MS-A0503 First course in probability and statistics

4A Parameter estimation

Jukka Kohonen

Deparment of mathematics and systems analysis
Aalto SCI

Academic year 2021–2022
Period III

Contents

Statistical inference

Goal: To infer what kind of process created the observed data.

1. Choose a suitable stochastic model for the process.
   family of probability distributions, e.g. "all normal distributions" or "all uniform distributions $[0, m]$"
2. Fit the model to the data (estimate the model parameters)
3. Perform calculations based on the fitted model
4. Make inference and decisions

We try to "guess" the truth out there.

- What is the true distance of a star, when four measurements gave 4.0, 4.2, 4.3 and 6.0 astronomical units?
- How many Finns will vote for party X, when in the latest poll 140 out of 1000 said they would do so?
- Will the price of crude oil rise or fall during this year?

Contents

Knowing a distribution, except for its parameters

We know/assume our data comes from a distribution with density $f(x)$, from known family but some parameters are unknown.

E.g. (only one unknown parameter):

- Bernoulli distribution: $f_p(1) = p$ and $f_p(0) = 1 - p$
- Exponential distribution: $f_\lambda(x) = \lambda e^{-\lambda x}, x > 0$
- Uniform over interval $[0, b]$: $f_b(x) = \frac{1}{b}$

E.g. (2 unknown parameters):

- Uniform over interval $[a, b]$: $f_{a,b}(x) = \frac{1}{b-a}$
- Normal distribution: $f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Having observed data $(x_1, \ldots, x_n)$, what is the best guess for the value of the unknown parameter?

Notation: Here a subscript contains parameters that specify one particular density function from a family (and not the name of a random variable like $f_X(x)$). Another notation (e.g. Ross) is with vertical bar: $f(x \mid \lambda)$.

Parameter estimation

We know/assume our data comes from a distribution with density $f_\theta(x)$, from known family but with unknown parameter(s) θ.

We have obtained n independent observations $x_1, \dots, x_n$, each from that same distribution f_θ.

For the parameter θ:

- an estimate is a guess of the value of θ, calculated from data $\vec{x} = (x_1, \dots, x_n)$ by some rule.
- an estimator is a function (calculating rule) $(x_1, \dots, x_n) \mapsto g(x_1, \dots, x_n)$ that gives an estimate.

For a given parameter, there might not be a unique "best" estimator.

We can form several desirable properties that an estimator should have. On this lecture: maximum likelihood and unbiasedness. But these might be contradictory.

Example: Proportion of defectives

A factory is producing components, and each has (independently) probability p of being defective. We have inspected 200 components and observed 22 to be defective. How should we estimate the unknown parameter p?

One natural choice is the *observed* proportion

$$\hat{p} = \frac{22}{200} = 11\%$$

But is this the best estimate, in some sense? Are there other possibilities?

Notation: Hatted letters $\hat{p}$ usually denote *estimated* values, and hatless letters p might denote the true value in the generating distribution or population.

Example: Parameter for discrete uniform distribution

We assume the enemy has n battle tanks with serial numbers $1, 2, \ldots, n$. We have captured three tanks whose serial numbers were $x_1 = 63$, $x_2 = 17$, $x_3 = 203$. How should we estimate n, which is an unknown parameter?

Assuming each captured tank is randomly one of the n tanks, its serial number has discrete uniform distribution

$$f_n(k) = \begin{cases} \frac{1}{n}, & k = 1, \ldots, n, \\ 0, & \text{otherwise.} \end{cases}$$

Here, after some thought, we will find at least two *different* “natural” estimators $\hat{n}(\vec{x})$. Each has some nice properties but they give different numerical values. More about this in Exercise 4B.

See also Wikipedia: German tank problem.

Contents

Likelihood function

Stochastic model: n independent observations $(X_1, \ldots, X_n)$, each from density f_θ.

According to the model, the probability of obtaining the values $(x_1, \ldots, x_n)$ (which we observed) is

$$P(X_1 = x_1,\ \ldots,\ X_n = x_n) \ = \ f_\theta(x_1) \cdots f_\theta(x_n)$$

in the discrete case. For continuous case (with ε small)

$$P(X_1 = x_1 \pm \frac{\varepsilon}{2},\ \ldots,\ X_n = x_n \pm \frac{\varepsilon}{2}) \ \approx \ \varepsilon^n f_\theta(x_1) \cdots f_\theta(x_n).$$

We define the likelihood function $L(\theta) = f_\theta(x_1) \cdots f_\theta(x_n)$, which indicates how probable our observed data was, according to the model f_θ, *if* the parameter had value θ.

Maximum likelihood estimate (ML estimate)

Likelihood function $L(\theta) = f_\theta(x_1) \cdots f_\theta(x_n)$ indicates how probable our observed data was, according to the model f_θ, *if* the parameter had value θ.

We would like to find a value of θ that assigns high probability for our observed data, because that makes it easy to believe that f_θ can actually have produced such data.

(More about this on later lectures about Bayesian inference.)

In fact we want the θ that maximizes the likelihood function. We call it the maximum likelihood estimate $\hat{\theta} = \hat{\theta}(\vec{x})$.

To find the point where a function is maximized ... is a typical problem solved in differential calculus!

Note that data x is given — we cannot change that. The only thing we can change is θ.

Example: Proportion of defectives

A factory is producing components, and each has (independently) probability p of being defective. We have inspected 200 components and observed 22 to be defective.
But p is unknown. Find its ML estimate.

