Decision problems

September 1, 2020

Uncertainty

- We cannot perfectly predict the future.
- We cannot know for sure what happened in the past.
- How can we quantify this uncertainty?
- Probabilities!

Axioms of probability

A probability measure P on (Ω, Σ) has the following properties:

Uncertainty

- We cannot perfectly predict the future.
- We cannot know for sure what happened in the past.
- How can we quantify this uncertainty?
- Probabilities!

Axioms of probability

A probability measure P on (Ω, Σ) has the following properties:

1. The probability of the certain event is $P(\Omega) = 1$

Uncertainty

- We cannot perfectly predict the future.
- We cannot know for sure what happened in the past.
- How can we quantify this uncertainty?
- Probabilities!

Axioms of probability

A probability measure P on (Ω, Σ) has the following properties:

1. The probability of the certain event is $P(\Omega) = 1$
2. The probability of the impossible event is $P(\emptyset) = 0$

Uncertainty

- We cannot perfectly predict the future.
- We cannot know for sure what happened in the past.
- How can we quantify this uncertainty?
- Probabilities!

Axioms of probability

A probability measure P on (Ω, Σ) has the following properties:

1. The probability of the certain event is $P(\Omega) = 1$
2. The probability of the impossible event is $P(\emptyset) = 0$
3. The probability of any event $A \in \Sigma$ is $0 \leq P(A) \leq 1$.

Uncertainty

- We cannot perfectly predict the future.
- We cannot know for sure what happened in the past.
- How can we quantify this uncertainty?
- Probabilities!

Axioms of probability

A probability measure P on (Ω, Σ) has the following properties:

1. The probability of the certain event is $P(\Omega) = 1$
2. The probability of the impossible event is $P(\emptyset) = 0$
3. The probability of any event $A \in \Sigma$ is $0 \leq P(A) \leq 1$.
4. If A, B are disjoint, i.e. $A \cap B = \emptyset$, meaning that they cannot happen at the same time, then

$$P(A \cup B) = P(A) + P(B)$$

Definition 1 (Conditional probability)

The probability of A happening if we know that B has happened is defined to be:

$$P(A \mid B) \triangleq \frac{P(A \cap B)}{P(B)}.$$

Conditional probabilities obey the same rules as probabilities.

Bayes's theorem

For $P(A_1 \cup A_2) = 1$, $A_1 \cap A_2 = \emptyset$,

$$P(A_i \mid B)$$

Definition 1 (Conditional probability)

The probability of A happening if we know that B has happened is defined to be:

$$P(A \mid B) \triangleq \frac{P(A \cap B)}{P(B)}.$$

Bayes's theorem

For $P(A_1 \cup A_2) = 1$, $A_1 \cap A_2 = \emptyset$,

$$P(A_i \mid B) = \frac{P(B \mid A_i) P(A_i)}{P(B)}$$

Definition 1 (Conditional probability)

The probability of A happening if we know that B has happened is defined to be:

$$P(A \mid B) \triangleq \frac{P(A \cap B)}{P(B)}.$$

Bayes's theorem

For $P(A_1 \cup A_2) = 1$, $A_1 \cap A_2 = \emptyset$,

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2)}$$

Definition 1 (Conditional probability)

The probability of A happening if we know that B has happened is defined to be:

$$P(A \mid B) \triangleq \frac{P(A \cap B)}{P(B)}.$$

Bayes's theorem

For $P(A_1 \cup A_2) = 1$, $A_1 \cap A_2 = \emptyset$,

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2)}$$

Example 2 (probability of rain)

What is the probability of rain given a forecast x_1 or x_2?

ω_1: rain	$P(\omega_1) = 80\%$
ω_2: dry	$P(\omega_2) = 20\%$

Table : Prior probability of rain tomorrow

Definition 1 (Conditional probability)

The probability of A happening if we know that B has happened is defined to be:

$$P(A \mid B) \triangleq \frac{P(A \cap B)}{P(B)}.$$

Bayes's theorem

For $P(A_1 \cup A_2) = 1$, $A_1 \cap A_2 = \emptyset$,

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2)}$$

Example 2 (probability of rain)

What is the probability of rain given a forecast x_1 or x_2?

ω_1: rain	$P(\omega_1) = 80\%$
ω_2: dry	$P(\omega_2) = 20\%$

Table : Prior probability of rain tomorrow

x_1: rain	$P(x_1 \mid \omega_1) = 90\%$
x_2: dry	$P(x_2 \mid \omega_2) = 50\%$

Table : Probability the forecast is correct

Definition 1 (Conditional probability)

The probability of A happening if we know that B has happened is defined to be:

$$P(A \mid B) \triangleq \frac{P(A \cap B)}{P(B)}.$$

Bayes's theorem

For $P(A_1 \cup A_2) = 1$, $A_1 \cap A_2 = \emptyset$,

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2)}$$

Example 2 (probability of rain)

What is the probability of rain given a forecast x_1 or x_2?

ω_1: rain	$P(\omega_1) = 80\%$
ω_2: dry	$P(\omega_2) = 20\%$

Table : Prior probability of rain tomorrow

x_1: rain	$P(x_1 \mid \omega_1) = 90\%$
x_2: dry	$P(x_2 \mid \omega_2) = 50\%$

Table : Probability the forecast is correct

$$P(\omega_1 \mid x_1) = 87.8\%$$
$$P(\omega_1 \mid x_2) = 44.4\%$$

Table : Probability that it will rain given the forecast

Classification in terms of conditional probabilities

- Features $x_t \in \mathcal{X}$.
- Class label $y_t \in \mathcal{Y}$.
- Probability model $P_\mu(x_t \mid y_t)$.
- Prior class probability $P_\mu(y_t = c)$.

$$P_\mu(y_t = c \mid x_t) = \frac{P_\mu(x_t \mid y_t = c) P_\mu(y_t = c)}{\sum_{c' \in \mathcal{Y}} P_\mu(x_t \mid y_t = c') P_\mu(y_t = c')}$$

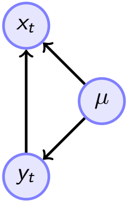
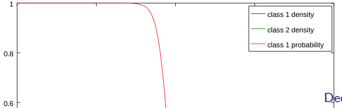
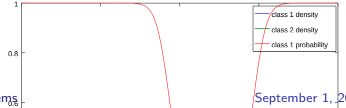
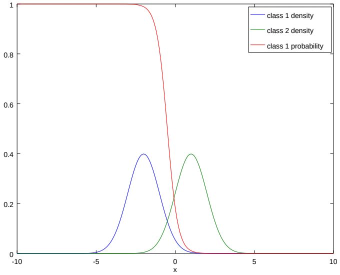


Figure : A generative classification model. μ identifies the model (paramter). x_t are the features and y_t the class label of the t-th example.

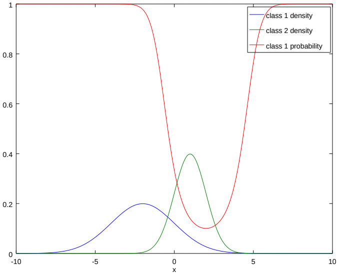




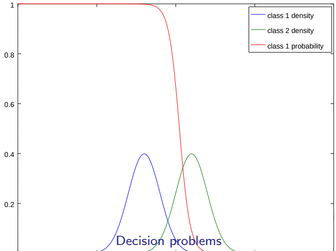
Classification in terms of conditional probabilities



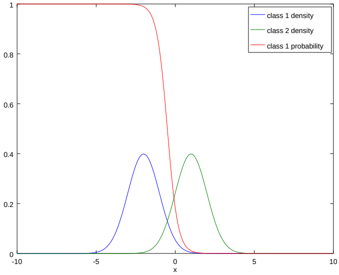
(a) Equal prior and variance



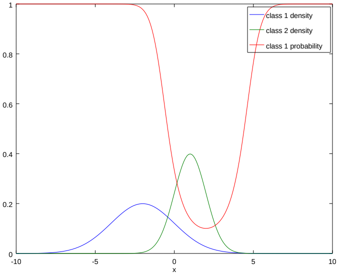
(b) Unequal variance



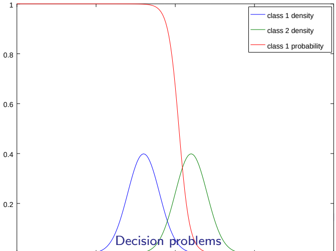
Classification in terms of conditional probabilities



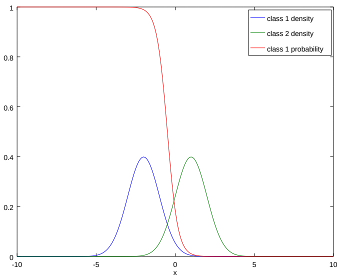
(a) Equal prior and variance



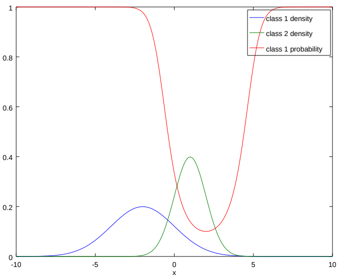
(b) Unequal variance



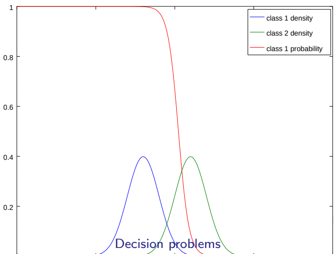
Classification in terms of conditional probabilities



