Supervised Learning: Exercise 1

Yawei Li

yawei.li@stat.uni-muenchen.de

Date: TODO

Exercise 1: Risk Minimizers for Generalized L2-Loss

Consider the regression learning setting, i.e., $\mathcal{Y} = \mathbb{R}$, and assume that your loss function of interest is $L(y, f(\mathbf{x})) = (m(y) - m(f(\mathbf{x})))^2$, where: $m : \mathbb{R} \to \mathbb{R}$ is a continuous strictly monotnone function.

Disclaimer: In the following we always assume that $\text{Var}(m(Y))$ **exists.**

(a) Consider the hypothesis space of a constant models $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R} | f(\mathbf{x}) = \boldsymbol{\theta}\ \forall \mathbf{x} \in \mathcal{X}\}$, where $\mathcal{X}$ is the feature space. Show that

$$\hat{f}(\mathbf{x}) = m^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)$$

is the optimal constant model for the loss function above, where m^{-1} is the inverse function of m.

Solution to Question (a)

1. f is a **constant model**: $f(\mathbf{x}) = \boldsymbol{\theta}$ for all $\mathbf{x}$.
2. The empirical risk can be formulated as:

$$\mathcal{R}_{\text{emp}}(f) = \sum_{i=1}^{n} \left(m(y^{(i)}) - m(f(\mathbf{x}^{(i)})) \right)^2 = \sum_{i=1}^{n} \left(m(y^{(i)}) - m(\boldsymbol{\theta}) \right)^2 .$$

3. $\mathcal{R}_{\text{emp}}(f)$ is **strictly convex** (because MSE loss and m is strictly monotone). So the minimum is unique, and can be computed by solving $\partial \mathcal{R}_{\text{emp}}(f) / \partial \boldsymbol{\theta} = \mathbf{0}$.

Solution to Question (a)

1. f is a **constant model**: $f(\mathbf{x}) = \boldsymbol{\theta}$ for all $\mathbf{x}$.
2. The empirical risk can be formulated as:

$$\mathcal{R}_{\text{emp}}(f) = \sum_{i=1}^{n} \left(m(y^{(i)}) - m(f(\mathbf{x}^{(i)})) \right)^2 = \sum_{i=1}^{n} \left(m(y^{(i)}) - m(\boldsymbol{\theta}) \right)^2 .$$

3. $\mathcal{R}_{\text{emp}}(f)$ is **strictly convex** (because MSE loss and m is strictly monotone). So the minimum is unique, and can be computed by solving $\partial \mathcal{R}_{\text{emp}}(f) / \partial \boldsymbol{\theta} = \mathbf{0}$.

Solution to Question (a)

1. f is a **constant model**: $f(\mathbf{x}) = \boldsymbol{\theta}$ for all $\mathbf{x}$.
2. The empirical risk can be formulated as:

$$\mathcal{R}_{\text{emp}}(f) = \sum_{i=1}^{n} \left(m(y^{(i)}) - m(f(\mathbf{x}^{(i)})) \right)^2 = \sum_{i=1}^{n} \left(m(y^{(i)}) - m(\boldsymbol{\theta}) \right)^2 .$$

3. $\mathcal{R}_{\text{emp}}(f)$ is **strictly convex** (because MSE loss and m is strictly monotone). So the minimum is unique, and can be computed by solving $\partial \mathcal{R}_{\text{emp}}(f) / \partial \boldsymbol{\theta} = \mathbf{0}$.

Solution to Question (a): Continued

Goal: Compute the optimal $\boldsymbol{\theta}$ by solving $\partial\mathcal{R}_{\text{emp}}(f)/\partial\boldsymbol{\theta} = \mathbf{0}$.

1. Compute the derivative:

$$\frac{\partial\mathcal{R}_{\text{emp}}(f)}{\partial\boldsymbol{\theta}} = 2\sum_{i=1}^{n}(m(y^{(i)}) - m(\boldsymbol{\theta})) \cdot \frac{\partial m(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = 0$$

2. Using the fact that $\frac{\partial m(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$ is constant for all i, we obtain:

$$\sum_{i=1}^{n}(m(y^{(i)}) - m(\boldsymbol{\theta})) = 0$$

$$\Rightarrow \sum_{i=1}^{n} m(y^{(i)}) = \sum_{i=1}^{n} m(\boldsymbol{\theta})$$

$$\Rightarrow m(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})$$

$$\Rightarrow \boldsymbol{\theta}^* = m^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)$$

Solution to Question (a): Continued

Goal: Compute the optimal $\boldsymbol{\theta}$ by solving $\partial \mathcal{R}_{\text{emp}}(f)/\partial \boldsymbol{\theta} = \mathbf{0}$.

