Probabilistic Machine Learning:
3. Bayesian Concept Learning

Tomasz Kajdanowicz, Przemysław Kazienko

Department of Computational Intelligence
Wroclaw University of Technology

The presentation was inspired by Chapter 3 of Kevin Murphy book "Machine Learning A Probabilistic Perspective", 2012, MIT.
Apropriate agreements to propagate his ideas has been acquired.

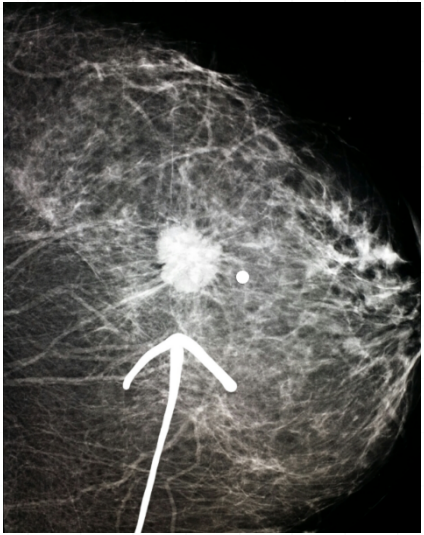
Bayes rule revisited

- Conditional pmf:

$$f_{X|Y}(x|y) = p(X = x|Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

- Bayes' theorem:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}$$



Example: medical diagnosis

- Problem:
  - What is the probability you have cancer, if the **mammogram** test is positive?
  - assume test has 80% sensitivity (if you have cancer, the test will be positive with probability 0.8)
- $p(x = 1|y = 1) = 0.8$, $x = 1$ - the mammogram is positive, $y = 1$ you have breast cancer
- How many breast cancers there are? Ans: the prior probability $p(y = 1) = 0.004$
- Test false alarm: $p(x = 1|y = 0) = 0.1$

$$p(y = 1|x = 1) = \frac{p(x = 1|y = 1)p(y = 1)}{p(x = 1|y = 1)p(y = 1) + p(x = 1|y = 0)p(y = 0)} = 0.031$$

- In other words, if your test is positive, you only have about a 3% chance of actually having breast cancer!

Lets generalize

- generative classifier
- specifies how to generate the data using the class-conditional density $p(x|y = c)$ and the class prior $p(y = c)$

$$p(y = c|x, \Theta) = \frac{p(y = c|\Theta)p(x|y = c, \Theta)}{\sum_{c'} p(y = c'|\Theta)p(x|y = c', \Theta)}$$

- $p(x|y = c, \Theta)$ tells what kind of data we expect to see in each class
- how to infer the unknown parameters Θ of such models?

Bayesian concept learning

- how a child learns to understand the meaning of a word, such as "dog"?
- parents: "look at the cute dog!", very unlikely negative examples

Bayesian concept learning

- "dog" learning example is equivalent to **concept learning**, which in turn is equivalent to binary classification
- we define $f(x) = 1$ if x is an example of the concept C, and $f(x) = 0$ otherwise
- the goal is to learn the indicator function f
- standard binary classification: requires **positive** and **negative** examples
- we will devise a way to learn from positive examples alone

The **number game**

Rules:

- choose some simple arithmetical concept C, such as "prime number" or "a number between 1 and 10"
- I then give you a series of randomly chosen positive examples $D = \{x_1, ..., x_N\}$ drawn from C, and ask you whether some new test case $\tilde{x}$ belongs to C
- for simplicity all numbers are integers between 1 and 100

The **number game**

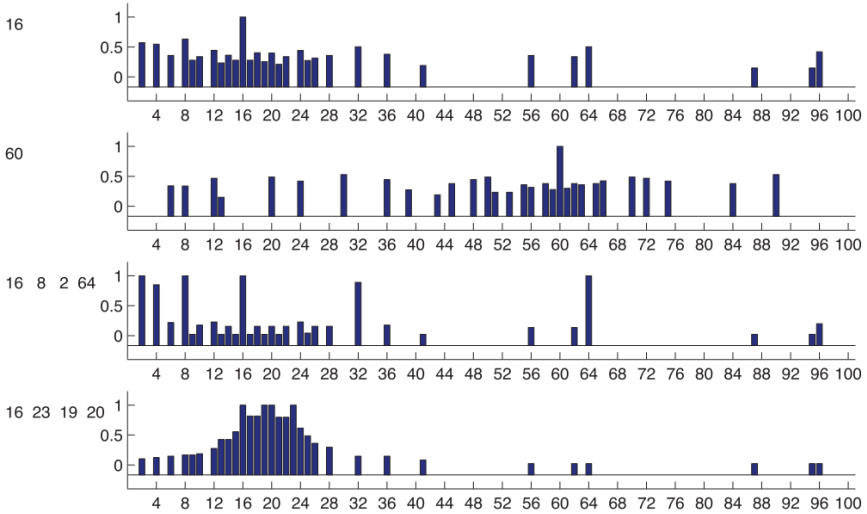
- what if I tell you "16" is a positive example of the concept?
- what other numbers do you think are positive? 17? 6? 32? 99?
- presumably numbers that are similar in some sense to 16 are more likely, but similar in what way?
- but some numbers are more likely than others!
- **posterior predictive distribution**:
  $p(\tilde{x}|D)$, which is the probability that $\tilde{x} \in C$ given the data D for any $\tilde{x} \in \{1, \ldots, 100\}$

