Introduction

Nipun Batra

August 3, 2023

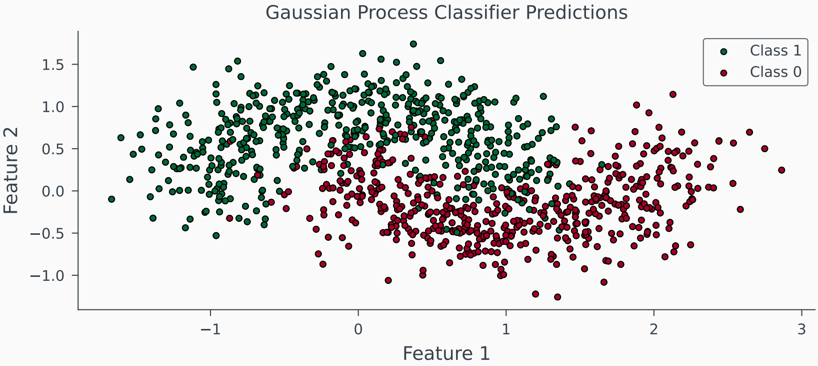
IIT Gandhinagar

What

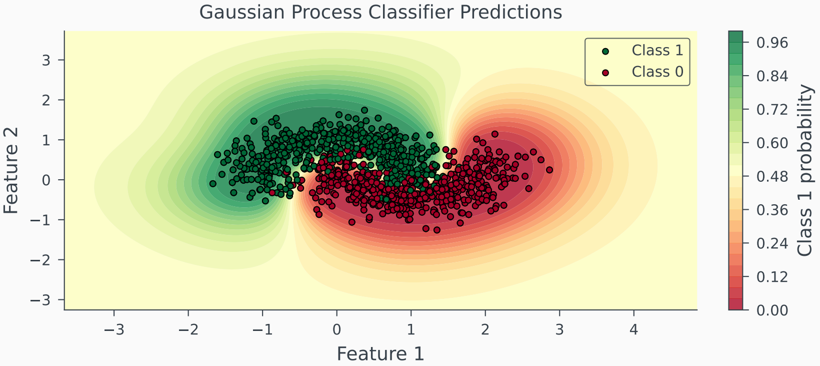
- Predict with uncertainty
- Optimize any black box function
- Efficiently create a training set
- Generative modelling

Dog or Cat?

