Please scan this code to register your presence on Ilias
for contact-tracing purposes. Your Ilias identity will be used.
– Only do this if you are *physically present*! –

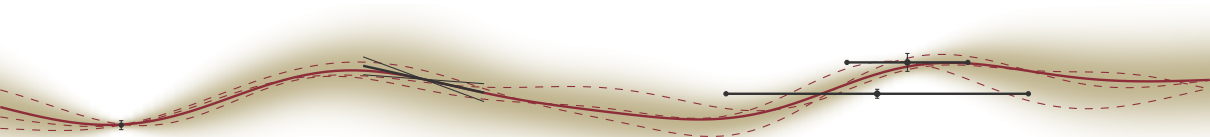
Data Literacy
Lecture 03
Estimation

Philipp Hennig

08 November 2021

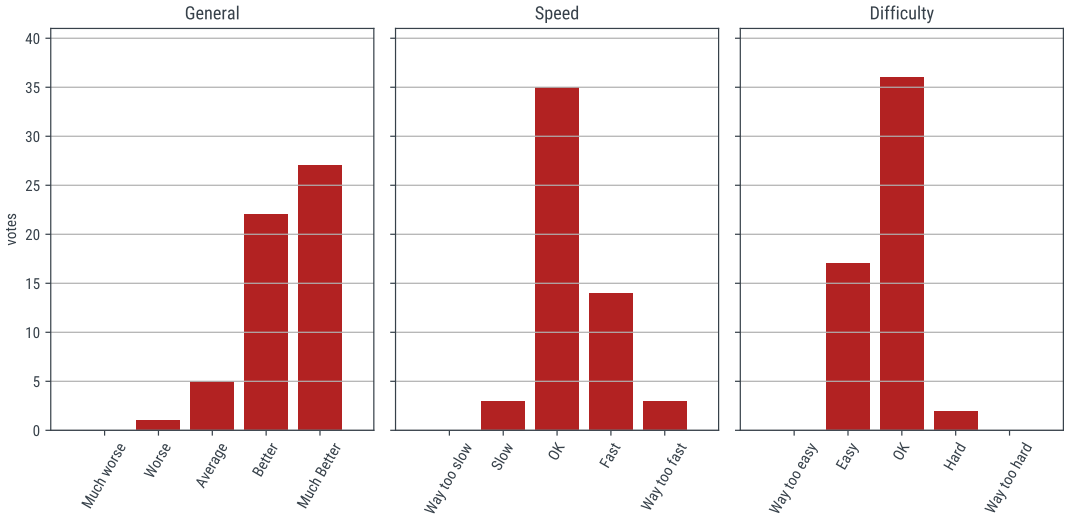
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning



