



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 04

ESTIMATING CONFIDENCE

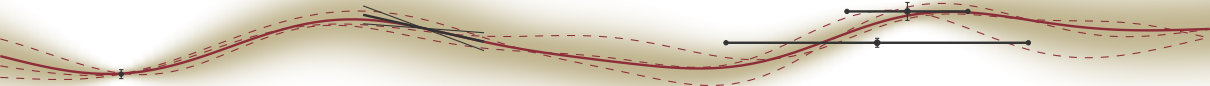
Philipp Hennig

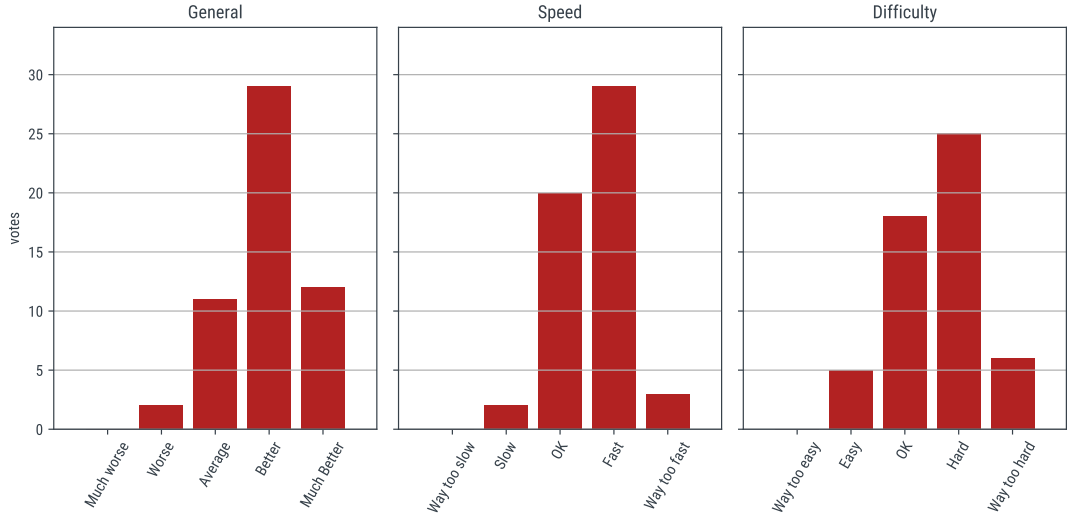
15 November 2021

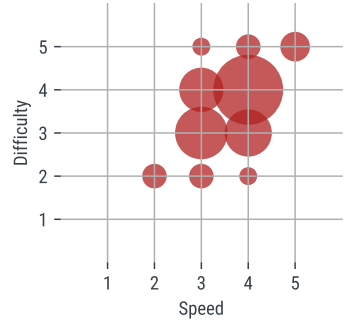
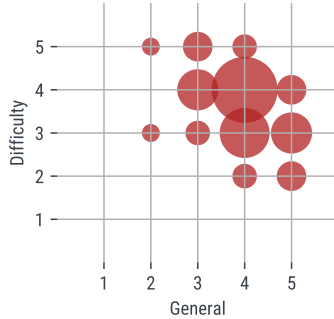
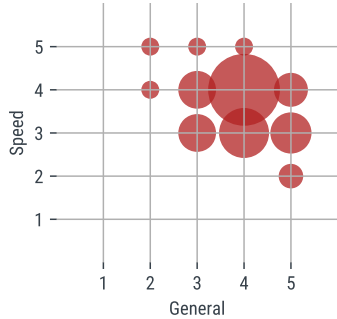
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



FACULTY OF SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
CHAIR FOR THE METHODS OF MACHINE LEARNING









Things you *didn't* like

- ▶ (conditional) entropy was too fast
- ▶ questions from the lecture hall were not repeated for the call
- ▶ zoom audio problems
- ▶ math too fast without blackboard, use iterative reveal instead
- ▶ too fast / too much

Things you liked

- ▶ The lighthouse example (8x)
- ▶ the odd ball example (10x)

Things you didn't understand

- ▶ KL divergence
- ▶ entropy
- ▶ the python code
- ▶ the maximum likelihood example
- ▶ can we have template solutions for the coding exercises?

