Exercise of Supervised Learning: Multiclass Classification

Yawei Li

yawei.li@stat.uni-muenchen.de

November 17, 2023

Exercise 4: Multiclass Hinge Loss

Consider the multiclass classification scenario consisting of a feature space $\mathcal{X}$ and a label space $\mathcal{Y} = \{1, \dots, g\}$ with $g \geq 2$ classes. Moreover, we consider the hypothesis space of models based on g discriminant/scoring functions:

$$\mathcal{H} = \{f = (f_1, \dots, f_g)^T : \mathcal{X} \to \mathbb{R}^g | f_k : \mathcal{X} \to \mathbb{R}, \ \forall k \in \mathcal{Y}\}.$$

A model f in $\mathcal{H}$ is used to make a prediction by means of transforming the scores into classes by choosing the class with the maximum score:

$$h(\mathbf{x}) = \arg\max_{k \in \{1, \dots, g\}} f_k(\mathbf{x}). \tag{1}$$

The multiclass hinge loss is defined by

$$L(y, f(\mathbf{x})) = \max_k (f_k(\mathbf{x}) - f_y(\mathbf{x}) + \boldsymbol{I}_{y \neq k}). \qquad \triangleright$$

(a) Show that 0-1-loss for a predictor h as in (1) based on a model $f \in \mathcal{H}$ is at most the multiclass hinge loss for f i.e.,

$$L_{0-1}(y, h(\mathbf{x})) = \boldsymbol{I}_{y \neq h(\mathbf{x})} \leq L(y, f(\mathbf{x})).$$

Solution to Question (a)

There are two cases: $y = \arg\max_k f_k(\mathbf{x})$ or $y \neq \arg\max_k f_k(\mathbf{x})$. If $y = \arg\max_k f_k(\mathbf{x})$, then

$$
\begin{aligned}
L(y, f(\mathbf{x})) &= \max_k (f_k(\mathbf{x}) - f_y(\mathbf{x}) + \boldsymbol{I}_{y \neq k}) \\
&= f_y(\mathbf{x}) - f_y(\mathbf{x}) + 0 \\
&= 0
\end{aligned}
$$

In this case, the 0-1-loss is

$$L_{0-1}(y, h(\mathbf{x})) = 0$$

because $y = \arg\max_k f_k(x)$. That is, the prediction via argmax is correct.
So $L(y, f(\mathbf{x})) = L_{0,1}(y, h(\mathbf{x}))$.

Solution to Question (a)

There are two cases: $y = \arg\max_k f_k(\mathbf{x})$ or $y \neq \arg\max_k f_k(\mathbf{x})$. If $y = \arg\max_k f_k(\mathbf{x})$, then

$$\begin{aligned} L(y, f(\mathbf{x})) &= \max_k (f_k(\mathbf{x}) - f_y(\mathbf{x}) + \boldsymbol{I}_{y \neq k}) \\ &= f_y(\mathbf{x}) - f_y(\mathbf{x}) + 0 \\ &= 0 \end{aligned}$$

In this case, the 0-1-loss is

$$L_{0-1}(y, h(\mathbf{x})) = 0$$

because $y = \arg\max_k f_k(x)$. That is, the prediction via argmax is correct.
So $L(y, f(\mathbf{x})) = L_{0,1}(y, h(\mathbf{x}))$.

Solution to Question (a)

There are two cases: $y = \arg\max_k f_k(\mathbf{x})$ or $y \neq \arg\max_k f_k(\mathbf{x})$. If $y = \arg\max_k f_k(\mathbf{x})$, then

$$
\begin{aligned}
L(y, f(\mathbf{x})) &= \max_k (f_k(\mathbf{x}) - f_y(\mathbf{x}) + \boldsymbol{I}_{y \neq k}) \\
&= f_y(\mathbf{x}) - f_y(\mathbf{x}) + 0 \\
&= 0
\end{aligned}
$$

In this case, the 0-1-loss is

$$L_{0-1}(y, h(\mathbf{x})) = 0$$

because $y = \arg\max_k f_k(x)$. That is, the prediction via argmax is correct.
So $L(y, f(\mathbf{x})) = L_{0,1}(y, h(\mathbf{x}))$.