1. Compute the derivative:

$$\frac{\partial \mathcal{R}_{\text{emp}}(f)}{\partial \boldsymbol{\theta}} = 2 \sum_{i=1}^{n} (m(y^{(i)}) - m(\boldsymbol{\theta})) \cdot \frac{\partial m(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

2. Using the fact that $\frac{\partial m(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is constant for all i, we obtain:

$$\sum_{i=1}^{n} (m(y^{(i)}) - m(\boldsymbol{\theta})) = 0$$

$$\Rightarrow \sum_{i=1}^{n} m(y^{(i)}) = \sum_{i=1}^{n} m(\boldsymbol{\theta})$$

$$\Rightarrow m(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} m(y^{(i)})$$

$$\Rightarrow \boldsymbol{\theta}^* = m^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} m(y^{(i)}) \right)$$

Question (b)

(b) Verify that the risk of the optimal constant model is $\mathcal{R}_L\left(\hat{f}\right) = (1 + \frac{1}{n})\text{Var}(m(y))$.

Recall that the risk of $\hat{f}$ is defined as

$$\mathcal{R}_L(\hat{f}) = \mathbb{E}_{xy}[L(y, \hat{f}(\mathbf{x})]$$

Solution to Question (b)

$$
\begin{aligned}
\mathcal{R}_L(\hat{f}) &= \mathbb{E}_{xy}[L(y, \hat{f}(\mathbf{x}))] \\
&= \mathbb{E}_{xy}[(m(y) - m(\hat{f}(\mathbf{x}))^2] \\
&= \mathbb{E}_{xy}\left[\left(m(y) - \frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)^2\right] \\
&= \mathbb{E}_{xy}[m(y)^2] - 2 \cdot \mathbb{E}_{xy}\left[m(y) \cdot \frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right] + \mathbb{E}_{xy}\left[\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\right]
\end{aligned}
$$

Solution to Question (b): Continued

Now take a look at the second term: $-2 \cdot \mathbb{E}_{xy}\left[m(y) \cdot \frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right]$.

Because $y, y^{(1)}, \ldots, y^{(n)}$ are i.i.d. with $\mathbb{E}_{xy}[m(y^{(i)})] = \mathbb{E}_{xy}[m(y)]$, we have

$$\begin{aligned}
\mathbb{E}_{xy}\left[m(y) \cdot \frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right] &= \frac{1}{n} \cdot \mathbb{E}_{xy}\left[m(y)\sum_{i=1}^{n} m(y^{(i)})\right] \\
&= \frac{1}{n} \cdot \mathbb{E}_{xy}[m(y)]\mathbb{E}_{xy}\left[\sum_{i=1}^{n} m(y^{(i)})\right] \\
&= \frac{1}{n} \cdot \mathbb{E}_{xy}[m(y)] \cdot n \cdot \mathbb{E}_{xy}[m(y)] \\
&= \mathbb{E}_{xy}[m(y)]^2.
\end{aligned}$$

Solution to Question (b): Continued

Now take a look at the third term: $\mathbb{E}_{xy}\left[\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\right]$.

Similarly, we have

$$
\begin{aligned}
&\mathbb{E}_{xy}\left[\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\right] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}_{xy}[m(y^{(i)})^2] + \sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}_{xy}[m(y^{(i)})m(y^{(j)})]\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}_{xy}[m(y^{(i)})^2] + \sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}_{xy}[m(y^{(i)})]\cdot\mathbb{E}_{xy}[m(y^{(j)})]\right) \quad \triangleright \\
&= \frac{1}{n}\left(n\mathbb{E}_{xy}[m(y)^2] + n(n-1)\mathbb{E}_{xy}[m(y)]^2\right) \quad \triangleright \\
&= \frac{1}{n}\mathbb{E}_{xy}[m(y)^2] + (1-\frac{1}{n})\mathbb{E}_{xy}[m(y)]^2
\end{aligned}
$$

