Decision theory

Kevin P. Murphy: Probabilistic Machine Learning Ch5

Nikolai Chow

RSE ML and Econometrics Reading Group

Monday 13th November, 2023

Table of Contents

Decision Theory

In decision theory, we assume:

- A decision maker or agent.
- A set of possible actions, $\mathcal{A}$, to choose from.
- A set of possible states $\mathcal{H}$.
- A set of possible outcomes, $\mathcal{X}$.
- An act $a \in \mathcal{A}$, is a function from $\mathcal{A}$ to $\mathcal{X}$.

Consider a hypothetical doctor treating a patient who may have COVID-19.

- Action 1: Do nothing.
- Action 2: Administer an expensive drug with potential side effects.

Covid Example

- State is defined by:
  1. Age of the patient (young vs. old).
  2. Presence of COVID-19 (yes vs no).
- Age is directly observable.
- Disease state is inferred from noisy observations.
- Therefore, the state is partially observed.

Loss Function

Actions have associated costs and benefits that depend on the state of nature $h \in \mathcal{H}$. We define a loss function $\ell(h, a)$:

- We define a loss function $\ell(h, a)$.
- Specifies the loss for taking action $a \in \mathcal{A}$.
- Given the state of nature $h \in \mathcal{H}$.

Optimal policy

The optimal policy (also called the **Bayes estimator**) specifies what action to take for each possible observation $\mathbf{x}$ so as to minimize the expected loss:

$$\pi^*(\mathbf{x}) = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{p(h|\mathbf{x})}[\ell(h, a)]$$

Bayesian Decision Theory is a statistical approach to decision-making that uses Bayes' Theorem to update the probability estimate for a hypothesis as additional evidence is obtained. It assists in making optimal decisions by considering prior knowledge and observed data.

Loss Table

State	Nothing	Drugs
No COVID-19, young	0	8
COVID-19, young	60	8
No COVID-19, old	0	8
COVID-19, old	10	8

Expected Loss

Test	Age	Pr(Covid)	Cost-Noop	Cost-Drugs	Action
0	young	0.014	0.84	8.00	0
0	old	0.014	0.14	8.00	0
1	young	0.795	47.73	8.00	1
1	old	0.795	7.95	9.00	0

your-table-label

Classification problems: Zero-one loss

Consider using Bayesian decision theory to decide the optimal class label to predict given an observed input $\mathbf{x} \in \mathcal{X}$. Suppose the states of nature correspond to class labels, so $\mathcal{H} = \mathcal{Y} = \{1, \ldots, C\}$.

Furthermore, suppose the actions also correspond to class labels, so $\mathcal{A} = \mathcal{Y}$. A commonly used loss function is the zero-one loss $\ell_{01}(y^*, \hat{y})$, defined as follows:

$$\ell_{01}(y^*, \hat{y}) = \begin{cases} 0 & \text{if } y^* = \hat{y} \\ 1 & \text{if } y^* \neq \hat{y} \end{cases}$$

In this case, the posterior expected loss is

$$R(\hat{y}|\mathbf{x}) = p(\hat{y} \neq y^*|\mathbf{x})$$

Classification problems: The "reject" option

In some cases, we may able to say "I don't know" instead of returning an answer that we don't really trust; this is called picking the **reject option**. Suppose now $\mathcal{A} = \mathcal{Y} \cup \{0\}$, where action 0 represents the reject action.

Now define the following loss function:

$$\ell(y, a) = \begin{cases} 0 & \text{if } y = a \text{ and } a \in \{1, \ldots, C\} \\ \lambda_r & \text{if } a = 0 \\ \lambda_e & \text{otherwise} \end{cases}$$

where λ_r is the cost of the reject action, and λ_e is the cost of a classification error.

Classification problems: The "reject" option

We have to choose between rejecting, with risk λ_r, and choosing the most probable class, $y^* = \arg\max_y p(\mathcal{H} = y|\mathbf{x})$, which has expected loss

$$\lambda_s(1 - p(\mathcal{H} = y^*|\mathbf{x}))$$

Hence, we should pick y^* if

$$\lambda_r \geq \lambda_s(1 - p(\mathcal{H} = y^*|\mathbf{x}))$$

$$\frac{\lambda_r}{\lambda_s} \geq (1 - p(\mathcal{H} = y^*|\mathbf{x}))$$

$$p(\mathcal{H} = y^*|\mathbf{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}$$

otherwise, we should reject.