Solution to Question (a): Continued

If $y \neq \arg\max_k f_k(\mathbf{x})$, then

$$L(y, f(\mathbf{x})) = \max_k(\underbrace{f_k(\mathbf{x}) - f_y(\mathbf{x})}_{>0} + \underbrace{\boldsymbol{I}_{y \neq k}}_{=1})$$
$$> 1.$$

The 0-1-loss is

$$L_{0-1}(y, h(\mathbf{x})) = 1$$

because $y \neq \arg\max_k f_k(x)$. That is, the prediction via argmax is incorrect. So $L(y, f(\mathbf{x})) > L_{0,1}(y, h(\mathbf{x}))$.

Combining the two cases, we have proved that

$$L_{0-1}(y, h(\mathbf{x})) = \boldsymbol{I}_{y \neq h(\mathbf{x})} \leq L(y, f(\mathbf{x})).$$

Solution to Question (a): Continued

If $y \neq \arg\max_k f_k(\mathbf{x})$, then

$$L(y, f(\mathbf{x})) = \max_k(\underbrace{f_k(\mathbf{x}) - f_y(\mathbf{x})}_{>0} + \underbrace{\boldsymbol{I}_{y\neq k}}_{=1})$$
$$> 1.$$

The 0-1-loss is

$$L_{0-1}(y, h(\mathbf{x})) = 1$$

because $y \neq \arg\max_k f_k(x)$. That is, the prediction via argmax is incorrect. So $L(y, f(\mathbf{x})) > L_{0,1}(y, h(\mathbf{x}))$.

Combining the two cases, we have proved that

$$L_{0-1}(y, h(\mathbf{x})) = \boldsymbol{I}_{y\neq h(\mathbf{x})} \leq L(y, f(\mathbf{x})).$$

Solution to Question (a): Continued

If $y \neq \arg\max_k f_k(\mathbf{x})$, then

$$L(y, f(\mathbf{x})) = \max_k(\underbrace{f_k(\mathbf{x}) - f_y(\mathbf{x})}_{>0} + \underbrace{\boldsymbol{I}_{y \neq k}}_{=1})$$
$$> 1.$$

The 0-1-loss is

$$L_{0-1}(y, h(\mathbf{x})) = 1$$

because $y \neq \arg\max_k f_k(x)$. That is, the prediction via argmax is incorrect. So $L(y, f(\mathbf{x})) > L_{0,1}(y, h(\mathbf{x}))$.

Combining the two cases, we have proved that

$$L_{0-1}(y, h(\mathbf{x})) = \boldsymbol{I}_{y \neq h(\mathbf{x})} \leq L(y, f(\mathbf{x})).$$

Question (b)

(b) Verify that the multiclass hinge loss of $f \in \mathcal{H}$ on a data point $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ is bounded from above by $\sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\}$. *Hint*: Note that this upper bound is

sometimes referred to as the multiclass hinge loss.

Solution to Question (b)

Case 1: $y = \arg\max_k f_k(\mathbf{x})$, then the hinge loss:

$$L(y, f(\mathbf{x})) = \max_k (f_k(\mathbf{x}) - f_y(\mathbf{x}) + \boldsymbol{I}_{y \neq k}) = 0$$

and

$$\sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} \geq \sum_{k \neq y} 0 = 0$$

So the $\sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\}$ is a upper bound in this case.

Solution to Question (b): Continued

Case 2: $y \neq \arg\max_k f_k(\mathbf{x})$, then the hinge loss:

$$\begin{aligned} L(y, f(\mathbf{x})) &= \max_k (f_k(\mathbf{x}) - f_y(\mathbf{x}) + \boldsymbol{I}_{y \neq k}) \\ &= \max_{k \neq y} (f_k(\mathbf{x}) - f_y(\mathbf{x}) + 1) \end{aligned}$$

Let's say $k^* = \arg\max_k f_k(\mathbf{x}) \neq y$, then it follows that

$$\begin{aligned} L(y, f(\mathbf{x})) &= \underbrace{f_{k^*}(\mathbf{x}) - f_y(\mathbf{x}) + 1}_{>1} \\ &= \max\{0, f_{k^*}(\mathbf{x}) - f_y(\mathbf{x}) + 1\} \\ &\leq \max\{0, f_{k^*}(\mathbf{x}) - f_y(\mathbf{x}) + 1\} + \underbrace{\max\{0, f_1(\mathbf{x}) - f_y(\mathbf{x}) + 1\} + \ldots + \max\{0, f_g(\mathbf{x}) - f_y(\mathbf{x}) + 1\}}_{\forall j \in \mathcal{Y} \text{ and } j \neq y \text{ and } j \neq k^*} \\ &= \sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} \end{aligned}$$

So, combining the two cases, $\sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\}$ is an upper bound for $L(y, f(\mathbf{x}))$.

