Decision making under uncertainty

Michael U. Gutmann

Probabilistic Modelling and Reasoning (INFR11134)
School of Informatics, The University of Edinburgh

Autumn Semester 2025

Recap

- We modelled actions as interventions in the data generating process.
- For DAGs, intervening on a node removes all incoming arrows into the node.
- We discussed how to compute/predict the effect of interventions.
- However, we have not yet discussed which action to choose in the face of uncertainty.
- This is the topic of decision theory, discussed here.

Program

1. Brief introduction to decision theory
2. Its use in statistics and machine learning

Program

1. Brief introduction to decision theory
   - Loss and decision principles
   - Connection to causality
   - Mild cognitive impairment example
2. Its use in statistics and machine learning

Loss

- We frame decision making under uncertainty as a game: I first take an action $\mathbf{a}$. Then a quantity $\mathbf{h}$ that was previously hidden to me is revealed, after which I incur the loss $\ell(\mathbf{h}, \mathbf{a})$.
- Since $\mathbf{h}$ is unknown/unobserved when we take the action $\mathbf{a}$, we are dealing with a case of decision making under uncertainty.
- Cognitive impairment example:
  - Hidden/unobserved quantity: cognitive impairment
  - Action: possible life style changes and cognitive training
- Classification example:
  - Hidden/unobserved quantity: true class label
  - Action: estimate of the class label
- Lots of further examples in statistics and beyond (e.g. finance or healthcare)

Decision principles

- Popular decision principles are the minimax and the expected loss (utility) principle.
- Minimax principle:
  - Choose the action that minimises the maximum loss $\max_{\mathbf{h}} \ell(\mathbf{h}, \mathbf{a})$.
  - Pro: Useful in case of adversaries. Does not depend on the distribution of $\mathbf{h}$.
  - Con: Often too pessimistic, resulting in overly conservative actions
- Expected loss principle:
  - Choose the action that minimises the expected loss $\mathbb{E}_{\mathbf{h}}[\ell(\mathbf{h}, \mathbf{a})]$, called the risk.
  - Pro: Broadly useful, combines impact of action (loss) with chance of occurrence (expectation).
  - Con: Action depends on the distribution of $\mathbf{h}$.
- We here focus on the expected loss principle.

Expected loss

- Risk $\mathbb{E}_{\mathbf{h}}[\ell(\mathbf{h}, \mathbf{a})]$ depends on the distribution of $\mathbf{h}$.
- Provides considerable flexibility:
  - It can be past data on $\mathbf{h}$ that we have collected.
  - It can be our personal/prior belief of what $\mathbf{h}$ may be.
  - It can be derived from some indirect evidence about $\mathbf{h}$ in the form of $\mathbf{x} \sim p(\mathbf{x}|\mathbf{h})$.
- Leads to different schools of decision making.
- In the last case, we compute the expectation with respect to the conditional

$$p(\mathbf{h}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{h})p(\mathbf{h})}{p(\mathbf{x})} \tag{1}$$

- Resulting expected loss

$$R(\mathbf{a}|\mathbf{x}) = \mathbb{E}_{p(\mathbf{h}|\mathbf{x})}[\ell(\mathbf{h}, \mathbf{a})] \tag{2}$$

is sometimes called the posterior expected loss / posterior risk.

From action to policy

- Denote by $f(\mathbf{h})$ the distribution of $\mathbf{h}$ that we average over and by $R(\mathbf{a}; f)$ the corresponding risk.
- The optimal action is

$$\mathbf{a}^*(f) = \operatorname*{argmin}_{\mathbf{a}} R(\mathbf{a}; f) \tag{3}$$

$$= \operatorname*{argmin}_{\mathbf{a}} \mathbb{E}_{f(\mathbf{h})} [\ell(\mathbf{h}, \mathbf{a})] \tag{4}$$

- In case of posterior risk, the optimal action depends on $\mathbf{x}$

$$\mathbf{a}^*(p(\mathbf{h}|\mathbf{x})) = \operatorname*{argmin}_{\mathbf{a}} R(\mathbf{a}; p(\mathbf{h}|\mathbf{x})) \tag{5}$$

$$= \operatorname*{argmin}_{\mathbf{a}} \mathbb{E}_{p(\mathbf{h}|\mathbf{x})} [\ell(\mathbf{h}, \mathbf{a})] \tag{6}$$

- We will overload notation and denote $\mathbf{a}^*(p(\mathbf{h}|\mathbf{x}))$ by $\mathbf{a}^*(\mathbf{x})$.
- Called a policy, mapping evidence $\mathbf{x}$ to actions.

