Discussion 4
Logistic Regression and Coordinate Descent
Machine Learning, Spring 2019

1 Interpretation of logistic regression

a

$$\begin{aligned}
\log \frac{p(+1 \mid \mathbf{x};\, \boldsymbol{\theta})}{p(-1 \mid \mathbf{x};\, \boldsymbol{\theta})} &= \log \frac{\sigma(\theta^\top x)}{\sigma(-\theta^\top x)} \\
&= \log \frac{1 + e^{\theta^\top x}}{1 + e^{-\theta^\top x}} \\
&= \log \frac{e^{.5\theta^\top x}(e^{-.5\theta^\top x} + e^{.5\theta^\top x})}{e^{-.5\theta^\top x}(e^{.5\theta^\top x} + e^{-.5\theta^\top x})} \\
&= .5\theta^\top x - (.5\theta^\top x) \\
&= \theta^\top x.
\end{aligned}$$

b

θ_i is the additive change in log odds under a unit marginal increase in the i-th component of x.

2 On the loss of logistic function

Both are correct depending on whether $y \in \{0, 1\}$ or $\{-1, +1\}$.
If $y \in \{0, 1\}$,

$$\mathbf{P}(y_i = 1 | \mathbf{w}, \mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}_i}}$$

and

$$\begin{aligned}
\mathbf{P}(y_i = 0 | \mathbf{w}, \mathbf{x}_i) &= 1 - \mathbf{P}(y_i = 1 | \mathbf{w}, \mathbf{x}_i) \\
&= 1 - \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}_i}}.
\end{aligned}$$

The cross-entropy loss (negated *log-likelihood*) is

$$
\begin{aligned}
l(\mathbf{w}) &= -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\left(\frac{1}{1+e^{-\mathbf{w}^\top\mathbf{x}_i}}\right) + (1-y_i)\log\left(1-\frac{1}{1+e^{-\mathbf{w}^\top\mathbf{x}_i}}\right)\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\left(\frac{e^{\mathbf{w}^\top\mathbf{x}_i}}{1+e^{\mathbf{w}^\top\mathbf{x}_i}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{\mathbf{w}^\top\mathbf{x}_i}}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\left(1+e^{\mathbf{w}^\top\mathbf{x}_i}\right) + (1-y_i)\log\left(1+e^{\mathbf{w}^\top\mathbf{x}_i}\right) - y_i\log\left(e^{\mathbf{w}^\top\mathbf{x}_i}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[-y_i\mathbf{w}^\top\mathbf{x}_i + \log\left(1+e^{\mathbf{w}^\top\mathbf{x}_i}\right)\right]
\end{aligned}
$$

On the other hand, if $y_i \in \{-1,+1\}$, again

$$
\mathbf{P}(y_i = 1|\mathbf{w},\mathbf{x}_i) = \frac{1}{1+e^{-\mathbf{w}^\top\mathbf{x}_i}}
$$

and

$$
\begin{aligned}
\mathbf{P}(y_i = 0|\mathbf{w},\mathbf{x}_i) &= 1-\mathbf{P}(y_i = 1|\mathbf{w},\mathbf{x}_i) \\
&= 1-\frac{1}{1+e^{-\mathbf{w}^\top\mathbf{x}_i}}.
\end{aligned}
$$

The cross-entropy loss (negated *log-likelihood*) is

$$
\begin{aligned}
l(\mathbf{w}) &= -\sum_{i=1}^{n}\left[\frac{\mathbb{1}[y_i=+1]}{n}\log\left(\frac{1}{1+e^{-\mathbf{w}^\top\mathbf{x}_i}}\right) + \frac{\mathbb{1}[y_i=-1]}{n}\log\left(1-\frac{1}{1+e^{-\mathbf{w}^\top\mathbf{x}_i}}\right)\right] \\
&= -\frac{1}{n}\sum_{y_i=+1}\log\left(\frac{1}{1+e^{-\mathbf{w}^\top\mathbf{x}_i}}\right) - \frac{1}{n}\sum_{y_i=-1}\log\left(\frac{1}{1+e^{\mathbf{w}^\top\mathbf{x}_i}}\right) \\
&= \frac{1}{n}\sum_{y_i=+1}\log\left(1+e^{-\mathbf{w}^\top\mathbf{x}_i}\right) + \frac{1}{n}\sum_{y_i=-1}\log\left(1+e^{\mathbf{w}^\top\mathbf{x}_i}\right) \\
&= \frac{1}{n}\sum_{y_i=+1}\log\left(1+e^{-y_i\mathbf{w}^\top\mathbf{x}_i}\right) + \frac{1}{n}\sum_{y_i=-1}\log\left(1+e^{-y_i\mathbf{w}^\top\mathbf{x}_i}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\log\left(1+e^{-y_i\mathbf{w}^\top\mathbf{x}_i}\right)
\end{aligned}
$$

3 ℓ_2-regularization and Coordinate Descent

1

Based on the Lemma a function f is convex if and only if $\nabla f \succeq 0$. Given that Hessian of $\frac{\lambda}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}$ is positive semi definite and a sum of two convex functions is convex we have that regularized loss is convex.

