Supervised Learning: Exercise 3

Yawei Li

yawei.li@stat.uni-muenchen.de

November 17, 2023

Exercise 3: Connection between between MLE and ERM

Suppose we are facing a regression task, i.e., $\mathcal{Y} = \mathbb{R}$, and the feature space $\mathcal{X} \subseteq \mathbb{R}^p$. Let us assume that the relationship between the features and labels is specified by

$$y = m^{-1}(m(f_{\text{true}}(\mathbf{x})) + \epsilon), \qquad \triangleleft$$

where $m : \mathbb{R} \to \mathbb{R}$ is a continuous strictly monotone function with m^{-1} being its inverse function, and the errors are Gaussian, i.e., $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In particular, for the data points $(\mathbf{x}^{(1)}, y^{(1)}, \ldots, \mathbf{x}^{(n)}, y^{(n)})$ it holds that

$$y^{(i)} = m^{-1}(m(f_{\text{true}}(\mathbf{x}^{(i)})) + \epsilon^{(i)}),$$

where $\epsilon^{(1)}, \ldots, \epsilon^{(n)}$ are i.i.d. with distribution $\mathcal{N}(0, \epsilon^2)$.

Disclaimer: We assume in the following that $m(y)$ and $m(f(\mathbf{x}))$ is well-defined for any $y \in \mathcal{Y}, f \in \mathcal{H}$ and $\mathbf{x} \in \mathcal{X}$.

(a) How can we transform the labels $y^{(1)}, \ldots, y^{(n)}$ to “new” labels $z^{(1)}, \ldots, z^{(n)}$ such that $z^{(i)} \mid \mathbf{x}$ is normally distributed? What are the parameters of this normal distribution?

Solution: Question (a)

We can use

$$z^{(i)} = m(y^{(i)}).$$

We know that $y^{(i)} = m^{-1}(m(f_{\text{true}}(\mathbf{x}^{(i)})) + \epsilon^{(i)})$, therefore $m(y^{(i)}) = m(f_{\text{true}}(\mathbf{x}^{(i)})) + \epsilon^{(i)}$.

Hence,

$$z^{(i)} = m(f_{\text{true}}(\mathbf{x}^{(i)})) + \epsilon^{(i)}, \quad \text{with } \epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2).$$

Therefore, $z^{(i)} \mid \mathbf{x}^{(i)}$ is normally distributed, i.e.,

$$z^{(i)} \mid \mathbf{x}^{(i)} \sim \mathcal{N}(m(f_{\text{true}}(\mathbf{x}^{(i)})), \sigma^2)$$

.

Solution: Question (a)

We can use
$$z^{(i)} = m(y^{(i)}).$$
We know that $y^{(i)} = m^{-1}(m(f_{\text{true}}(\mathbf{x}^{(i)})) + \epsilon^{(i)})$, therefore $m(y^{(i)}) = m(f_{\text{true}}(\mathbf{x}^{(i)})) + \epsilon^{(i)}$.
Hence,
$$z^{(i)} = m(f_{\text{true}}(\mathbf{x}^{(i)})) + \epsilon^{(i)}, \quad \text{with } \epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2).$$
Therefore, $z^{(i)} \mid \mathbf{x}^{(i)}$ is normally distributed, i.e.,
$$z^{(i)} \mid \mathbf{x}^{(i)} \sim \mathcal{N}(m(f_{\text{true}}(\mathbf{x}^{(i)})), \sigma^2)$$

.

Question (b)

Assume that the hypothesis space is

$\mathcal{H} = \{f(\cdot, \boldsymbol{\theta}) : \mathcal{X} \to \mathbb{R} | f(\cdot|\boldsymbol{\theta})$ belongs to a certain functional family parameterized by $\boldsymbol{\theta} \in \Theta\}$,

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d))$ is a parameter vector, which is an element of a **parameter space** Θ. Based on your findings (a), establish a relationship between minimizing the negative log-likelihood for $(\mathbf{x}^{(1)}, z^{(1)}), \ldots (\mathbf{x}^{(n)}, z^{(n)})$ and empirical loss minimizing over $\mathcal{H}$ of the generalized L2-loss function of Exercise sheet 1, i.e., $L(y, f(\mathbf{x}))) = (m(y) - m(f(\mathbf{x})))^2$.