Summary of last lecture:


- ▶ to solve *inference problems* (make statements about variables that are not directly observed, but are only related to the data through a stochastic or non-invertible function), use **Bayes' theorem**.

$$p(x | D) = \frac{p(D | x)p(x)}{p(D)} = \frac{p(D | x)p(x)}{\int p(D | x)p(x) dx}$$

- ▶ when *point estimates* are required, they are invariably related to the posterior $p(x | D)$ (or, in the special case where the prior doesn't matter, the likelihood $p(D | x)$).
- ▶ A particularly common estimator is the **maximum likelihood** estimate

$$\hat{x} = \arg \max_x p(D | x).$$

- ▶ But unless the posterior is a point mass, any estimator is almost surely wrong. How wrong, though? Quantifying this error is the goal of *confidence* estimates ("error bars"), which we will address today. We will encounter two ways to do this:
 - ▶ A rigorous mathematical one that requires some integration skills
 - ▶ A simple empirical one that is easy to implement



An running example:
`BioNTech_Confidence.ipynb`



What is the probability $\pi \in [0, 1]$ of a positive result in a Bernoulli experiment?

What is the probability $\pi \in [0, 1]$ of a positive result in a Bernoulli experiment?

- Likelihood of n positive and $N - n = m$ negative observations:

$$p(n, m \mid \pi) = \binom{n+m}{n} \pi^n \cdot (1 - \pi)^m$$

- Inference? Bayes' theorem!

$$p(\pi \mid n, m) = \frac{p(n, m \mid \pi) p(\pi)}{p(n, m)} = \frac{p(n, m \mid \pi) p(\pi)}{\int p(n, m \mid \pi) p(\pi) d\pi}$$

What is the probability $\pi \in [0, 1]$ of a positive result in a Bernoulli experiment?

- Likelihood of n positive and $N - n = m$ negative observations:

$$p(n, m \mid \pi) = \binom{n+m}{n} \pi^n \cdot (1 - \pi)^m$$

- Inference? Bayes' theorem!

$$p(\pi \mid n, m) = \frac{p(n, m \mid \pi) p(\pi)}{p(n, m)} = \frac{p(n, m \mid \pi) p(\pi)}{\int p(n, m \mid \pi) p(\pi) d\pi}$$

- uniform prior $p(\pi) = 1 = \pi^0 \cdot (1 - \pi)^0$.

What is the probability $\pi \in [0, 1]$ of a positive result in a Bernoulli experiment?

- Likelihood of n positive and $N - n = m$ negative observations:

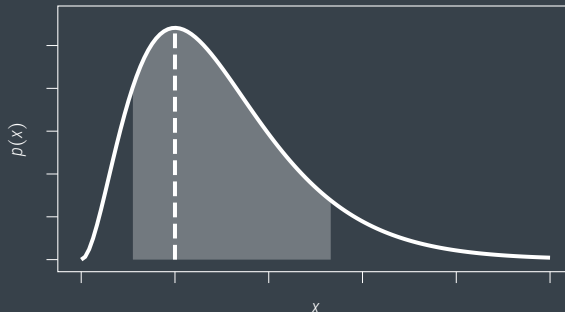
$$p(n, m \mid \pi) = \binom{n+m}{n} \pi^n \cdot (1 - \pi)^m$$

- Inference? Bayes' theorem!

$$p(\pi \mid n, m) = \frac{p(n, m \mid \pi) p(\pi)}{p(n, m)} = \frac{p(n, m \mid \pi) p(\pi)}{\int p(n, m \mid \pi) p(\pi) d\pi}$$