Solution to Question (b): Continued

Combining the results so far, we get

$$\begin{aligned}
\mathcal{R}_L(\hat{f}) &= \mathbb{E}_{xy}[m(y)^2] - 2 \cdot \mathbb{E}_{xy}\left[m(y) \cdot \frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right] + \mathbb{E}_{xy}\left[\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)\right] \\
&= \mathbb{E}_{xy}[m(y)^2] - 2\mathbb{E}_{xy}[m(y)]^2 + \frac{1}{n}\mathbb{E}_{xy}[m(y)^2] + (1 - \frac{1}{n})\mathbb{E}_{xy}[m(y)]^2 \\
&= \left(1 + \frac{1}{n}\right)\left(\mathbb{E}_{xy}[m(y)^2] - \mathbb{E}_{xy}[m(y)]^2\right) \\
&= \left(1 + \frac{1}{n}\right)\mathrm{Var}(m(y)).
\end{aligned}$$

Question (c)

Derive that the risk minimizer f^* is given by $f^*(\mathbf{x}) = m^{-1}(\mathbb{E}_{y|\mathbf{x}}[m(y)|\mathbf{x}])$.

Hints:

- Consider unstricted hypothesis space $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R}\}$.
- Since $\mathcal{H}$ is unrestricted, for each $\mathbf{x}$, we can predict any value $c \in \mathbb{R}$ we want. $\rightsquigarrow$ Point-wise prediction.
- Point-wise prediction: given unlimited space, we can use a look-up table to store $f^*(\mathbf{x})$ for all $\mathbf{x}$.

Question (c)

Derive that the risk minimizer f^* is given by $f^*(\mathbf{x}) = m^{-1}(\mathbb{E}_{y|\mathbf{x}}[m(y)|\mathbf{x}])$.

Hints:

- Consider unstricted hypothesis space $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R}\}$.
- Since $\mathcal{H}$ is unrestricted, for each $\mathbf{x}$, we can predict any value $c \in \mathbb{R}$ we want. $\rightsquigarrow$ Point-wise prediction.
- Point-wise prediction: given unlimited space, we can use a look-up table to store $f^*(\mathbf{x})$ for all $\mathbf{x}$.

Question (c)

Derive that the risk minimizer f^* is given by $f^*(\mathbf{x}) = m^{-1}(\mathbb{E}_{y|\mathbf{x}}[m(y)|\mathbf{x}])$.

Hints:

- Consider unstricted hypothesis space $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R}\}$.
- Since $\mathcal{H}$ is unrestricted, for each $\mathbf{x}$, we can predict any value $c \in \mathbb{R}$ we want. $\rightsquigarrow$ Point-wise prediction.
- Point-wise prediction: given unlimited space, we can use a look-up table to store $f^*(\mathbf{x})$ for all $\mathbf{x}$.

Question (c)

Derive that the risk minimizer f^* is given by $f^*(\mathbf{x}) = m^{-1}(\mathbb{E}_{y|\mathbf{x}}[m(y)|\mathbf{x}])$.

Hints:

- Consider unstricted hypothesis space $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R}\}$.
- Since $\mathcal{H}$ is unrestricted, for each $\mathbf{x}$, we can predict any value $c \in \mathbb{R}$ we want. $\rightsquigarrow$ Point-wise prediction.
- Point-wise prediction: given unlimited space, we can use a look-up table to store $f^*(\mathbf{x})$ for all $\mathbf{x}$.

Solution to Question (c)

By the law of total expectation,

$$\begin{aligned}\mathcal{R}_L(f) &= \mathbb{E}_{xy}[L(y, f(\mathbf{x}))] \\ &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}}[L(y, f(\mathbf{x})) \mid \mathbf{x}]] \\ &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}}[(m(y) - m(f(\mathbf{x})))^2 \mid \mathbf{x}]].\end{aligned}$$

Since we consider a point-wise prediction, we can omit the $\mathbb{E}_{\mathbf{x}}$, and we focus on computing $f^*(\mathbf{x}) = c$ given a **fixed x**. In other words, we solve the optimal c for each $\mathbf{x}$ separately.