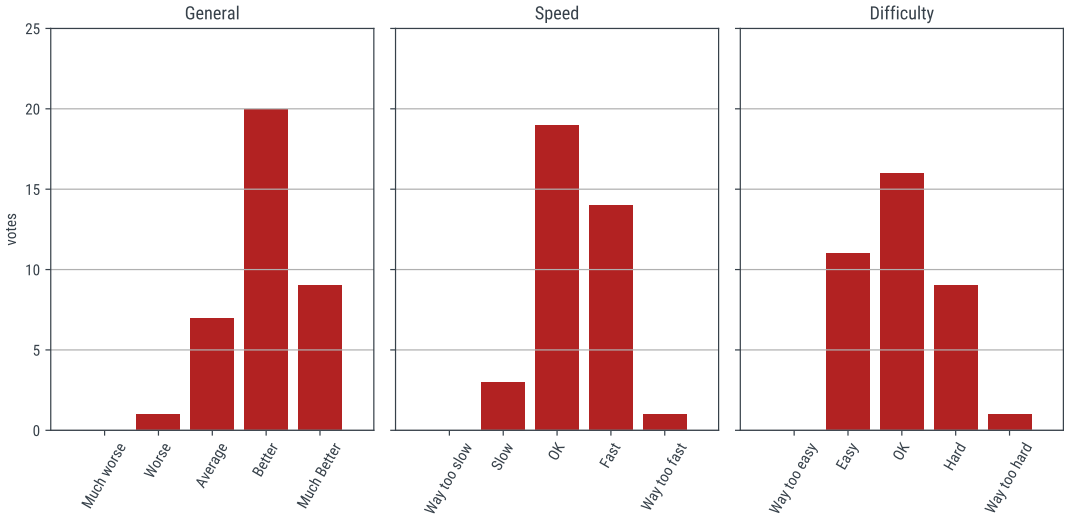
Feedback: Lecture 1

quantitative



Feedback: Lecture 2

quantitative



Detailed Feedback: Lecture 1

your answers

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Things you *didn't* like

- No lecturer in video
- using COVID as the only example, since vaccination is severely politicized
- I was not able to scan the QR codes, because I sat on the left side.
- The virtual questions in the chat were not considered at the end
- too fast

Things you liked

- The math refresher with COVID numbers was fun and illustrative
- The real world example of a possible Corona infection, was a really good introduction!

Things you didn't understand

- It wasn't quite clear how we are going to be graded on the assignments- whether just Coding question is mandatory (like PML), or all 3 parts are compulsory to get a "sufficient" grade

Rejoinder

from last week

Last week:

- In studies in which you want to identify *expected values* $\mathbb{E}_p[f(x)]$ of observables in the *population* $p(x)$ from a *small number* N of samples $x \sim p$, the "ideal" strategy is to sample **independent and identically distributed** from the whole population.

$$\hat{f} = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

- This is "ideal" in the sense that it produces the **unbiased** sampling estimator with known (up to bias) error and "good" (best possible for unbiased estimators) convergence rate.

But not everything is a random number, and, we can't always observe the quantity of interest x!

What kind of data do you have, and what do you want to know?

It's not always about avoiding bias

Some questions people answer(ed) with the help of data:

- What is the mass of the Higgs Boson?
- Which citizens are currently infectious?
- How should the blades of a wind turbine be set to maximize output?
- Which set of weights maximizes accuracy of this deep network?

In these problems, we have a varying degree of control over which data to collect, and what it tells us about the quantity in question.

Whenever an inference machine *interacts* with its data source, it can perform a form of **active** learning. Whenever a *generative* model is available, we can use it to *maximize information content*.

Information Content and Entropy

the math

C.E. Shannon. A Mathematical Theory of Communication. 1949

Definition: The **Shannon information content** of an outcome is

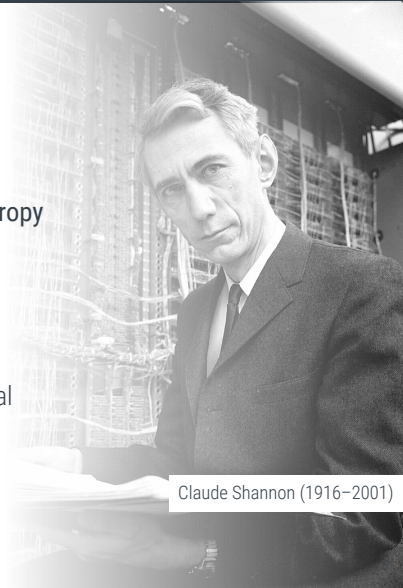
$$h(X = x_i) = \log_2 \frac{1}{p(X = x_i)} \quad \text{bits.}$$

When x_i happens, we have "gained that many bits of information". The **Entropy** $\mathbb{H}(X)$ of X is the *expected* Shannon information content of X under p

$$\mathbb{H}(X) = -\sum_{x_i \in \mathcal{X}} p(X = x_i) \cdot \log_2 p(X = x_i)$$

The **Conditional Entropy** of X given Y (nb. not the Entropy of the conditional $p(x \mid y)$!) is

$$\mathbb{H}(X \mid Y) = -\sum_{x_i \in \mathcal{X}, y_j \in \mathcal{Y}} p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(y_j)}.$$



Claude Shannon (1916–2001)