The **number game**

- what if I tell you "16", "8", "2" and "64" are positive examples of the concept?
- you say: the concept is "powers of two"
- this is an example of **induction**

- what if I tell you the data is $D = \{16, 23, 19, 20\}$?

The **number game**



How to capture it in machine?

- classic approach: **induction** - suppose a hypothesis space of concepts H (e.g. odd numbers, even numbers, all numbers between 1 and 100, powers of two, all numbers ending in j (for $0 \leq j \leq 9$), etc.)
- **version space**: subset of H that is consistent with the data D
- BUT after seeing $D = \{16\}$, there are many consistent rules, how to combine them to predict if $\tilde{x} \in C$
- good explanation: the Bayesian explanation

Likelihood

- explain why after seeing $D = \{16, 8, 2, 64\}$ we choose $h_{two} \doteq$ "powers of two", and not, say, $h_{even} \doteq$ "even numbers"
- lets assume that examples are sampled uniformly at random from the **extension of a concept** (set of numbers that belong to it)
- then the probability of independently sampling N items (with replacement) from h is

$$p(D|h) = \left[\frac{1}{size(h)}\right]^N = \left[\frac{1}{|h|}\right]^N$$

- REMARK: favors the simplest (smallest) hypothesis consistent with the data: **Occam's razor**

Wiliam Occam

How likelihood works

- let $D = \{16, 8, 2, 64\}$
- then $p(D|h_{two}) = \frac{1}{6}$ (only 6 powers of two less than 100)

  $p(D|h_{even}) = \frac{1}{50}$ (50 even numbers)

- after 4 examples (N=4)

  $h_{two} = (\frac{1}{6})^4 = 0.00077$

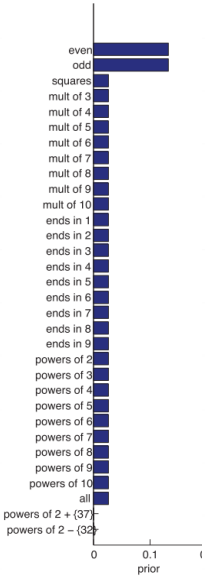
  $h_{even} = (\frac{1}{50})^4 = 0.00000016$

Prior

- let $D = \{16\}$
- $h' \doteq$ "powers of two except 32" is more likely than $h \doteq$ "powers of two"
- h' conceptually unnatural
- to capture such intuition: assign low prior probability to unnatural concepts
- is **subjective**

- in our example lets use a simple prior:
  uniform probability on 30 simple arithmetical concepts, ("prime numbers", "numbers ending in 9", etc.), "even numbers", "odd numbers" more likely and "unnatural" concepts, ("powers of 2, plus 37", "powers of 2, except 32") - low prior weight

Prior in the number game

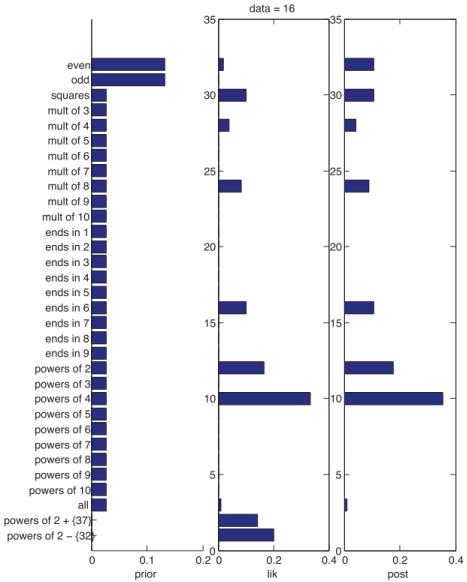


Posterior

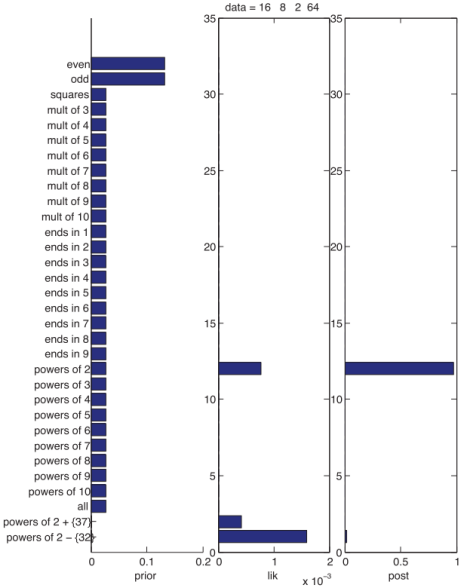
- simply: the likelihood times the prior, normalized

