Week 5: Normal Distribution MLE
Machine Learning
CS1675
Fall 2023

February 7, 2024

Normal likelihood with unknown μ: MLE The normal, or Gaussian distribution (*alias* Bell curve) is a probability density function fully described by the values μ (mean) and σ^2 (variance). The square root of variance, σ, is also known as the standard deviation.

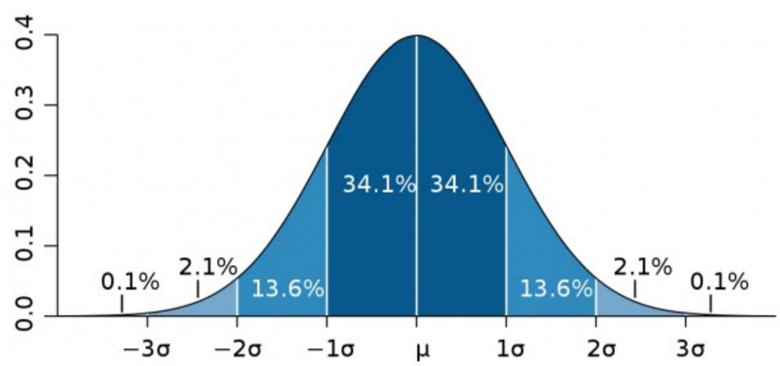

Figure 1: Normal distribution for one-dimensional x_n with mean $\mu = 0$ and standard deviation $\sigma = 1$. The filled in areas represent the probability of a random sample x to lie within 1 ($\mu \pm \sigma$), 2 ($\mu \pm 2\sigma$), or 3 standard deviations of the mean. The curve can be shifted left or right by changing the mean, and squeezed taller or smashed shorter by changing the variance.

For N observations of the random variable x, x is drawn from a normal distribution with known variance σ^2 and unknown mean, μ. Our goal is to estimate

μ given some vector **x** of inputs. All $x \in \mathbf{x}$ are drawn independently, meaning the overall likelihood of **x** is given by:

$$p(\mathbf{x}|\mu, \sigma) = \prod_{n=1}^{N} p(x_n|\mu, \sigma) = \prod_{n=1}^{N} \mathtt{normal}(x_n|\mu, \sigma)$$

for:

$$\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$$

Estimate μ by maximizing the log-likelihood.

$$\begin{aligned} p(\mathbf{x}|\mu, \sigma) &= \prod_{n=1}^{N} \mathtt{normal}(x_n|\mu, \sigma) \\ \hat{\mu} = \mu_{ML} &= \arg\max_{\mu} p(\mathbf{x}|\mu, \sigma) \\ &= \arg\max_{\mu} \log[p(\mathbf{x}|\mu, \sigma)] \end{aligned}$$

To find this maximum value for μ, we will have to calculate the derivative of the log-likelihood and find its zero-point. First, we will expand the expression for the log-likelihood:

$$
\begin{aligned}
\log[p(\mathbf{x}|\mu,\sigma)] &= \log\left[\prod_{n=1}^{N} \texttt{normal}(x_n|\mu,\sigma)\right] \\
&= \sum_{n=1}^{N} \log\left[\texttt{normal}(x_n|\mu,\sigma)\right] \\
&= \sum_{n=1}^{N} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2}\right] \\
&= \sum_{n=1}^{N} \left(\log\left[e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2}\right] - \log\left[\sqrt{2\pi\sigma^2}\right]\right) \\
&= \sum_{n=1}^{N} \left(\log\left[e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2}\right]\right) - \sum_{n=1}^{N} \left(\log\left[\sqrt{2\pi\sigma^2}\right]\right) \\
&= \sum_{n=1}^{N} \left(-\frac{1}{2\sigma^2}(x_n-\mu)^2\right) - N\log\left[\sqrt{2\pi\sigma^2}\right] \\
&= \sum_{n=1}^{N} \left(-\frac{1}{2\sigma^2}(x_n-\mu)^2\right) - \frac{N}{2}\log\left[2\pi\sigma^2\right]
\end{aligned}
$$

Next, we will calculate the derivative. Note the second term from the above equation disappears because we are taking the derivative with respect to μ and so that constant term always becomes zero:

$$
\begin{aligned}
\frac{\partial}{\partial\mu}\log[p(\mathbf{x}|\mu,\sigma)] &= \sum_{n=1}^{n} \frac{\partial}{\partial\mu}\left(-\frac{1}{2\sigma^2}(x_n-\mu)^2\right) \\
&= \sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n-\mu) \\
&= \frac{1}{\sigma^2}\left(\sum_{n=1}^{N} x_n - \sum_{n=1}^{N} \mu\right) \\
&= \frac{1}{\sigma^2}\left(\sum_{n=1}^{N} x_n - \mu N\right)
\end{aligned}
$$

We can get rid of the summation symbol over x_n by substituting in the mean

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{N} x_n$$

Rearranging, and substituting into the last equation:

$$\begin{aligned}\frac{\partial}{\partial\mu}\log[p(\mathbf{x}|\mu,\sigma)] &= \frac{1}{\sigma^2}\left(\bar{\mathbf{x}}N - \mu N\right)\\ &= \frac{N}{\sigma^2}(\bar{\mathbf{x}} - \mu)\end{aligned}$$

Now, in order to find $\arg\max_\mu$ we must solve for the derivative equal to zero:

$$\frac{\partial}{\partial\mu}\log[p(\mathbf{x}|\mu,\sigma)] = \frac{N}{\sigma^2}(\bar{\mathbf{x}} - \mu_{ML}) = 0$$

$$\mu_{ML} = \bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

In order to check whether $\frac{\partial \log p}{\partial\mu} = 0$ is a maximum or minimum, we can test the second derivative for the following conditions:

- if $f''(x) < 0$ then f has a local maximum at x
- if $f''(x) > 0$ then f has a local minimum at x
- if $f''(x) = 0$ then x is *possibly* an inflection point. Further analysis required.

The second derivative of the log-likelihood is a negative constant, thus μ_{ML} is a local maximum:

$$\begin{aligned}\frac{\partial^2}{\partial\mu^2}\log\left[p(\mathbf{x}|\mu,\sigma^2)\right] &= \frac{\partial}{\partial\mu}\left(\frac{\partial}{\partial\mu}\log\left[p(\mathbf{x}|\mu,\sigma^2)\right]\right)\\ &= \frac{\partial}{\partial\mu}\left(\frac{N}{\sigma^2}(\bar{\mathbf{x}} - \mu_{ML})\right)\\ &= -\frac{N}{\sigma^2}\end{aligned}$$

The maximum-likelihood estimate μ_{ML} for a normal distribution is equal to the mean of the observed data points. ■