Confidence Intervals

Christos Dimitrakakis

October 26, 2023

Outline

Simple hypothesis tests

- Consider n hypotheses, $H_1, \ldots, H_n \in \mathcal{H}$
- Each hypothesis corresponds to a model $P(x|H_i)$ giving a probability value to every possible data $x \in X$.
- Given specific data x, we want to select the most likely model.

Maximum Likelihood

Pick the model with the highest likelihood:

- $\hat{H} = \arg\max_{H_i} P(x|H_i)$

Maximum A Posteriori

- Given prior $P(H_i)$
- Pick $\hat{H} = \arg\max_{H_i} P(H_i|x)$
- We use Bayes's theorem to calculate the posterior $P(H_i|x)$.
- When $P(H_i)$ is uniform, it is the same as maximum likelihood.

The Theorem of Bayes

- Given some probability space (P, Ω, Σ).
- P is a probability measure on Ω
- Ω is the outcome space.
- Σ is a collection of subsets of Ω, corresponding to events.
- Let $\{H_i\}$ be a partition of Ω, i.e.

$$H_i \cup H_j = \emptyset \ \forall i \neq j, \qquad \bigcup_i H_i = \Omega.$$

Then, for any event $A \in \Sigma$, $A \subset \Omega$,

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{\sum_j P(A|H_j)P(H_j)}$$

Proof of Bayes's theorem

Note that $P(H_i \cap A) = P(H_i|A)P(A) = P(A|H_i)P(H_i)$. Rearranging,

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{P(A)}.$$

Since $\{H_i\}$ is a partition,

$$P(A) = P\left(\bigcup_i A \cap H_i\right) = \sum_i P(A \cap H_i) = \sum_i P(A|H_i)P(H_i)$$

Extensions

- We can use any non-negative scoring function $f_h(x)$:

$$P(h|x) = \frac{f_h(x)P(h)}{\sum_{h' \in \mathcal{H}} f_{h'}(x)P(h')}$$

- For infinite $\mathcal{H}$ we can use this notation:

$$P(B|x) = \frac{\int_B f_h(x)dP(h)}{\int_{\mathcal{H}} f_h(x)dP(h)}, \qquad B \subset \mathcal{H}.$$

Null Hypothesis Tests

- Consider a model H_0 such that $P(x|H_0)$ is known.
- We need to compare against an unknown alternative.
- We calculate a statistic $s : X \to \mathbb{R}$ to partition X in S_0, S_1 i.e.

$$S_0 = \{x : s(x) \leq \theta\}, \qquad S_1 = \{x : s(x) > \theta\}$$

- Then $P(S_0|H_0) = 1 - \alpha$, $P(S_1|H_0) = \alpha$ for some α
- We tune θ to achieve the desired α.
- If $x \in S_0$, we accept H_0, otherwise we reject it.

Example statistics

- Likelihood test: $s(x) = P(x|H_0)$,
- Mean test: $s(x) = |x - \mathbb{E}[x|H_0]|^2$.

Likelihood test

- We can use $s(x) = P(x|H_0)$.
- Now we can choose a threshold θ so that:

$$S_1 = \{x : s(x) \geq \theta\}$$

Example: Laplace distributiotan

- Density: $f(x|\mu, \lambda) = \frac{1}{2\lambda} e^{-\frac{1}{\lambda}|x-\mu|}$
- H_0: $x \sim \mathrm{Laplace}(0, 1)$.
- $f(x|0, 1) \geq \theta$ means $|x| \leq \ln(1/2\theta)$. So

$$P(S_1|H_0) = \int_{-infty}^{-} \ln(1/2\theta) e^{-x} dx = 1/2\theta.$$

- Consequently, $\theta = 1/2(1-\alpha)$, i.e. we accept H_0 if $|x| \leq \ln(4 - 4\alpha)$

Bernoulli test

- H_0: The coin tosses are fair
- Then the probability of any sequence $x = x_1, \ldots, x_T$ is 2^{-T}.
- The expected number of heads is $T/2$.
- Statistic $s(x) = \sum_t x_t$.
- Select interval $S = [cT, (1-c)T]$.
- There is some $c \in [0, 1/2]$ so that $P(S|H_0) = 1 - \alpha$
- To calculate c we can use the inverse CDF of s.

p values

How to use p values

- First select a significance threshold α.
- Collect the data, obtain the p value
- If $p \leq \alpha$, reject the null hypothesis H_0.
- This ensures that, if H_0 is true, the probability of rejecting it is exactly α!

Problems with p values

- They do not 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).