- uniform prior $p(\pi) = 1 = \pi^0 \cdot (1 - \pi)^0$.
- posterior after n positive, m negative observations:

$$p(\pi \mid n, m) = \frac{\pi^n (1 - \pi)^m \cdot 1}{\int \pi^n (1 - \pi)^m \cdot 1 d\pi} = \frac{\pi^n (1 - \pi)^m}{B(n+1, m+1)}$$



Confidence – The Frequentist Answer

The following slides show that the MLE is “decent”. Vaguely speaking, “for sufficiently large n , the MLE approximates an Gaussian random variable with mean θ and a variance identified by (only) the local shape of ℓ_n at $\hat{\theta}_{\text{ML}}$.” That local shape is described by a quantity known as the *Fisher information*. In large-sample settings, this provides a license to work with Gaussian approximations, for example to construct confidence sets and (so-called *Wald*) tests.

Consistency of the MLE (Recap from last lecture)

properties of the MLE



[after Wasserman, §9]

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

Theorem (Consistency of MLE)

Assume the data is truly generated from $p(\mathbf{x} \mid \theta_)$. Further assume the model is **identifiable**; that is, $\psi \neq \theta \Rightarrow D_{\text{KL}}(p(\mathbf{x} \mid \theta) \parallel p(\mathbf{x} \mid \psi)) > 0$. 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 $M_n(\theta) = \frac{1}{n} \sum_i \log \left(\frac{p(x_i \mid \theta)}{p(x_i \mid \theta_*)} \right)$
- ▶ This is an MC estimator. For large n , it converges to its expected value

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

which is maximized at $\theta = \theta_*$ (proof: Prob. ML, SoSe 2021). \square



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

Definition (score function & Fisher information)

The **score function** is defined to be the derivative of the log likelihood,

$$s(X, \theta) := \frac{\partial \log p(X; \theta)}{\partial \theta}.$$

The **Fisher information** is defined to be the variance of the score function (we write $I(\theta) := I_1(\theta)$),

$$I_n(\theta) := \text{var}_{p(\mathbf{x}|\theta)} \left(\sum_{i=1}^n s(X_i; \theta) \right) = \sum_{i=1}^n \text{var}_{p(x|\theta)} (s(X_i; \theta))$$



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

$$s(X, \theta) := \frac{\partial \log p(X; \theta)}{\partial \theta}, \text{ and } I_n(\theta) := \text{var}_{p(\mathbf{x} \mid \theta)} \left(\sum_{i=1}^n s(X_i; \theta) \right) = \sum_{i=1}^n \text{var}_{p(\mathbf{x} \mid \theta)}(s(X_i; \theta))$$

The Fisher Information is the expected curvature of the log likelihood under the true model.

Theorem

- ▶ $\mathbb{E}_{\theta}(s(X; \theta)) = 0$ (proof: homework).
- ▶ Thus, $\text{var}(s(X, \theta)) = \mathbb{E}_{\theta}(s^2(X; \theta))$ and $I_n(\theta) = nI(\theta)$
- ▶ Also, (proof: homework)

$$I(\theta) = -\mathbb{E}_{p(\mathbf{x} \mid \theta)} \left(\frac{\partial^2 \log p(X \mid \theta)}{\partial \theta^2} \right) = - \int \left(\frac{\partial^2 \log p(X \mid \theta)}{\partial \theta^2} \right) p(\mathbf{x} \mid \theta) d\mathbf{x}$$

Careful with details, though. See Küstner et al., NeurIPS 2019



Theorem (Asymptotic Normality of the MLE, proof in Wasserman, Thm. 9.18)

Let $e := \sqrt{\text{var}_{p(x|\theta)}(\hat{\theta}_n)}$. Under appropriate regularity conditions, it holds that

1. $e \approx \sqrt{1/l_n(\theta)}$ and

$$\frac{(\hat{\theta} - \theta)}{e} \rightsquigarrow \mathcal{N}(0; 1).$$

2. Let $\hat{e} := \sqrt{1/l_n(\hat{\theta})}$. Then

$$\frac{(\hat{\theta} - \theta)}{\hat{e}} \rightsquigarrow \mathcal{N}(0; 1).$$

(where \rightsquigarrow denotes convergence in distribution, meaning that the cumulative distribution function (CDF) $F_n(t)$ of the LHS approaches that of the RHS, $F(t)$, for $n \rightarrow \infty$ and for all t for which F is continuous.)