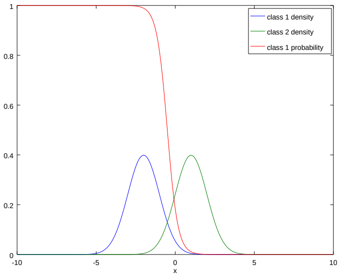
(a) Equal prior and variance



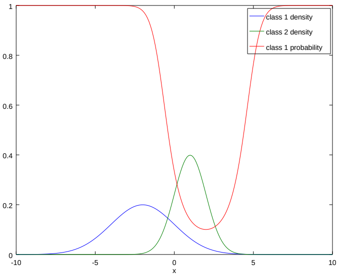
(b) Unequal variance



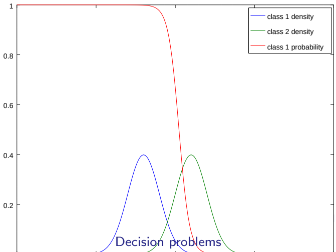
Classification in terms of conditional probabilities



(a) Equal prior and variance



(b) Unequal variance

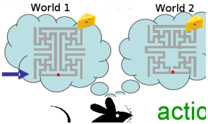


Subjective probability

Subjective probability measure ξ

- If we think event A is more likely than B, then $\xi(A) > \xi(B)$.
- Usual rules of probability apply:
  1. $\xi(A) \in [0, 1]$.
  2. $\xi(\emptyset) = 0$.
  3. If $A \cap B = \emptyset$, then $\xi(A \cup B) = \xi(A) + \xi(B)$.

Bayesian inference illustration



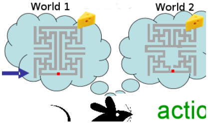
Use a subjective belief $\xi(\mu)$ on $\mathcal{M}$

- Prior belief $\xi(\mu)$ represents our initial uncertainty.

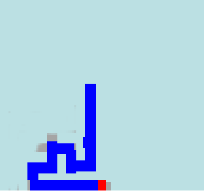
Bayesian inference illustration

Use a subjective belief $\xi(\mu)$ on $\mathcal{M}$

- Prior belief $\xi(\mu)$ represents our initial uncertainty.
- We observe history h.

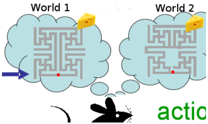


prior

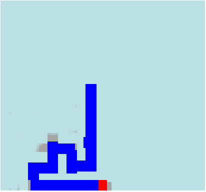


evidence

Bayesian inference illustration



prior



evidence

Use a subjective belief $\xi(\mu)$ on $\mathcal{M}$

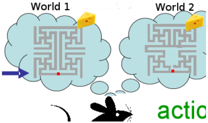
- Prior belief $\xi(\mu)$ represents our initial uncertainty.
- We observe history h.
- Each possible μ assigns a probability $P_\mu(h)$ to h.

Bayesian inference illustration

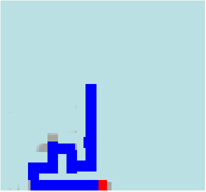
Use a subjective belief $\xi(\mu)$ on $\mathcal{M}$

- Prior belief $\xi(\mu)$ represents our initial uncertainty.
- We observe history h.
- Each possible μ assigns a probability $P_\mu(h)$ to h.
- We can use this to update our belief via Bayes' theorem to obtain the posterior belief:

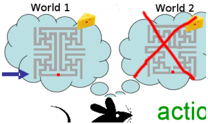
$$\xi(\mu \mid h) \propto P_\mu(h)\xi(\mu) \quad \text{(conclusion = evidence} \times \text{prior)}$$



prior



evidence



conclusion

Some examples

Example 4

John claims to be a medium. He throws a coin n times and predicts its value always correctly. Should we believe that he is a medium?

- μ_1: John is a medium.
- μ_0: John is not a medium.

The answer depends on what we expect a medium to be able to do, and how likely we thought he'd be a medium in the first place.

Bayesian inference

Family of models $\mathcal{M} = \{\mu_1, \ldots, \mu_k\}$

Defines a family of probabilities for any data x:

$$\{P_\mu | \mu \in \mathcal{M}\}, \qquad P_\mu(x) \equiv \mathbb{P}(x \mid \mu).$$

Bayesian inference

Family of models $\mathcal{M} = \{\mu_1, \ldots, \mu_k\}$

Defines a family of probabilities for any data x:

$$\{P_\mu | \mu \in \mathcal{M}\}, \qquad P_\mu(x) \equiv \mathbb{P}(x \mid \mu).$$

Prior belief ξ over models $\mathcal{M}$

$\xi(\mu)$ is a distribution how much we believe it is correct before seeing any data.

Bayesian inference

Family of models $\mathcal{M} = \{\mu_1, \ldots, \mu_k\}$

Defines a family of probabilities for any data x:

$$\{P_\mu | \mu \in \mathcal{M}\}, \qquad P_\mu(x) \equiv \mathbb{P}(x \mid \mu).$$

Prior belief ξ over models $\mathcal{M}$

$\xi(\mu)$ is a distribution how much we believe it is correct before seeing any data.

Posterior belief $\xi(\mu \mid x)$ over models after seeing x

$$\xi(\mu \mid x) = \frac{P_\mu(x)\xi(\mu)}{\sum_{\mu' \in \mathcal{M}} P_{\mu'}(x)\xi(\mu')}$$

Bayesian inference

Family of models $\mathcal{M} = \{\mu_1, \ldots, \mu_k\}$

Defines a family of probabilities for any data x:

$$\{P_\mu | \mu \in \mathcal{M}\}, \qquad P_\mu(x) \equiv \mathbb{P}(x \mid \mu).$$

Prior belief ξ over models $\mathcal{M}$

$\xi(\mu)$ is a distribution how much we believe it is correct before seeing any data.

Posterior belief $\xi(\mu \mid x)$ over models after seeing x

$$\xi(\mu \mid x) = \frac{P_\mu(x)\xi(\mu)}{\sum_{\mu' \in \mathcal{M}} P_{\mu'}(x)\xi(\mu')} = \frac{\mathbb{P}(x \mid \mu)\xi(\mu)}{\sum_{\mu' \in \mathcal{M}} \mathbb{P}(x \mid \mu')\xi(\mu')}.$$

Bayesian inference

Family of models $\mathcal{M} = \{\mu_1, \ldots, \mu_k\}$

Defines a family of probabilities for any data x:

$$\{P_\mu | \mu \in \mathcal{M}\}, \qquad P_\mu(x) \equiv \mathbb{P}(x \mid \mu).$$

Prior belief ξ over models $\mathcal{M}$

$\xi(\mu)$ is a distribution how much we believe it is correct before seeing any data.

Posterior belief $\xi(\mu \mid x)$ over models after seeing x

$$\xi(\mu \mid x) = \frac{P_\mu(x)\xi(\mu)}{\sum_{\mu' \in \mathcal{M}} P_{\mu'}(x)\xi(\mu')} = \frac{\mathbb{P}(x \mid \mu)\xi(\mu)}{\sum_{\mu' \in \mathcal{M}} \mathbb{P}(x \mid \mu')\xi(\mu')}.$$

Interpretation

- $\mathcal{M}$: Set of all possible models that could describe the data.
- $P_\mu(x)$: Probability of x under model μ.
- Alternative notation $\mathbb{P}(x \mid \mu)$: Probability of x given that model μ is correct.
- $\xi(\mu)$: Our belief, before seeing the data, that μ is correct.
- $\xi(\mu \mid x)$: Our belief, aftering seeing the data, that μ is correct.

Exercise 1 (Continued example for medium)

$$P_\mu(x) = \prod_{t=1}^{n} P_\mu(x_t). \qquad \text{(independence property)}$$

If a classmate correctly predicts 4 coin tosses, what is your belief they are a medium?

Exercise 1 (Continued example for medium)

$$P_\mu(x) = \prod_{t=1}^{n} P_\mu(x_t). \qquad \text{(independence property)}$$

$$P_{\mu_1}(x_t = 1) = 1, \qquad P_{\mu_1}(x_t = 0) = 0. \qquad \text{(true medium model)}$$

$$P_{\mu_0}(x_t = 1) = 1/2, \qquad P_{\mu_0}(x_t = 0) = 1/2. \qquad \text{(non-medium model)}$$

If a classmate correctly predicts 4 coin tosses, what is your belief they are a medium?

Exercise 1 (Continued example for medium)

$$P_\mu(x) = \prod_{t=1}^{n} P_\mu(x_t). \qquad \text{(independence property)}$$

$$P_{\mu_1}(x_t = 1) = 1, \qquad P_{\mu_1}(x_t = 0) = 0. \qquad \text{(true medium model)}$$

$$P_{\mu_0}(x_t = 1) = 1/2, \qquad P_{\mu_0}(x_t = 0) = 1/2. \qquad \text{(non-medium model)}$$

$$\xi(\mu_0) = 1/2, \qquad \xi(\mu_1) = 1/2. \qquad \text{(prior belief)}$$

If a classmate correctly predicts 4 coin tosses, what is your belief they are a medium?