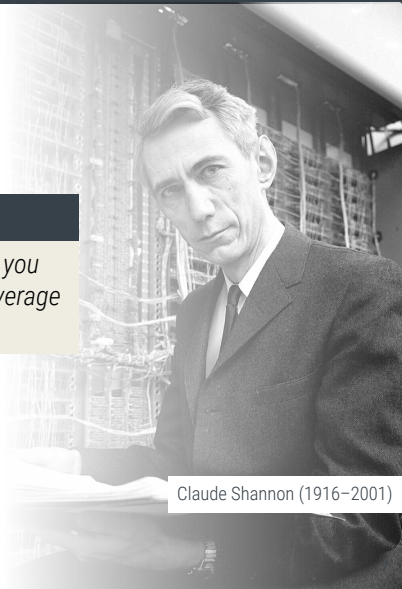
Information Content and Entropy

the math

C.E. Shannon. A Mathematical Theory of Communication. 1949

Theorem (Shannon Source Coding Theorem)

The length of a perfectly compressed file x [the number of binary questions you have to ask to reproduce it] from a source with distribution p is $h(x)$. The average number of binary questions to identify outcomes is thus ***at least*** $\mathbb{H}(x)$.



Claude Shannon (1916–2001)

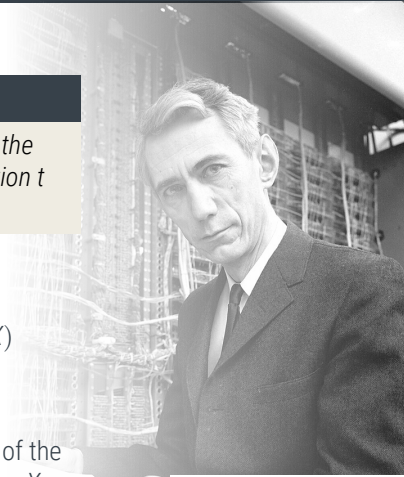
Corollary (Maximum Information content)

To minimize the number of questions necessary to identify outcomes from the source p, ask them about derived quantities $y_t = f(x)$ such that each question t has maximum entropy.

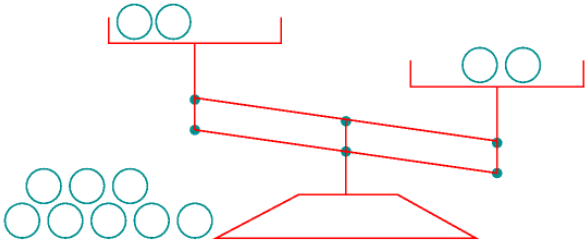
More formally, the posterior has the conditional entropy is

$$\begin{aligned}\mathbb{H}(\Theta \mid X) &= \int (\log_2(p(x \mid \theta)) + \log_2(p(\theta)) - \log_2(p(x)))\ \mathrm{d}p(\Theta, X) \\ &= \mathbb{H}(X \mid \Theta) + \mathbb{H}(\Theta) - \mathbb{H}(X)\end{aligned}$$

For deterministic quantities (i.e. if $\mathbb{H}(X \mid \Theta) = 0$), minimizing the entropy of the posterior (making Θ *easy* to predict) means *maximizing* $\mathbb{H}(X)$ (*choosing* the X that is *hardest* to predict).



Claude Shannon (1916–2001)

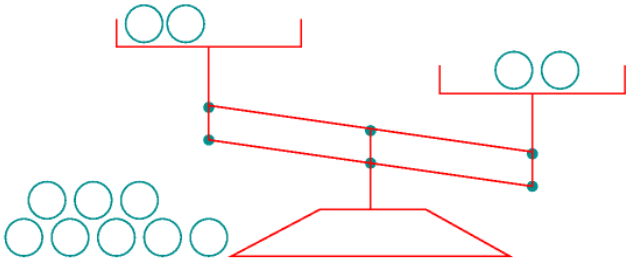


Example: Find the odd ball

You are given 12 balls, all equal in weight except for one that is either *heavier* or *lighter* than the others.