Causal decision theory

- A more recent decision principle is framed in terms of causality (Pearl, Causality, Ch 4).
- Choose the action that minimises

$$\mathcal{R}(\mathbf{a}) = \mathbb{E}_{p(\mathbf{y};do(\mathbf{a}))}\left[\ell_o(\mathbf{y})\right] \tag{7}$$

  where $\ell_o(\mathbf{y})$ is the loss for *outcome* $\mathbf{y}$.
- Note the use of the postinterventional distribution $p(\mathbf{y}; do(\mathbf{a}))$.
- It might incorporate evidence already.
- We may include an action-dependency in the loss, so that we have $\ell_o(\mathbf{y}, \mathbf{a})$
- In some cases, we can relate $\mathcal{R}(\mathbf{a})$ to the expected loss.

Rewriting it in terms of expected loss

- If outcome $\mathbf{y}$ depends on unobserved variables $\mathbf{h}$, we have

$$p(\mathbf{y}; do(\mathbf{a})) = \int p(\mathbf{y}, \mathbf{h}; do(\mathbf{a}))\mathrm{d}\mathbf{h} \tag{8}$$

$$= \int p(\mathbf{h}; do(\mathbf{a}))p(\mathbf{y}|\mathbf{h}; do(\mathbf{a}))\mathrm{d}\mathbf{h} \tag{9}$$

$$= \mathbb{E}_{p(\mathbf{h}; do(\mathbf{a}))}\left[p(\mathbf{y}|\mathbf{h}; do(\mathbf{a}))\right] \tag{10}$$

- Assume that the intervention happens after $\mathbf{h}$ is generated. Then $p(\mathbf{h}; do(\mathbf{a})) = p(\mathbf{h})(*)$.
- Under this assumption

$$\mathcal{R}(\mathbf{a}) = \mathbb{E}_{p(\mathbf{y},\mathbf{h}; do(\mathbf{a}))}\left[\ell_o(\mathbf{y}, \mathbf{a})\right] \tag{11}$$

$$= \mathbb{E}_{p(\mathbf{h}; do(\mathbf{a}))}\mathbb{E}_{p(\mathbf{y}|\mathbf{h}; do(\mathbf{a}))}\left[\ell_o(\mathbf{y}, \mathbf{a})\right] \tag{12}$$

$$\overset{(*)}{=} \mathbb{E}_{p(\mathbf{h})}\mathbb{E}_{p(\mathbf{y}|\mathbf{h}; do(\mathbf{a}))}\left[\ell_o(\mathbf{y}, \mathbf{a})\right] \tag{13}$$

- This has the form of the expected loss with

$$\ell(\mathbf{h}, \mathbf{a}) = \mathbb{E}_{p(\mathbf{y}|\mathbf{h}; do(\mathbf{a}))}\left[\ell_o(\mathbf{y}, \mathbf{a})\right] \tag{14}$$

Mild cognitive impairment (MCI) example

- 70 year old man: prior MCI probability 5/45, posterior MCI probability 2/3 after the online test.
- Loss function $L(h, a)$ with $h \in \{0, 1\}$ indicating whether he has MCI or not and $a \in \{0, 1\}$ whether he takes action or not.
- Assume taking action means life-style changes (less alcohol, more exercise) and attending cognitive training sessions for four months.
- Loss is measured in terms of "quality-adjusted life years" (QALY).
- 1 QALY means one year in perfect health. 0.5 QALY can mean one year which is only half enjoyable, e.g. due to pain.

MCI example: loss function

- Loss function as a table

Action/MCI	$h = 1$	$h = 0$
$a = 1$	c+(1-e)H	c
$a = 0$	H	0

- H is the harm of having MCI unmanaged over the decision horizon (1 year). Assume its 14 low quality days $\rightarrow H = 14/365 = 0.038$ QALY.
- e is the reduction of the harm. Assume $e = 0.25$.
- c is the loss in quality time due to taking the action (e.g. travel to training sessions, attending them instead of doing something fun, social stigma etc). Assume 0.002 QALY.

