T-61.5040

Learning Models and Methods, (5 cr) L

Spring 2007

Lecturer: **Petteri Pajunen**

Course Assistant: **Ville Viitaniemi**

LECTURE 13: 25.4.2007

MAKING DECISIONS

Decision Theory

- so far we have concentrated on computing a posterior over the unknown quantity θ
- the distribution correctly describes what is known about θ, as long as we believe our model
- but the distribution must be put to use: there is no reason to compute it otherwise
- a general way of quantifying the use of posterior is decision making

- implicit decisions are made even when computing the posterior and/or specifying the model
- otherwise, one should e.g. use infinitely many input variables and keep simulating the posterior indefinitely
- decisions: simplify model, stop simulation
- reasons: too many input variables cost, computation costs

- example: crossing a river
- you estimate that an old bridge fails with probability 0.01 if you try to cross it
- using a new bridge is possible, but takes more time
- what information do you need to make a decision?

- knowing the failure probability is not enough
- if failing bridge means certain death, you probably decide to use the new bridge
- but if failure means falling a meter or two into water, your decision may be different
- probabilities are the same in both cases

- in general, we need to express the "value" of an outcome following a decision
- in the bridge example, we need

	fail	not fail
old	u_1	u_2
new	u_3	u_4

- lets denote this as a function $U(a, \theta)$ where $a \in \{\text{old,new}\}$ and $\theta \in \{\text{fail, not fail}\}$

- the *choice* $a \in \{a_1, \ldots, a_k\}$ denotes the decision
- θ denotes the uncertain outcome
- if U has monetary value, then it is called *payoff*
- later, we use a nonlinear function of money called *utility*

Decision Criteria

- many heuristic criteria have been proposed
- easy to demonstate nonoptimality by counterexamples
- *nonstochastic criteria* ignore the posterior
- *stochastic criteria* use the information in the posterior

- examples of nonstochastic criteria:
- *maximin:*

$$\text{argmax}_a \min_\theta U(a,\theta)$$

- *maximax:*

$$\text{argmax}_a \max_\theta U(a,\theta)$$

- *minimax regret:*

$$\text{argmin}_a \left[\max_{a_i \neq a} \max_\theta \left(U(a_i,\theta) - U(a,\theta) \right) \right]$$

- stochastic criteria use the probability distribution over the outcomes
- some examples:
- *modal outcome:* choose the highest payoff of the most probable outcome
- *expected value/payoff:* choose the highest expected payoff

- including the posterior of the outcome should improve the decisions
- but there are several different stochastic criteria
- only one of them can be correct
- counterexamples suggest some problems with modal outcome and expected payoff

- consider following decisions, outcomes, and payoffs:

	θ_1	θ_2	θ_3
$p(\theta\|y)$	0.25	0.3	0.45
a_1	500	500	200
a_2	100	100	600

- *modal outcome*: mode is θ_3, and the payoff-maximizing decision is a_2

- modal outcome is essentially point estimation
- therefore it suffers from most point estimation problems
- for example, consider combining θ_1 and θ_2
- now the combined outcome is the mode and the decision a_1 has the highest payoff

- expected payoff is a special case of the optimal decision rule, but is not generally optimal
- consider the decision between receiving 1000 €, or flipping a coin and either receiving 1.1 M€ or having to pay 1 M€
- the first choice has a smaller expected value
- but you would probably choose it anyway

Utility

- *utility* fixes the expected value criterion so that it becomes optimal
- given a few axioms, there exists a numerical measure called utility:
  - the optimal decision is one that maximizes *expected utility*
  - utility is *subjective*: different decision-makers can have different utilities
  - generally a nonlinear function $U(C)$ of monetary payoff C

- $U(C)$ can be found using *certainty equivalents*
- example: cointoss with payoffs 0 and 1 EUR
- certainty equivalent is the maximum amount C you are willing to pay for the bet
- then utility $U(C) = 0.5U(0) + 0.5U(1)$
- by a series of such bets, one can approximate the function $U(C)$

