Deep Learning

Chapter 10: Maximum Likelihood Estimation

Mina Rezaei

Department of Statistics – LMU Munich

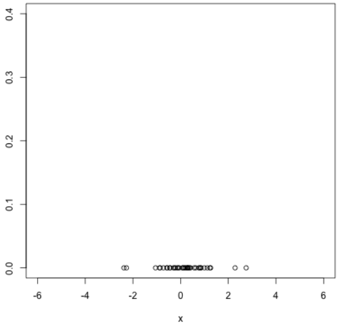
Winter Semester 2020

LECTURE OUTLINE

Maximum Likelihood

Maximum Likelihood

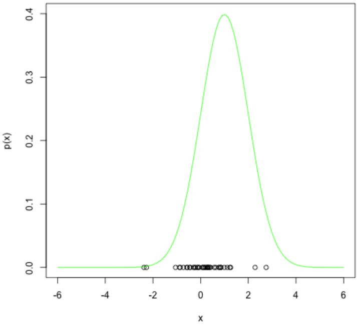
MAXIMUM LIKELIHOOD



We choose the model distribution p_θ to be Gaussian, that is

$$p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{\left(\mathbf{x}^{(i)} - \mu\right)^2}{\sigma^2}\right)$$

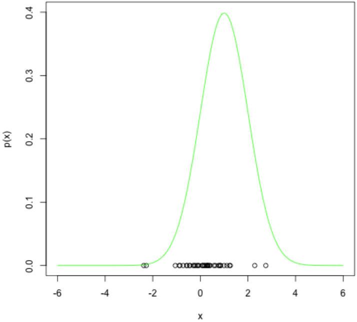
MAXIMUM LIKELIHOOD



We choose the model distribution p_θ to be Gaussian, that is

$$p_{\boldsymbol{\theta}}\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}\left(\mathbf{x}^{(i)}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(\mathbf{x}^{(i)} - \mu)^2}{\sigma^2}\right)$$

MAXIMUM LIKELIHOOD

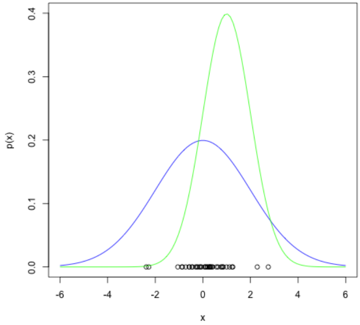


We choose the model distribution p_θ to be Gaussian, that is

$$p_{\boldsymbol{\theta}}\left(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}\right)=\prod_{i=1}^{n} p_{\boldsymbol{\theta}}\left(\mathbf{x}^{(i)}\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(\mathbf{x}^{(i)}-\mu)^2}{\sigma^2}\right)$$

Given $\left\{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}\right\}$, how should we estimate $\boldsymbol{\theta}=\{\mu,\sigma^2\}$?

MAXIMUM LIKELIHOOD

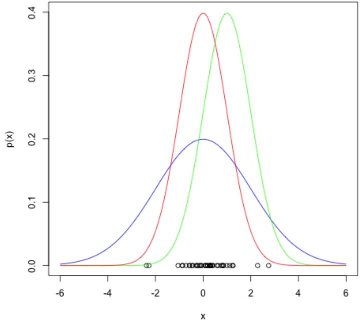


We choose the model distribution p_θ to be Gaussian, that is

$$p_{\boldsymbol{\theta}}\left(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}\right)=\prod_{i=1}^{n}p_{\boldsymbol{\theta}}\left(\mathbf{x}^{(i)}\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(\mathbf{x}^{(i)}-\mu)^2}{\sigma^2}\right)$$

Given $\left\{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}\right\}$, how should we estimate $\boldsymbol{\theta}=\{\mu,\sigma^2\}$?

MAXIMUM LIKELIHOOD



We choose the model distribution p_θ to be Gaussian, that is

$$p_{\boldsymbol{\theta}}\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}\left(\mathbf{x}^{(i)}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(\mathbf{x}^{(i)} - \mu)^2}{\sigma^2}\right)$$

Given $\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right\}$, how should we estimate $\boldsymbol{\theta} = \{\mu, \sigma^2\}$?

RECALL: MAXIMUM LIKELIHOOD ESTIMATION

The **likelihood function** is given by

$$L\left(\boldsymbol{\theta}|\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)}\right) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}\left(\mathbf{x}^{(i)}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(\mathbf{x}^{(i)}-\mu)^2}{\sigma^2}\right) \quad .$$

To maximize it, we often consider the **log-likelihood**

$$\begin{aligned}\log L(\boldsymbol{\theta}|\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)}) &= \log \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) = \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) \\ &= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\sum_{i=1}^{n} \frac{(\mathbf{x}^{(i)}-\mu)^2}{\sigma^2} \quad .\end{aligned}$$

RECALL: MAXIMUM LIKELIHOOD ESTIMATION

Setting derivatives equal to zero yields

$$\frac{\partial \log L\left(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right)}{\partial \mu} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} \mathbf{x}^{(i)} - n\mu\right)$$

and

$$\frac{\partial \log L(\boldsymbol{\theta}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})}{\partial \sigma} = \frac{1}{2\sigma^2}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - \mu)^2 - n\right) \quad .$$

Leading to

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}^{(i)} \quad \text{and} \quad \hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - \hat{\mu})^2 \quad .$$

NOTES ON MAXIMUM LIKELIHOOD LEARNING

- For a model p with visible variables $\vec{x}$ and hidden variables $\vec{z}$, the likelihood computation involves

$$p(\vec{\mathbf{x}^{(i)}} \mid \vec{\theta}) = \sum_{\vec{z}} p(\vec{\mathbf{x}^{(i)}}, \vec{z} \mid \vec{\theta}) \ .$$

  This is difficult, especially because of the sum which prevents the logarithm to act directly on the joint distribution.

NOTES ON MAXIMUM LIKELIHOOD LEARNING

- For a model p with visible variables $\vec{x}$ and hidden variables $\vec{z}$, the likelihood computation involves

$$p(\mathbf{x}^{\vec{(i)}} \mid \vec{\theta}) = \sum_{\vec{z}} p(\mathbf{x}^{\vec{(i)}}, \vec{z} \mid \vec{\theta}) \ .$$

  This is difficult, especially because of the sum which prevents the logarithm to act directly on the joint distribution.

- If we can not find the maximum likelihood parameters analytically (i.e. by setting the derivative to zero) one can maximize the likelihood via SGD or related algorithms.

- If p_{data} is the true distribution underlying S, maximizing the logarithmic likelihood function corresponds to minimizing an empirical estimate of the Kullback-Leibler divergence $KL(p_{\text{data}} \,\|\, p)$.