Exercise 1 (Continued example for medium)

$$P_\mu(x) = \prod_{t=1}^{n} P_\mu(x_t). \qquad \text{(independence property)}$$

$$P_{\mu_1}(x_t = 1) = 1, \qquad P_{\mu_1}(x_t = 0) = 0. \qquad \text{(true medium model)}$$

$$P_{\mu_0}(x_t = 1) = 1/2, \qquad P_{\mu_0}(x_t = 0) = 1/2. \qquad \text{(non-medium model)}$$

$$\xi(\mu_0) = 1/2, \qquad \xi(\mu_1) = 1/2. \qquad \text{(prior belief)}$$

$$\xi(\mu_1 \mid x) = \frac{P_{\mu_1}(x)\xi(\mu_1)}{\mathbb{P}_\xi(x)} \qquad \text{(posterior belief)}$$

$$\mathbb{P}_\xi(x) \triangleq P_{\mu_1}(x)\xi(\mu_1) + P_{\mu_0}(x)\xi(\mu_0). \qquad \text{(marginal distribution)}$$

If a classmate correctly predicts 4 coin tosses, what is your belief they are a medium?

Sequential update of beliefs

	M	T	W	T	F	S	S
CNN	0.5	0.6	0.7	0.9	0.5	0.3	0.1
SMHI	0.3	0.7	0.8	0.9	0.5	0.2	0.1
YR	0.6	0.9	0.8	0.5	0.4	0.1	0.1
Rain?	Y	Y	Y	N	Y	N	N

Table : Predictions by three different entities for the probability of rain on a particular day, along with whether or not it actually rained.

Exercise 2

- n meteorological stations $\{\mu_i \mid i = 1, \ldots, n\}$
- The i-th station predicts rain $P_{\mu_i}(x_t \mid x_1, \ldots, x_{t-1})$.
- Let $\xi_t(\mu)$ be our belief at time t. Derive the next-step belief $\xi_{t+1}(\mu) \triangleq \xi_t(\mu|y_t)$ in terms of the current belief ξ_t.
- Write a python function that computes this posterior

Sequential update of beliefs

	M	T	W	T	F	S	S
CNN	0.5	0.6	0.7	0.9	0.5	0.3	0.1
SMHI	0.3	0.7	0.8	0.9	0.5	0.2	0.1
YR	0.6	0.9	0.8	0.5	0.4	0.1	0.1
Rain?	Y	Y	Y	N	Y	N	N

Table : Predictions by three different entities for the probability of rain on a particular day, along with whether or not it actually rained.

Exercise 2

- n meteorological stations $\{\mu_i \mid i = 1, \ldots, n\}$
- The i-th station predicts rain $P_{\mu_i}(x_t \mid x_1, \ldots, x_{t-1})$.
- Let $\xi_t(\mu)$ be our belief at time t. Derive the next-step belief $\xi_{t+1}(\mu) \triangleq \xi_t(\mu|y_t)$ in terms of the current belief ξ_t.
- Write a python function that computes this posterior

$$\xi_{t+1}(\mu) \triangleq \xi_t(\mu|x_t) = \frac{P_\mu(x_t \mid x_1, \ldots, x_{t-1})\xi_t(\mu)}{\sum_{\mu'} P_{\mu'}(x_t \mid x_1, \ldots, x_{t-1})\xi_t(\mu')}$$

Bayesian inference for Bernoulli distributions

Estimating a coin's bias

A fair coin comes heads 50% of the time. We want to test an unknown coin, which we think may not be completely fair.

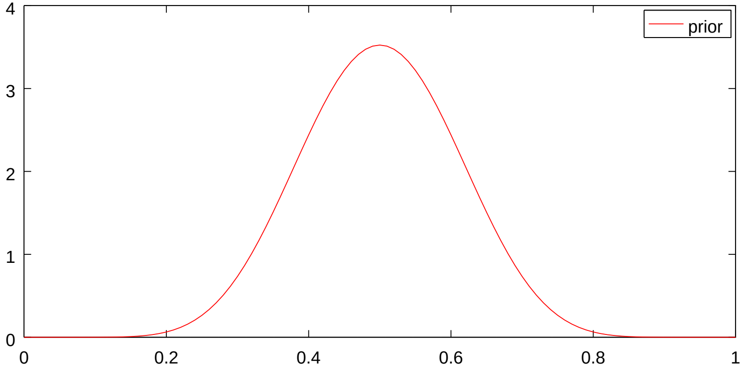


Figure : Prior belief ξ about the coin bias θ.

Bayesian inference for Bernoulli distributions

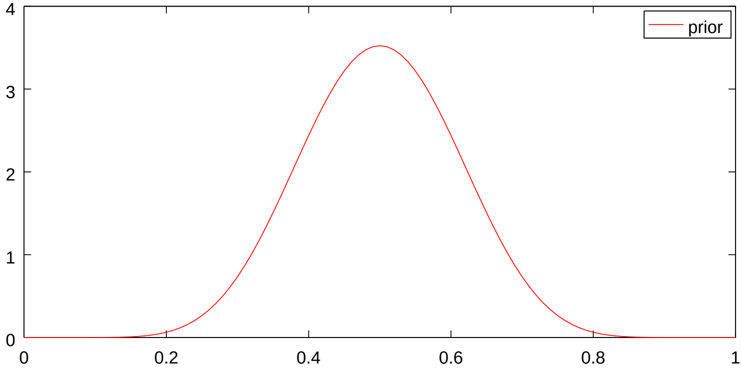


Figure : Prior belief ξ about the coin bias θ.

For a sequence of throws $x_t \in \{0, 1\}$,

$$P_\theta(x) \propto \prod_t \theta^{x_t} (1-\theta)^{1-x_t} = \theta^{\#\text{Heads}} (1-\theta)^{\#\text{Tails}}$$

Bayesian inference for Bernoulli distributions

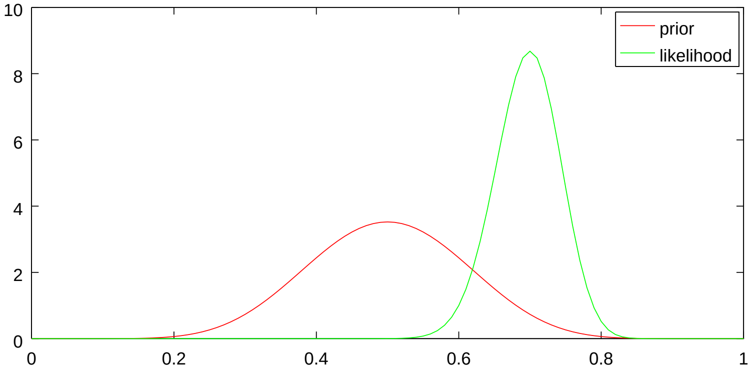


Figure : Prior belief ξ about the coin bias θ and likelihood of θ for the data.

Say we throw the coin 100 times and obtain 70 heads. Then we plot the likelihood $P_\theta(x)$ of different models.

Bayesian inference for Bernoulli distributions

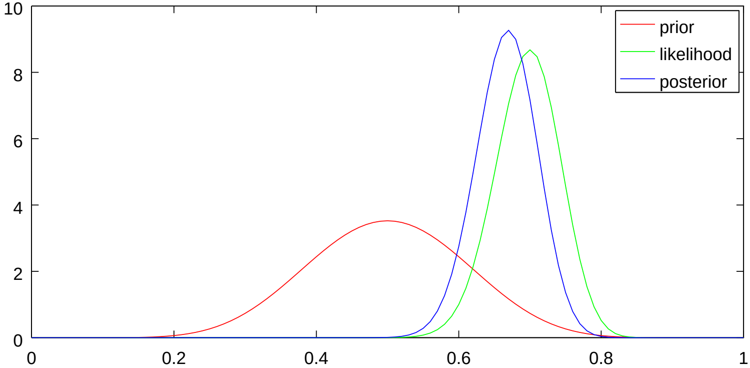


Figure : Prior belief $\xi(\theta)$ about the coin bias θ, likelihood of θ for the data, and posterior belief $\xi(\theta \mid x)$

From these, we calculate a posterior distribution over the correct models. This represents our conclusion given our prior and the data.

Learning outcomes

Understanding

- The axioms of probability, marginals and conditional distributions.
- The philosophical underpinnings of Bayesianism.
- The simple conjugate model for Bernoulli distributions.

Skills

- Be able to calculate with probabilities using the marginal and conditional definitions and Bayes rule.
- Being able to implement a simple Bayesian inference algorithm in Python.

Reflection

- How useful is the Bayesian representation of uncertainty?
- How restrictive is the need to select a prior distribution?
- Can you think of another way to explicitly represent uncertainty in a way that can incorporate new evidence?

Preferences

Example 5

Food

A McDonald's cheeseburger

B Surstromming

C Oatmeal

Money

A 10,000,000 SEK

B 10,000,000 USD

C 10,000,000 BTC

Entertainment

A Ticket to Liseberg

B Ticket to Rebstar

C Ticket to Nutcracker

Rewards and utilities

- Each choice is called a reward $r \in \mathcal{R}$.
- There is a utility function $U : \mathcal{R} \to \mathbb{R}$, assigning values to reward.
- We (weakly) prefer A to B iff $U(A) \geq U(B)$.