- an optimal decision out of a finite set of choices $a_1, \ldots, a_k$ is the one that maximizes expected utility
- denote the utility of decision a_i given the outcome θ by $U(a_i, \theta)$
- the expected utility for choice a_i is

$$\mathrm{E}(U(a_i, \theta)) = \int U(a_i, \theta) p(\theta|y) d\theta$$

- in practice, one may want to define the monetary payoff $C(a_i, \theta)$ and then use $U(C(a_i, \theta))$

- important points:
  - optimal decisions require *subjective probabilities* and *subjective utility*
  - given certain axioms, it is optimal to maximize expected utility
  - approximate utility obtained by certainty equivalents

- bridge example, failing probability is 0.01:

	fail	not fail	EU
$p(\theta\|y)$	0.01	0.99	
old	-10	10	9.8
new	9	9	9

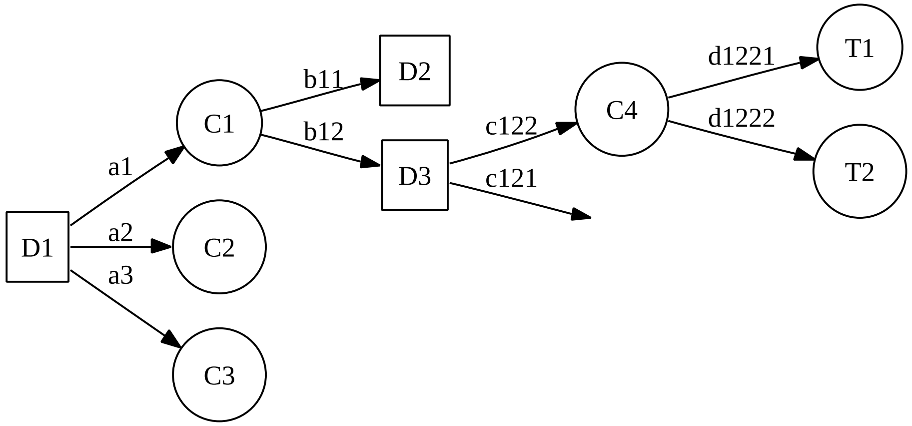
- what if failing is less pleasant?

	fail	not fail	EU
$p(\theta\|y)$	0.01	0.99	
old	-100	10	8.9
new	9	9	9

- maximizing expected utility is optimal in theory
- but decision theory can be difficult to apply
- *sequential decisions* (decisions depend on earlier decisions) are computationally very heavy
- a corresponding *decision tree* explodes in size as a function of number of decisions (e.g. playing chess)

- small enough decision trees can be solved
- a decision tree alternates between *decision and chance nodes*
- decisions are made at decision nodes, leading to chance nodes
- chance node has random outcomes leading to another decision node

- example: first choose from a_1, a_2, a_3
- choice a_i is followed by an outcome b_{i1} or b_{i2}
- outcome b_{ij} follows a choice of c_{ij1} or c_{ij2}
- choice c_{ijk} leads to *terminal outcomes* d_{ijk1}, d_{ijk2}
- assume that the utilities of terminal outcomes are known



- how to make the decision $D1$? We don't know the utilities of $C1, C2, C3$
- the solution requires *dynamic programming*
- start by computing the expected utility on lowest chance nodes (e.g. $C4$)
- use terminal outcomes to do this

- using expected utilities such as $\mathrm{E}(U(C4))$, we can make optimal decisions at nodes $(D2, D3)$
- replace $D2, D3$ by the maximum expected utility of the lowest chance nodes
- continue by computing the EU of $C1, C2, C3$
- this allows us to make the decision $D1$ optimally

Model Selection

- not needed in principle, since the only correct model $p(\theta, D)$ is defined by the subjective prior uncertainty
- one cannot have several conflicting prior uncertainties, so a unique model is obtained in theory
- but there are various reasons why selecting a model from a set of candidates may be useful in practice
  - if θ is point estimated, then risk of overfitting may be reduced by model selection
  - a simple model is required for practical reasons (computational and data collection costs)

- not using correct $p(\theta, D)$ is a compromise between
  - cost of obtaining the wrong posterior uncertainty and predictive distribution
  - benefits such as less computation and smaller data collection costs
- probabilities cannot tell you how much computation costs, or how much having the wrong posterior costs
- decision theory solves the problem in principle