Mean estimation

- Data $D = x_1, \ldots, x_T$
- i.i.d samples $x_t \sim P$
- Expectation $\mathbb{E}_P(x_t) = \mu$,
- Empirical mean:

$$\hat{\mu}(D) = \frac{1}{T} \sum_{t=1}^{T} x_t.$$

The error of the empirical mean

Since the data D is random, what is the probability that our estimate is far away from μ?

$$\mathbb{P}[|\hat{\mu}(D) - \mu| > \epsilon] \leq \delta.$$

This means that the probability that our error is larger than ϵ is at most δ, with s $\epsilon, \delta > 0$.

Two methods:

- Distribution-specific confidence intervals
- Concentration inequalities

Distribution-specific intervals

Bernoulli

If $x_t \sim \mathrm{Bernoulli}(\mu)$, then the distribution of $\hat{\mu}$ is given by the Binomial distribution.

Binomial

Let $n_t = \sum_{i=1}^{t} x_i$, where $x_t \sim \mathrm{Bernoulli}(\mu)$. Then n_t has a binomial distribution with parameter μ and t trials, i.e. $n_t \sim \mathrm{Binomial}(\mu, t)$, and its probability function is

$$\mathbb{P}(n_t = k) = \binom{t}{k} \mu^k (1 - \mu)^{1-k}$$

Markov's Inequality

If $x \geq 0$

$$\mathbb{P}(x \geq u) \leq \frac{\mathbb{E}[x]}{u}$$

Proof

$$\begin{aligned}
\mathbb{E}[x] &= \int_0^\infty xp(x)dx && (1)\\
&= \int_0^u xp(x)dx + \int_u^\infty xp(x)dx && (2)\\
&\geq \int_u^\infty up(x)dx && (3)\\
&= u\,\mathbb{P}(x \geq u) && (4)
\end{aligned}$$

Chernoff bound

$$\mathbb{P}(x - \mu \geq u) = \mathbb{P}(e^{\lambda(x-\mu)} \geq e^{\lambda u}) \leq \frac{\mathbb{E}[e^{\lambda(x-\mu)}]}{e^{\lambda u}}$$

- This follows directly from Markov's inequality.
- Tuning λ gives us the tightest bound.

Normal tail bound

Moment generating function

If $x \sim \mathrm{Normal}(\mu, \sigma^2)$ then

$$\mathbb{E}[e^{\lambda x}] = e^{\mu\lambda + \sigma^2\lambda^2/2}$$

Proof

$$\mathbb{E}[e^{\lambda x}] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda x} e^{-\frac{|x-\mu|^2}{2\sigma^2}}\, dx \qquad = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda x - \frac{|x-\mu|^2}{2\sigma^2}}\, dx \tag{5}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda(\sqrt{2}\sigma y + \mu) - y^2}\, dy \tag{6}$$

where $y = (x - \mu)/(\sqrt{2}\sigma)$, so $x = \sqrt{2}\sigma y + \mu$.

Normal tail bound

If $x_t \sim \mathrm{Normal}(\mu, 1)$, then

$$\mathbb{P}(|x_t - \mu| > \epsilon) \le 2e^{-\epsilon^2/2}$$

Normal bound

- $\hat{\mu} \sim \mathrm{Normal}(\mu, 1/T)$.
- For any $c > 0$, $\mathbb{V}[cx] = c\,\mathbb{V}[x] \Rightarrow T\hat{\mu} \sim \mathrm{Normal}(T\mu, 1)$. So:

$$\begin{aligned}
\mathbb{P}(|T\hat{\mu} - T\mu| \geq \epsilon) &\leq 2e^{-\epsilon^2/2} && \text{from the tail bound} && (7)\\
\mathbb{P}(|\hat{\mu} - \mu| \geq \epsilon/T) &\leq 2e^{-\epsilon^2/2} && \text{as } a \geq b \Leftrightarrow ca \geq cb \text{ for } c > 0 && (8)\\
\mathbb{P}(|\hat{\mu} - \mu| \geq u) &\leq 2e^{-T^2u^2/2} && \text{where } u = \epsilon/T && (9)
\end{aligned}$$

Bayesian Reasoning

You are tested for COVID are found negative. The doctor says that the probability of a false positive (i.e. that the probability that the test is positive if you do not have COVID) is $1/10$ and the probability of a negative test if you have COVID is $1/5$. The prevalence of COVID in the population in the population $1/10$. What is the probability that you actually have COVID?

Exercise

A statistical test

Consider the null hypothesis H_0 that $x_t \sim (1/2)$ and the sample mean $\hat{\mu_T} = \frac{1}{T}\sum_{t=1}^{T} x_t$. The probability of making an error of more than ϵ is

$$1 - \sum_{k=T\epsilon} T\epsilon$$