Design a strategy to determine which is the odd ball **and** whether it's heavier or lighter, **in as few uses of the balance as possible.**

Sir David JC MacKay (1967–2016)



- Shannon says: Don't come up with a complicated theory, just choose a strategy that **maximizes the Entropy of the observations**!
- The distribution over $12 \times 2 = 24$ possible outcomes is uniform.
- Thus we need to collect $\log_2 24 = 4.58$ bits.
- Each question with three possible outcomes reveals at best $= \log_2 3 = 1.59$ bits
- We thus have to ask at least $\lceil 4.58/1.59 \rceil = \lceil 2.88 \rceil = 3$ questions.

Designing the Right Questions:

- If your goal is to reveal the variable Θ from data X and you have *control* over the experiments, consider that the conditional entropy of the posterior is

$$\mathbb{H}(\Theta \mid X) = \mathbb{H}(X \mid \Theta) + \mathbb{H}(\Theta) - \mathbb{H}(X)$$

- and *try* to find experiments that minimize this entropy. If the result is deterministic ($\mathbb{H}(X \mid \Theta) = 0$) or *homoscedastic* ($\mathbb{H}(X \mid \Theta) = \text{const.}$), this means *maximizing* $\mathbb{H}(X)$.

Inference

What to do once you have your data,
but it doesn't contain the quantity of interest directly.

The Coming Lectures

Outlook

Given Data D and a *generative model* (likelihood) $p(D \mid \theta)$ under some unknown parameters/variables θ, we might bei interested in

Estimation Give a *good guess* $\hat{\theta}$ for θ

Confidence Estimate the *error* $\hat{\theta} - \theta$

Test There may be a "boring" / "expected" / standard **null hypothesis** θ_0. We want to know how likely it is that the data is in fact generated by this value (how *likely* the null is). If the null is highly unlikely, we can declare that something noteworthy has *probably* happened — we *reject the null hypothesis*. This is called a **hypothesis test**.

The operational object in all of the above is the likelihood $p(D \mid \theta)$. The **Bayesian** approach is to try and operate with the likelihood for as long as possible.

A well-kept secret

It's all about the likelihood! (More precisely: the generative model)

- Statistics is often taught as a laundry list. But a large chunk of it boils down to
  1. invent a likelihood
  2. transform the shape of the likelihood into a binary statement
- note how there is no step 1b: Question the likelihood!
- these correspond to two high-level tasks
  1. informal: phrase your idea/thoughts in math
  2. formal: turn the likelihood *function* into a (continuous or binary) *number*: **Estimation**, **Confidence**, **Testing**.
- it is often a good idea to question whether step 2. is even necessary. Sometimes, just plotting the likelihood is more powerful, because it does not remove information.

A Running Example

Estimating positions

Foto by Andrew Charney (via unsplash)

- You are on a ship in the sea, trying to identify your position y
- You can see **three lighthouses** on the horizon
- For each lighthouse, you know the location x_i
- You can measure the *bearing* (angle vs. true north) for each lighthouse with a measurement error.

Where are you?