Classification problems: The "reject" option

The probability below $\lambda^* = 1 - \frac{\lambda_r}{\lambda_e}$; otherwise, you should just pick the most probable class. In other words, the optimal policy is as follows:

$$a^* = \begin{cases} y^* & \text{if } p(y^*|x) > \lambda^* \\ \text{reject} & \text{otherwise} \end{cases}$$

where

$$y^* = \underset{y \in \{1,\ldots,C\}}{\arg\max}\ p(y|x)$$

$$\lambda^* = 1 - \frac{\lambda_r}{\lambda_e}$$

Classification problems: The "reject" option

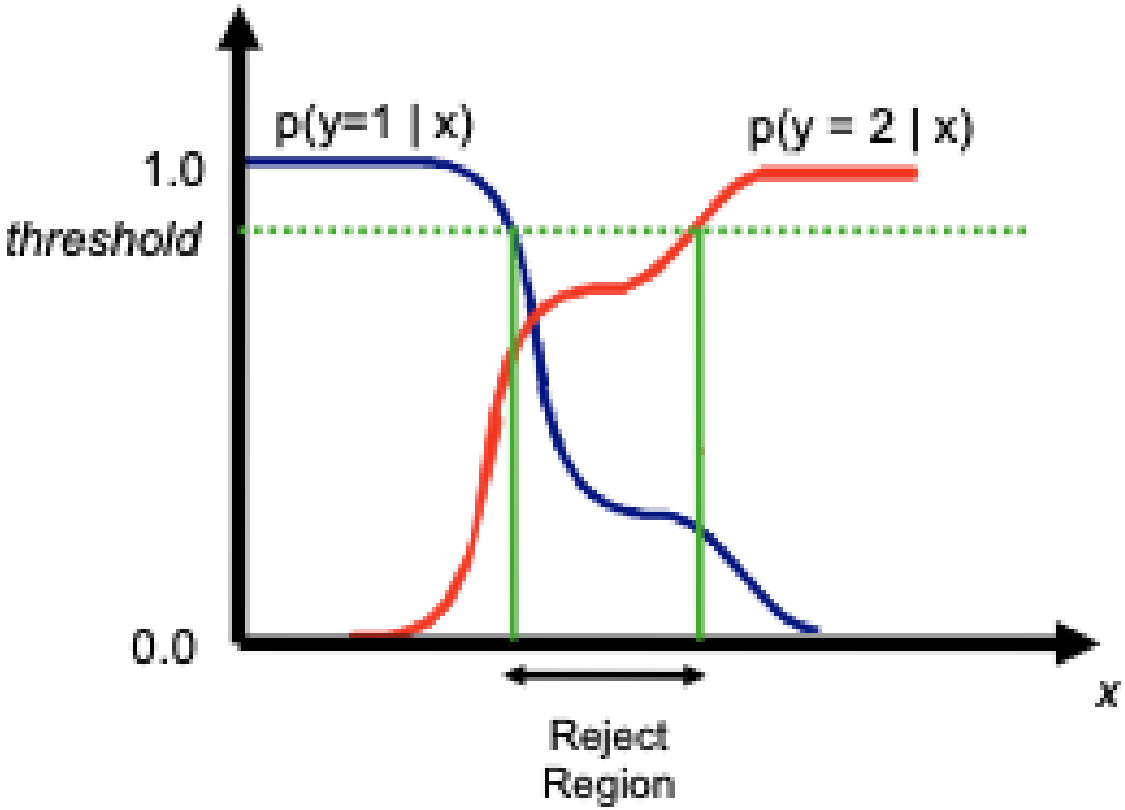


Figure 1: A two labels case

A Problem that Stumped Milton Friedman[1]

The Navy has two alternative designs for a projectile, designated as **Design A** and **Design B**. The objective is to determine which design is superior. The Navy consults a statistician on:

- How to conduct the test effectively.
- Approaches to analyze the results statistically.

[1] Based on Lectures in QuantEcon

A Problem that Stumped Milton Friedman

A decision-maker observes a sequence of random variable draws denoted by z. The main objective is to identify which of the two probability distributions, f_0 or f_1, is governing z.

