Introduction

Nipun Batra

July 19, 2023

IIT Gandhinagar

What

- 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

Univariate Normal Distribution

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

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

Let us assume we have a dataset $D = \{x_1, x_2, \ldots, x_n\}$, 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:

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

Log Likelihood Function

Log-likelihood function:

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

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:

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

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.

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

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

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

Bias of an Estimator

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 $\mathcal{D}^1 = \{x_1^1, x_2^1, \ldots, x_n^1\}$ 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 $\mathcal{D}^2 = \{x_1^2, x_2^2, \ldots, x_n^2\}$ 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 $\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^k$, we would have k different estimates of μ.

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

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

Simplifying further, we have:

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

This expectation is calculated over multiple datasets $\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^k$, 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:

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

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

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

Bias of σ^2_{MLE}

The MLE of σ^2 is given by

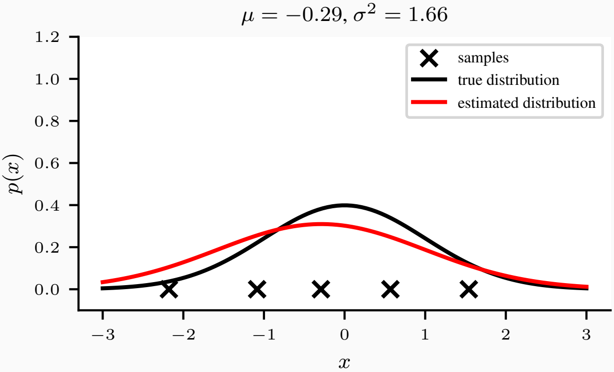
$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ where μ is the MLE of the mean.

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

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

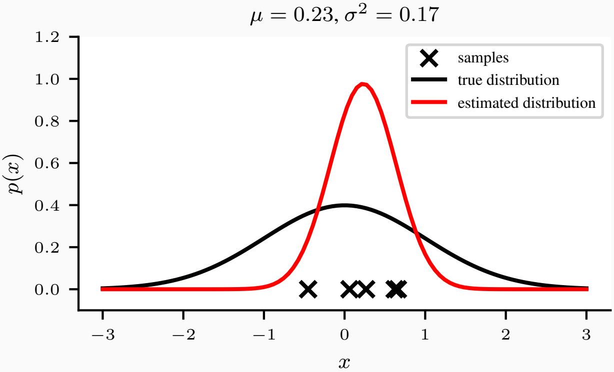
$\mathbb{E}(\hat{\sigma}^2_{MLE}) = \frac{n-1}{n}\sigma^2$

Bias



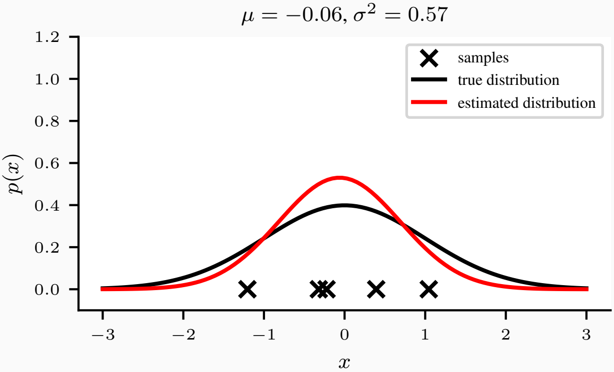
$\mu = -0.29, \sigma^2 = 1.66$
1.2
1.0
0.8
0.6
0.4
0.2
0.0
$p(x)$
samples
true distribution
estimated distribution
−3
−2
−1
0
1
2
3
x

Bias

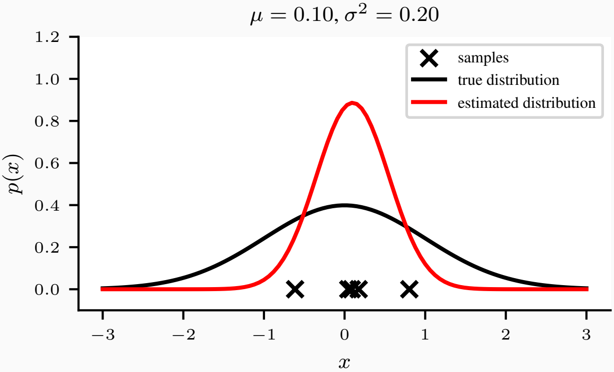


$\mu = 0.23, \sigma^2 = 0.17$
samples
true distribution
estimated distribution
$p(x)$
x
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

Bias

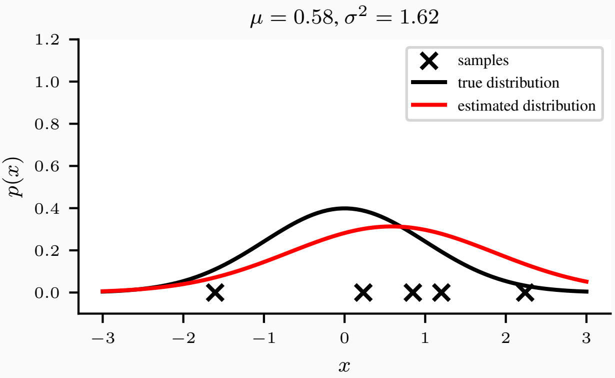

$\mu = -0.06, \sigma^2 = 0.57$
1.2
1.0
0.8
0.6
0.4
0.2
0.0
$p(x)$
samples
true distribution
estimated distribution
−3
−2
−1
0
1
2
3
x

Bias

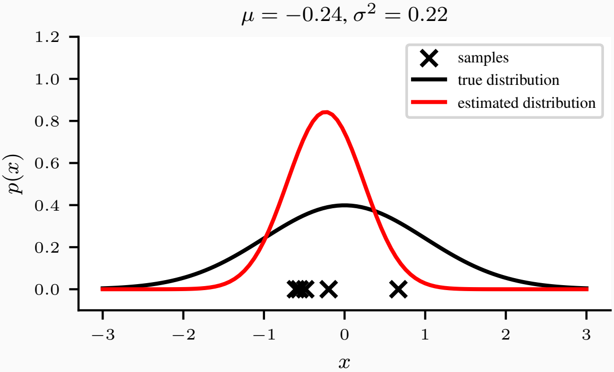


$\mu = 0.10, \sigma^2 = 0.20$
1.2
1.0
0.8
0.6
0.4
0.2
0.0
$p(x)$
samples
true distribution
estimated distribution
−3
−2
−1
0
1
2
3
x

Bias


$\mu = 0.58, \sigma^2 = 1.62$
samples
true distribution
estimated distribution
$p(x)$
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
x

Bias


$\mu = -0.24, \sigma^2 = 0.22$
samples
true distribution
estimated distribution
$p(x)$
x
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3