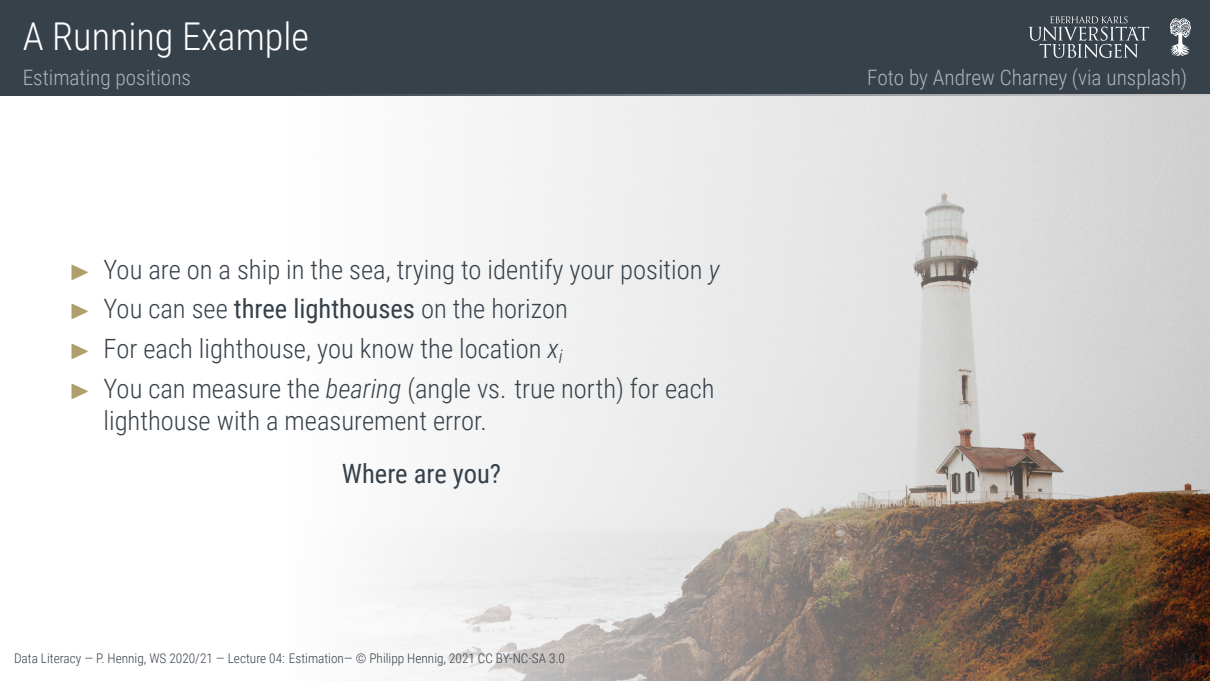


Foto by Andrew Charney (via unsplash)

- For simplicity, we will assume that all measurements are independent (correlated errors are an easy extension). We will also assume the error is *Gaussian* (normal) distributed.
- This amounts to the *likelihood*

$$p(\boldsymbol{\theta} \mid y, x) = \prod_{i=1}^{3} \mathcal{N}\left(\theta_i; \arctan\left(\frac{(y - x_i)_2}{(y - x_i)_1}\right), \sigma_i\right)$$

$$= \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2\sigma_i^2}\left(\theta_i - \arctan\left(\frac{(y - x_i)_2}{(y - x_i)_1}\right)^2\right)\right)$$

(up to the correct sign in the arctan and multiples of π…)

Objects of Interest

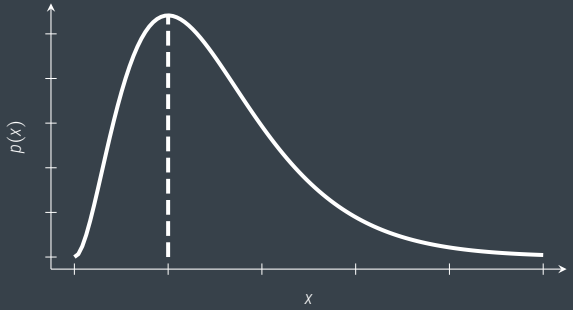
in Data Analysis

Given Data D and a *generative model* M (likelihood) $p(D \mid \theta, M)$ under some unknown parameters/variables θ, we might be interested in

$$p(\theta \mid D, M) = \frac{p(D \mid \theta, M) p(\theta \mid M)}{p(D \mid M)}$$

Estimation	Give a *good guess* $\hat{\theta}$ for θ	**location** of $p(\theta \mid D)$
Confidence	Estimate the *error* $\hat{\theta} - \theta$	**shape** of $p(\theta \mid D)$
Test	Ideally: "guess" whether the *model* $p(D, \theta \mid M)$ is correct. In reality: provide evidence *against* established model.	**evidence / mass** $p(D \mid M)$

The operational object in all of the above is the likelihood $p(D \mid \theta)$. The **Bayesian** approach is to try and operate with the likelihood for as long as possible. The **frequentist** approach is to construct derived random variables as *estimators*, then analyse their properties.



Estimation

Sometimes, you have to make a guess

Consider data $\boldsymbol{x} = [x_1, \ldots, x_n]$ drawn *identically and independently distributed (iid)* from $p(\boldsymbol{x} \mid \theta)$.