$$p(h|D) = \frac{p(D|h)p(h)}{\sum_{h' \in H} p(D, h')} = \frac{p(h)\mathbb{1}(D \in h)/|h|^N}{\sum_{h' \in H} p(h')\mathbb{1}(D \in h)/|h'|^N}$$

Prior, Likelihood, Posterior in the number game



Prior, Likelihood, Posterior in the number game



MAP estimate

- when we have enough data, the posterior $p(h|D)$ becomes peaked on a single concept, namely the **MAP estimate**

$$p(h|D) \rightarrow \delta_{\hat{h}^{MAP}}(h)$$

$$\hat{h}^{MAP} = argmax_h \; p(h|D)$$

- δ - Dirac measure

MAP estimate

- MAP estimate can be written as

$$\hat{h}^{MAP} = argmax_h \; p(D|h)p(h) = argmax_h \; [\log p(D|h) + \log p(h)]$$

- δ - Dirac measure

MLE

We know, that:

- likelihood term depends exponentially on N
- the prior stays constant

- so, as we get more and more data, the MAP estimate converges towards the **maximum likelihood estimate (MLE)**

$$\hat{h}^{mle} = argmax_h \; p(D|h) = argmax_h \; \log p(D|h)$$

MLE recap

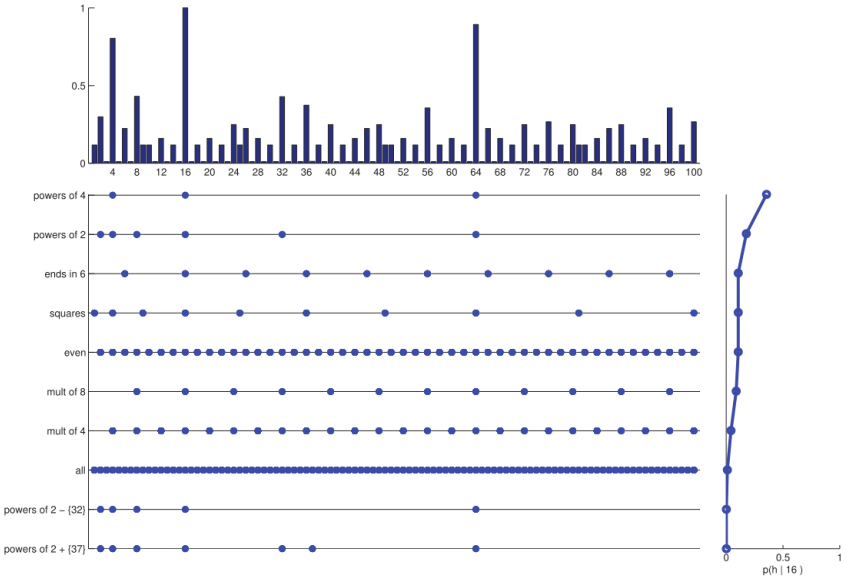
- if the true hypothesis is in the hypothesis space, then the MAP/ ML estimate will converge upon this hypothesis
- hypothesis space is identifiable in the limit, meaning we can recover the truth in the limit of infinite data
- if our hypothesis set is not rich enough to represent the "truth" (which will usually be the case), we will converge on the hypothesis that is as close as possible to the truth

Posterior predictive distribution

$$p(\tilde{x} \in C|D) = \sum_h p(y = 1|\tilde{x}, h)p(h|D)$$

- weighted average of the predictions of each individual hypothesis
- called Bayes model averaging

Posterior predictive distribution



Overfitting

- MAP learning is simple, it cannot explain the gradual shift from similarity-based reasoning (with uncertain posteriors) to rule-based reasoning (with certain posteriors)
- having $D = \{16\}$ and we use the such simple prior, the minimal consistent hypothesis is "all powers of 4" $\rightarrow$ only 4 and 16 get a non-zero probability of being predicted: **overfitting**.
- in Bayesian approach: observing $D = \{16\}$ there are many hypotheses with non-zero posterior $\rightarrow$ predictive distribution is broad, it narrows when more data arrives