T-61.5040

Learning Models and Methods, (5 cr) L

Spring 2007

Lecturer: **Petteri Pajunen**

Course Assistant: **Ville Viitaniemi**

LECTURE 3: 31.1.2007

PROBABILITIES IN LEARNING

Bayesian Inference

- we have seen that data cannot be generalized without any information (NFL thms)
- vague assumptions (e.g. weak regularity) do not help much
- even knowing the correct model may lead to problems such as overfitting
- solution: represent all information about the learning problem correctly

- we will use *probabilistic modeling* to represent the information
- can be justified, though not in a universally acceptable way
- also suggested by situations, where probabilistic models are obviously correct
- seems that no practical alternative exists (learning methods with demonstrable shortcomings do not count)

Learning Problem Again

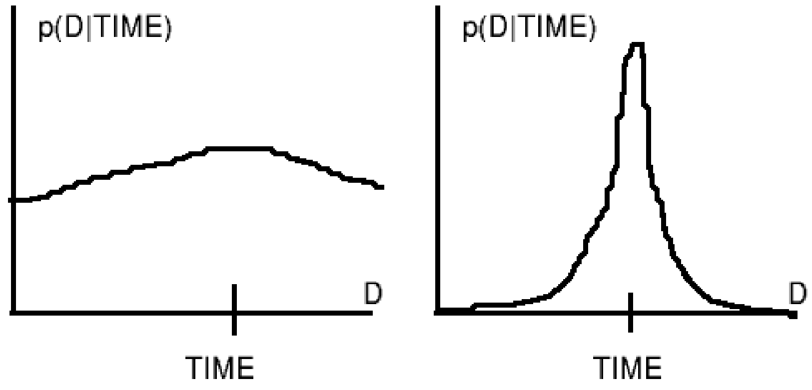
- assume there is a variable whose value is unknown
- denote this by θ, and the set of all possible θ's by Θ
- goal of learning is to get information about θ
- what kind of information we need in addition to data?

- lets go through an example that suggests certain important concepts
- example: what time is it?
- correct but unknown time is θ
- data D is what you read from your watch

- after looking at your watch, do you know what time it is?
- not exactly, because no watch can keep infinitely accurate time
- most of us would expect that the time shown by the watch is on average θ, but can deviate a little from it
- note that different people probably have a different description for the possible deviation

- we must have some information on how D depends on θ
- this *does not come from data* and *is not objective*
- lets pick just one watch: now there is exactly one learning problem
- the owner of the watch should have the best information about the accuracy of the watch

- for example, owner may know that the watch is very accurate
- then others would underestimate the accuracy
- distribution of D is sharp for an accurate watch and wide for a less accurate watch



- what if we just estimate $\hat{\theta} = D$?
- this may be reasonable if the watch is assumed to be accurate enough
- but what if the watch says "3 am"?
- we could all go home and get some sleep if we accept the above solution

- no one really thinks its 3 am, but why?
- because we have information about θ *before looking at the watch*
- this information is also *different for each person*
- to solve the problem properly, one must make a compromise between information about the watch and information about θ

- information about θ is called *prior information*
- it should be obvious that such information exists
- past experience about similar problems, perhaps based on data which is not available anymore
- in the example, several sources of information (having looked at the watch a while ago, knowing the time the lecture is given etc..)

- another problem with the solution $\hat{\theta} = D$
- if the watch is assumed to be inaccurate, then θ may be far from D
- as a single number, $\hat{\theta} = D$ carries no information about this inaccuracy
- in general, result of learning is *not a single value for the unknown* θ

- e.g. you want to catch a bus which departs every hour
- if D is "14:57" you may decide to try to catch the bus
- but if the watch is very likely to be five minutes late, it has a strong influence on the decision
- result of learning should describe the remaining uncertainty in θ

- many learning methods have some of these elements
- for example, a set of solutions $\{f(x|\theta)\}$ contains some information about how D depends on θ
- often no "3 am" safeguards: best-fitting θ is usually selected
- no proper description of remaining uncertainty, at most confidence intervals

- in the watch example, we had to come up with two descriptions of uncertainty
- *prior uncertainty*: what we know about θ
- *how data depends on θ*: what we know about D if θ is given
- in what way should we quantify the uncertainties?

- probability turns out to be a good measure for uncertainty
- probabilities will be used (also) for "one-time events"
- example: what is the probability that it rains tomorrow?
- it is possible to consider such probabilities since the source of uncertainty is *subjective lack of information*
- i.e. use of probabilities is not restricted to events repeatable infinitely many times

Probability as a Measure of Uncertainty

- probability is the *unique* measure of uncertainty, given certain axioms or assumptions
- this can be understood or demonstrated in several ways:
  - by analogy: if something is random when repeated (e.g. coin toss), its uncertainty is represented by a probability distribution
  - Dutch Book Theorem (Ramsey, de Finetti)
  - Cox's Axioms

Betting and Subjective Probability

- what happens if one insists using a non-probability measure for uncertainty?
- non-probability means that your measure does not follow the rules of computing with probabilities
- a betting argument shows that a non-probability measure leads to trouble

- consider a bet T_A which pays the owner 1 EUR if A happens, otherwise it pays nothing
- you would certainly take the bet for free
- also, you would not pay more than 1 EUR for it in any situation
- assume that there is a unique limit price (buy below it and sell above it)

- denote your limit price for T_A as p
- p depends on the actual event A
- lets choose p as an arbitrary non-probability measure $q(A)$
- Ramsey and de Finetti showed that using q leads to irrational behaviour