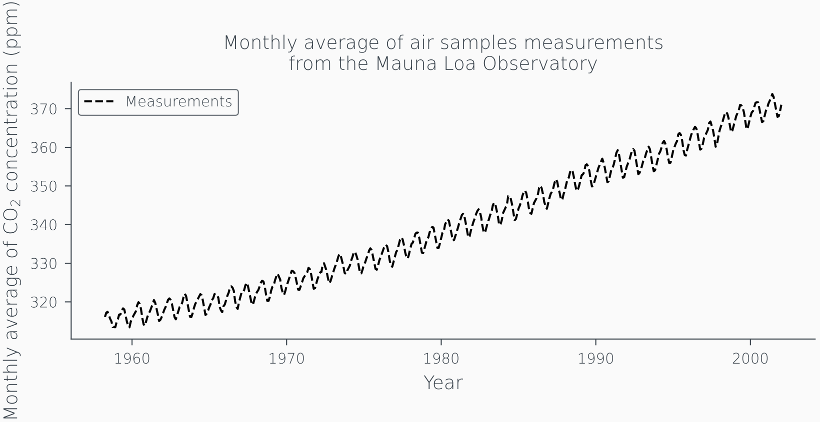
Far from the Moon



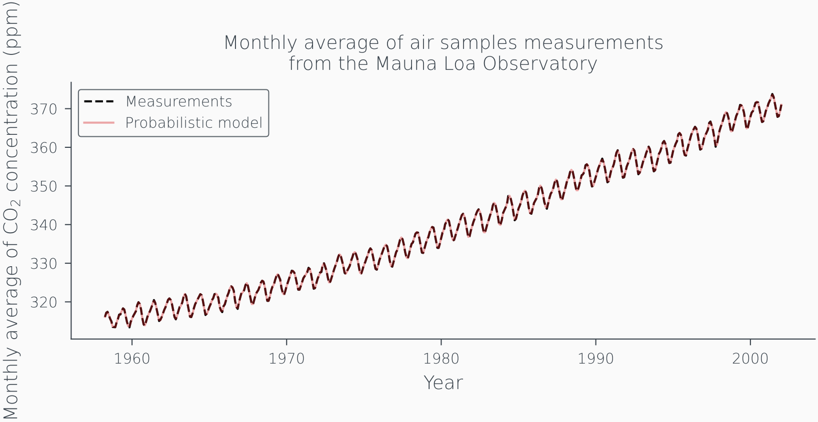
Far from the Moon



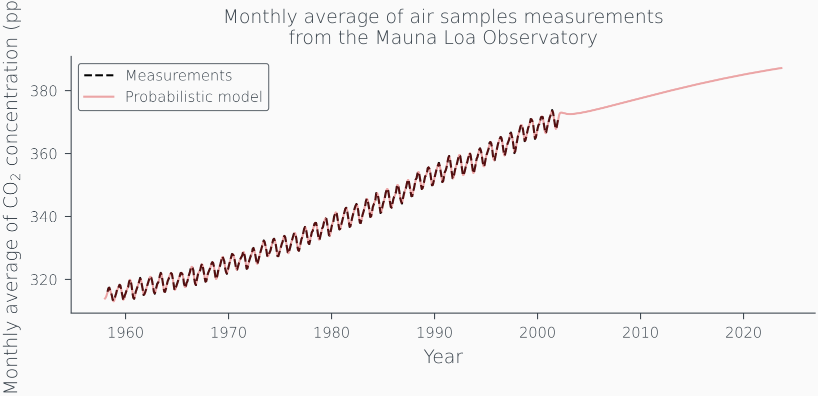
Global Warming?



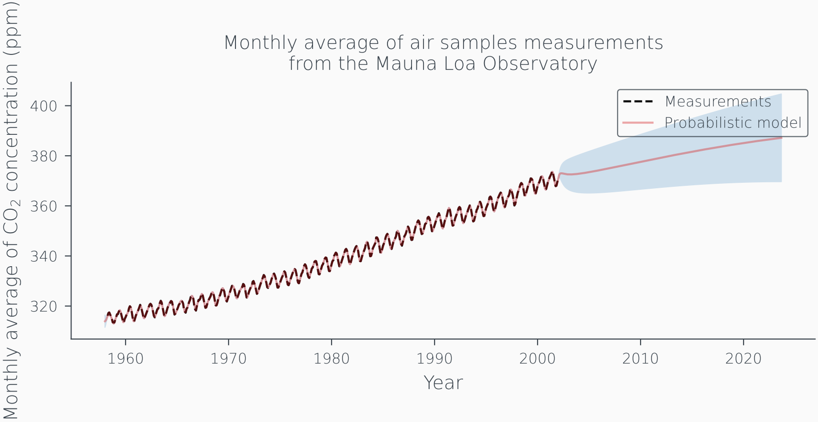
Global Warming?



Global Warming?



Global Warming?



Identify the person

Identify the person

Questions

- We used squared error loss function for linear regression. Why?
- We used cross entropy loss function for logistic regression. Why?
- How does np.random.randn work?
- np.std(x) and pd.std(x) give different results. Why?