Exercise 3

From your individual preferences, derive a common utility function that reflects everybody's preferences in the class for each of the three examples. Is there a simple algorithm for deciding this? Would you consider the outcome fair?

Preferences among random outcomes

Example 6

Would you rather ...

- A Have 100 EUR now?
- B Flip a coin, and get 200 EUR if it comes heads?

Risk and monetary rewards

Preferences among random outcomes

Example 6

Would you rather ...

- A Have 100 EUR now?
- B Flip a coin, and get 200 EUR if it comes heads?

The expected utility hypothesis

Rational decision makers prefer choice A to B if

$$\mathbb{E}(U|A) \geq \mathbb{E}(U|B),$$

where the expected utility is

$$\mathbb{E}(U|A) = \sum_r U(r)\,\mathbb{P}(r|A).$$

In the above example, $r \in \{0, 100, 200\}$ and $U(r)$ is increasing, and the coin is fair.

Risk and monetary rewards

Preferences among random outcomes

Example 6

Would you rather ...

A Have 100 EUR now?

B Flip a coin, and get 200 EUR if it comes heads?

The expected utility hypothesis

Rational decision makers prefer choice A to B if

$$\mathbb{E}(U|A) \geq \mathbb{E}(U|B),$$

where the expected utility is

$$\mathbb{E}(U|A) = \sum_r U(r)\,\mathbb{P}(r|A).$$

In the above example, $r \in \{0, 100, 200\}$ and $U(r)$ is increasing, and the coin is fair.

Risk and monetary rewards

- If U is convex, we are risk-seeking.

Preferences among random outcomes

Example 6

Would you rather ...

A Have 100 EUR now?

B Flip a coin, and get 200 EUR if it comes heads?

The expected utility hypothesis

Rational decision makers prefer choice A to B if

$$\mathbb{E}(U|A) \geq \mathbb{E}(U|B),$$

where the expected utility is

$$\mathbb{E}(U|A) = \sum_r U(r)\,\mathbb{P}(r|A).$$

In the above example, $r \in \{0, 100, 200\}$ and $U(r)$ is increasing, and the coin is fair.

Risk and monetary rewards

- If U is convex, we are risk-seeking.
- If U is concave, we are risk-averse.

Preferences among random outcomes

Example 6

Would you rather ...

A Have 100 EUR now?

B Flip a coin, and get 200 EUR if it comes heads?

The expected utility hypothesis

Rational decision makers prefer choice A to B if

$$\mathbb{E}(U|A) \geq \mathbb{E}(U|B),$$

where the expected utility is

$$\mathbb{E}(U|A) = \sum_r U(r)\,\mathbb{P}(r|A).$$

In the above example, $r \in \{0, 100, 200\}$ and $U(r)$ is increasing, and the coin is fair.

Risk and monetary rewards

- If U is convex, we are risk-seeking.
- If U is linear, we are risk neutral.
- If U is concave, we are risk-averse.

Uncertain rewards

- Decisions $a \in \mathcal{A}$
- Each choice is called a reward $r \in \mathcal{R}$.
- There is a utility function $U : \mathcal{R} \to \mathbb{R}$, assigning values to reward.
- We (weakly) prefer A to B iff $U(A) \geq U(B)$.

Example 7

You are going to work, and it might rain. What do you do?

- a_1: Take the umbrella.
- a_2: Risk it!
- ω_1: rain
- ω_2: dry

$\rho(\omega, a)$	a_1	a_2
ω_1	dry, carrying umbrella	wet
ω_2	dry, carrying umbrella	dry

$U[\rho(\omega, a)]$	a_1	a_2
ω_1	0	-10
ω_2	0	1

Table : Rewards and utilities.

Uncertain rewards

- Decisions $a \in \mathcal{A}$
- Each choice is called a reward $r \in \mathcal{R}$.
- There is a utility function $U : \mathcal{R} \to \mathbb{R}$, assigning values to reward.
- We (weakly) prefer A to B iff $U(A) \geq U(B)$.

Example 7

You are going to work, and it might rain. What do you do?

- a_1: Take the umbrella.
- a_2: Risk it!
- ω_1: rain
- ω_2: dry

$\rho(\omega, a)$	a_1	a_2
ω_1	dry, carrying umbrella	wet
ω_2	dry, carrying umbrella	dry

$U[\rho(\omega, a)]$	a_1	a_2
ω_1	0	-10
ω_2	0	1

Table : Rewards and utilities.

- $\max_a \min_\omega U = 0$

Uncertain rewards

- Decisions $a \in \mathcal{A}$
- Each choice is called a reward $r \in \mathcal{R}$.
- There is a utility function $U : \mathcal{R} \to \mathbb{R}$, assigning values to reward.
- We (weakly) prefer A to B iff $U(A) \geq U(B)$.

Example 7

You are going to work, and it might rain.
What do you do?

- a_1: Take the umbrella.
- a_2: Risk it!
- ω_1: rain
- ω_2: dry

$\rho(\omega, a)$	a_1	a_2
ω_1	dry, carrying umbrella	wet
ω_2	dry, carrying umbrella	dry

$U[\rho(\omega, a)]$	a_1	a_2
ω_1	0	-10
ω_2	0	1

Table : Rewards and utilities.

- $\max_a \min_\omega U = 0$
- $\min_\omega \max_a U = 0$

Expected utility

$$\mathbb{E}(U \mid a) = \sum_r U[\rho(\omega, a)] \, \mathbb{P}(\omega \mid a)$$

Example 8

You are going to work, and it might rain. The forecast said that the probability of rain (ω_1) was 20%. What do you do?

- a_1: Take the umbrella.
- a_2: Risk it!

$\rho(\omega, a)$	a_1	a_2
ω_1	dry, carrying umbrella	wet
ω_2	dry, carrying umbrella	dry
$U[\rho(\omega, a)]$	a_1	a_2
ω_1	0	-10
ω_2	0	1
$\mathbb{E}_P(U \mid a)$	0	-1.2

Table : Rewards, utilities, expected utility for 20% probability of rain.

Bayes decision rules

Consider the case where outcomes are independent of decisions:

$$U(\xi, a) \triangleq \sum_{\mu} U(\mu, a)\xi(\mu)$$

This corresponds e.g. to the case where $\xi(\mu)$ is the belief about an unknown world.

Definition 9 (Bayes utility)

The maximising decision for ξ has an expected utility equal to:

$$U^*(\xi) \triangleq \max_{a \in \mathcal{A}} U(\xi, a). \tag{2.1}$$

The n-meteorologists problem

Exercise 4

- Meteorological models $\mathcal{M} = \{\mu_1, \ldots, \mu_n\}$
- Rain predictions at time t: $p_{t,\mu} \triangleq P_\mu(x_t = \mathrm{rain})$.
- Prior probability $\xi(\mu) = 1/n$ for each model.
- Should we take the umbrella?

	M	T	W	T	F	S	S
CNN	0.5	0.6	0.7	0.9	0.5	0.3	0.1
SMHI	0.3	0.7	0.8	0.9	0.5	0.2	0.1
YR	0.6	0.9	0.8	0.5	0.4	0.1	0.1
Rain?	Y	Y	Y	N	Y	N	N

Table : Predictions by three different entities for the probability of rain on a particular day, along with whether or not it actually rained.

The n-meteorologists problem

Exercise 4

	M	T	W	T	F	S	S
CNN	0.5	0.6	0.7	0.9	0.5	0.3	0.1
SMHI	0.3	0.7	0.8	0.9	0.5	0.2	0.1
YR	0.6	0.9	0.8	0.5	0.4	0.1	0.1
Rain?	Y	Y	Y	N	Y	N	N

Table : Predictions by three different entities for the probability of rain on a particular day, along with whether or not it actually rained.

1. What is your belief about the quality of each meteorologist after each day?

The n-meteorologists problem

Exercise 4

	M	T	W	T	F	S	S
CNN	0.5	0.6	0.7	0.9	0.5	0.3	0.1
SMHI	0.3	0.7	0.8	0.9	0.5	0.2	0.1
YR	0.6	0.9	0.8	0.5	0.4	0.1	0.1
Rain?	Y	Y	Y	N	Y	N	N

Table : Predictions by three different entities for the probability of rain on a particular day, along with whether or not it actually rained.

1. What is your belief about the quality of each meteorologist after each day?
2. What is your belief about the probability of rain each day?

$$P_\xi(x_t = \mathrm{rain} \mid x_1, x_2, \dots x_{t-1}) = \sum_{\mu \in \mathcal{M}} P_\mu(x_t = \mathrm{rain} \mid x_1, x_2, \dots x_{t-1}) \xi(\mu \mid x_1, x_2, \dots x_{t-1})$$

The *n*-meteorologists problem

Exercise 4

	M	T	W	T	F	S	S
CNN	0.5	0.6	0.7	0.9	0.5	0.3	0.1
SMHI	0.3	0.7	0.8	0.9	0.5	0.2	0.1
YR	0.6	0.9	0.8	0.5	0.4	0.1	0.1
Rain?	Y	Y	Y	N	Y	N	N

Table : Predictions by three different entities for the probability of rain on a particular day, along with whether or not it actually rained.