- an **estimator** of θ is a function $\hat{\theta}_n = g(\boldsymbol{x})$. Note that this is a random number, because $\boldsymbol{x}$ is random.
- the distribution $P_{\hat{\theta}}$ of $\hat{\theta}$, induced by $p(\boldsymbol{x} \mid \theta)$ is called the **sampling distribution**.
- the **bias** of $\hat{\theta}_n$ is (note: purely formal concept, cf. also last lecture)

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}_{p(\boldsymbol{x}|\theta)}(\hat{\theta}_n) - \theta.$$

- the estimator is **consistent** if it *converges to θ in probability*. That is, for all $\varepsilon > 0$ and all θ,

$$\lim_{n \to \infty} P_{\hat{\theta}} \left(|\hat{\theta}_n - \theta| > \varepsilon \right) = 0.$$

The Maximum Likelihood Estimate

A characterisation of the most popular estimator

[after Wasserman, §9]

A common – though not necessarily the best – estimator is the **maximum likelihood estimator**

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} p(\boldsymbol{x} \mid \theta).$$

With a prior $p(\theta)$ we can define the **maximum a posteriori** estimator $\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\boldsymbol{x} \mid \theta) p(\theta)$.

Consider an **iid. dataset** $\boldsymbol{x} = [x_1, \ldots, x_n]$ with

$$p(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta) \quad \text{and thus the } \textbf{log likelihood } \ell_n(\theta) = \log p(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{n} \log p(x_i \mid \theta)$$

The **maximum likelihood estimator (MLE)** is

$$\hat{\theta}_{\mathrm{ML}} := \arg\max_{\theta} p(\boldsymbol{x} \mid \theta) = \arg\max_{\theta} \ell_n(\theta).$$

Remark: For invertible functions $\tau = g(\theta)$, the MLE of τ is $\hat{\tau} = g(\hat{\theta})$ (the MLE is *equivariant*).

Interlude: KL divergence

The most mis-spelled names in statistics

Definition (Kullback-Leibler divergence)

Let $p(x)$ and $q(x)$ be pdfs. The **KL-divergence from** q *to* p (actually, from the distribution identified by p to that identified by q) is defined as

$$D_{\mathrm{KL}}(p\|q) := \int \log\left(\frac{p(x)}{q(x)}\right) dp(x)$$

Some properties:

- $D_{\mathrm{KL}}(P\|Q) \neq D_{KL}(Q\|P)$
- $D_{\mathrm{KL}}(P\|Q) \geq 0, \forall P, Q$ (**Gibbs' inequality**), and
- $D_{\mathrm{KL}}(P\|Q) = 0 \Leftrightarrow p \equiv q$ almost everywhere

Solomon Kullback
(1907–1994)

Richard Leibler
(1914–2003)

Reminder: the MLE is $\hat{\theta}_{\mathrm{ML}} := \arg\max_\theta p(\boldsymbol{x} \mid \theta) = \arg\max_\theta \ell_n(\theta) = \arg\max \sum_{i=1}^n \log p(x_i \mid \theta)$.

Theorem (Consistency of MLE)

*Assume the data is truly generated from $p(\boldsymbol{x} \mid \theta_*)$. Further assume the model is **identifiable**; that is, $\psi \neq \theta \Rightarrow D_{KL}(p(\boldsymbol{x} \mid \theta) \| p(\boldsymbol{x} \mid \psi)) > 0$. Then the MLE is **consistent**. This means that it converges to the true value θ_* (in probability).*

Proof sketch (formal proof in Wasserman, Thm 9.13):

- Maximizing ℓ_n is equivalent to maximizing $M_n(\theta) = \frac{1}{n} \sum_i \log \left(\frac{p(x_i|\theta)}{p(x_i|\theta_*)} \right)$
- This is an MC estimator. For large n, it converges to its expected value

$$\mathbb{E}(M_n(\theta)) = \int p(x \mid \theta_*) \log \left(\frac{p(x \mid \theta)}{p(x \mid \theta_*)} \right) dx = -D_{\mathrm{KL}}(p(x \mid \theta_*) \| p(x \mid \theta))$$

which is maximized at $\theta = \theta_*$ (proof: Prob. ML, SoSe 2022). □

- Consider data $\boldsymbol{x} = [x_1, \ldots, x_n] \in \mathbb{R}^n$ drawn iid. from $p(x \mid \mu, \sigma^2) = \mathcal{N}(x; \mu, \sigma^2)$
  We know $\mu = \mathbb{E}_p[x]$ and $\sigma^2 = \mathbb{E}_p[x^2] - \mu^2$. What is the maximum likelihood estimate of $\theta = (\mu, \sigma^2)$?