To solve $f^*(\mathbf{x}) = \arg\min_c \mathbb{E}_{y|\mathbf{x}}[(m(y) - m(f(\mathbf{x})))^2 \mid \mathbf{x}]$, we adopt the same way as the solution of Question (a), obtaining

$$f^*(\mathbf{x}) = m^{-1}(\mathbb{E}_{y|\mathbf{x}}[m(y) \mid \mathbf{x}]).$$

Solution to Question (c)

By the law of total expectation,

$$\begin{aligned}\mathcal{R}_L(f) &= \mathbb{E}_{xy}[L(y, f(\mathbf{x}))] \\ &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}}[L(y, f(\mathbf{x})) \mid \mathbf{x}]] \\ &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}}[(m(y) - m(f(\mathbf{x})))^2 \mid \mathbf{x}]].\end{aligned}$$

Since we consider a point-wise prediction, we can omit the $\mathbb{E}_{\mathbf{x}}$, and we focus on computing $f^*(\mathbf{x}) = c$ given a **fixed x**. In other words, we solve the optimal c for each $\mathbf{x}$ separately.

To solve $f^*(\mathbf{x}) = \arg\min_c \mathbb{E}_{y|\mathbf{x}}[(m(y) - m(f(\mathbf{x})))^2 \mid \mathbf{x}]$, we adopt the same way as the solution of Question (a), obtaining

$$f^*(\mathbf{x}) = m^{-1}(\mathbb{E}_{y|\mathbf{x}}[m(y) \mid \mathbf{x}]).$$

Solution to Question (c)

By the law of total expectation,

$$\begin{aligned}\mathcal{R}_L(f) &= \mathbb{E}_{xy}[L(y, f(\mathbf{x}))] \\ &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}}[L(y, f(\mathbf{x})) \mid \mathbf{x}]] \\ &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}}[(m(y) - m(f(\mathbf{x})))^2 \mid \mathbf{x}]].\end{aligned}$$

Since we consider a point-wise prediction, we can omit the $\mathbb{E}_{\mathbf{x}}$, and we focus on computing $f^*(\mathbf{x}) = c$ given a **fixed x**. In other words, we solve the optimal c for each $\mathbf{x}$ separately.

To solve $f^*(\mathbf{x}) = \arg\min_c \mathbb{E}_{y|\mathbf{x}}[(m(y) - m(f(\mathbf{x})))^2 \mid \mathbf{x}]$, we adopt the same way as the solution of Question (a), obtaining

$$f^*(\mathbf{x}) = m^{-1}(\mathbb{E}_{y|\mathbf{x}}[m(y) \mid \mathbf{x}]).$$

Question (d)

(d): What is the optimal **constant** model in terms of the (theoretical) risk for the loss above and what is the risk?

Note: in Question (c), we allow f outputs different values c for different $\mathbf{x}$. In Question (d), we aim to search an optimal $\bar{f}(\mathbf{x}) = c$ for all $\mathbf{x}$.

Question (d)

(d): What is the optimal **constant** model in terms of the (theoretical) risk for the loss above and what is the risk?

Note: in Question (c), we allow f outputs different values c for different $\mathbf{x}$. In Question (d), we aim to search an optimal $\bar{f}(\mathbf{x}) = c$ for all $\mathbf{x}$.

Solution to Question (d)

The (theoretical) risk for a constant model $\bar{f}(\mathbf{x}) = c$ is:

$$
\begin{aligned}
\mathcal{R}_L(\bar{f}) &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{y|\mathbf{x}}\left[(m(y) - m(\bar{f}(\mathbf{x})))^2\right]\right] \\
&= \int_y \int_{\mathbf{x}} (m(y) - m(\bar{f}(\mathbf{x})))^2 p(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y \\
&= \int_y \int_{\mathbf{x}} (m(y) - m(c))^2 p(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y \\
&= \int_y (m(y) - m(c))^2 p(y)\mathrm{d}y \quad \triangleright \\
&= \mathbb{E}_y\left[(m(y) - m(c))^2\right]
\end{aligned}
$$