So what?

Informally speaking, the results above say that, *if the model (the likelihood) is correct, identifiable, and n is large enough (i.e. there is enough data)*, the maximum likelihood estimator $\hat{\theta}$ can be approximately thought of as a Gaussian “measurement” of the *true* parameter θ , with “noise” identified by the (inverse) Fisher information:

$$\hat{\theta} \sim \mathcal{N}(\hat{\theta}; \theta, I_n^{-1}(\theta))$$

There same statement holds, with minor adaptations, for $\theta \in \mathbb{R}^d$ and the **Fisher information matrix** $I_n(\theta) \in \mathbb{R}^{d \times d}$ and

$$[I_n(\theta)]_{ij} = -\mathbb{E}_{f(x|\theta)} \left[\frac{\partial^2 \log f(x | \theta)}{\partial \theta_i \partial \theta_j} \right]$$



564

THÉORIE ANALYTIQUE

L'intégrale du numérateur étant prise depuis $x=\theta$ jusqu'à $x=\theta'$, et celle du dénominateur étant prise depuis $x=0$ jusqu'à $x=1$.

La valeur de x la plus probable, est celle qui rend y un *maximum*. Nous la désignerons par a . Si aux limites de x , y est nul, alors chaque valeur de y a une valeur égale correspondante de l'autre côté du *maximum*.

Quand les valeurs de x , considérées indépendamment du résultat observé, ne sont pas également possibles : en nommant x la fonction

DES PROBABILITÉS.

565

ϕ est égal à $\frac{x-a}{\sqrt{\log Y - \log y}}$, et $U, \frac{dU}{dx}, \frac{d^2U}{dx^2}$, etc. sont ce que deviennent $\phi, \frac{d\phi}{dx}, \frac{d^2\phi}{dx^2}$, etc., lorsqu'on y change après les différentiations, x en a , a étant la valeur de x qui rend y un *maximum*. T est égal à ce que devient la fonction $\sqrt{\log Y - \log y}$, lorsqu'on change x en $a - \theta$ dans y , et T' est ce que devient la même fonction, lorsqu'on y change x dans $a + \theta$. L'expression précédente de ϕdx donne la valeur de cette intégrale, dans les limites

It can be hard to compute the Fisher in practice. A simpler approximation is then to just compute the Hessian at the mode itself, and build the **Laplace approximation**

$$q(\theta) \approx \mathcal{N} \left(\theta; \hat{\theta}, - \left(\mathbb{E}_{p(x|\hat{\theta})} \frac{\partial^2 \log p(x | \theta)}{\partial \theta_i \partial \theta_j} \right)^{-1} \right) \approx \mathcal{N} \left(\theta; \hat{\theta}, - \left(\frac{\partial^2 \log p(x | \theta)}{\partial \theta_i \partial \theta_j} \Big|_{\theta=\hat{\theta}} \right)^{-1} \right)$$

les formules nécessaires pour déterminer par des approximations convergentes, les intégrales du numérateur et du dénominateur de la formule (1), lorsque les événements simples dont se compose l'événement observé, sont répétés un très-grand nombre de fois ; car alors y a pour facteurs, des fonctions de x élevées à de grandes puissances. Nous allons, au moyen de ces formules, déterminer la loi de probabilité des valeurs de x , à mesure qu'elles s'éloignent de la valeur a , la plus probable ; ou qui rend y un *maximum*. Pour cela, reprenons la formule (c) du n° 27 du premier Livre,

