Bayesian Inference

Toy example: Chance of finding life in the universe if bacteria are found on mars.

Prior belief:
- Denote the chance of a life form on a planet as $0 \le \eta \le 1$
- Our prior is $P(\eta) = \frac{1}{\eta_0} e^{-\eta/\eta_0}$ with $\eta_0 = 10^{-10}$ (truncated for simplicity)
- chance that a random planet has life: $\int d\eta \, \overbrace{P(\text{life}|\eta)}^{\eta} P(\eta)$

Posterior: Say NASA measures biological methane, how does that change our prior?

$$P(\eta \mid \text{Evidence for life}) \underset{\text{Bayes Rule}}{=} \frac{P(\text{Evidence for life} \mid \eta)\, P(\eta)}{P(\text{Evidence for life})}$$

$$\underset{\text{(As a func. of } \eta)}{\propto} P(\text{Evidence for life} \mid \eta)\, P(\eta)$$

$$= \frac{\eta}{\eta_0} e^{-\eta/\eta_0} / Z \rightarrow \text{normalization}$$

Inference:

$$P(\text{life} \mid \text{Evidence}) = \int d\eta \, P(\text{life}|\eta) P(\eta|\text{Evidence})$$

$$= \int d\eta \, \frac{\eta^2}{\eta_0} e^{-\eta/\eta_0} / Z = (\ldots)$$

$$Z = \int_0^\infty d\eta \frac{\eta}{\eta_0} e^{-\eta/\eta_0} = \left(\frac{1}{\eta_0}\right) \partial_{(\eta_0^{-1})} \int d\eta \, e^{-\eta/\eta_0} = \frac{-1}{\eta_0} \partial_{\eta_0^{-1}} \cdot \frac{1}{\eta_0^{-1}}$$

$$= +\frac{1}{\eta_0} \frac{1}{\eta_0^{-2}} = \eta_0$$

Some Algebra

$$(\ldots) = \frac{1}{\eta_0^2} \partial^2_{\eta_0^{-1}} \int d\eta \, e^{-\eta/\eta_0} = \frac{1}{\eta_0^2} \frac{2}{\eta_0^{-3}} = 2\eta_0$$

Chance of life: following N evidence

$$P(\eta \mid N\text{-observations}) = \frac{\eta^{N}}{\eta_0} e^{-\frac{\eta}{\eta_0}} \Big/ \eta_0^{N} \cdot N!$$

$$p(\text{life} \mid N\text{-observations}) = (N+1)\,\eta_0$$

A paradox: • We have "found" N life forms, but still our estimate of how likely it is to find life is just $(N+1)\eta_0 = (N+1)10^{-10} \sim 10^{-10}$ → Prior

• This is the result of having a very strong prior against life.

• Exercise: Repeat the above with the prior $p(\eta) = (1-2\eta_0)\delta(\eta) + 2\eta_0$ $\quad \eta_0 = 10^{-10}$

Bayesian ML:

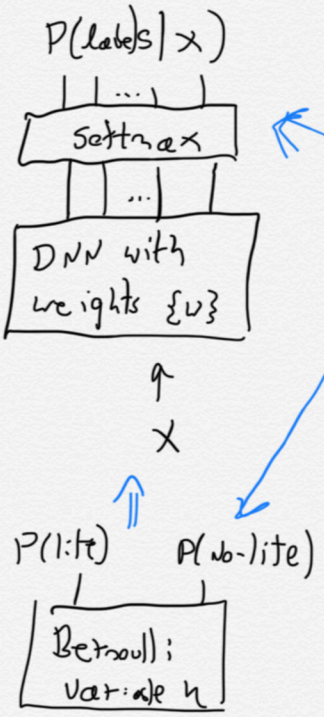
• life example ⟹ Machine Learning using Deep Neural Networks

η ⟹ $\{w\}$ - weights of a DNN

• Evidence ⟹ $(x, y) \equiv$ (datapoint, label)

• model $P(\text{life} \mid \eta)$ ⟹

• $P((x,y) \mid \{w\})$ = chance of predicting y given input x with network patrons $\{w\}$



Prior

• $P(\eta)$ ⟹ Prior distribution on weights $P(\{w\}) = e^{-\frac{ww^T}{2\sigma^2}} / Z$ typically (Xavier) init.