MCI example: optimal action

- Loss function

Action/MCI	$h = 1$	$h = 0$
$a = 1$	0.0305	0.002
$a = 0$	0.038	0

- Prior risk (units of QALY):

$$R(a = 1; p(h)) = 5/45 \cdot 0.0305 + 40/45 \cdot 0.002 = 0.0051$$
$$R(a = 0; p(h)) = 5/45 \cdot 0.038 + 40/45 \cdot 0 = 0.0042$$

No action $a = 0$ has lower risk and is thus better.

- Posterior risk (units of QALY):

$$R(a = 1; p(h|x)) = 2/3 \cdot 0.0305 + 1/3 \cdot 0.002 = 0.0209$$
$$R(a = 0; p(h|x)) = 2/3 \cdot 0.038 + 1/3 \cdot 0 = 0.0253$$

Taking action $a = 1$ has lower risk and is thus better.

Program

1. Brief introduction to decision theory
   - Loss and decision principles
   - Connection to causality
   - Mild cognitive impairment example
2. Its use in statistics and machine learning

Program

1. Brief introduction to decision theory

2. Its use in statistics and machine learning
   - Common loss functions
   - Applications of the 0-1 loss
   - Applications of the log-loss

Common loss functions

- Loss functions $\ell(\mathbf{h}, \mathbf{a})$ are best tailored to the problem at hand.
- There are, however, a number of "standard" loss functions that are widely used in statistics and machine learning.
- They include
  - Quadratic loss
  - Absolute error loss
  - Zero-one loss
  - Log-loss
- For each loss, we next derive the optimal action/policy.
- We denote the distribution of $\mathbf{h}$ by f.

Quadratic loss

- For simplicity, assume h and a are one-dimensional.
- The quadratic loss is

$$\ell(h, a) = (h - a)^2 \tag{15}$$

- Optimal action

$$a^*(f) = \operatorname*{argmin}_a \mathbb{E}_{f(h)} \left[(h - a)^2 \right] \tag{16}$$

- May be continuous-valued even if h is discrete.

Quadratic loss

- To solve the optimisation problem, let $m = \mathbb{E}_f[h]$ be the expected value of h under f. The loss is

$$\begin{aligned}
\mathbb{E}_{f(h)}\left[(h-a)^2\right] &= \mathbb{E}_{f(h)}\left[(h-m+m-a)^2\right] && (17)\\
&= \mathbb{E}_{f(h)}\Big[(h-m)^2 + 2(h-m)(m-a) + \\
&\quad + (m-a)^2\Big] && (18)\\
&= \mathbb{E}_{f(h)}\left[(h-m)^2\right] + 0 + (m-a)^2 && (19)\\
&= \mathbb{V}(h) + (m-a)^2 && (20)
\end{aligned}$$

- We thus have

$$a^*(f) = \operatorname*{argmin}_a (m-a)^2 = m$$

- The optimal action is the expected value of h wrt f: $\mathbb{E}_f[h]$.
- Generalises to multidimensional case: $\mathbf{a}^*(f) = \mathbb{E}_f[\mathbf{h}]$.

Absolute error loss

- We assume that h and a are one-dimensional.
- We here choose a loss that increases linearly as a deviates from h, rather than quadratically.
- The easiest is $\ell(h, a) = |h - a|$.
- Optimal action

$$a^*(f) = \operatorname*{argmin}_a \mathbb{E}_{f(h)}\left[|h - a|\right] \tag{21}$$

- We can show that $a^*(f)$ is the median of $f(x)$, i.e. any point at which probability mass is equally split between the left and the right.

  More formally, let $F(\alpha) = \mathbb{P}(h \leq \alpha)$ be the cumulative distribution function of f. The median is any number m that satisfies $F(m^-) \leq 0.5 \leq F(m)$ where m^- is the lower limit of F at m. Not unique in case of discrete random variables.