- most model selection heuristics use only probabilities
- examples: MDL, MML, AIC, Evidence framework
- they generally disagree on the same data and model!
- most can be thought as giving a “penalty” to complex models

- model selection should in principle be treated as a decision
- therefore one should pick the model which maximizes expected utility
- this is easier said than done, but in some applications costs are important enough to warrant this approach
- the key element is the utility corresponding to each candidate model

- the utility depends on the problem, but may include for example
  - data collection costs: simpler models might use only part of the data, so the unused data does not have to be collected at all
  - computational costs: solving a complicated model is generally costly
  - model accuracy: how well the model performs in making predictions

- a full-flegded decision approach is rarely done explicitly
- but most of the above costs are implicitly considered in practice
- nobody uses a model that:

  ... takes forever to solve
  ... requires data that is much too expensive
  ... could be made much more accurate with little extra cost

- example: assessing quality of healthcare (Fouskakis and Draper, 2005)
- hospitals are modeled as processes which map input variables (results of clinical tests) to an outcome (death within 30 days)
- collecting variables is expensive (36 variables initially considered)
- utility-based model selection was used to select a subset of input variables

- this requires choosing the utility of each subset
- it includes a negative term, which can be obtained from actual costs of performing the necessary clinical tests
- the utility of the model accuracy was obtained by testing the performance, and eliciting a monetary payoff from health experts for the different outcomes

Posterior Approximation

- also a decision (which wrong posterior to compute?)
- often the decision to use the correct posterior has a very small utility (e.g. high computational cost)
- approximations are a compromise between low computational costs and wrong results
- if the cost of wrong results can be high, utilities should be considered in posterior approximation

- example: point estimation, decide which θ_0 approximates the posterior $p(\theta|y)$
- the expected utility is

$$\mathrm{E}(U(\theta_0,\theta)) = \int U(\theta_0,\theta)p(\theta|y)d\theta$$

- certain utilities lead to familiar point estimates

- for example, $U(\theta_0, \theta) = -(\theta_0 - \theta)^2$ leads to least-squares estimate
- $U = \delta(\theta_0 - \theta)$ leads to the MAP estimate, as before
- in this sense choosing a point estimate means making a decision
- note that none of the point estimates are generally better than others: they arise as a function of $U(\theta_0, \theta)$ and the posterior

- in general, choose $q(\theta)$ to approximate $p(\theta|y)$
- a utility $U(q(\theta), \theta)$ is called
  - *proper* if $\mathrm{E}(U(q, \theta))$ is maximized by $q = p(\theta|y)$
  - *local* if $U(q(\theta), \theta_i) = u_i(q(\theta_i))$ (utility at θ_i depends only on $q(\theta_i)$)
- Theorem (see Bernardo's book for details): a continuosly differentiable, local, and proper utility (for distributions) is

$$U(q, \theta) = A \log q + B(\theta)$$

- what is lost in utility by choosing q instead of p?
- if the utility is local, proper, and smooth, then we lose

$$\begin{aligned} &\mathrm{E}[U(p, \theta) - U(q, \theta)] = \\ &\quad = \mathrm{E}[A \log p + B(\theta) - A \log q - B(\theta)] = \\ &\quad = \int [A(\log p - \log q)]\, p(\theta|y) d\theta = \\ &\quad = A \int p \log \frac{p}{q} d\theta = AD(p\|q) \end{aligned}$$

- this is Kullback-Leibler divergence, up to multiplication by A

- example: true/false statements in T-61.5040 exam
- a set of statements is given, and your answer to each is a probability $q(\text{TRUE})$
- your *subjective probability* $p(\text{TRUE})$ can be different!
- you will get $1 - 4(1-q)^2$ points if the statement is true, and $1 - 4q^2$ points if it is false

- some properties:
  - to maximize expected number of points, choose $q = p$
  - answering $q = 0.5$ gives zero points (as does not answering at all)
  - answering $q \in \{0, 1\}$ gives 1 or -3 points
  - guessing $q \in \{0, 1\}$ gives on average -1 points