Solution to Question (b)

The likelihood for $(\mathbf{x}^{(1)}, z^{(1)}), \dots (\mathbf{x}^{(n)}, z^{(n)})$ is

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} p\left(z^{(i)} \mid f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}), \sigma^2\right)$$

So the negative log-likelihood (NLL) is

Therefore, the NLL is propotional to the empirical risk of $f(\cdot|\theta)$ w.r.t. the generalized L2-Loss.

Solution to Question (b)

The likelihood for $(\mathbf{x}^{(1)}, z^{(1)}), \dots (\mathbf{x}^{(n)}, z^{(n)})$ is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p\left(z^{(i)} \mid f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}), \sigma^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[z^{(i)} - m(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))\right]^2\right).\end{aligned}$$

So the negative log-likelihood (NLL) is

Therefore, the NLL is propotional to the empirical risk of $f(\cdot|\theta)$ w.r.t. the generalized L2-Loss.

Solution to Question (b)

The likelihood for $(\mathbf{x}^{(1)}, z^{(1)}), \ldots (\mathbf{x}^{(n)}, z^{(n)})$ is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p\left(z^{(i)} \mid f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}), \sigma^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[z^{(i)} - m(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))\right]^2\right).\end{aligned}$$

So the negative log-likelihood (NLL) is

Therefore, the NLL is propotional to the empirical risk of $f(\cdot|\theta)$ w.r.t. the generalized L2-Loss.

Solution to Question (b)

The likelihood for $(\mathbf{x}^{(1)}, z^{(1)}), \ldots (\mathbf{x}^{(n)}, z^{(n)})$ is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p\left(z^{(i)} \mid f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}), \sigma^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[z^{(i)} - m(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))\right]^2\right).\end{aligned}$$

So the negative log-likelihood (NLL) is

$$-\ell(\boldsymbol{\theta}) = -\log(\mathcal{L}(\boldsymbol{\theta}))$$

Therefore, the NLL is propotional to the empirical risk of $f(\cdot|\theta)$ w.r.t. the generalized L2-Loss.

Solution to Question (b)

The likelihood for $(\mathbf{x}^{(1)}, z^{(1)}), \dots (\mathbf{x}^{(n)}, z^{(n)})$ is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p\left(z^{(i)} \mid f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}), \sigma^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[z^{(i)} - m(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))\right]^2\right).\end{aligned}$$

So the negative log-likelihood (NLL) is

$$\begin{aligned}-\ell(\boldsymbol{\theta}) &= -\log(\mathcal{L}(\boldsymbol{\theta})) \\ &\propto \sum_{i=1}^{n}\left[z^{(i)} - m(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))\right]^2\end{aligned}$$

Therefore, the NLL is propotional to the empirical risk of $f(\cdot|\theta)$ w.r.t. the generalized L2-Loss.

Solution to Question (b)

The likelihood for $(\mathbf{x}^{(1)}, z^{(1)}), \dots (\mathbf{x}^{(n)}, z^{(n)})$ is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p\left(z^{(i)} \mid f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}), \sigma^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[z^{(i)} - m(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))\right]^2\right).\end{aligned}$$

So the negative log-likelihood (NLL) is

$$\begin{aligned}-\ell(\boldsymbol{\theta}) &= -\log(\mathcal{L}(\boldsymbol{\theta})) \\ &\propto \sum_{i=1}^{n} \left[z^{(i)} - m(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))\right]^2 \\ &= \sum_{i=1}^{n} \left[m(y^{(i)}) - m(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))\right]^2\end{aligned}$$

Therefore, the NLL is propotional to the empirical risk of $f(\cdot|\theta)$ w.r.t. the generalized L2-Loss.

Solution to Question (b)

The likelihood for $(\mathbf{x}^{(1)}, z^{(1)}), \dots (\mathbf{x}^{(n)}, z^{(n)})$ is

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} p\left(z^{(i)} \mid f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}), \sigma^2\right)$$
$$\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[z^{(i)} - m(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))\right]^2\right).$$

So the negative log-likelihood (NLL) is

$$-\ell(\boldsymbol{\theta}) = -\log(\mathcal{L}(\boldsymbol{\theta}))$$
$$\propto \sum_{i=1}^{n}\left[z^{(i)} - m(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))\right]^2$$
$$= \sum_{i=1}^{n}\left[m(y^{(i)}) - m(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))\right]^2$$

Therefore, the NLL is propotional to the empirical risk of $f(\cdot|\boldsymbol{\theta})$ w.r.t. the generalized L2-Loss.