• Inference

$$\int d\eta\, p(\text{life} \mid \eta)\, P(\eta \mid \text{evidence}) \Longrightarrow \int \underbrace{dw_1 \ldots dw_N}_{\text{multidimensional integral}} \underbrace{P((x,y) \mid \{w\})}_{\text{Highly non-linear func. of } w_1, \ldots w_N} P(\{w\} \mid \text{Dataset})$$

ML/Bayesian Inference using GPs:

- $P(\eta) \to P_0[f]$
- Evidence $\longrightarrow$ $f(x_n) = y_n + \xi$ (datapoint, label, Gaussian Noise $\langle \xi\xi \rangle = \sigma^2$)
- Posterior $\to$ $P[f \mid \text{Dataset}] \propto P_0[f] \cdot P[\text{Dataset} \mid f]$

$$\propto P_0[f] \cdot e^{-\frac{(f(x_n)-y_n)^2}{2\sigma^2}}$$

- Note some qualitative differences from previous setup:

I. Usually model (finite num of params) yields predictions

e.g. DNN(weights) yields a prediction $DNN(x) = y$

and a prior is given on params which yields a prior on predictions.

In GPs the model is defined indirectly through its predictions. Prior is given immediatly on predictions.

II. Number of parameters is infinite as even nice smooth functions are defined by a (discrete) infinite set of parameters

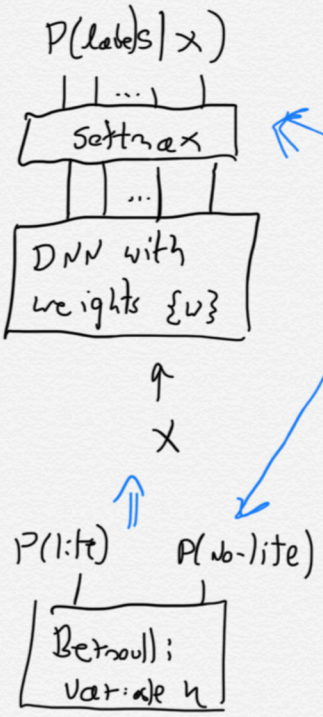
Analytic expression for Bayesian Inference on a GP:

- $$P[f \mid \text{Dataset}] \propto e^{-\frac{1}{2}\int dx\,dx'\, f(x)\, S(x,x')\, f(x') + \sum_{n=1}^{N} \frac{(f(x_n)-y_n)^2}{2\sigma^2}} = e^{-(S[f]+D[f])}$$

- $y_* \equiv$ Prediction on an unseen data point x_*

$$= \int D[f]\, P[f \mid \text{Dataset}]\, f(x_*) \underset{\text{Gaussian}}{=} f_0(x_*) \quad \text{where} \quad \frac{\delta}{\delta f(x)}(S+D)\Big|_{f_0} = 0$$

Bayesian ML:



- life example $\Longrightarrow$ Machine Learning using Deep Neural Networks

  $\eta \Longrightarrow \{w\}$ - weights of a DNN

- Evidence $\Longrightarrow$ $(x, b) \equiv$ (datapoint, label)
- model $P(\text{life}|\eta)$ $\Longrightarrow$
- $P((x,y)|\{w\})$ = chance of predicting y given input x with network params $\{w\}$

- Prior $P(\eta)$ $\Longrightarrow$ Prior distribution on weights $P(\{w\}) \underset{\substack{\text{typically} \\ \text{(Xavier)} \\ \text{init.}}}{=} e^{-\frac{ww^T}{2\sigma^2}}/Z$

- Inference

$$\int d\eta\, P(\text{life}|\eta)P(\eta|\text{evidence}) \Longrightarrow \int \underbrace{dw_1 \ldots dw_J}_{\text{multidimensional integral}} \underbrace{P((x,b)|\{w\})}_{\text{Highly non-linear func. of } w_1 \ldots w_J} P(\{w\}|\text{Dataset})$$

! typically impossible numerically

Gaussian Processes:

- A GP is a probability distribution on functions

$$P_0[f(x)] \propto \exp\left(-\frac{1}{2}\int dx\,dx'\, f(x)f(x')\,\mathcal{K}(x,x')\right)$$

I add "0" to denote priors

Square Brackets denote functionals

ML/Bayesian Inference using GPs:

