

Introduction

Nipun Batra

July 29, 2023

IIT Gandhinagar

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

Predict with Uncertainty: Classification

Predict with Uncertainty: Regression

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?

How: 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.

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

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

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

V. 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 $P(\theta|D)$.

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

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. Bayesian Inference with Sampling Methods

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

Main idea is to evaluate the density of the unnormalized posterior $P(\theta|D)$ at any point θ up to a constant of proportionality. This is called **unnormalized density**.

One Equation Throughout the Course

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

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

Univariate Normal Distribution

The probability density function of a univariate normal distribution is given by:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad (1)$$

Let us assume we have a dataset $D = \{x_1, x_2, \dots, x_n\}$, where each x_i is an independent sample from the above distribution. We want to estimate the parameters $\theta = \{\mu, \sigma\}$ from the data.

Our likelihood function is given by:

$$P(D|\theta) = \mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^n f(x_i|\mu, \sigma^2) \quad (2)$$

Log Likelihood Function

Log-likelihood function:

$$\log \mathcal{L}(\mu, \sigma^2) = \sum_{i=1}^n \log f(x_i | \mu, \sigma^2) \quad (3)$$

Simplifying the above equation, we get:

$$\begin{aligned} \log \mathcal{L}(\mu, \sigma^2) &= \sum_{i=1}^n \log f(x_i | \mu, \sigma^2) \\ &= \sum_{i=1}^n \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right) \\ &= \sum_{i=1}^n \left(\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \log \left(\exp \left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right) \right) \end{aligned}$$

$$\begin{aligned}
 \log \mathcal{L}(\mu, \sigma^2) &= \sum_{i=1}^n \left(\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\
 &= \sum_{i=1}^n \left(-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\
 &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2
 \end{aligned}$$

Log Likelihood Function for Univariate Normal Distribution

Log-likelihood function for normally distributed data is:

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

Maximum Likelihood Estimate for μ

To find the MLE for μ , we differentiate the log-likelihood function with respect to μ and set it to zero:

$$\frac{\partial \log \mathcal{L}(\mu, \sigma^2)}{\partial \mu} = \frac{\partial}{\partial \mu} \left(-\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right) = 0$$
$$\frac{\partial}{\partial \mu} \left(\sum_{i=1}^n (x_i - \mu)^2 \right) = 0$$

Maximum Likelihood Estimate for μ

MLE of μ , denoted as $\hat{\mu}_{\text{MLE}}$, is given by:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n x_i$$

MLE for σ for normally distributed data

Recall that the log-likelihood function is given by:

$$\log \mathcal{L}(\mu, \sigma^2) = \sum_{i=1}^n \log f(x_i | \mu, \sigma^2) \quad (4)$$

Let us find the maximum likelihood estimate of σ^2 now. We can do this by taking the derivative of the log-likelihood function with respect to σ^2 and equating it to zero.

$$\frac{\partial \log \mathcal{L}(\mu, \sigma^2)}{\partial \sigma^2} = \sum_{i=1}^n \frac{\partial \log f(x_i | \mu, \sigma^2)}{\partial \sigma^2} = 0 \quad (5)$$

MLE for σ for normally distributed data

Log Likelihood Function for Univariate Normal Distribution

Log-likelihood function for normally distributed data is:

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

Now, we can differentiate the log-likelihood function with respect to σ and equate it to zero.

MLE for σ for normally distributed data

$$\frac{\partial}{\partial \sigma} \log \mathcal{L}(\mu, \sigma^2) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = 0$$

Multiplying through by σ^3 , we have:

$$-n\sigma^2 + \sum_{i=1}^n (x_i - \mu)^2 = 0$$

Maximum Likelihood Estimate for σ^2

MLE of σ^2 , denoted as $\hat{\sigma}_{\text{MLE}}^2$, is given by:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

Bias of an Estimator

The bias of an estimator $\hat{\theta}$ of a parameter θ is defined as:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

where $\mathbb{E}(\hat{\theta})$ is the expected value of the estimator $\hat{\theta}$.