Absolute error loss (proof, not examinable)

- We consider the more general case

$$\ell(h, a) = \begin{cases} k_1(h-a) & \text{if } h \geq a \\ k_2(a-h) & \text{if } h < a \end{cases} \quad \text{with } k_1, k_2 > 0 \qquad (22)$$

  This loss is called the quantile or pinball loss.

- The posterior expected loss is (assuming pdfs)

$$R(a; f) = \int_a^\infty k_1(h-a) f(h) dh + \int_{-\infty}^a k_2(a-h) f(h) dh$$

- Taking derivatives gives (using Leibniz rule)

$$\begin{aligned} \frac{d}{da} R(a; f(h)) &= -k_1 \int_a^\infty f(h) dh + k_2 \int_{-\infty}^a f(h) dh \\ &= -k_1 (1 - \int_{-\infty}^a f(h) dh) + k_2 \int_{-\infty}^a f(h) dh \end{aligned}$$

Absolute error loss (proof, not examinable)

- *continued from previous slide:*

$$\begin{aligned}\frac{d}{da}R(a;f) &= -k_1 + (k_1+k_2)\int_{-\infty}^{a} f(h)dh \\ &= -k_1 + (k_1+k_2)\mathbb{P}(h \leq a)\end{aligned}$$

- Setting the derivative to zero gives the necessary condition for an optimum:

$$\mathbb{P}(h \leq a^*) = \frac{k_1}{k_1+k_2} \tag{23}$$

- The second derivative of the expected loss is $(k_1+k_2)f(a)$. Assuming that $f(a^*) > 0$, a^* is a unique minimum. (For a^* where $f(a^*) = 0$, we have multiple solutions, which is the case of discrete-valued random variables).
- This means that $a^*(f)$ is the $k_1/(k_1+k_2)$-th quantile of $f(h)$.
- For $k_1 = k_2$, we obtain the 0.5 quantile, i.e. the median.

Zero-one loss

- The unobserved variable $\mathbf{h}$ and $\mathbf{a}$ may be multivariate.
- For discrete $\mathbf{h}$ and $\mathbf{a}$, the zero-one loss is

$$\ell(\mathbf{h}, \mathbf{a}) = \mathbb{1}(\mathbf{h} \neq \mathbf{a}) = 1 - \mathbb{1}(\mathbf{h} = \mathbf{a}) = \begin{cases} 1 & \text{if } \mathbf{h} \neq \mathbf{a} \\ 0 & \text{if } \mathbf{h} = \mathbf{a} \end{cases} \tag{24}$$

- For continuous $\mathbf{h}$ and $\mathbf{a}$, it is defined as $\ell(\mathbf{h}, \mathbf{a}) = 1 - \delta(\mathbf{h} - \mathbf{a})$ where $\delta(.)$ is the Dirac-delta distribution (picture as Gaussian with vanishingly small variance)
- Optimal action

$$\mathbf{a}^*(f) = \underset{\mathbf{a}}{\operatorname{argmin}} \, \mathbb{E}_{f(\mathbf{h})} \left[\ell(\mathbf{h}, \mathbf{a}) \right] \tag{25}$$

Zero-one loss

- To solve the optimisation problem, note that

$$\mathbb{E}_{f(\mathbf{h})}\left[\ell(\mathbf{h}, a)\right] = 1 - f(\mathbf{h} = \mathbf{a}) \tag{26}$$

  This holds for continuous or discrete quantities.
- Since
  $\operatorname{argmin}_{\mathbf{a}} 1 - f(\mathbf{h} = \mathbf{a}) = \operatorname{argmax}_{\mathbf{a}} f(\mathbf{h} = \mathbf{a}) = \operatorname{argmax}_{\mathbf{h}} f(\mathbf{h})$
- The optimal action is to choose the mode of f

$$\mathbf{a}^*(f) = \operatorname*{argmax}_{\mathbf{h}} f(\mathbf{h}) \tag{27}$$

$$= \operatorname*{argmax}_{\mathbf{h}} \log f(\mathbf{h}) \tag{28}$$

Log-loss

- So far, the action $\mathbf{a}$ was a scalar or vector.
- Here, the action is a distribution over $\mathbf{h}$; denote it by $q(\mathbf{h})$.
- The game is as follows: before seeing $\mathbf{h}$, you are asked to name a distribution $q(\mathbf{h})$.
- Then $\mathbf{h}$ is revealed, and you incur the loss $\ell(q, \mathbf{h}) = \log(1/q(\mathbf{h})) = -\log q(\mathbf{h})$.
- If your chosen q assigns a small probability to the $\mathbf{h}$ that occurs, you incur a large penality.
- The log loss $\ell(q, \mathbf{h}) = \log(1/q(\mathbf{h}))$ is called the "surprise" in information theory. Note that the surprise is 0 for the certain event, i.e. if $q(\mathbf{h}) = 1$.
- First used for pmfs, but later also for pdfs.