First we form the stochastic model. If we inspect $n = 200$ components, we will see K defectives, where K follows the binomial distribution with parameters n and p:

$$f_p(k) \;=\; P(K = k) \;=\; \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, 200$$

So which value of p maximizes this likelihood function?

$$L(p) \;=\; \binom{200}{22} p^{22} (1-p)^{200-22}$$

We only have one free variable p, so we are maximizing a one-variable function. (The quantities $n = 200$ and $k = 22$ are given and fixed, we cannot change them.)

Example: Proportion of defectives

$$L(p) \;=\; \binom{200}{22} p^{22}(1-p)^{178}$$

attains its maximum when $\ell(p) = \log L(p)$ attains its maximum, and

$$\ell(p) \;=\; \log f_p(22) \;=\; \log \binom{200}{22} + 22 \log p + 178 \log(1-p)$$

$$\ell'(p) \;=\; 22 \frac{1}{p} - 178 \frac{1}{1-p}$$

$$\ell''(p) \;=\; -22 \frac{1}{p^2} - 178 \frac{1}{(1-p)^2} \;\leq\; 0$$

Thus the ML estimate for p is found where ℓ' is zero:

$$\ell'(p) = 0 \quad \Longleftrightarrow \quad \frac{22}{p} \;=\; \frac{178}{1-p} \quad \Longleftrightarrow \quad p = \frac{22}{200}$$

Taking the logaritm was just a trick for getting a nicer derivative. Alternatively, we could have tried to maximize the function L directly.

ML estimate for the binomial probability parameter

Fact

If K follows $\mathrm{Bin}(n, p)$, with n known but p unknown, and we observed $K = k$, then the ML estimate for p is

$$\hat{p} \;=\; \frac{k}{n}.$$

Proof.

Repeat the previous calculation with $200 \mapsto n$ and $22 \mapsto k$. $\square$

ML estimates for the two parameters of normal

The density function for a normal distribution

$$f_{(\mu,\sigma)}(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

has two parameters μ and σ. What if both are unknown?

Fact

Having observed $\vec{x} = (x_1, \ldots, x_n)$, the ML estimates for (μ, σ) are

$$\hat{\mu} = m(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \textit{and} \qquad \hat{\sigma} = \mathrm{sd}(\vec{x}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m(\vec{x}))^2}$$

that is, the average and standard deviation of the observed data $\vec{x}$ (note: using divisor n, not $n-1$).

Proof: Take both partial derivatives (w.r.t. both parameters), set them to zero and solve. See e.g. Ross p. 242.

Contents

Statistical inference

Parametric probability distributions

Maximum Likelihood (ML) estimators

Some desirable properties of estimators

Unbiased estimator

Suppose the data $\vec{X} = (X_1, \dots, X_n)$ are coming from distribution f_θ, with θ unknown. We are using an estimator $\vec{x} \mapsto \hat{\theta}(\vec{x})$. So the estimate we compute is a random variable $\hat{\theta}(\vec{X})$.

We say our estimator is unbiased if

$$\mathbb{E}\hat{\theta}(\vec{X}) \;=\; \theta$$

that is, if the *expectation* of our estimator is "correct".

Long-run interpretation: If we took many such n-element samples, we would get a series of (varying) estimates $\hat{\theta}$, but at least *on average* they would equal θ.

Example: Proportion of defectives

Recall that the ML-estimate for the p parameter of $\mathrm{Bin}(n, p)$ having seen k defectives in n components, is

$$\hat{p}(k) \ = \ \frac{k}{n}.$$

Now suppose p is the true probability (for each component to be defective). Then K follows $\mathrm{Bin}(n, p)$, and we the *expectation* of the estimate that we compute is

$$\mathbb{E}\left(\hat{p}(K)\right) \ = \ \mathbb{E}\left(\frac{K}{n}\right) \ = \ \frac{1}{n}\mathbb{E}(K) \ = \ \frac{1}{n} \times np \ = \ p.$$

Thus the function we are using,

$$k \mapsto \hat{p}(k)$$

is an unbiased estimator for the parameter k.

Example: Normal distribution, ML-estimator of μ

Recall that the ML-estimate for the μ parameter of normal distribution is

$$m(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

If the data X_i are normal with mean μ, then

$$\mathbb{E}[m(\vec{X})] \;=\; \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \;=\; \mu,$$

so the function m is an unbiased estimator for μ.

Example: Normal distribution, ML-estimator for σ^2

The value of σ^2 (variance parameter) that maximizes the likelihood is the variance of the empirical distribution,

$$\mathrm{var}(\vec{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - m(x))^2.$$

If the data X_i are normal with mean μ and variance σ^2, then

$$\mathbb{E}[\mathrm{var}(\vec{X})] \;=\; \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - m(\vec{X})^2)\right) \;=\; \cdots \;=\; \frac{n-1}{n}\sigma^2,$$

thus our ML-estimator $\mathrm{var}(\vec{x})$ is biased. On average it is too small!

Since we know the bias, we could *correct* it by multiplying by $n/(n-1)$. We get the so called (Bessel-)corrected sample variance

$$\mathrm{var}_s(\vec{x}) \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m(\vec{x}))^2.$$

which is unbiased, but no longer ML-estimator!
(If n is large, there is not much difference.)

On next lecture, we form “confidence intervals” for our parameters.