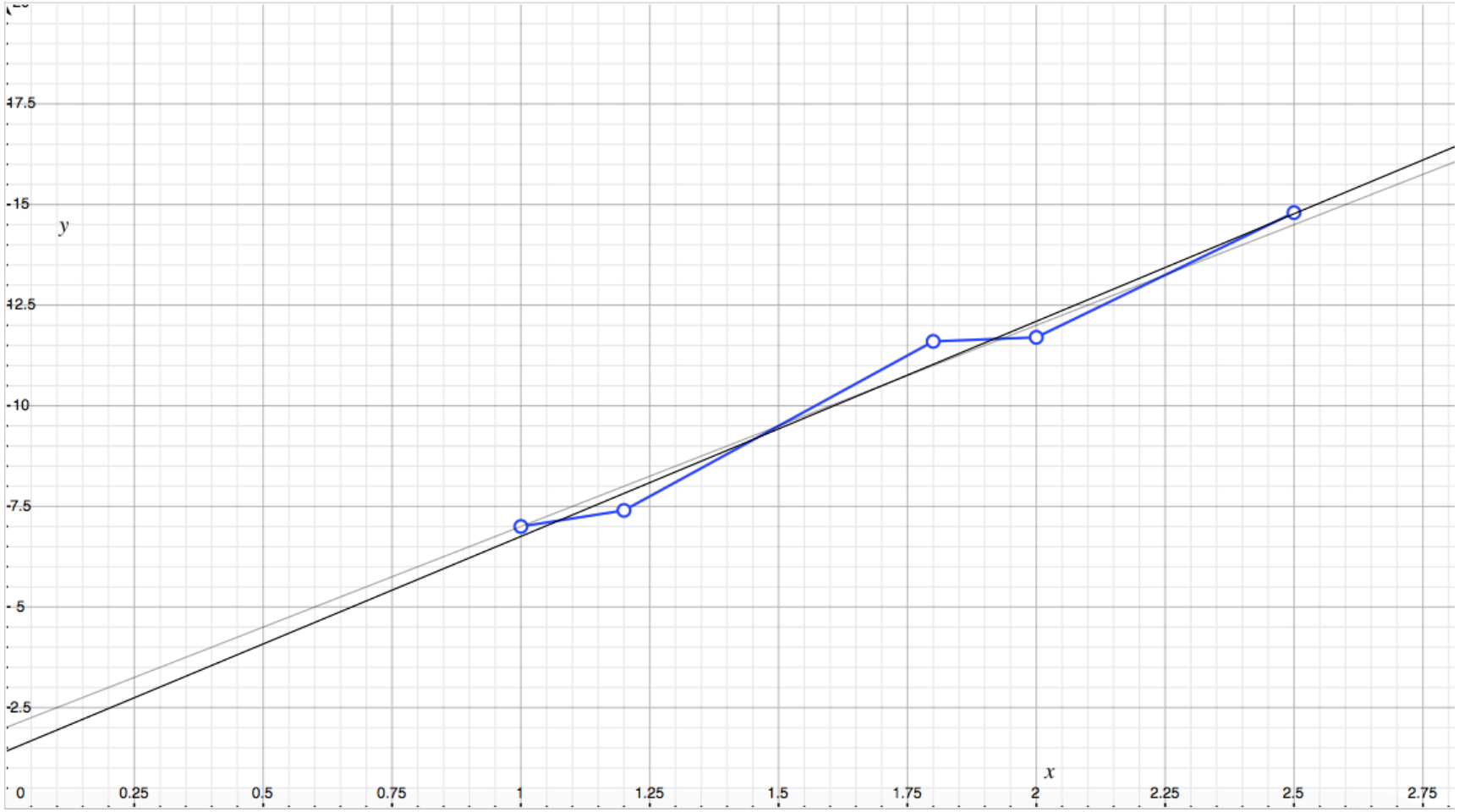
Toy regression example

2D data (x, y):

(1.0, 7.0), (1.2, 7.4), (1.8, 11.6), (2.0, 11.7), (2.5, 14.8)

xy:	7.00,	8.88,	20.88,	23.40,	37.00
x^2:	1.00,	1.44,	3.24,	4.00,	6.25

Toy regression illustration



Exercises

1. Memorize:
   1. classification = anything in, discrete out
   2. clustering = classification into machine-made groups
   3. regression = anything in, continuous out
   4. supervised = example answers are given
   5. knowledge-based = unsupervised with a task-general resource

2. Find a regression line for the 2D data on slide 20.

3. Find a regression line for our Height versus Shoe Size data (data file will be posted online).