1. What is your belief about the quality of each meteorologist after each day?
2. What is your belief about the probability of rain each day?

$$P_\xi(x_t = \text{rain} \mid x_1, x_2, \ldots x_{t-1}) = \sum_{\mu \in \mathcal{M}} P_\mu(x_t = \text{rain} \mid x_1, x_2, \ldots x_{t-1})\xi(\mu \mid x_1, x_2, \ldots x_{t-1})$$

3. Assume you can decide whether or not to go running each day. If you go running and it does not rain, your utility is 1. If it rains, it's -10. If you don't go running, your utility is 0. What is the decision maximising utility in expectation (with respect to the posterior) each day?

Deciding a class given a model

- Features $x_t \in \mathcal{X}$.
- Label $y_t \in \mathcal{Y}$.
- Decisions $a_t \in \mathcal{A}$.
- Decision rule $\pi(a_t \mid x_t)$ assigns probabilities to actions.

Standard classification problem

$$\mathcal{A} = \mathcal{Y}, \qquad U(a, y) = \mathbb{I}\{a = y\}$$

Exercise 5

If we have a model $P_\mu(y_t \mid x_t)$, and a suitable U, what is the optimal decision to make?

Deciding a class given a model

- Features $x_t \in \mathcal{X}$.
- Label $y_t \in \mathcal{Y}$.
- Decisions $a_t \in \mathcal{A}$.
- Decision rule $\pi(a_t \mid x_t)$ assigns probabilities to actions.

Standard classification problem

$$\mathcal{A} = \mathcal{Y}, \qquad U(a, y) = \mathbb{I}\{a = y\}$$

Exercise 5

If we have a model $P_\mu(y_t \mid x_t)$, and a suitable U, what is the optimal decision to make?

$$a_t \in \arg\max_{a \in \mathcal{A}} \sum_y P_\mu(y_t = y \mid x_t) U(a, y)$$

Deciding a class given a model

- Features $x_t \in \mathcal{X}$.
- Label $y_t \in \mathcal{Y}$.
- Decisions $a_t \in \mathcal{A}$.
- Decision rule $\pi(a_t \mid x_t)$ assigns probabilities to actions.

Standard classification problem

$$\mathcal{A} = \mathcal{Y}, \qquad U(a, y) = \mathbb{I}\{a = y\}$$

Exercise 5

If we have a model $P_\mu(y_t \mid x_t)$, and a suitable U, what is the optimal decision to make?

$$a_t \in \arg\max_{a \in \mathcal{A}} \sum_y P_\mu(y_t = y \mid x_t) U(a, y)$$

For standard classification,

$$a_t \in \arg\max_{a \in \mathcal{A}} P_\mu(y_t = a \mid x_t)$$

Deciding the class given a model family

- Training data $D_T = \{(x_i, y_i) \mid i = 1, \ldots, T\}$
- Models $\{P_\mu \mid \mu \in \mathcal{M}\}$.
- Prior ξ on $\mathcal{M}$.

Posterior over classification models

$$\xi(\mu \mid D_T) = \frac{P_\mu(y_1, \ldots, y_T \mid x_1, \ldots, x_T)\xi(\mu)}{\sum_{\mu' \in \mathcal{M}} P_{\mu'}(y_1, \ldots, y_T \mid x_1, \ldots, x_T)\xi(\mu')}$$

Deciding the class given a model family

- Training data $D_T = \{(x_i, y_i) \mid i = 1, \ldots, T\}$
- Models $\{P_\mu \mid \mu \in \mathcal{M}\}$.
- Prior ξ on $\mathcal{M}$.

Posterior over classification models

$$\xi(\mu \mid D_T) = \frac{P_\mu(y_1, \ldots, y_T \mid x_1, \ldots, x_T)\xi(\mu)}{\sum_{\mu' \in \mathcal{M}} P_{\mu'}(y_1, \ldots, y_T \mid x_1, \ldots, x_T)\xi(\mu')}$$

If not dealing with time-series data, we assume independence between x_t:

$$P_\mu(y_1, \ldots, y_T \mid x_1, \ldots, x_T) = \prod_{i=1}^{T} P_\mu(y_i \mid x_i)$$

Deciding the class given a model family

- Training data $D_T = \{(x_i, y_i) \mid i = 1, \ldots, T\}$
- Models $\{P_\mu \mid \mu \in \mathcal{M}\}$.
- Prior ξ on $\mathcal{M}$.

Posterior over classification models

$$\xi(\mu \mid D_T) = \frac{P_\mu(y_1, \ldots, y_T \mid x_1, \ldots, x_T)\xi(\mu)}{\sum_{\mu' \in \mathcal{M}} P_{\mu'}(y_1, \ldots, y_T \mid x_1, \ldots, x_T)\xi(\mu')}$$

The Bayes rule for maximising $\mathbb{E}_\xi(U \mid a, x_t, D_T)$

The decision rule simply chooses the action:

$$a_t \in \arg\max_{a \in \mathcal{A}} \sum_y \sum_{\mu \in \mathcal{M}} P_\mu(y_t = y \mid x_t)\xi(\mu \mid D_T)U(a, y) \tag{3.1}$$

Deciding the class given a model family

- Training data $D_T = \{(x_i, y_i) \mid i = 1, \ldots, T\}$
- Models $\{P_\mu \mid \mu \in \mathcal{M}\}$.
- Prior ξ on $\mathcal{M}$.

Posterior over classification models

$$\xi(\mu \mid D_T) = \frac{P_\mu(y_1, \ldots, y_T \mid x_1, \ldots, x_T)\xi(\mu)}{\sum_{\mu' \in \mathcal{M}} P_{\mu'}(y_1, \ldots, y_T \mid x_1, \ldots, x_T)\xi(\mu')}$$

The Bayes rule for maximising $\mathbb{E}_\xi(U \mid a, x_t, D_T)$

The decision rule simply chooses the action:

$$a_t \in \arg\max_{a \in \mathcal{A}} \sum_y \sum_{\mu \in \mathcal{M}} P_\mu(y_t = y \mid x_t)\xi(\mu \mid D_T) U(a, y) \tag{3.1}$$

We can rewrite this by calculating the posterior marginal marginal label probability

$$\mathbb{P}_{\xi \mid D_T}(y_t \mid x_t) \triangleq \mathbb{P}_\xi(y_t \mid x_t, D_T) = \sum_{\mu \in \mathcal{M}} P_\mu(y_t \mid x_t)\xi(\mu \mid D_T).$$

Deciding the class given a model family

- Training data $D_T = \{(x_i, y_i) \mid i = 1, \ldots, T\}$
- Models $\{P_\mu \mid \mu \in \mathcal{M}\}$.
- Prior ξ on $\mathcal{M}$.

Posterior over classification models

$$\xi(\mu \mid D_T) = \frac{P_\mu(y_1, \ldots, y_T \mid x_1, \ldots, x_T)\xi(\mu)}{\sum_{\mu' \in \mathcal{M}} P_{\mu'}(y_1, \ldots, y_T \mid x_1, \ldots, x_T)\xi(\mu')}$$

The Bayes rule for maximising $\mathbb{E}_\xi(U \mid a, x_t, D_T)$

The decision rule simply chooses the action:

$$a_t \in \arg\max_{a \in \mathcal{A}} \sum_y \sum_{\mu \in \mathcal{M}} P_\mu(y_t = y \mid x_t)\xi(\mu \mid D_T)U(a, y) \tag{3.1}$$

$$= \arg\max_{a \in \mathcal{A}} \sum_y \mathbb{P}_{\xi \mid D_T}(y_t \mid x_t)U(a, y) \tag{3.2}$$

We can rewrite this by calculating the posterior marginal marginal label probability

$$\mathbb{P}_{\xi \mid D_T}(y_t \mid x_t) \triangleq \mathbb{P}_\xi(y_t \mid x_t, D_T) = \sum_{\mu \in \mathcal{M}} P_\mu(y_t \mid x_t)\xi(\mu \mid D_T).$$

Approximating the model

Full Bayesian approach for infinite $\mathcal{M}$

Here ξ can be a probability density function and

$$\xi(\mu \mid D_T) = P_\mu(D_T)\xi(\mu) / \mathbb{P}_\xi(D_T), \qquad \mathbb{P}_\xi(D_T) = \int_{\mathcal{M}} P_\mu(D_T)\xi(\mu)\,\mathrm{d},$$

can be hard to calculate.

Approximating the model

Full Bayesian approach for infinite $\mathcal{M}$

Here ξ can be a probability density function and

$$\xi(\mu \mid D_T) = P_\mu(D_T)\xi(\mu) / \mathbb{P}_\xi(D_T), \qquad \mathbb{P}_\xi(D_T) = \int_{\mathcal{M}} P_\mu(D_T)\xi(\mu)\,\mathrm{d},$$

can be hard to calculate.