Log-loss

- The log-loss is said to be local since it only evaluates the quoted q at the value $\mathbf{h}$ that actually occurs.
- Since we do not know the value of $\mathbf{h}$ when we are asked to report q, we choose q such that the expected loss is minimized,

$$\mathbf{a}^*(f) = \operatorname*{argmin}_q \mathbb{E}_{f(\mathbf{h})}\left[\ell(\mathbf{h}, q)\right] \tag{29}$$

$$= \operatorname*{argmin}_q \mathbb{E}_{f(\mathbf{h})}\left[-\log q(\mathbf{h})\right] \tag{30}$$

- Note that $\mathbf{a}^*(f)$ is a distribution.
- Expected log loss $\mathbb{E}_{f(\mathbf{h})}\left[-\log q(\mathbf{h})\right]$ is called cross-entropy of q relative to f.

Log-loss

- To solve the optimisation problem, note that

$$\begin{aligned}\mathbb{E}_{f(\mathbf{h})}\left[-\log q(\mathbf{h})\right] &= \mathbb{E}_{f(\mathbf{h})}\left[-\log q(\mathbf{h}) + \log f(\mathbf{h}) - \log f(\mathbf{h})\right] \\ &= \underbrace{\mathbb{E}_{f(\mathbf{h})}\left[\log \frac{f(\mathbf{h})}{q(\mathbf{h})}\right]}_{\mathrm{KL}(f||q)} \underbrace{-\mathbb{E}_{f(\mathbf{h})}\left[\log f(\mathbf{h})\right]}_{\text{entropy}} \qquad (31)\end{aligned}$$

- The Kullback-Leibler (KL) divergence $\mathrm{KL}(f||q)$ is non-negative and zero iff $f = q$ (see later lectures)
- The entropy indicates the average surprise of f, it is a measure of the randomness of $\mathbf{h} \sim f(\mathbf{h})$. It does not depend on q.
- The optimal q thus equals f, i.e.

$$\mathbf{a}^*(f) = \operatorname*{argmin}_q \mathbb{E}_{f(\mathbf{h})}\left[-\log q(\mathbf{h})\right] = f(\mathbf{h}) \qquad (32)$$

Summary

- We derived the optimal action

$$\mathbf{a}^*(f) = \operatorname*{argmin}_{\mathbf{a}} \mathbb{E}_{f(\mathbf{h})} \left[\ell(\mathbf{h}, \mathbf{a}) \right] \tag{33}$$

  for different loss functions $\ell(\mathbf{h}, \mathbf{a})$
- The optimal actions correspond to different aspects of f

loss	optimal action
quadratic	expected value of f
absolute error	median of f
pinball	any quantile of f
0-1	mode of f
log	f itself

- Allows us to obtain properties of f, and f itself, by solving an optimisation problem.

Their use in machine learning

By playing with different choices of $\mathbf{h}$, $f(\mathbf{h})$ and the loss $\ell(\mathbf{h}, \mathbf{a})$, we can frame classification, regression, inference, and density estimation as a decision problem.

loss	$f(\mathbf{h})$	unknown $\mathbf{h}$	action $\mathbf{a}$	field
quadratic	$p(\mathbf{h}\|\mathbf{x})$	continuous label	label	regression
absolute error	$p(h\|\mathbf{x})$	continuous label	label	robust regression
pinball	$p(h\|\mathbf{x})$	continuous label	label	quantile regression
0-1	$p(h\|\mathbf{x})$	discrete label	label	classification
0-1	$p(\mathbf{h}\|\mathbf{x})$	model parameters	parameters	parameter estimation
log	data	a data point	data distribution	density estimation
log	$p(\mathbf{h}\|\mathbf{x})$	latent variable	distribution of $\mathbf{h}$	inference