The ultimate goal for the observer is to ascertain if the observations originate from f_0 or f_1. Using a Bayesian approach, the observer begins with a prior probability:

$$\pi_{-1} = \mathbb{P}\{f = f_0 \mid \text{no observations}\} \in (0, 1)$$

A Problem that Stumped Milton Friedman

After $k+1$ observations, $z_k, z_{k-1}, \ldots, z_0$, the observer updates the probability that the observations conform to distribution f_0 to:

$$\pi_k = \mathbb{P}\{f = f_0 \mid z_k, z_{k-1}, \ldots, z_0\}$$

This is recursively updated via Bayes' law:

$$\pi_{k+1} = \frac{\pi_k f_0(z_{k+1})}{\pi_k f_0(z_{k+1}) + (1-\pi_k) f_1(z_{k+1})}, \quad k = -1, 0, 1, \ldots$$

A Problem that Stumped Milton Friedman

Following the observation of $z_k, z_{k-1}, \ldots, z_0$, the observer believes the probability distribution of z_{k+1} is:

$$f_{\pi_k}(v) = \pi_k f_0(v) + (1 - \pi_k) f_1(v)$$

This is a blend of the distributions f_0 and f_1 where the weight on f_0 is the posterior probability that $f = f_0$.

A Problem that Stumped Milton Friedman

After observing a sequence, the decision-maker can:

1. Conclude $f = f_0$ and refrain from drawing more z values.
2. Conclude $f = f_1$ and refrain from drawing more z values.
3. Defer the decision and opt to draw z_{k+1}.

There are associated potential losses:

- L_0: Deciding $f = f_0$ when in fact $f = f_1$.
- L_1: Deciding $f = f_1$ when in fact $f = f_0$.
- c: Cost of postponing the decision and choosing to draw another z.

A Problem that Stumped Milton Friedman

Define $J(\pi)$ as the total optimal loss for a decision-maker with current belief π. It satisfies the Bellman equation:

$$J(\pi) = \min\left\{(1-\pi)L_0, \pi L_1, c + \mathbb{E}[J(\pi')]\right\}$$

where π' is given by Bayes' Law:

$$\pi' = \frac{\pi f_0(z')}{\pi f_0(z') + (1-\pi) f_1(z')}$$

A Problem that Stumped Milton Friedman

The optimal decision rule is characterized by two numbers $\alpha, \beta \in (0,1) \times (0,1)$ that satisfy

$$(1-\pi)L_0 < \min\{\pi L_1, c + \mathbb{E}[J(\pi')]\} \text{ if } \pi \geq \alpha$$

and

$$\pi L_1 < \min\{(1-\pi)L_0, c + \mathbb{E}[J(\pi')]\} \text{ if } \pi \leq \beta.$$

The optimal decision rule is then:

- Accept $f = f_0$ if $\pi \geq \alpha$.
- Accept $f = f_1$ if $\pi \leq \beta$.
- Draw another z if $\beta \leq \pi \leq \alpha$.