Quiz time

Farmer or Librarian? (3Blue1Brown)

Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Rewriting it using the ML notation:

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

- $P(\theta|D)$ is called the posterior
- $P(D|\theta)$ is called the likelihood
- $P(\theta)$ is called the prior
- $P(D)$ is called the evidence

One Equation Throughout the Course

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int_\theta P(D|\theta) \cdot P(\theta) d\theta}$$

I. Maximum Likelihood Estimation

Given a dataset D, find the parameters θ that maximize the likelihood of the data.

$$\theta_{\text{MLE}} = \arg\max_\theta P(D|\theta)$$

For example, given a linear regression problem setup, we set the likelihood as normal distribution and find the parameters θ that maximize the likelihood of the data.

One Equation Throughout the Course

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int_\theta P(D|\theta) \cdot P(\theta) d\theta}$$

II. Maximum A Posteriori Estimation

Given a dataset D, find the parameters θ that maximize the posterior of the data considering both the likelihood and the prior.

$$\theta_{\text{MAP}} = \arg\max_\theta P(\theta|D) = \arg\max_\theta P(D|\theta) \cdot P(\theta)$$

For example, given a linear regression problem, we assume prior over the parameters θ and find the parameters θ that maximize the posterior of the data.

One Equation Throughout the Course

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int_\theta P(D|\theta) \cdot P(\theta) d\theta}$$

III. Bayesian Inference with Conjugate Priors

Find full posterior: $P(\theta|D)$ given likelihood $P(D|\theta)$ and prior $P(\theta)$ where the prior and the posterior belong to the same family of distributions.

One Equation Throughout the Course

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int_\theta P(D|\theta) \cdot P(\theta) d\theta}$$

IV. Main Challenge in Bayesian Inference

Compute the evidence $P(D)$ is intractable in most cases. It involves integrating over all possible values of θ. Thus, computing the posterior $P(\theta|D)$ is intractable in most cases.

One Equation Throughout the Course

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int_\theta P(D|\theta) \cdot P(\theta) d\theta}$$

Va. Approx. Bayesian Inference with Variational Inference

Approximate the posterior $P(\theta|D)$ with a tractable distribution $Q_\phi(\theta)$ characterized by a set of parameters ϕ. Our goal is to find the parameters ϕ that minimize the KL divergence between the approximate posterior $Q_\phi(\theta)$ and the true posterior $P(\theta|D)$.

$$\phi_{\text{VI}} = \arg\min_\phi \text{KL}\left(Q_\phi(\theta) || P(\theta|D)\right)$$

One Equation Throughout the Course

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int_\theta P(D|\theta) \cdot P(\theta) d\theta}$$

Vb. Approx. Bayesian Inference with Laplace Approximation

Approximate the posterior $P(\theta|D)$ with a Gaussian distribution centered at the MAP estimate θ_{MAP} and the covariance matrix is the inverse of the Hessian matrix of the negative log posterior evaluated at θ_{MAP}.

$$P(\theta|D) \approx \mathcal{N}\left(\theta|\theta_{\text{MAP}}, H^{-1}\right)$$

$$H = -\nabla^2 \log P(\theta|D)\Big|_{\theta=\theta_{\text{MAP}}}$$

One Equation Throughout the Course

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int_\theta P(D|\theta) \cdot P(\theta) d\theta}$$

Vc. Approx. Bayesian Inference with Sampling Methods

It is intractable to compute the posterior $P(\theta|D)$ in most cases. But, we can instead get samples from the posterior $P(\theta|D)$.

One Equation Throughout the Course

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int_\theta P(D|\theta) \cdot P(\theta) d\theta}$$

VI. Approx. Integrals with Monte Carlo Integration

Aim: predict the model's output y^* at a new input x^*.

$$P(y^*|x^*, D) = \int_\theta P(y^*|x^*, \theta) \cdot P(\theta|D) d\theta$$

We can instead use Monte Carlo integration to approximate the above integral as follows:

$$P(y^*|x^*, D) \approx \frac{1}{S} \sum_{s=1}^{S} P(y^*|x^*, \theta_s)$$

where $\theta_s \sim P(\theta|D)$.