- $P(\eta) \longrightarrow P_0[f]$
- Evidence $\longrightarrow$ $f(x_n) = y_n + \xi$

datapoint, label, Gaussian Noise $\langle \xi\xi \rangle = \sigma^2$

$$= \int D[f]\, P[f|\text{Dataset}]\, f(x_*) \underset{\text{Gaussian}}{=} f_0(x_*) \quad \text{where} \quad \frac{\delta}{\delta f(x)}(S+D)\Big|_{f_0} = 0$$

extremum

$$\cdot \frac{\delta}{\delta f(x')}(S+D) = \int dx\, S(x',x) f(x) + \sum_{n=1}^{N} \delta(x'-x_n)\big(f(x_n)-y_n\big)/\sigma^2 = 0$$

$$\Rightarrow \quad \left(S^{-1} + \frac{\mathbb{1}}{\sigma^2}\,\text{Data Set}\right)\circ f = \frac{\mathbb{1}}{\sigma^2}\cdot y \qquad [S^{-1}]_{xx'} = S(x,x')$$

$$\left[\frac{\mathbb{1}}{\text{Data Set}}\right]_{xx'} = \sum_{n=1}^{N} \delta(x-x_n)\delta(x'-x_n)$$

$$f = \left(S^{-1} + \mathbb{1}/\sigma^2\right)^{-1} \circ \mathbb{1}/\sigma^2 \cdot y = \left[\left(I + \mathbb{1}/\sigma^2 \circ S^{-1}\right)\circ S\right]^{-1} \circ \mathbb{1}/\sigma^2 \cdot y$$

$$= S^{-1}\cdot\left(I + \frac{\mathbb{1}}{\sigma^2}S^{-1}\right)^{-1}\cdot \mathbb{1}/\sigma^2 \cdot y = S^{-1}\cdot \sum_{m=0}^{\infty}\left(\frac{\mathbb{1}}{\sigma^2}S^{-1}\right)^{m}\frac{\mathbb{1}}{\sigma^2}\cdot y$$

$$= S^{-1}\circ\left(\frac{\mathbb{1}}{\sigma^2} + \frac{\mathbb{1}}{\sigma^2}S^{-1}\frac{\mathbb{1}}{\sigma^2} + \frac{\mathbb{1}}{\sigma^2}S^{-1}\frac{\mathbb{1}}{\sigma^2}S^{-1}\frac{\mathbb{1}}{\sigma^2} + \dots\right) y$$

Note that

$$\left[\frac{\mathbb{1}}{\sigma^2}S^{-1}\frac{\mathbb{1}}{\sigma^2}\right]_{x_1 x_4} = \int dx_2\, dx_3 \sum_{nm} \delta(x_1-x_n)\delta(x_2-x_n)\, S^{-1}(x_2,x_3)\, \delta(x_3-x_m)\delta(x_4-x_m)$$

$$= \sum_{nm} \delta(x_1-x_n)\, S^{-1}(x_n,x_m)\, \delta(x_4-x_m)$$

$$\left[\frac{\mathbb{1}}{\sigma^2}S^{-1}\frac{\mathbb{1}}{\sigma^2}S^{-1}\frac{\mathbb{1}}{\sigma^2}\right] = \sum_{n,m,q} \delta(x_1-x_n)\, S^{-1}(x_n,x_q)\, S^{-1}(x_q,x_m)\, \delta(x_m-x_2)$$

Exercise show that:

1. $$f(x_*) = S^{-1}(x_*,x_n)\left[\frac{1}{K^{-1}+\sigma^2}\right]_{nm} y_m \quad \text{where} \quad K_{nm} = S^{-1}(x_n,x_m)$$

$$S^{-1}(x,x') = \int Df\, P_0[f]\, f(x) f(x') \qquad \text{and} \int dx'\, S^{-1}(x,x')S(x',x'') = \delta(x-x'')$$

- Closed expression for prediction!

upsides ☺

- Require knowledge of $S^{-1}(x,x') = \int Df\, P_0[f]\, f(x) f(x')$!

downsides ☹

- Requires inverting and storing a matrix of size $N \times N$ [$10^7 \times 10^7$ for Imagenet]
- Analytical but not insightful.

Example: RBF kernel on a long ring.