A Problem that Stumped Milton Friedman

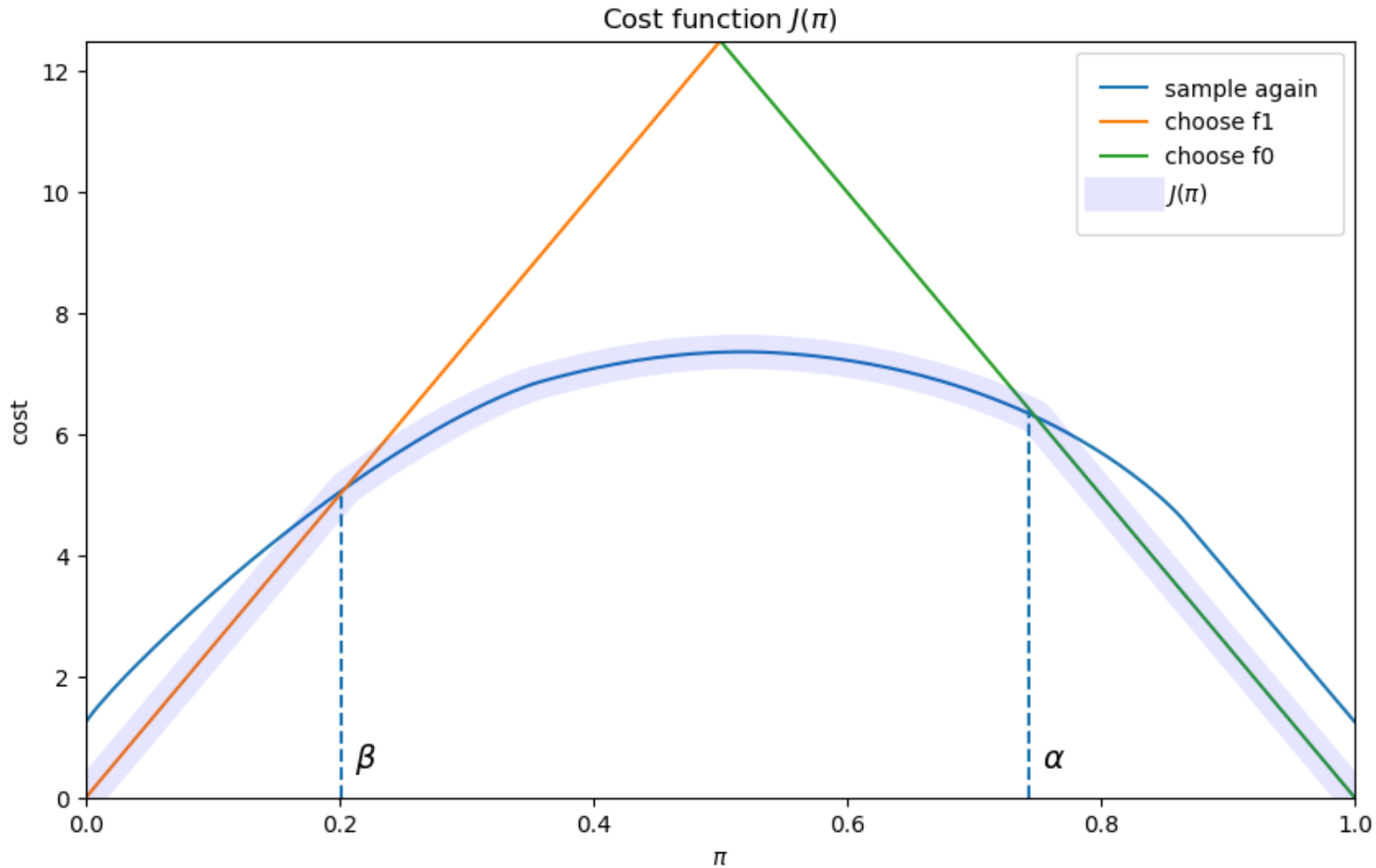
To make our computations manageable, we can write the continuation cost $h(\pi)$ as

$$\begin{aligned} h(\pi) &= c + \mathbb{E}[J(\pi')] \\ &= c + \mathbb{E}_{\pi'}\left[\min\{(1-\pi')L_0, \pi' L_1, h(\pi')\}\right] \\ &= c + \int \min\{(1-\kappa(z',\pi))L_0, \kappa(z',\pi)L_1, h(\kappa(z',\pi))\} f_\pi(z')\, dz' \end{aligned}$$

Thus, we iterate with an operator h denoted by Q, where

$$Qh(\pi) = c + \int \min\{(1-\kappa(z',\pi))L_0, \kappa(z',\pi)L_1, h(\kappa(z',\pi))\} f_\pi(z')\, dz'$$

A Problem that Stumped Milton Friedman



Regression problems: L2 loss

Consider the case where the set of actions and states are both equal to the real line, $\mathcal{A} = \mathcal{H} = \mathbb{R}$.

The most common loss for continuous states and actions is the ℓ_2 loss, also called squared error or quadratic loss, which is defined as follows:

$$\ell_2(h, a) = (h - a)^2$$

Probabilistic prediction problems (PPP)

Consider the true "state of nature" is a distribution, $h = p(Y|\mathbf{x})$, the action is another distribution, $a = q(Y|\mathbf{x})$, and we want to pick q to minimize $\mathbb{E}[\ell(p, q)]$ for a given $\mathbf{x}$.

PPP: KL, cross-entropy and log loss

A common form of loss functions for comparing two distributions is the Kullback-Leibler divergence, or KL divergence, which is defined as follows:

$$D_{KL}(p \parallel q) = \sum_{y \in Y} p(y) \log \frac{p(y)}{q(y)}$$

To find the optimal distribution to use when predicting future data, we can minimize $D_{KL}(p \parallel q)$. It is equivalent to minimize the cross-entropy:

$$q^*(Y|x) = \arg\min_q - \sum_y p(Y|x) \log q(Y|x)$$

Table of Contents

Choosing the "right" model

Consider the setting in which we have several candidate (parametric) models (e.g., neural networks with different numbers of layers), and we want to choose the "right" one. This can be tackled using tools from Bayesian decision theory.

Bayesian hypothesis testing

Suppose we have two hypotheses, the null hypothesis, M_0, and the alternative hypothesis, M_1, and we want to know which one is more likely to be true for dataset $\mathcal{D}$. This is called hypothesis testing.