$$y dx = Y \cdot \left\{ U + \frac{1}{2} \cdot \frac{d^2 U^2}{dx^2} + \frac{1.5}{2} \cdot \frac{d^3 U^3}{dx^3} + \text{etc.} \right\} \cdot \frac{f dx \cdot e^{-x^2}}{\sqrt{\pi}} + \frac{1}{2} \cdot e^{-T^2} \cdot \left\{ \frac{d^2 U^2}{dx^2} - T \cdot \frac{d^3 U^3}{dx^3} + (T^2 + 1) \cdot \frac{d^4 U^4}{dx^4} - \text{etc.} \right\} \cdot \frac{f dx \cdot e^{-x^2}}{\sqrt{\pi}} \quad (2)$$

$$\frac{f dx \cdot e^{-x^2}}{\sqrt{\pi}} + \frac{1}{2} \cdot e^{-T^2} \cdot \left\{ \frac{d^2 U^2}{dx^2} - T \cdot \frac{d^3 U^3}{dx^3} + (T^2 + 1) \cdot \frac{d^4 U^4}{dx^4} + \text{etc.} \right\} \cdot \frac{f dx \cdot e^{-x^2}}{\sqrt{\pi}} \quad (3)$$

On voit par le n° 25 du premier Livre, que dans le cas où y a pour facteurs, des fonctions de x élevées à de grandes puissances de l'ordre $\frac{1}{\alpha}$, α étant une fraction extrêmement petite, alors U est le plus souvent de l'ordre $\sqrt{\alpha}$, ainsi que ses différences successives ; $U, \frac{dU}{dx}, \frac{d^2U}{dx^2}$, etc. sont respectivement des ordres $\sqrt{\alpha}, \alpha, \alpha^{\frac{3}{2}}$, etc. ; d'où il suit que la convergence des séries de la for-

- ▶ Consider a dataset $X := [x_1, \dots, x_n]$ assumed to be drawn iid. from some distribution $p(x \mid \theta)$. We have constructed an estimator $\hat{\theta} = T(X)$ for θ from it. (Btw., $T : X \mapsto \hat{\theta}$ is known as a *test statistic*).
- ▶ A simple alternative to confidence estimation is to
 1. **simulate additional experiments** $X^b = [x_1^b, \dots, x_n^b]$ for $b = 1, \dots, B$ by *drawing x_j^b with replacement from the original dataset*.
 2. Use the samples to compute statistics of the estimator $\hat{\theta}$. For example, estimate the variance of $\hat{\theta}$ as

$$v_{\text{boot}} = \frac{1}{B-1} \sum_{b=1}^B \left(T(X^b) - \frac{1}{B} \sum_{r=1}^B T(X^r) \right)^2$$

This is known as *the bootstrap*. It works because one can think of

$$p(x \mid \theta) \approx q(x) = \frac{1}{n} \sum_i^n \delta(x - x_i)$$

so sampling from X with replacement is an approximation of sampling from p . But this obviously only works if n is large.

But if you can, always try computing

$$p(\theta | X) = \frac{p(D | \theta)p(\theta)}{\int p(D | \theta)p(\theta) d\theta}$$

Summary

- ▶ If the model is correct, the posterior fully captures the uncertainty over latent quantities
- ▶ If tracking the full posterior is too cumbersome, the *Fisher information matrix*, or even simply the Hessian of the log likelihood (i.e. a *Laplace* approximation) *asymptotically* captures the uncertainty of the maximum likelihood / maximum-a-posteriori estimate
- ▶ The *bootstrap* estimate is even easier to implement: just re-sample from the training data with replacement. But it creates an outer loop around the entire training process, which is potentially costly.
- ▶ All these methods are approximations that only really work in the big-data regime. For low-sample cases, do Bayesian inference.

Please provide feedback:





Methods of Machine Learning Research Seminar
Wednesday, 17 November, 15:00 (st)
Dr. Martin Trapp

Synergies between Bayesian Nonparametrics and Deep Architectures

I will discuss recent work on the intersection of deep architectures, such as neural networks and deep tractable models, and Bayesian nonparametrics and highlight the synergies that arise in the intersection from two different perspectives...

details at <https://talks.tue.ai/talks/talk/id=25>