0-1 loss for classification and parameter estimation

- For the 0-1 loss the optimal action is

$$\mathbf{a}^*(f) = \operatorname*{argmax}_{\mathbf{h}} f(\mathbf{h}) \tag{34}$$

- If we use $p(\mathbf{h}|\mathbf{x})$ for $f(\mathbf{h})$, we obtain

$$\mathbf{a}^*(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{h}} p(\mathbf{h}|\mathbf{x}) \tag{35}$$

$$= \operatorname*{argmax}_{\mathbf{h}} \log p(\mathbf{h}|\mathbf{x}) \tag{36}$$

  which is called the maximum a-posteriori (MAP) estimate.

- Corresponds to the most probable class given $\mathbf{x}$ in case of classification (discrete $\mathbf{h}$).

0-1 loss for classification and parameter estimation

- Using Bayes rule, we further obtain

$$\mathbf{a}^*(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{h}} \log p(\mathbf{x}|\mathbf{h}) + \log p(\mathbf{h}) \tag{37}$$

- Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and assume they are independent so that $p(\mathbf{x}|\mathbf{h}) = \prod_i p(\mathbf{x}_i|\mathbf{h})$. Then

$$\mathbf{a}^*(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{h}} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\mathbf{h}) + \log p(\mathbf{h}) \tag{38}$$

- If $\mathbf{h}$ corresponds to parameters $\boldsymbol{\theta}$ of the model, the first term is the log-likelihood and the second the prior over the parameters.
- If the prior is uniform, then the MAP estimate equals the maximum likelihood estimate.

Log-loss for density estimation

$$\mathbf{a}^*(f) = \operatorname{argmin}_q \mathbb{E}_{f(\mathbf{h})} [-\log q(\mathbf{h})]$$

- Let $\mathbf{h}$ be a data point that is being revealed to us. Before we see $\mathbf{h}$, we are asked to report a distribution $q(\mathbf{h})$ over it.
- Let $f(\mathbf{h})$ be the data distribution and assume that we have samples $\mathbf{h}_1, \ldots, \mathbf{h}_n$, with $\mathbf{h}_i \sim f(\mathbf{h})$.
- We can then approximate the expectation over f with a sample average. This gives the so-called empirical risk

$$\hat{R}_n(q) = \frac{1}{n} \sum_{i=1}^{n} -\log q(\mathbf{h}_i) \tag{39}$$

- This is considered a "frequentist" approach since we do not update our belief about $\mathbf{h}$ in light of the data $\mathbf{h}_1, \ldots, \mathbf{h}_n$.
- Taking the expectation with respect to $p(\mathbf{h}|\mathbf{h}_1, \ldots, \mathbf{h}_n)$ would make the approach "Bayesian".

Log-loss for density estimation

- The action that minimises the empirical risk is

$$\hat{\mathbf{a}}^*(f) = \operatorname*{argmin}_q \hat{R}_n(q) = \operatorname*{argmin}_q \frac{1}{n} \sum_{i=1}^{n} -\log q(\mathbf{h}_i) \qquad (40)$$

- Most of the time, q is restricted to be a member of a parametric family $\mathcal{Q} = \{q(\mathbf{h}; \boldsymbol{\theta})\}$. The problem then becomes

$$\hat{\mathbf{a}}^*(f) = \operatorname*{argmin}_{q \in \mathcal{Q}} \frac{1}{n} \sum_{i=1}^{n} -\log q(\mathbf{h}_i) = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} -\log q(\mathbf{h}_i; \boldsymbol{\theta})$$

- This is the same as maximising the log-likelihood.
- We have seen two approaches that lead to maximum likelihood estimation.
  - Bayesian: minimising the posterior expected 0-1 loss with a flat prior
  - Frequentist: minimising empirical risk under log-loss

Program recap

1. Brief introduction to decision theory
   - Loss and decision principles
   - Connection to causality
   - Mild cognitive impairment example

2. Its use in statistics and machine learning
   - Common loss functions
   - Applications of the 0-1 loss
   - Applications of the log-loss