Solution to Question (b): Continued

Case 2: $y \neq \arg\max_k f_k(\mathbf{x})$, then the hinge loss:

$$\begin{aligned} L(y, f(\mathbf{x})) &= \max_k (f_k(\mathbf{x}) - f_y(\mathbf{x}) + \boldsymbol{I}_{y \neq k}) \\ &= \max_{k \neq y} (f_k(\mathbf{x}) - f_y(\mathbf{x}) + 1) \end{aligned}$$

Let's say $k^* = \arg\max_k f_k(\mathbf{x}) \neq y$, then it follows that

$$\begin{aligned} L(y, f(\mathbf{x})) &= \underbrace{f_{k^*}(\mathbf{x}) - f_y(\mathbf{x}) + 1}_{>1} \\ &= \max\{0, f_{k^*}(\mathbf{x}) - f_y(\mathbf{x}) + 1\} \\ &\leq \max\{0, f_{k^*}(\mathbf{x}) - f_y(\mathbf{x}) + 1\} + \underbrace{\max\{0, f_1(\mathbf{x}) - f_y(\mathbf{x}) + 1\} + \ldots + \max\{0, f_g(\mathbf{x}) - f_y(\mathbf{x}) + 1\}}_{\forall j \in \mathcal{Y} \text{ and } j \neq y \text{ and } j \neq k^*} \\ &= \sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} \end{aligned}$$

So, combining the two cases, $\sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\}$ is an upper bound for $L(y, f(\mathbf{x}))$.

Question (c)

In the case of binary classification, i.e., $g = 2$ and $\mathcal{Y} = \{-1, +1\}$, we use a single discriminant model $f(\mathbf{x}) = f_1(\mathbf{x}) - f_{-1}(\mathbf{x})$ based on two scoring functions: $f_1, f_{-1} : \mathcal{X} \to \mathbb{R}$ for the prediction by means of $h(\mathbf{x}) = \mathrm{sgn}(f(\mathbf{x}))$. Here, f_1 is the score for the positive class and f_{-1} is the score for the negative class. Show that the upper bound in (b) coincide with the binary hinge loss $L(y, f(\mathbf{x})) = \max\{0, 1 - yf(\mathbf{x})\}$.

Solution to Question (c)

Case 1: $y = +1$. In this case,

$$
\begin{aligned}
\sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} &= \sum_{k \neq +1} \max\{0, f_k(\mathbf{x}) - f_1(\mathbf{x}) + 1\} \\
&= \max\{0, \underbrace{f_{-1}(\mathbf{x}) - f_1(\mathbf{x})}_{:= -f(\mathbf{x})} + 1\} \\
&= \max\{0, 1 - f(\mathbf{x})\} \\
&= \max\{0, 1 - y \cdot f(\mathbf{x})\}
\end{aligned}
$$

So the equation holds in this case.

Solution to Question (c): Continued

Case 2: $y = -1$. In this case,

$$\begin{aligned}
\sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} &= \sum_{k \neq -1} \max\{0, f_k(\mathbf{x}) - f_{-1}(\mathbf{x}) + 1\} \\
&= \max\{0, f_1(\mathbf{x}) - f_{-1}(\mathbf{x}) + 1\} \\
&= \max\{0, 1 + f(\mathbf{x})\} \\
&= \max\{0, 1 - y \cdot f(\mathbf{x})\}
\end{aligned}$$

Therefore, it is proven that the equation holds in two cases.