- An estimator is said to be unbiased if $\text{Bias}(\hat{\theta}) = 0$.
- An estimator is said to be biased if $\text{Bias}(\hat{\theta}) \neq 0$.

Bias of an Estimator: $\hat{\mu}_{MLE}$

Question: What is the expectation of $\hat{\mu}_{MLE}$ calculated over? What is the source of randomness?

Let us assume that the true underlying distribution is $\mathcal{N}(\mu, \sigma^2)$.

Let $\mathcal{D}^1 = \{x_1^1, x_2^1, \dots, x_n^1\}$ be a dataset obtained from this distribution.

The MLE of μ based on \mathcal{D}^1 is given by:

$$\hat{\mu}_{MLE}^1 = \frac{1}{n} \sum_{i=1}^n x_i^1$$

If we obtained another dataset $\mathcal{D}^2 = \{x_1^2, x_2^2, \dots, x_n^2\}$ from the same distribution, the MLE of μ based on \mathcal{D}^2 would be:

$$\hat{\mu}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n x_i^2$$

Bias of an Estimator: $\hat{\mu}_{MLE}$

If we repeat this process and obtain datasets $\mathcal{D}^1, \mathcal{D}^2, \dots, \mathcal{D}^k$, we would have k different estimates of μ .

Taking the expectation of these k estimates gives us the expected value of $\hat{\mu}_{MLE}$:

$$\mathbb{E}(\hat{\mu}_{MLE}) = \frac{1}{k} \sum_{i=1}^k \hat{\mu}_{MLE}^i$$

Simplifying further, we have:

$$\mathbb{E}(\hat{\mu}_{MLE}) = \frac{1}{kn} \sum_{i=1}^k \sum_{j=1}^n x_j^i$$

This expectation is calculated over multiple datasets $\mathcal{D}^1, \mathcal{D}^2, \dots, \mathcal{D}^k$, where each dataset represents a different realization of the random variables from the underlying distribution.

Bias of an Estimator: $\hat{\mu}_{MLE}$

To show that the estimator $\hat{\mu}_{MLE}$ is unbiased, we need to demonstrate that $\mathbb{E}(\hat{\mu}_{MLE}) = \mu$.

Recall that each x_j^i is a random variable following $\mathcal{N}(\mu, \sigma^2)$. Therefore, the sum $\sum_{i=1}^k x_j^i$ follows $\mathcal{N}(k\mu, k\sigma^2)$.

Thus, we can write:

$$\begin{aligned}\mathbb{E}(\hat{\mu}_{MLE}) &= \frac{1}{kn} \sum_{i=1}^k \sum_{j=1}^n x_j^i = \frac{1}{kn} \sum_{j=1}^n \left(\sum_{i=1}^k x_j^i \right) \\ &= \frac{1}{kn} \sum_{j=1}^n (k\mu) = \frac{1}{kn} (kn\mu) = \mu\end{aligned}$$

Estimator $\hat{\mu}_{MLE}$ is unbiased

$$\mathbb{E}(\hat{\mu}_{MLE}) = \mu$$

Bias of σ_{MLE}^2

The MLE of σ^2 is given by

$\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$ where μ is the MLE of the mean.

$$\begin{aligned}\mathbb{E}(\hat{\sigma}_{MLE}^2) &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(x_i - \mu)^2] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[x_i^2] - 2\mu \mathbb{E}[x_i] + \mu^2 = \frac{1}{n} \sum_{i=1}^n \sigma^2 + \mu^2 - 2\mu\mu \\ &= \frac{n-1}{n} \sigma^2 + \mu^2 - \mu^2 = \frac{n-1}{n} \sigma^2\end{aligned}$$

Estimator $\hat{\sigma}_{MLE}^2$ is biased

$$\mathbb{E}(\hat{\sigma}_{MLE}^2) = \frac{n-1}{n} \sigma^2$$