Question (c)

(c) In many practical applications, one often observed statisical property is that the label y given $\mathbf{x}$ follows a *log-normal distribution*. Note that we can obtain such a relationship by using $m(x) = \log(x)$ above. In the following we want to consider the conjecture of James D. Forbes, who conjectured in the year 1857 that the relationship between the air pressure y and the boiling point of water x is given by

$$y = \theta_1 \exp(\theta_2 x + \epsilon),$$

for some specific values $\theta_1 \in \mathbb{R}_+, \theta_2 \in \mathbb{R}$ and some error terms ϵ (of course, we assume that this error term is stochastic and normally distributed).
What would be a suitable hypothesis space $\mathcal{H}$ if this conjecture holds?

Solution to Question (c)

- The goal is to introduce a proper form for $f(\cdot|\boldsymbol{\theta})$ so that $y = m^{-1}(m(f(x)) + \epsilon)$.
- We use the transformation $m(\cdot) = \log(\cdot)$. So $m^{-1}(\cdot) = \exp(\cdot)$.
- The Forbes' conjectured model $y = \theta_1 \exp(\theta_2 x + \epsilon)$ can be written as

Therefore, $f(x \mid \theta) = \theta_1 \exp(\theta_2 x)$ is a suitable functional form for the hypothesis. So

$$\mathcal{H} = \{f(x \mid \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x) \mid \boldsymbol{\theta} \in \Theta\}.$$

(The standard solution uses $\mathcal{X} = 1 \times \mathbb{R}$, i.e., $\mathbf{x} = (x1, x2)^T = (1, x_2)^T$.)

Solution to Question (c)

- The goal is to introduce a proper form for $f(\cdot|\boldsymbol{\theta})$ so that $y = m^{-1}(m(f(x)) + \epsilon)$.
- We use the transformation $m(\cdot) = \log(\cdot)$. So $m^{-1}(\cdot) = \exp(\cdot)$.
- The Forbes' conjectured model $y = \theta_1 \exp(\theta_2 x + \epsilon)$ can be written as

  Therefore, $f(x \mid \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x)$ is a suitable functional form for the hypothesis. So

$$\mathcal{H} = \{f(x \mid \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x) \mid \boldsymbol{\theta} \in \Theta\}.$$

  (The standard solution uses $\mathcal{X} = 1 \times \mathbb{R}$, i.e., $\mathbf{x} = (x1, x2)^T = (1, x_2)^T$.)

Solution to Question (c)

- The goal is to introduce a proper form for $f(\cdot|\boldsymbol{\theta})$ so that $y = m^{-1}(m(f(x)) + \epsilon)$.
- We use the transformation $m(\cdot) = \log(\cdot)$. So $m^{-1}(\cdot) = \exp(\cdot)$.
- The Forbes' conjectured model $y = \theta_1 \exp(\theta_2 x + \epsilon)$ can be written as

Therefore, $f(x \mid \theta) = \theta_1 \exp(\theta_2 x)$ is a suitable functional form for the hypothesis. So

$$\mathcal{H} = \{f(x \mid \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x) \mid \boldsymbol{\theta} \in \Theta\}.$$

(The standard solution uses $\mathcal{X} = 1 \times \mathbb{R}$, i.e., $\mathbf{x} = (x1, x2)^T = (1, x_2)^T$.)

Solution to Question (c)

- The goal is to introduce a proper form for $f(\cdot|\boldsymbol{\theta})$ so that $y = m^{-1}(m(f(x)) + \epsilon)$.
- We use the transformation $m(\cdot) = \log(\cdot)$. So $m^{-1}(\cdot) = \exp(\cdot)$.
- The Forbes' conjectured model $y = \theta_1 \exp(\theta_2 x + \epsilon)$ can be written as

$$y = \theta_1 \exp(\theta_2 x + \epsilon)$$

Therefore, $f(x \mid \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x)$ is a suitable functional form for the hypothesis. So

$$\mathcal{H} = \{f(x \mid \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x) \mid \boldsymbol{\theta} \in \Theta\}.$$

(The standard solution uses $\mathcal{X} = 1 \times \mathbb{R}$, i.e., $\mathbf{x} = (x1, x2)^T = (1, x_2)^T$.)