If we use 0-1 loss, the optimal decision is to pick the alternative hypothesis iff $p(M_1|\mathcal{D})/p(M_0|\mathcal{D}) > 1$. If we use $p(M_0) = p(M_1)$, the decision rule is equivalent to select M_1 iff $p(M_1|\mathcal{D})/p(M_0|\mathcal{D}) > 1$. This quantity is known as the **Bayes factor**:

$$B_{1,0} = \frac{p(D|M_1)}{p(D|M_0)}$$

Bayesian hypothesis testing

Bayes factor BF(1, 0)	**Interpretation**
$BF < 1/100$	Decisive evidence for M_0
$BF < 1/10$	Strong evidence for M_0
$1/10 < BF < 1/3$	Moderate evidence for M_0
$1/3 < BF < 1$	Weak evidence for M_0
$1 < BF < 3$	Weak evidence for M_1
$3 < BF < 10$	Moderate evidence for M_1
$BF > 10$	Strong evidence for M_1
$BF > 100$	Decisive evidence for M_1

Table 1: Jeffreys scale of evidence for interpreting Bayes factors.

Bayesian hypothesis testing: Example

Suppose we observe some coin tosses, and want to decide if the data was generated by a fair coin, $\theta = 0.5$, or a potentially biased coin, where θ could be any value in $[0, 1]$. The marginal likelihood under M_0 is simply

$$p(\mathcal{D}|M_0) = \left(\frac{1}{2}\right)^N$$

where N is the number of coin tosses. The marginal likelihood under M_1, using a Beta prior, is

$$p(\mathcal{D}|M_1) = \int p(\mathcal{D}|\theta)p(\theta)d\theta = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)}$$