Maximum a posteriori model

We only choose a single model through the following optimisation:

$$\mu_{\mathrm{MAP}}(\xi, D_T) = \arg\max_{\mu \in \mathcal{M}} P_\mu(D_T)\xi(\mu)$$

Approximating the model

Full Bayesian approach for infinite $\mathcal{M}$

Here ξ can be a probability density function and

$$\xi(\mu \mid D_T) = P_\mu(D_T)\xi(\mu) / \mathbb{P}_\xi(D_T), \qquad \mathbb{P}_\xi(D_T) = \int_{\mathcal{M}} P_\mu(D_T)\xi(\mu)\,\mathrm{d},$$

can be hard to calculate.

Maximum a posteriori model

We only choose a single model through the following optimisation:

$$\mu_{\mathrm{MAP}}(\xi, D_T) = \arg\max_{\mu \in \mathcal{M}} \overbrace{\ln P_\mu(D_T)}^{\text{goodness of fit}} + \underbrace{\ln \xi(\mu)}_{\text{regulariser}} .$$

Learning outcomes

Understanding

- Preferences, utilities and the expected utility principle.
- Hypothesis testing and classification as decision problems.
- How to interpret p-values Bayesian tests.
- The MAP approximation to full Bayesian inference.

Skills

- Being able to implement an optimal decision rule for a given utility and probability.
- Being able to construct a simple null hypothesis test.

Reflection

- When would expected utility maximisation not be a good idea?
- What does a p value represent when you see it in a paper?
- Can we prevent high false discovery rates when using p values?
- When is the MAP approximation good?

Simple hypothesis testing

The simple hypothesis test as a decision problem

- $\mathcal{M} = \{\mu_0, \mu_1\}$
- a_0: Accept model μ_0
- a_1: Accept model μ_1

U	μ_0	μ_1
a_0	1	0
a_1	0	1

Table : Example utility function for simple hypothesis tests.

Example 10 (Continuation of the medium example)

- μ_1: that John is a medium.
- μ_0: that John is not a medium.

$$\mathbb{E}_\xi(U \mid a_0) = 1 \times \xi(\mu_0 \mid \boldsymbol{x}) + 0 \times \xi(\mu_1 \mid \boldsymbol{x}), \qquad \mathbb{E}_\xi(U \mid a_1) = 0 \times \xi(\mu_0 \mid \boldsymbol{x}) + 1 \times \xi(\mu_1 \mid \boldsymbol{x})$$

Null hypothesis test

Many times, there is only one model under consideration, μ_0, the so-called null hypothesis.

The null hypothesis test as a decision problem

- a_0: Accept model μ_0
- a_1: Reject model μ_0

Example 11

Construction of the test for the medium

Null hypothesis test

Many times, there is only one model under consideration, μ_0, the so-called null hypothesis.

The null hypothesis test as a decision problem

- a_0: Accept model μ_0
- a_1: Reject model μ_0

Example 11

Construction of the test for the medium

- μ_0 is simply the $\mathcal{Bernoulli}(1/2)$ model: responses are by chance.

Null hypothesis test

Many times, there is only one model under consideration, μ_0, the so-called null hypothesis.

The null hypothesis test as a decision problem

- a_0: Accept model μ_0
- a_1: Reject model μ_0

Example 11

Construction of the test for the medium

- μ_0 is simply the $\mathcal{Bernoulli}(1/2)$ model: responses are by chance.
- We need to design a policy $\pi(a \mid x)$ that accepts or rejects depending on the data.

Null hypothesis test

Many times, there is only one model under consideration, μ_0, the so-called null hypothesis.

The null hypothesis test as a decision problem

- a_0: Accept model μ_0
- a_1: Reject model μ_0

Example 11

Construction of the test for the medium

- μ_0 is simply the $\mathcal{Bernoulli}(1/2)$ model: responses are by chance.
- We need to design a policy $\pi(a \mid x)$ that accepts or rejects depending on the data.
- Since there is no alternative model, we can only construct this policy according to its properties when μ_0 is true.

Null hypothesis test

Many times, there is only one model under consideration, μ_0, the so-called null hypothesis.

The null hypothesis test as a decision problem

- a_0: Accept model μ_0
- a_1: Reject model μ_0

Example 11

Construction of the test for the medium

- μ_0 is simply the $\mathcal{Bernoulli}(1/2)$ model: responses are by chance.
- We need to design a policy $\pi(a \mid \boldsymbol{x})$ that accepts or rejects depending on the data.
- Since there is no alternative model, we can only construct this policy according to its properties when μ_0 is true.
- In particular, we can fix a policy that only chooses a_1 when μ_0 is true a proportion δ of the time.

Null hypothesis test

Many times, there is only one model under consideration, μ_0, the so-called null hypothesis.

The null hypothesis test as a decision problem

- a_0: Accept model μ_0
- a_1: Reject model μ_0

Example 11

Construction of the test for the medium

- μ_0 is simply the $\mathcal{B}ernoulli(1/2)$ model: responses are by chance.
- We need to design a policy $\pi(a \mid \boldsymbol{x})$ that accepts or rejects depending on the data.
- Since there is no alternative model, we can only construct this policy according to its properties when μ_0 is true.
- In particular, we can fix a policy that only chooses a_1 when μ_0 is true a proportion δ of the time.
- This can be done by construcing a threshold test from the inverse-CDF.

Using p-values to construct statistical tests

Definition 12 (Null statistical test)

The statistic $f : \mathcal{X} \to [0, 1]$ is designed to have the property:

$$P_{\mu_0}(\{x \mid f(x) \leq \delta\}) = \delta.$$

If our decision rule is:

$$\pi(a \mid x) = \begin{cases} a_0, & f(x) \geq \delta \\ a_1, & f(x) < \delta, \end{cases}$$

the probability of rejecting the null hypothesis when it is true is exactly δ.

The value of the statistic $f(x)$, otherwise known as the p-value, is uninformative.

Issues with p-values

- They only measure quality of fit on the data.
- Not robust to model misspecification.
- They ignore effect sizes.
- They do not consider prior information.
- They do not represent the probability of having made an error.
- The null-rejection error probability is the same irrespective of the amount of data (by design).

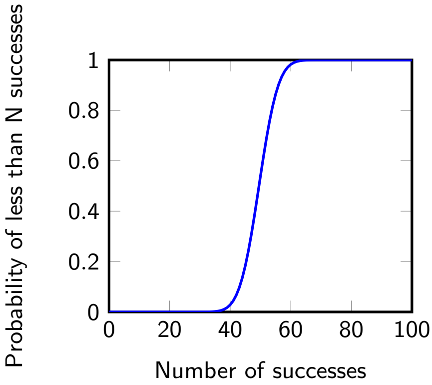
p-values for the medium example

p-values for the medium example

- μ_0 is simply the $\mathcal{Bernoulli}(1/2)$ model: responses are by chance.

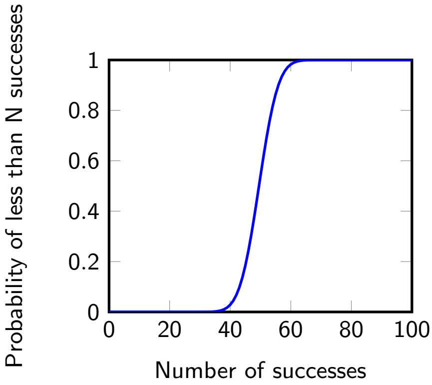
p-values for the medium example

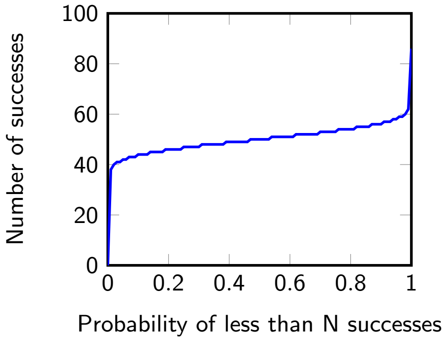
- μ_0 is simply the $\mathcal{Bernoulli}(1/2)$ model: responses are by chance.
- CDF: $P_{\mu_0}(N \leq n \mid K = 100)$



p-values for the medium example

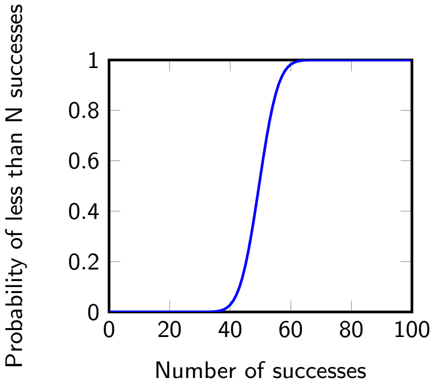
- μ_0 is simply the $\mathcal{B}ernoulli(1/2)$ model: responses are by chance.
- CDF: $P_{\mu_0}(N \leq n \mid K = 100)$
- ICDF: the number of successes that will happen with probability at least δ

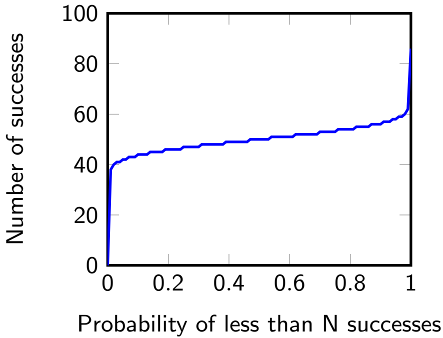




p-values for the medium example

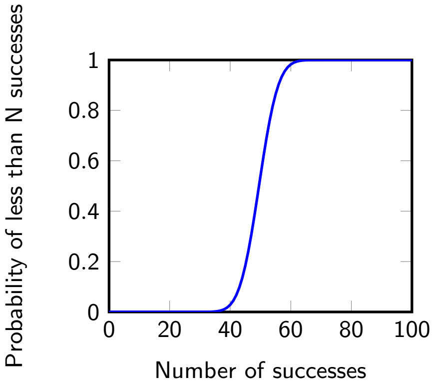
- μ_0 is simply the $\mathcal{B}ernoulli(1/2)$ model: responses are by chance.
- CDF: $P_{\mu_0}(N \leq n \mid K = 100)$
- ICDF: the number of successes that will happen with probability at least δ
- e.g. we'll get at most 50 successes a proportion $\delta = 1/2$ of the time.

