COMP7015 Artificial Intelligence **Semester 1, 2022/23**

Derivation of Logistic Loss and its Gradient in Lecture 5

Instructor: Dr. Kejing Yin Oct. 7, 2022

1 Derivation of Logistic Loss from the Likelihood

Given:

- $p(y=+1|x) = \sigma(\mathbf{w}^\top \mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^\top \mathbf{x})}$,
- $p(y=-1|x) = 1-\sigma(\mathbf{w}^\top \mathbf{x}) = \frac{\exp(-\mathbf{w}^\top \mathbf{x})}{1+\exp(-\mathbf{w}^\top \mathbf{x})} = \frac{1}{1+\exp(+\mathbf{w}^\top \mathbf{x})}$, and
- the *log likelihood*:

$$\ell(\mathbf{w}) = \sum_{(\mathbf{x}_i, y_i)\in\mathcal{D}} \mathbb{1}[y_i=+1]\log p(y_i=+1|\mathbf{x}) + \mathbb{1}[y_i=-1]\log p(y_i=-1|\mathbf{x})$$

We would like to derive the following *logistic loss function* (the negative likelihood):

$$\text{Loss}(\mathbf{x}, y, \mathbf{w}) = \log(1+\exp(-(\mathbf{w}^\top \mathbf{x})y))$$

We can derive this as follows:

The label can only be positive $(+1)$ or negative (-1). For a sample with positive label, the term in red will be zero, and the term in blue will become:

$$\begin{aligned}
\ell^{+}(\mathbf{x}, y, \mathbf{w}) &= \underbrace{\mathbb{1}[y_i=+1]}_{1}\log p(y_i=+1|\mathbf{x}) \\
&= -\log\left(1+\exp(-\mathbf{w}^\top \mathbf{x})\right) \\
&= -\log\left(1+\exp(-(\mathbf{w}^\top \mathbf{x})y)\right) \quad (\textit{since } y=+1)
\end{aligned}$$

Similarly, for samples with negative label, the term in blue will be zero and the term in red will become:

$$\begin{aligned}
\ell^{-}(\mathbf{x}, y, \mathbf{w}) &= \underbrace{\mathbb{1}[y_i=-1]}_{1}\log p(y_i=-1|\mathbf{x}) \\
&= -\log\left(1+\exp(+\mathbf{w}^\top \mathbf{x})\right) \\
&= -\log\left(1+\exp(-(\mathbf{w}^\top \mathbf{x})y)\right) \quad (\textit{since } y=-1)
\end{aligned}$$

By multiplying y in the two cases, we can have the same expression in either case. So we can just let the loss to be negative of it:

$$\text{Loss}(\mathbf{x}, y, \mathbf{w}) = \log(1+\exp(-(\mathbf{w}^\top \mathbf{x})y))$$

Remarks: we derive this loss function using $+1$ and -1 for denoting positive and negative labels, respectivley. You can also use 1 and 0 for positive and negative labels, which will give you another expression. Both are correct and the only difference is using $y=-1$ or $y=0$ for negative labels.

2 Derivation of the Gradient of the Logistic Loss

The logistic loss function is given by:

$$\mathrm{TrainLoss}_{\text{logistic}}(\mathbf{w}) = \sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{train}}} \log(1+\exp(-(\mathbf{w}^\top\mathbf{x})y))$$

Applying the chian rule, we can compute its gradient as follows:

$$\begin{aligned}
\nabla_{\mathbf{w}}\,\mathrm{TrainLoss}(\mathbf{w}) &= \sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{train}}} \frac{\partial}{\partial\mathbf{w}} \log(1+\exp(-(\mathbf{w}^\top\mathbf{x})y)) \\
&= \sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{train}}} \frac{1}{1+\exp(-(\mathbf{w}^\top\mathbf{x})y)} \frac{\partial}{\partial\mathbf{w}} \exp(-(\mathbf{w}^\top\mathbf{x})y) \\
&= \sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{train}}} \frac{\exp(-(\mathbf{w}^\top\mathbf{x})y)}{1+\exp(-(\mathbf{w}^\top\mathbf{x})y)} \frac{\partial}{\partial\mathbf{w}} (-(\mathbf{w}^\top\mathbf{x})y) \\
&= -\sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{train}}} y\frac{\exp(-(\mathbf{w}^\top\mathbf{x})y)}{1+\exp(-(\mathbf{w}^\top\mathbf{x})y)}\mathbf{x}
\end{aligned}$$