Bayesian hypothesis testing: Example

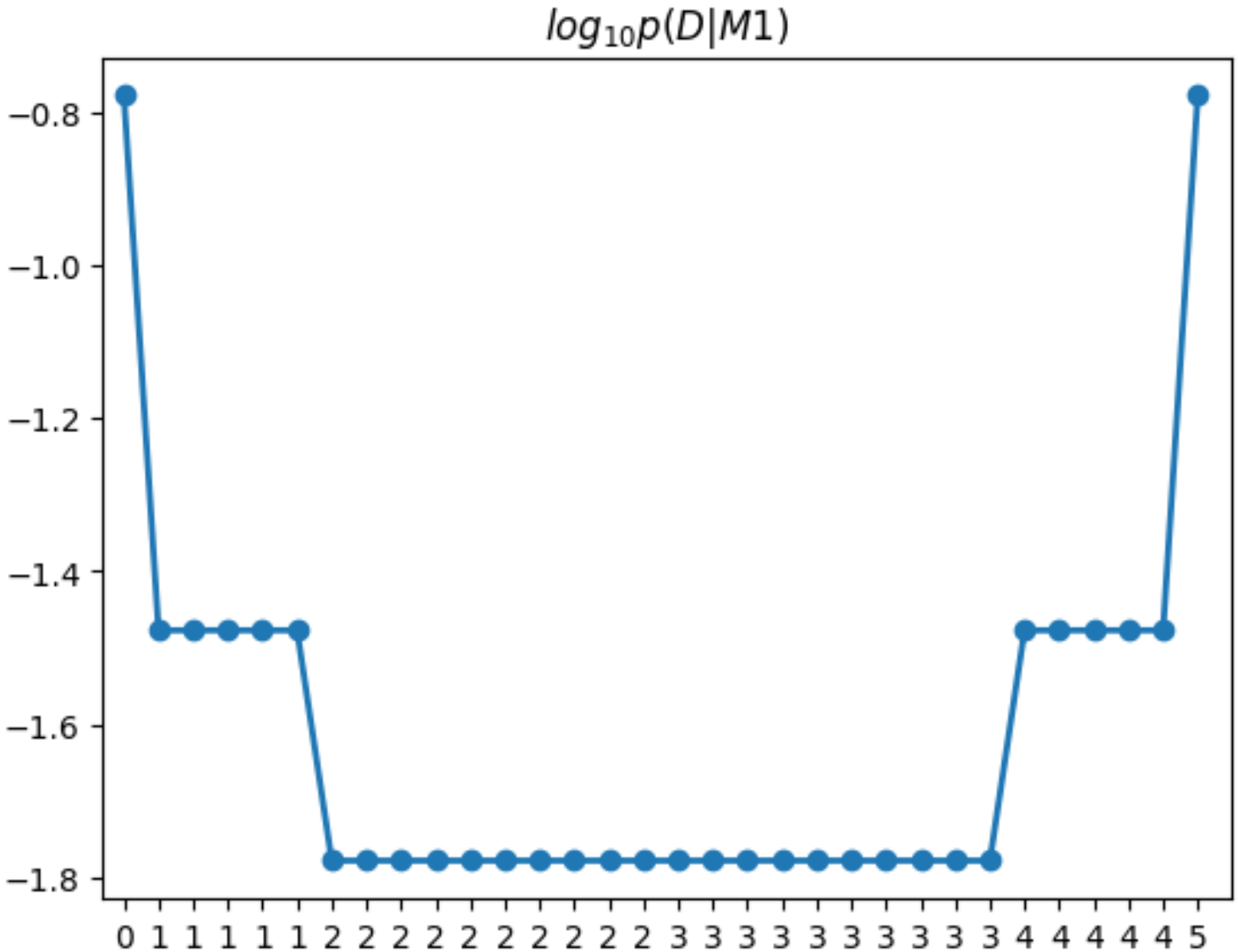


Figure 2: Log marginal likelihood vs number of heads

Hypothesis testing without Bayes factors

Computing the necessary marginal likelihoods can be computationally difficult. Now, suppose we have two classifiers, m_1 and m_2, and we want to know which one is better.

1. μ_1 and μ_2 be their average accuracies.
2. $\delta = \mu_1 - \mu_2$ be the difference in their accuracies.

The probability that model 1 is more accurate, on average, than model 2 is given by $p(\delta > \varepsilon|\mathcal{D})$ or $p(|\delta| > \varepsilon|\mathcal{D})$, where ε represents the minimal magnitude of effect size that is meaningful for the problem at hand. This is called a one-sided test or two-sided test.

Bayesian t-test for difference in means

Suppose we have two classifiers, m_1 and m_2, which are evaluated on the same set of N test examples. Let e_i^m be the error of method m on test example i.

1. We will work with the differences between methods, $x_i = e_i^1 - e_i^2$.
2. We assume $x_i \sim \mathcal{N}(\delta, \sigma^2)$.
3. We are interested in $p(\delta|\mathbf{x})$, where $\mathbf{x} = (x_1, \ldots, x_N)$.

Bayesian t-test for difference in means

If we use an uninformative prior for the unknown parameters (δ, σ), one can show that the posterior marginal for the mean is given by a Student distribution:

$$p(\delta|\mathbf{x}) = T_{N-1}(\delta|\bar{x}, s^2/N)$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the sample mean, and $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$. Hence, we can easily compute $p(|\delta| > \varepsilon|\mathbf{x})$, with a ROPE of $\varepsilon = 0.01$ (say).

Table of Contents

Issues related to Bayesian interpretation

Bayesian probability: a measure of belief or certainty?

Issues related to Bayesian interpretation

Consider the following information about Linda:

"Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations."

Rank the following statements from most (1) to least (8) probable.

1. Linda is a teacher in elementary school.
2. Linda works in a bookstore and takes Yoga classes.
3. Linda is active in the feminist movement.
4. Linda is a psychiatric social worker.
5. Linda is a member of the League of Women Voters.
6. Linda is a bank teller.
7. Linda is an insurance salesperson.
8. Linda is a bank teller and is active in the feminist movement.

Issues related to Bayesian interpretation

Consider the following information about Linda:

"Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations."

Rank the following statements from most (1) to least (8) probable.

1. (5.2) Linda is a teacher in elementary school.
2. (3.3) Linda works in a bookstore and takes Yoga classes.
3. (2.1) Linda is active in the feminist movement.
4. (3.1) Linda is a psychiatric social worker.
5. (5.4) Linda is a member of the League of Women Voters.
6. (6.2) Linda is a bank teller.
7. (6.4) Linda is an insurance salesperson.
8. (4.1) Linda is a bank teller and is active in the feminist movement.

Issues related to Bayesian interpretation

Consider the following information about Linda:

"Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations."

Rank the following statements from most (1) to least (8) probable.

1. (5.2) Linda is a teacher in elementary school.
2. (3.3) Linda works in a bookstore and takes Yoga classes.
3. (2.1) Linda is active in the feminist movement.
4. (3.1) Linda is a psychiatric social worker.
5. (5.4) Linda is a member of the League of Women Voters.
6. (6.2) Linda is a bank teller.
7. (6.4) Linda is an insurance salesperson.
8. (4.1) Linda is a bank teller and is active in the feminist movement.