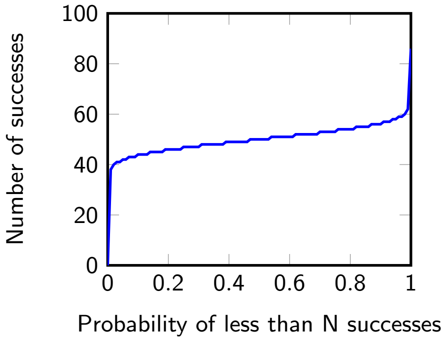




p-values for the medium example

- μ_0 is simply the $\mathcal{Bernoulli}(1/2)$ model: responses are by chance.
- CDF: $P_{\mu_0}(N \leq n \mid K = 100)$
- ICDF: the number of successes that will happen with probability at least δ
- e.g. we'll get at most 50 successes a proportion $\delta = 1/2$ of the time.
- Using the (inverse) CDF we can construct a policy π that selects a_1 when μ_0 is true only a δ portion of the time, for any choice of δ.





Building a test

The test statistic

We want the test to reflect that we don't have a significant number of failures.

$$f(x) = 1 - \mathrm{binocdf}(\sum_{t=1}^{n} x_t, n, 0.5)$$

What $f(x)$ is and is not

- It is a **statistic** which is $\leq \delta$ a δ portion of the time when μ_0 is true.
- It is **not** the probability of observing x under μ_0.
- It is **not** the probability of μ_0 given x.

Exercise 6

- Let us throw a coin 8 times, and try and predict the outcome.

Exercise 6

- Let us throw a coin 8 times, and try and predict the outcome.
- Select a p-value threshold so that $\delta = 0.05$. For 8 throws, this corresponds to

.

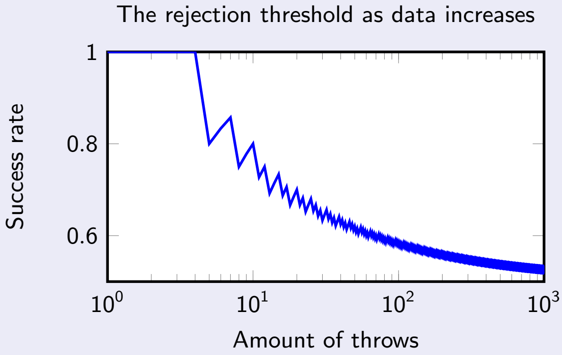


Figure : Here we see how the rejection threshold, in terms of the success rate, changes with the number of throws to achieve an error rate of $\delta = 0.05$.

Exercise 6

- Let us throw a coin 8 times, and try and predict the outcome.
- Select a p-value threshold so that $\delta = 0.05$. For 8 throws, this corresponds to > 6 successes or $\geq 87.5\%$ success rate.
- Let's calculate the p-value for each one of you

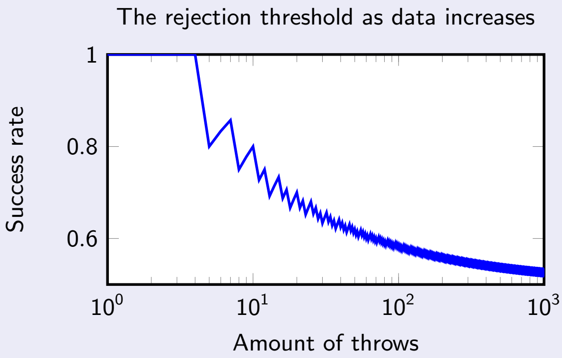


Figure : Here we see how the rejection threshold, in terms of the success rate, changes with the number of throws to achieve an error rate of $\delta = 0.05$.

Exercise 6

- Let us throw a coin 8 times, and try and predict the outcome.
- Select a p-value threshold so that $\delta = 0.05$. For 8 throws, this corresponds to > 6 successes or $\geq 87.5\%$ success rate.
- Let's calculate the p-value for each one of you
- What is the rejection performance of the test?

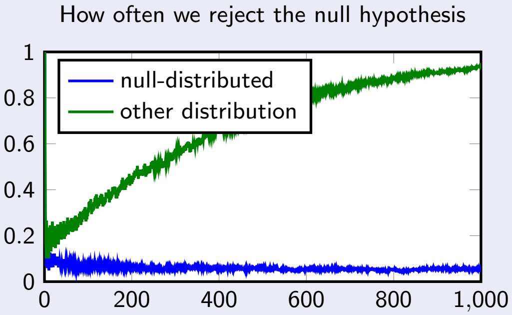


Figure : Here we see the rejection rate of the null hypothesis (μ_0) for two cases. Firstly, for the case when μ_0 is true. Secondly, when the data is generated from $\mathcal{Bernoulli}(0.55)$.

Statistical power and false discovery.

Beyond not rejecting the null when it's true, we also want:

- High power: Rejecting the null when it is false.
- Low false discovery rate: Accepting the null when it is true.

Power

The power depends on what hypothesis we use as an alternative.

False discovery rate

False discovery depends on how likely it is a priori that the null is false.

The Bayesian version of the test

Example 13

1. Set $U(a_i, \mu_j) = \mathbb{I}\{i = j\}$.
2. Set $\xi(\mu_i) = 1/2$.
3. μ_0: $\mathcal{B}ernoulli(1/2)$.
4. μ_1: $\mathcal{B}ernoulli(\theta)$, $\theta \sim \mathcal{U}nif([0, 1])$.
5. Calculate $\xi(\mu \mid x)$.
6. Choose a_i, where $i = \arg\max_j \xi(\mu_j \mid x)$.

Bayesian model averaging for the alternative model μ_1

$$P_{\mu_1}(x) = \int_{\Theta} B_\theta(x) \, \mathrm{d}\beta(\theta) \tag{3.3}$$

$$\xi(\mu_0 \mid x) = \frac{P_{\mu_0}(x)\xi(\mu_0)}{P_{\mu_0}(x)\xi(\mu_0) + P_{\mu_1}(x)\xi(\mu_1)} \tag{3.4}$$

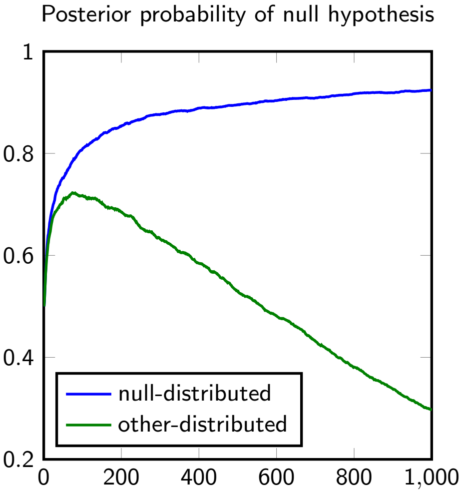


Figure : Here we see the convergence of the posterior probability.

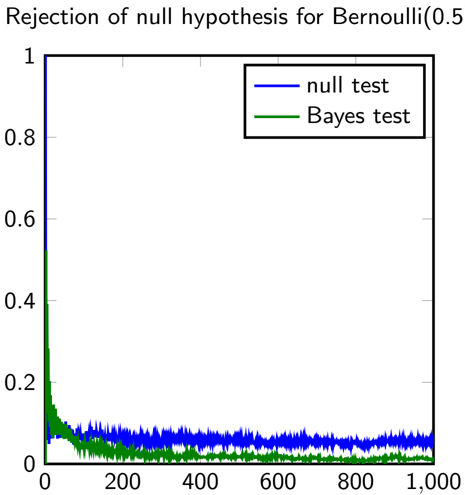


Figure : Comparison of the rejection probability for the null and the Bayesian test when μ_0 is true.

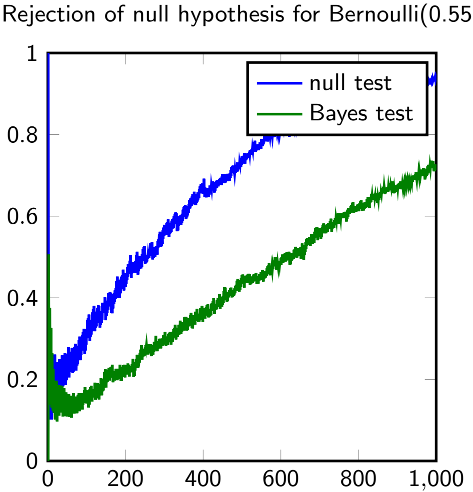


Figure : Comparison of the rejection probability for the null and the Bayesian test when μ_1 is true.