Solution to Question (c)

- The goal is to introduce a proper form for $f(\cdot|\boldsymbol{\theta})$ so that $y = m^{-1}(m(f(x)) + \epsilon)$.
- We use the transformation $m(\cdot) = \log(\cdot)$. So $m^{-1}(\cdot) = \exp(\cdot)$.
- The Forbes' conjectured model $y = \theta_1 \exp(\theta_2 x + \epsilon)$ can be written as

$$\begin{aligned} y &= \theta_1 \exp(\theta_2 x + \epsilon) \\ &= \exp(\log(\theta_1 \cdot \exp(\theta_2 x + \epsilon))) \end{aligned}$$

Therefore, $f(x \mid \theta) = \theta_1 \exp(\theta_2 x)$ is a suitable functional form for the hypothesis. So

$$\mathcal{H} = \{f(x \mid \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x) \mid \boldsymbol{\theta} \in \Theta\}.$$

(The standard solution uses $\mathcal{X} = 1 \times \mathbb{R}$, i.e., $\mathbf{x} = (x1, x2)^T = (1, x_2)^T$.)

Solution to Question (c)

- The goal is to introduce a proper form for $f(\cdot|\boldsymbol{\theta})$ so that $y = m^{-1}(m(f(x)) + \epsilon)$.
- We use the transformation $m(\cdot) = \log(\cdot)$. So $m^{-1}(\cdot) = \exp(\cdot)$.
- The Forbes' conjectured model $y = \theta_1 \exp(\theta_2 x + \epsilon)$ can be written as

$$\begin{aligned} y &= \theta_1 \exp(\theta_2 x + \epsilon) \\ &= \exp(\log(\theta_1 \cdot \exp(\theta_2 x + \epsilon))) \\ &= \exp[\log \theta_1 + \log(\exp(\theta_2 x)) + \log(\exp(\epsilon))] \end{aligned}$$

Therefore, $f(x \mid \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x)$ is a suitable functional form for the hypothesis. So

$$\mathcal{H} = \{f(x \mid \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x) \mid \boldsymbol{\theta} \in \Theta\}.$$

(The standard solution uses $\mathcal{X} = 1 \times \mathbb{R}$, i.e., $\mathbf{x} = (x1, x2)^T = (1, x_2)^T$.)

Solution to Question (c)

- The goal is to introduce a proper form for $f(\cdot|\boldsymbol{\theta})$ so that $y = m^{-1}(m(f(x)) + \epsilon)$.
- We use the transformation $m(\cdot) = \log(\cdot)$. So $m^{-1}(\cdot) = \exp(\cdot)$.
- The Forbes' conjectured model $y = \theta_1 \exp(\theta_2 x + \epsilon)$ can be written as

$$\begin{aligned} y &= \theta_1 \exp(\theta_2 x + \epsilon) \\ &= \exp(\log(\theta_1 \cdot \exp(\theta_2 x + \epsilon))) \\ &= \exp[\log \theta_1 + \log(\exp(\theta_2 x)) + \log(\exp(\epsilon))] \\ &= \exp[\log \theta_1 + \log(\exp(\theta_2 x)) + \epsilon] \end{aligned}$$

Therefore, $f(x \mid \theta) = \theta_1 \exp(\theta_2 x)$ is a suitable functional form for the hypothesis. So

$$\mathcal{H} = \{f(x \mid \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x) \mid \boldsymbol{\theta} \in \Theta\}.$$

(The standard solution uses $\mathcal{X} = 1 \times \mathbb{R}$, i.e., $\mathbf{x} = (x1, x2)^T = (1, x_2)^T$.)