- exercise problem shows that a non-probability measure $q(A)$ leads to accepting a set of bets *guaranteed to make you lose money*
- this is called a Dutch Book due to bookmakers (persons taking bets e.g. in horseracing) who attempt to set their odds so that whatever happens, they will win money
- Dutch Book Theorem: the possibility of a set of bets guaranteed to make you lose money is equivalent to a non-probability uncertainty system

Cox's Axioms

- Cox derived probability as a measure of uncertainty from a short list of axioms
- further work has reduced and/or changed some of the axioms and as a result, there are a number of different derivations
- some key axioms (but not all):
  1. uncertainties are real numbers $p(A) \in \mathbb{R}$
  2. if A and B are equivalent, then $p(A) = p(B)$
  3. for a certain event A, $p(A) = 1$

- Dutch Book and Cox's Axioms support the use of probability as a measure of uncertainty
- other measures must disagree with at least one of the axioms
- the main axiom one might not accept is the "single real number" axiom
- is there uncertainty about uncertainty?

- there are certain approaches to uncertainty allowing for "sets of probabilities"
- from a practical point of view, one can add new unknowns which are used to give information about θ
- then one can use a probability distribution over all unknowns
- results in a probability over θ anyway

- example: coin toss
- unbiased coin: $p(\text{heads}) = 0.5$ and $p(\text{tails}) = 0.5$
- biased coin: "heads" or "tails" has probability 0.75
- suppose your probabilities are 1/3 for "balanced coin", 1/3 for "heads more probable" and 1/3 for "tails more probable"

- lets compute the total distribution:

$$\begin{aligned} p(\text{tails}) &= \frac{1}{3}p(\text{tails} \mid \text{balance}) \\ &+ \frac{1}{3}p(\text{tails} \mid \text{tails probable}) \\ &+ \frac{1}{3}p(\text{tails} \mid \text{heads probable}) \\ &= \frac{1}{3} * \frac{1}{2} + \frac{1}{3} * \frac{3}{4} + \frac{1}{3} * \frac{1}{4} = \frac{1}{2} \end{aligned}$$

- similarly we obtain $p(\text{heads}) = \frac{1}{2}$
- averaging over the "extra" uncertainty effecticely removes it

- same probabilities for unbiased and possibly biased coin
- do we lose some information here?
- no, if we are to make a decision based on the probabilities
- any decision, which depends on the outcome of the toss must be the same in both cases

- Dutch Book and Cox's Axioms tell us to use only probabilities for quantifying uncertainty
- the learning problem is to find out what is known about θ after data D is observed
- denote this distribution as

$$p(\theta|D, I)$$

- the "extra" variable I means the information that was available before looking at the data

- we omit the variable I from now on, it has no relevance to any calculations
- it simply denotes the source of subjective information that will not be used explicitly
- since $p(\theta|D)$ is the proper description of uncertainty about θ, nothing more can be done using only data and information I
- Bayesian Inference is mainly about computing this distribution

- the notation $p(A|B)$ means a *conditional probability*
- it should be understood as "probability of A, given that B is known"
- computing with conditional probabilities is technically easy
- but thinking in terms of them is important but not always easy (see exercise problems this week)

- why not use traditional statistics?
- there the emphasis is on the distribution $p(D|\theta)$ (likelihood as a function of θ)
- this is the "uncertainty" of data, given θ
- but data is observed and θ is unknown: $p(D|\theta)$ cannot be used on its own

- ad-hoc methods are used which are based on taking averages over all possible data
- e.g. confidence intervals: some interval $[f(D), g(D)]$ has 95 percent confidence if it contains θ 95 percent of the time, averaged over all possible D
- this does *not* mean that θ is in $[f(D), g(D)]$ with 95 percent probability for a given D!
- also, $p(D|\theta)$ *as a function of* θ is not a proper description of uncertainty in θ

- notice that $p(\theta|D)$ is a probability distribution as a function of θ
- an interval can easily be defined using $p(\theta|D)$ so that it will contain θ with a given probability
- the probability $p(D|\theta)$ is very important part in computing $p(\theta|D)$, but it is not enough on its own
- the remaining part is $p(\theta)$ which describes the uncertainty in θ *before* seeing data

- since $p(\theta)$ can and should be chosen based on subjective information, its use has caused some critisism
- but choosing $p(D|\theta)$ is also subjective
- avoiding subjective choices is impossible, since the model comes from information available before seeing data
- Bayesian inference has some ways of attempting to obtain objective results (more on this later)

- example: line fitting
- model: $y = \mu x + \beta + n$ where $n \sim N(0, \sigma^2)$
- in words, fit a line to data assuming Normally distributed errors
- unknown $\theta = (\mu, \beta)$: it might include σ^2, but for now we assume that it is known

- what is $p(D|\theta)$?
- data is a set of pairs $(x_i, y_i), i = 1, \ldots, n$
- if they are independently generated, then

$$\begin{aligned} p(D|\theta) &= \prod_{i=1}^{n} p(x_i, y_i|\theta) \\ &= \prod_{i} C e^{-\frac{1}{2\sigma^2}(y_i - \mu x_i - \beta)^2} \end{aligned}$$

- the prior $p(\theta)$ depends on what else you assume, and could be any probability distribution
- note that standard regression methods don't use $p(\theta)$
- the posterior $p(\theta|D)$ is computed next week

- summary:
  - learning requires information about uncertain quantities
  - probability is the right way to quantify this (according to Dutch Book, Cox's Axioms)
  - both $p(\theta)$ and $p(D|\theta)$ are subjective
  - result of learning is $p(\theta|D)$, which is not an estimate of θ as a single number