$$\ell_n(\mu, \sigma^2) = \sum_{i=1}^{n} \left(-\frac{(x_i - \mu)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 \right)$$

$$\frac{\partial \ell_n(\mu, \sigma^2)}{\partial \mu} = \left(\sum_{i=1}^{n} x_i \right) - N\mu \qquad \Rightarrow \hat{\mu} = \frac{1}{N}\sum_i x_i =: \bar{x}$$

$$\left.\frac{\partial \ell_n(\mu, \sigma^2)}{\partial \sigma^2}\right|_{\partial \ell_n / \partial \mu = 0} = \frac{1}{2\sigma^4}\left(\sum_i (x_i - \bar{x})^2 \right) - \frac{N}{2\sigma^2} \qquad \Rightarrow \hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i - \bar{x})^2 =: s^2$$

$$\mathbb{E}_{\mathcal{N}(x;\mu,\sigma^2)}(\bar{x}) = \frac{1}{N}\sum_i \mathbb{E}(x_i) = \frac{N}{N}\mu = \mu \qquad \Rightarrow \hat{\mu} \text{ is } \textbf{unbiased}, \text{ but}$$

$$\begin{aligned}
\mathbb{E}_{\mathcal{N}(x;\mu,\sigma^2)}(s^2) &= \frac{1}{N}\sum_i \mathbb{E}(x_i - \bar{x})^2 = \frac{1}{N}\sum_i \mathbb{E}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\
&= \frac{1}{N}\sum_i \left((\sigma^2 + \mu^2) - 2\left(\frac{1}{N}\sum_j \mathbb{E}(x_j x_i)\right) + \frac{1}{N^2}\sum_j\sum_k \mathbb{E}(x_j x_k)\right) \\
&= \frac{1}{N}\sum_i \left((\sigma^2 + \mu^2) - 2\left(\frac{1}{N}((N-1)\mu^2 + (\mu^2 + \sigma^2))\right) + \frac{1}{N^2}N((N-1)\mu^2 + (\mu^2 + \sigma^2))\right) \\
&= \sigma^2\left(1 - \frac{1}{N}\right) = \sigma^2\frac{N-1}{N}
\end{aligned}$$

The (simultaneous) maximum likelihood estimate for mean and variance is *biased*. But the estimate $(\hat{\mu}, \hat{\sigma}^2) = (\bar{x}, N/(N-1)s^2)$ is unbiased.

Summary

- To actively identify a deterministic (or homoscedastic) quantity, maximize the entropy of your data.
- If you need to estimate an unknown quantity, try Bayesian inference!
- A lot of classic statistical estimation, maximum likelihood in particular, has inadequate computing as a hidden origin. With modern computers, we can often afford to evaluate the whole likelihood, or at least a Taylor-approximation of it
- But when things are really expensive, or you really just need a point estimate, you could do worse than maximum likelihood
- There are also some nasty pitfalls, though, especially for asymmetric, non-smooth or even diverging likelihoods (see exercises)
- In reality, you never have infinite data ($n \nrightarrow \infty$)!

Please provide feedback:

Always write down the probability of everything.

A date for your calendar?

Methods of Machine Learning Research Seminar

Methods of Machine Learning Research Seminar
Wednesday, 10 November, 15:00 (st)
Dr. Toni Karvonen

This talk is concerned with sample path properties of Gaussian processes. We use reproducing kernel Hilbert space techniques to identify a "small" set of functions in which the samples from a given Gaussian process reside with probability one. Then we discuss the implications that this sample characterisation has on model selection in applications such as probabilistic numerics where Gaussian processes are used to model deterministic functions which are observed without noise.

details at https://talks.tue.ai/talks/talk/id=18