Solution to Question (c)

- The goal is to introduce a proper form for $f(\cdot|\boldsymbol{\theta})$ so that $y = m^{-1}(m(f(x)) + \epsilon)$.
- We use the transformation $m(\cdot) = \log(\cdot)$. So $m^{-1}(\cdot) = \exp(\cdot)$.
- The Forbes' conjectured model $y = \theta_1 \exp(\theta_2 x + \epsilon)$ can be written as

$$\begin{aligned}
y &= \theta_1 \exp(\theta_2 x + \epsilon) \\
&= \exp(\log(\theta_1 \cdot \exp(\theta_2 x + \epsilon))) \\
&= \exp[\log \theta_1 + \log(\exp(\theta_2 x)) + \log(\exp(\epsilon))] \\
&= \exp[\log \theta_1 + \log(\exp(\theta_2 x)) + \epsilon] \\
&= \exp[\log(\theta_1 \cdot \exp(\theta_2 x)) + \epsilon]
\end{aligned}$$

Therefore, $f(x \mid \theta) = \theta_1 \exp(\theta_2 x)$ is a suitable functional form for the hypothesis. So

$$\mathcal{H} = \{f(x \mid \theta) = \theta_1 \exp(\theta_2 x) \mid \theta \in \Theta\}.$$

(The standard solution uses $\mathcal{X} = 1 \times \mathbb{R}$, i.e., $\mathbf{x} = (x1, x2)^T = (1, x_2)^T$.)

Solution to Question (c)

- The goal is to introduce a proper form for $f(\cdot|\boldsymbol{\theta})$ so that $y = m^{-1}(m(f(x)) + \epsilon)$.
- We use the transformation $m(\cdot) = \log(\cdot)$. So $m^{-1}(\cdot) = \exp(\cdot)$.
- The Forbes' conjectured model $y = \theta_1 \exp(\theta_2 x + \epsilon)$ can be written as

$$
\begin{aligned}
y &= \theta_1 \exp(\theta_2 x + \epsilon) \\
&= \exp(\log(\theta_1 \cdot \exp(\theta_2 x + \epsilon))) \\
&= \exp[\log \theta_1 + \log(\exp(\theta_2 x)) + \log(\exp(\epsilon))] \\
&= \exp[\log \theta_1 + \log(\exp(\theta_2 x)) + \epsilon] \\
&= \exp[\log(\theta_1 \cdot \exp(\theta_2 x)) + \epsilon] \\
&= m^{-1}(m(\theta_1 \cdot \exp(\theta_2 x)) + \epsilon).
\end{aligned}
$$

Therefore, $f(x \mid \theta) = \theta_1 \exp(\theta_2 x)$ is a suitable functional form for the hypothesis. So

$$\mathcal{H} = \{f(x \mid \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x) \mid \boldsymbol{\theta} \in \Theta\}.$$

(The standard solution uses $\mathcal{X} = 1 \times \mathbb{R}$, i.e., $\mathbf{x} = (x1, x2)^T = (1, x_2)^T$.)

Solution to Question (c)

- The goal is to introduce a proper form for $f(\cdot|\boldsymbol{\theta})$ so that $y = m^{-1}(m(f(x)) + \epsilon)$.
- We use the transformation $m(\cdot) = \log(\cdot)$. So $m^{-1}(\cdot) = \exp(\cdot)$.
- The Forbes' conjectured model $y = \theta_1 \exp(\theta_2 x + \epsilon)$ can be written as

$$
\begin{aligned}
y &= \theta_1 \exp(\theta_2 x + \epsilon) \\
&= \exp(\log(\theta_1 \cdot \exp(\theta_2 x + \epsilon))) \\
&= \exp[\log \theta_1 + \log(\exp(\theta_2 x)) + \log(\exp(\epsilon))] \\
&= \exp[\log \theta_1 + \log(\exp(\theta_2 x)) + \epsilon] \\
&= \exp[\log(\theta_1 \cdot \exp(\theta_2 x)) + \epsilon] \\
&= m^{-1}(m(\theta_1 \cdot \exp(\theta_2 x)) + \epsilon).
\end{aligned}
$$

Therefore, $f(x \mid \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x)$ is a suitable functional form for the hypothesis. So

$$
\mathcal{H} = \{f(x \mid \boldsymbol{\theta}) = \theta_1 \exp(\theta_2 x) \mid \boldsymbol{\theta} \in \Theta\}.
$$

(The standard solution uses $\mathcal{X} = 1 \times \mathbb{R}$, i.e., $\mathbf{x} = (x1, x2)^T = (1, x_2)^T$.)

Solution to Question (c): Continued

Show the code in the standard solution.