Question (d)

Recall the statement of the lecture regarding the binary hinge loss:

“... the hinge loss only equals zero for a margin ≥ 1 *encouraging confident (correct) predictions.”*

Can we say something similar for the altenative multiclass hinge loss in (b)?
Hint: multiclass hinge loss: $\sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\}$.

Solution to Question (d)

Yes, we can say that *it is only zero if all the* $g-1$ *margins are greater than 1.*

- Margins: $m_{y,k}(\mathbf{x}) = f_y(\mathbf{x}) - f_k(\mathbf{x})$, where $k \in \mathcal{Y} \setminus \{y\}$.
- Mathematically: $m_{y,k}(\mathbf{x}) \geq 1 \; \forall k \neq y \Leftrightarrow \sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} = 0$.

Proof:

$$
\begin{aligned}
m_{y,k}(\mathbf{x}) \geq 1 \; \forall k \neq y &\Rightarrow f_y(\mathbf{x}) - f_k(\mathbf{x}) \geq 1 \; \forall k \neq y \\
&\Rightarrow f_k(\mathbf{x}) - f_y(\mathbf{x}) \geq -1 \; \forall k \neq y \\
&\Rightarrow \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} = 0 \; \forall k \neq y \\
&\Rightarrow \sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} = 0
\end{aligned}
$$

Solution to Question (d)

Yes, we can say that *it is only zero if all the $g-1$ margins are greater than 1.*

- Margins: $m_{y,k}(\mathbf{x}) = f_y(\mathbf{x}) - f_k(\mathbf{x})$, where $k \in \mathcal{Y} \setminus \{y\}$.
- Mathematically: $m_{y,k}(\mathbf{x}) \geq 1 \ \forall k \neq y \Leftrightarrow \sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} = 0$.

Proof:

$$\begin{aligned} m_{y,k}(\mathbf{x}) \geq 1 \ \forall k \neq y &\Rightarrow f_y(\mathbf{x}) - f_k(\mathbf{x}) \geq 1 \ \forall k \neq y \\ &\Rightarrow f_k(\mathbf{x}) - f_y(\mathbf{x}) \geq -1 \ \forall k \neq y \\ &\Rightarrow \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} = 0 \ \forall k \neq y \\ &\Rightarrow \sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} = 0 \end{aligned}$$

Solution to Question (d)

Yes, we can say that *it is only zero if all the* $g-1$ *margins are greater than 1.*

- Margins: $m_{y,k}(\mathbf{x}) = f_y(\mathbf{x}) - f_k(\mathbf{x})$, where $k \in \mathcal{Y} \setminus \{y\}$.
- Mathematically: $m_{y,k}(\mathbf{x}) \geq 1 \; \forall k \neq y \Leftrightarrow \sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} = 0$.

Proof:

$$
\begin{aligned}
m_{y,k}(\mathbf{x}) \geq 1 \; \forall k \neq y &\Rightarrow f_y(\mathbf{x}) - f_k(\mathbf{x}) \geq 1 \; \forall k \neq y \\
&\Rightarrow f_k(\mathbf{x}) - f_y(\mathbf{x}) \geq -1 \; \forall k \neq y \\
&\Rightarrow \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} = 0 \; \forall k \neq y \\
&\Rightarrow \sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} = 0
\end{aligned}
$$

Solution to Question (d)

Yes, we can say that *it is only zero if all the $g-1$ margins are greater than 1.*

- Margins: $m_{y,k}(\mathbf{x}) = f_y(\mathbf{x}) - f_k(\mathbf{x})$, where $k \in \mathcal{Y} \setminus \{y\}$.
- Mathematically: $m_{y,k}(\mathbf{x}) \geq 1 \; \forall k \neq y \Leftrightarrow \sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} = 0$.

Proof:

$$
\begin{aligned}
m_{y,k}(\mathbf{x}) \geq 1 \; \forall k \neq y &\Rightarrow f_y(\mathbf{x}) - f_k(\mathbf{x}) \geq 1 \; \forall k \neq y \\
&\Rightarrow f_k(\mathbf{x}) - f_y(\mathbf{x}) \geq -1 \; \forall k \neq y \\
&\Rightarrow \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} = 0 \; \forall k \neq y \\
&\Rightarrow \sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\} = 0
\end{aligned}
$$

Question (e) and Solution to (e)

Show the standard solution.