Introduction to Probability Theory 5

Clayton Greenberg

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

October 31, 2014

UNIVERSITÄT DES SAARLANDES

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

UNIVERSITÄT
DES
SAARLANDES

Regression exercise

- Both: height = 1.985(shoe) + 91.518, r = 0.774
- Men: height = 2.653(shoe) + 62.247, r = 0.629
- Women: height = 1.435(shoe) + 112.730, r = 0.444

- r is the *correlation coefficient*. It expresses how well the data points fit into a line. +1 means a perfect positive correlation, 0 means no correlation, and -1 means a perfect negative correlation.

UNIVERSITÄT
DES
SAARLANDES

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

classification = anything in, discrete out

clustering = classification into machine-made groups

regression = anything in, continuous out

supervised = example answers are given

knowledge-based = unsupervised with a task-general resource

UNIVERSITÄT
DES
SAARLANDES

Probability Theory Jeopardy

Formulas	Examples	PCFGs	Entropy	Machine Learning
200	200	200	200	200
400	400	400	400	400
600	600	600	600	600
800	800	800	800	800
1000	1000	1000	1000	1000

Final Jeopardy

Formulas for 200

- This fraction gives the probability of a given event. If the outcomes are equiprobable, it becomes the size of the event set divided by size of the outcome set.

- What is "what you want" over "what is possible"?

Formulas for 400

- This expression describes entropy without using the words "expectation" or "surprisal", but still uses the definitions of "expectation" and "surprisal".

- What is the weighted average of the negative logarithm of probability?

Formulas for 600

- This fraction is equal to the probability of "spooks" given "Halloween".

- What is p(spooks, Halloween) / p(Halloween)?

Formulas for 800

- This expression gives the probability of Halloween given spooks using the probability of spooks given Halloween.

- What is

p(spooks | Halloween)*p(Halloween)/p(spooks)?

Formulas for 1000

- This is the result of applying the chain rule twice to p("are you scared").

- What is p(are)*p(you | are)*p(scared | are, you)?

Examples for 200

- Dracula wants an apartment in Saarbrücken that is in an old building AND doesn't have big windows AND has neighbors who do not cook with garlic. Edward wants an apartment in Saarbrücken that is in an old building OR doesn't have big windows OR has neighbors who do not cook with garlic. This person has a greater chance of finding an apartment.

- Who is Edward?

Examples for 400

- Of the words "id", "boo," "the," and "ghost," this word will have the lowest surprisal in a working language model.

- What is "the"?

Examples for 600

- If we have five coins (S, C, A, R, E) that land on heads with probability (0.4, 0.2, 0.9, 1.0, 0.7), this ordering gives the coins in increasing entropy.

- What is RACES?

Examples for 800

- Suppose you buy a lottery ticket for 1€. It has a 1 in 5 chance of winning 1€ and a 1 in 10,000,000 chance of winning 6,000,000€. These odds describe mutually exclusive lucky numbers. This number is the expected value of the ticket (cost included).

- What is -0.20€?

Examples for 1000

- Suppose p(black) = 3/32 and p(cat | black) = 1/24. This is the surprisal of "black cat" in bits.

  - What is 8 bits?

PCFGs for 200

- Upon applications of grammar rules, this symbol can be transformed into "on the hill."

- What is a PP?

PCFGs for 400

- For a string with two viable parses, each with 15 nodes, this is the number of numbers that must be multiplied to compute the probability of the string.

- What is 30?

PCFGs for 600

- This is the number of parses of probability 0.1 that a string would need in order to be more likely than a second string with 3 parses of probability 0.17.

- What is 6?

PCFGs for 800

- These are the assumptions made about rules and trees in order to make calculating the probability of strings possible with a PCFG.

- Rules are independent and trees are mutually exclusive.

PCFGs for 1000

- This is the result of decomposing p(V Det N P Det N | VP) into terms that can be found in a PCFG.

- What is p(V NP | VP)*p(NP PP | NP)*
p(Det N | NP)*p(P NP | PP)*p(Det N | NP) +
p(VP PP | VP)*p(V NP | VP)*
p(Det N | NP)*p(P NP | PP)*p(Det N | NP)?

Entropy for 200

- This value is lower bounded by entropy.

- What is expected symbol code length?

Entropy for 400

- In an encoding in which the expected symbol code length equals the entropy, this value is equal to the code length for each symbol.

- What is surprisal?

Entropy for 600

- This is a distribution with more than two symbols for which the expected symbol code length equals the entropy.

- Many answers possible.

Entropy for 800

- This is the difference between the expected symbol code length and entropy for the Huffman code for the symbols in boo! using the counts from this string.

- What is 0?

Entropy for 1000

- This number is strictly greater than the greatest possible difference between the expected symbol code length for a Huffman code and entropy.

- What is 1?

Machine Learning for 200

- Part-of-speech tagging is an example of this machine learning task.

- What is classification?

Machine Learning for 400

- Determining the relationship between surprisal and reading time is an example of this machine learning task.

- What is regression?

Machine Learning for 600

- These are 5 features that can be used for a food classification task.

- Many answers possible.

Machine Learning for 800

- These three options can be used in the case that data for a supervised task does not exist.

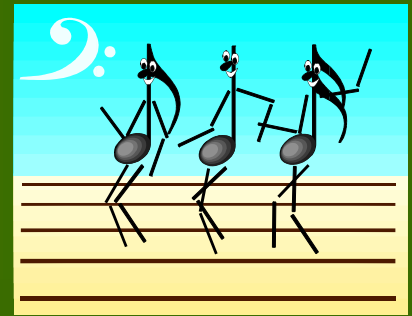- What are annotation, clustering, and regression?

Machine Learning for 1000

- This is an example of a knowledge-based task.

- Many answers possible.

Final Jeopardy

- This is a list of as many green statements as possible from our course. You will receive 200 points for each correct green statement.

- Up to 3600 points are possible.