- $f: S^1 \to \mathbb{R}$
- $\langle f(x) f(x') \rangle \equiv \int Df\, f(x) f(x')\, e^{-\frac{1}{2}\int dx\, dx'\, f(x) f(x') S(x,x')}$

$$= S^{-1}(x,x') = K(x,x') = e^{-\frac{|x-x'|^2}{2\ell^2}} \qquad \ell \ll R$$

$L = 2\pi R$

- What is $S(x,x')$?

$$S^{-1}(x,x') = \sum_n \phi_n(x) \phi_n^*(x') \lambda_n = \sum_n \frac{1}{L} e^{ik_n x} e^{-ik_n x'} e^{-\frac{(\ell k_n)^2}{2}}$$

$$S(x,x') = \sum_n \frac{1}{L} e^{ik_n x} e^{-ik_n x'} e^{+(\ell k_n)^2/2} \qquad k_n = \frac{2\pi n}{L}$$

- Dataset: is $\{(x_n, y_n)\}_{n=1}^{N-1}$ $\quad x_n = \frac{n}{N} \cdot L \quad y_n = 2\sin\left(2\pi \frac{n}{N}\right)$

- What does this GP predict for a point $x_* = x_n$ (for all x_n)?
- We need to invert $K_{nm} = \sum_{q=-\infty}^{\infty} \frac{1}{L} e^{ik_q(x_n - x_m)} e^{-\frac{(\ell k_q)^2}{2}}$

to do so let's find its eigs and eigenvals:

- eigenvec $|p\rangle$: $\langle n|p\rangle = e^{i\frac{2\pi}{N} p \cdot n} \frac{1}{\sqrt{N}}$ → Lattice wave

- **Dataset:** is $\{(x_n, y_n)\}_{n=1}^{N-1}$ $\quad x_n = \frac{n}{N} \cdot L \quad y_n = 2\sin\left(2\pi \frac{n}{N}\right)$

- What does this GP predict for a point $x_* = x_n$ (for all x_n) ?
- We need to invert $K_{nm} = \sum_{q=-\infty}^{\infty} \frac{1}{L} e^{ik_q(x_n - x_m)} e^{-\frac{(\ell k_q)^2}{2}}$

  to do so let's find its eigs and eigenvals:
- eigenvec $|p\rangle$: $\langle n|p\rangle = e^{i\frac{2\pi}{N} p \cdot n} \frac{1}{\sqrt{N}}$ → Lattice wave
- Eigs λ_p : $\langle p|K|p\rangle = \sum_{nm} \frac{1}{\sqrt{N}} e^{\frac{i2\pi p \cdot n}{N}} \sum_q \frac{e^{ik_q(x_n - x_m)}}{L} e^{-\frac{(\ell k_q)^2}{2}} e^{-\frac{i2\pi p m}{N}}$

$$= \frac{N}{L} \sum_{\tilde{q}} \exp\left(-\left(\ell k_{(p+N\tilde{q})}\right)^2/2\right)$$

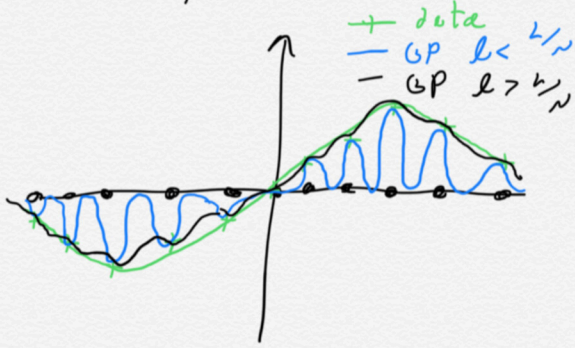
$$\boxed{\sum_{q=-\infty}^{\infty} e^{i\frac{2\pi q}{L} \cdot x_n - \frac{i2\pi p n}{N}} = \sum_{q=-\infty}^{\infty} e^{\frac{i2\pi q \cdot n}{N} - \frac{2\pi i p n}{N}} = \sum_{q=-\infty}^{\infty} e^{\frac{2\pi i (q-p) \cdot n}{N}} = N \cdot \delta_{q,p \bmod N}}$$

- Inference: Note that $y_n = (|p=m\rangle - |p=-m\rangle)$

  According to (1)

$$f(x_t) = \sum_{nm} k(x_t, x_n) K^{-1}_{nm} y_m = \frac{1}{\lambda_m} \sum_n k(x_t, x_n) y_n$$

point on the dataset

$$= \frac{\lambda_m}{\lambda_m} y_t$$

$$f(x_*) = \sum_n \frac{k(x_*, x_n) y_n}{\lambda_m}$$