Therefore, the optimal constant model is

$$
\bar{f}(\mathbf{x}) = c = m^{-1}(\mathbb{E}_y[m(y)])
$$

Solution to Question (d): Continued

The risk given $\bar{f}(\mathbf{x}) = c = m^{-1}(\mathbb{E}_y[m(y)])$ is:

$$\mathcal{R}_L(\bar{f}) = \mathbb{E}_{xy}[(m(y) - m(\bar{f}(\mathbf{x}))^2] = \mathbb{E}_y[(m(y) - \mathbb{E}_y[m(y)])^2] = \mathrm{Var}(m(y))$$

Question (e)

(e): Recall the decomposition of the Bayes regret into the estimation and the approximation error. Show that the former is $\frac{1}{n}\mathrm{Var}(m(y))$, while the latter is $\mathrm{Var}\left(\mathbb{E}_{y|\mathbf{x}}[m(y) \mid \mathbf{x}]\right)$ for the optimal constant model $\hat{f}(\mathbf{x})$ if the hypothesis space $\mathcal{H}$ consists of the constant models.

Solution to Question (e)

- Recall from Question (a) that $\hat{f}(\mathbf{x}) = m^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)$ and $\mathcal{R}_L(\hat{f}) = \left(1 + \frac{1}{n}\right)\mathrm{Var}(m(y))$.
- Recall from Question (d) that $\bar{f}(\mathbf{x}) = \arg\min_{f\in\mathcal{H}} \mathcal{R}_L(f)$ and $\mathcal{R}_L(\bar{f}) = \mathrm{Var}(m(y))$.
- Recall from (c) that the point-wise risk minimizer $f^*(\mathbf{x}) = m^{-1}(\mathbb{E}_{y|\mathbf{x}}[m(y)\,|\mathbf{x}]) = \arg\min_f R_L(f)$ for an **unstricted** function space.
- Remember to differentiate between these risk minimizers.
- The Bayes regret can be decomposed as:

$$\mathcal{R}_L(\hat{f}) - \mathcal{R}_L^* = \underbrace{\left[\mathcal{R}_L(\hat{f}) - \inf_{f\in\mathcal{H}} \mathcal{R}_L(f)\right]}_{\text{estimation error}} + \underbrace{\left[\inf_{f\in\mathcal{H}} \mathcal{R}_L(f) - \mathcal{R}_L^*\right]}_{\text{approximation error}}$$

Solution to Question (e): Continued

The estimation error is:

$$\begin{aligned}
\mathcal{R}_L(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{R}_L(f) &= \mathcal{R}_L(\hat{f}) - \mathcal{R}_L(\bar{f}) \\
&= \left(1 + \frac{1}{n}\right) \mathrm{Var}(m(y)) - \mathrm{Var}(m(y)) \\
&= \frac{1}{n} \mathrm{Var}(m(y)).
\end{aligned}$$

Solution to Question (e): Continued

The approximation error is:

$$
\begin{aligned}
\inf_{f \in \mathcal{H}} \mathcal{R}_L(f) - \mathcal{R}_L^* &= \mathcal{R}_L(\bar{f}) - \mathcal{R}_L(f^*) \\
&= \mathrm{Var}(m(y)) - \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{y|\mathbf{x}}[(m(y) - m(f^*(\mathbf{x}))^2 \mid \mathbf{x}]\right] \quad \text{(plug in (d))} \\
&= \mathrm{Var}(m(y)) - \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{y|\mathbf{x}}[(m(y) - m(m^{-1}(\mathbb{E}_{y|\mathbf{x}}[m(y) \mid \mathbf{x}])))^2 \mid \mathbf{x}]\right] \quad \text{(plug in (c))} \\
&= \mathrm{Var}(m(y)) - \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{y|\mathbf{x}}[(m(y) - \mathbb{E}_{y|\mathbf{x}}[m(y) \mid \mathbf{x}])^2 \mid \mathbf{x}]\right] \\
&= \mathrm{Var}(m(y)) - \mathbb{E}_{\mathbf{x}}\left[\mathrm{Var}(m(y) \mid \mathbf{x})\right] \\
&= \mathrm{Var}(\mathbb{E}_{y|\mathbf{x}}[m(y) \mid \mathbf{x}])
\end{aligned}
$$

The last step holds because of the law of total variation:

$$\mathrm{Var}(Y) = \mathbb{E}_X[\mathrm{Var}(Y \mid X)] + \mathrm{Var}[\mathbb{E}_{Y \mid X}[Y \mid X]]$$