Concentration inequalities and confidence intervals

Further reading

Points of significance (Nature Methods)

- Importance of being uncertain `https://www.nature.com/articles/nmeth.2613`
- Error bars `https://www.nature.com/articles/nmeth.2659`
- P values and the search for significance `https://www.nature.com/articles/nmeth.4120`
- Bayes' theorem `https://www.nature.com/articles/nmeth.3335`
- Sampling distributions and the bootstrap `https://www.nature.com/articles/nmeth.3414`

Classification as an optimisation problem.

The μ-optimal classifier

$$\max_{\theta \in \Theta} f(\pi_\theta, \mu, U), \qquad f(\pi_\theta, \mu, U) \triangleq \mathbb{E}_\mu^{\pi_\theta}(U) \tag{4.1}$$

$$f(\pi_\theta, \mu, U) = \sum_{x,y,a} U(a, y) \pi_\theta(a \mid x) P_\mu(y \mid x) P_\mu(x) \tag{4.2}$$

$$\approx \sum_{t=1}^{T} \sum_{a_t} U(a_t, y_t) \pi_\theta(a_t \mid x_t), \qquad (x_t, y_t)_{t=1}^{T} \sim P_\mu. \tag{4.3}$$

Stochastic gradient methdos

Gradient ascent

$$\theta_{i+1} = \theta_i + \alpha \nabla_\theta g(\theta_i).$$

Stochastic gradient ascent

$$g(\theta) = \int_{\mathcal{M}} f(\theta, \mu) \,\mathrm{d}\xi(\mu)$$

$$\theta_{i+1} = \theta_i + \alpha \nabla_\theta f(\theta_i, \mu_i), \qquad \mu_i \sim \xi.$$

Two views of neural networks

Neural network classification model $P_\theta(\boldsymbol{y} \mid \boldsymbol{x}_t)$

$$\begin{tikzcd} \boldsymbol{x}_t \arrow[r, bend right] & \boldsymbol{y}_t \end{tikzcd}$$

Objective: Find the best model for D_T.

Neural network classification policy $\pi(a_t \mid \boldsymbol{x}_t)$

$$\begin{tikzcd} \boldsymbol{x}_t \arrow[r, bend right] & a_t \end{tikzcd}$$

Objective: Find the best policy for $U(a, \boldsymbol{x})$.

Two views of neural networks

Neural network classification model $P_\theta(\boldsymbol{y} \mid \boldsymbol{x}_t)$

$$\begin{tikzcd} \boldsymbol{x}_t \arrow[r, bend right] & \boldsymbol{y}_t \end{tikzcd}$$

Objective: Find the best model for D_T.

Neural network classification policy $\pi(a_t \mid \boldsymbol{x}_t)$

$$\begin{tikzcd} \boldsymbol{x}_t \arrow[r, bend right] & a_t \end{tikzcd}$$

Objective: Find the best policy for $U(a, \boldsymbol{x})$.

Difference between the two views

- We can use standard probabilistic methods for P.
- Finding the optimal π is an optimisation problem.

Linear networks and the perceptron algorithm

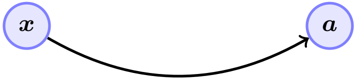


Figure : Abstract graphical model for a neural network

Definition 14 (Linear classifier)

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\theta}_1 & \cdots & \boldsymbol{\theta}_C \end{bmatrix} = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,C} \\ \vdots & \ddots & \vdots \\ \theta_N & \cdots & \theta_{N,C} \end{bmatrix}$$

$$\pi_{\Theta}(a \mid \boldsymbol{x}) = \exp\left(\boldsymbol{\theta}_a^\top \boldsymbol{x}\right) / \sum_{a'} \exp\left(\boldsymbol{\theta}_{a'}^\top \boldsymbol{x}\right)$$

Linear networks and the perceptron algorithm

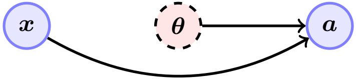


Figure : Abstract graphical model for a neural network

Definition 14 (Linear classifier)

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\theta}_1 & \cdots & \boldsymbol{\theta}_C \end{bmatrix} = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,C} \\ \vdots & \ddots & \vdots \\ \theta_N & \cdots & \theta_{N,C} \end{bmatrix}$$

$$\pi_{\Theta}(a \mid \boldsymbol{x}) = \exp\left(\boldsymbol{\theta}_a^\top \boldsymbol{x}\right) / \sum_{a'} \exp\left(\boldsymbol{\theta}_{a'}^\top \boldsymbol{x}\right)$$

Linear networks and the perceptron algorithm

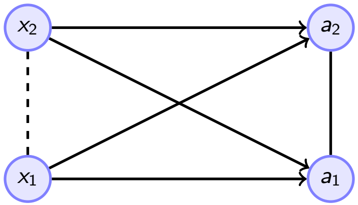


Figure : Graphical model for a linear neural network.

Definition 14 (Linear classifier)

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\theta}_1 & \cdots & \boldsymbol{\theta}_C \end{bmatrix} = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,C} \\ \vdots & \ddots & \vdots \\ \theta_N & \cdots & \theta_{N,C} \end{bmatrix}$$

$$\pi_{\Theta}(a \mid \boldsymbol{x}) = \exp\left(\boldsymbol{\theta}_a^\top \boldsymbol{x}\right) / \sum_{a'} \exp\left(\boldsymbol{\theta}_{a'}^\top \boldsymbol{x}\right)$$

Linear networks and the perceptron algorithm

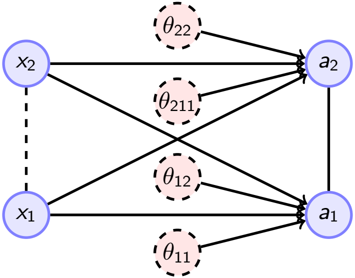


Figure : Graphical model for a linear neural network.

Definition 14 (Linear classifier)

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\theta}_1 & \cdots & \boldsymbol{\theta}_C \end{bmatrix} = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,C} \\ \vdots & \ddots & \vdots \\ \theta_N & \cdots & \theta_{N,C} \end{bmatrix}$$

$$\pi_{\Theta}(a \mid \boldsymbol{x}) = \exp\left(\boldsymbol{\theta}_a^\top \boldsymbol{x}\right) / \sum_{a'} \exp\left(\boldsymbol{\theta}_{a'}^\top \boldsymbol{x}\right)$$

Linear networks and the perceptron algorithm

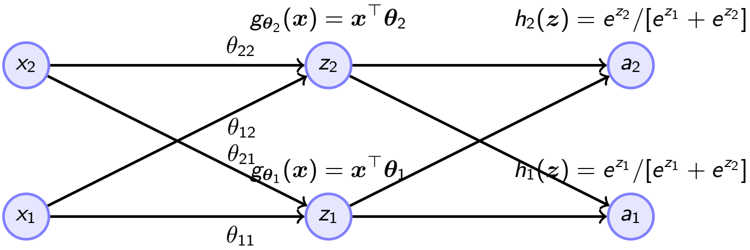


Figure : Architectural view of a linear neural network.

Definition 14 (Linear classifier)

$$\boldsymbol{\Theta} = \begin{bmatrix}\boldsymbol{\theta}_1 & \cdots & \boldsymbol{\theta}_C\end{bmatrix} = \begin{bmatrix}\theta_{1,1} & \cdots & \theta_{1,C} \\ \vdots & \ddots & \vdots \\ \theta_N & \cdots & \theta_{N,C}\end{bmatrix}$$

$$\pi_{\Theta}(a \mid \boldsymbol{x}) = \exp\left(\boldsymbol{\theta}_a^\top \boldsymbol{x}\right) / \sum_{a'} \exp\left(\boldsymbol{\theta}_{a'}^\top \boldsymbol{x}\right)$$

Gradient ascent for a matrix U

$$\max_\theta \sum_{t=1}^{T} \sum_{a_t} U(a_t, y_t) \pi_\theta(a_t \mid x_t) \qquad \text{(objective)}$$

$$\nabla_\theta \sum_{t=1}^{T} \sum_{a_t} U(a_t, y_t) \pi_\theta(a_t \mid x_t) \qquad \text{(gradient)}$$

$$= \sum_{t=1}^{T} \sum_{a_t} U(a_t, y_t) \nabla_\theta \pi_\theta(a_t \mid x_t) \qquad (4.4)$$

Chain Rule of Differentiation

$$f(z), z = g(x), \qquad \frac{df}{dx} = \frac{df}{dg}\frac{dg}{dx} \qquad \text{(scalar version)}$$

$$\nabla_\theta \pi = \nabla_g \pi \nabla_\theta g \qquad \text{(vector version)}$$

Learning outcomes

Understanding

- Classification as an optimisation problem.
- (Stochastic) gradient methods and the chain rule.
- Neural networks as probability models or classification policies.
- Linear neural netwoks.
- Nonlinear network architectures.

Skills

- Using a standard NN class in python.

Reflection

- How useful is the ability to have multiple non-linear layers in a neural network.
- How rich is the representational power of neural networks?
- Is there anything special about neural networks other than their allusions to biology?