2

Define $\theta_k = k\theta_*$. Observe that

$$\begin{aligned}
\mathcal{L}(\theta_k) &= \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{-y_i k\theta_*^\top x_i}\right)\\
&\leq \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{-k\delta}\right)\\
&= \log(1+e^{-k\delta}).
\end{aligned}$$

But

$$\lim_{k\to\infty}\log(1+e^{-k\delta}) = \log(1+0) = 0.$$

If we are not happy about commuting this limit to infinity with the continuous function, we can use the fact that $\log(1+x)\leq x$. Hence, we have

$$0 \leq \lim_{k\to\infty}\mathcal{L}(\theta_k) \leq \lim_{k\to\infty}\log(1+e^{-k\delta}) = 0.$$

3

The objective is minimized, however the learned parameters $\boldsymbol{\theta}^*$ can take very large values *e.g.,* ∞. This is bad, because $\boldsymbol{\theta}^*$ is sensitive to input $\boldsymbol{x}$, and can easily make overfit decisions on the testing dataset.

Therefore benefit of ℓ_2 regularization is that it forces the learned parameters $\boldsymbol{\theta}^*$ be close to $\mathbf{0}$.

4

If we do not use Taylor expansion, we have the linear search for coordinate descent as

$$\frac{\partial \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\boldsymbol{\theta}+\Delta\theta_j\mathbf{e}_j)+\lambda/2\|\boldsymbol{\theta}+\Delta\theta_j\mathbf{e}_j\|^2}{\partial\Delta\theta_j} = -\frac{1}{N}\sum_{i=1}^{N}\frac{e^{-y_i(\boldsymbol{\theta}+\Delta\theta_j\mathbf{e}_j)^\top\mathbf{x}_i}}{1+e^{-y_i(\boldsymbol{\theta}+\Delta\theta_j\mathbf{e}_j)^\top\mathbf{x}_i}}y_i x_{ij}+\lambda(\theta_j+\Delta\theta_j)=0$$

Thus, no closed form for $\Delta\theta_j$.

5

If we use Taylor approximation,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}+\Delta\boldsymbol{\theta}) \approx& \mathcal{L}(\boldsymbol{\theta})+\nabla\mathcal{L}(\boldsymbol{\theta})^\top\Delta\boldsymbol{\theta}+\frac{1}{2}\Delta\boldsymbol{\theta}^\top\nabla^2\mathcal{L}(\boldsymbol{\theta})\Delta\boldsymbol{\theta}\\
=&\frac{1}{N}\sum_{y_i=1}\left(\frac{1}{2}p(\mathbf{x}_i)(1-p(\mathbf{x}_i))\Delta\boldsymbol{\theta}^\top\mathbf{x}_i\mathbf{x}_i^\top\Delta\boldsymbol{\theta}-(1-p(\mathbf{x}_i))\mathbf{x}_i^\top\Delta\boldsymbol{\theta}\right)\\
&+\frac{1}{N}\sum_{y_i=-1}\left(\frac{1}{2}p(\mathbf{x}_i)(1-p(\mathbf{x}_i))\Delta\boldsymbol{\theta}^\top\mathbf{x}_i\mathbf{x}_i^\top\Delta\boldsymbol{\theta}-p(\mathbf{x}_i)\mathbf{x}_i^\top\Delta\boldsymbol{\theta}\right)\\
&+C(\boldsymbol{\theta})
\end{aligned}$$

where $C(\boldsymbol{\theta})$ is some constant independent of $\Delta\boldsymbol{\theta}$ which can be disregarded safely, and define

$$\begin{aligned}
w_i =& p(\mathbf{x}_i)(1-p(\mathbf{x}_i))\\
z_i =& \frac{(y_i+1)/2-p(\mathbf{x}_i)}{w_i}
\end{aligned}$$

We have

$$\mathcal{L}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) \approx \frac{1}{2N} \sum_i w_i \left(z_i - \Delta\boldsymbol{\theta}^\top \mathbf{x}_i\right)^2 + \tilde{C}(\boldsymbol{\theta})$$

Thus, we are supposed to optimize $\arg\min_{\Delta\boldsymbol{\theta}} \frac{1}{2N} \sum_i w_i \left(z_i - \Delta\boldsymbol{\theta}^\top \mathbf{x}_i\right)^2 + \frac{\lambda}{2}\|\boldsymbol{\theta} + \Delta\boldsymbol{\theta}\|^2$ by coordinate descent, i.e., setting $\Delta\boldsymbol{\theta} = (0, ..., \Delta\theta_j, ..., 0)^\top$. Thus, the optimization is reduced to

$$\arg\min_{\Delta\theta_j} \frac{1}{2N} \sum_i w_i \left(z_i - \Delta\theta_j \cdot x_{ij}\right)^2 + \frac{\lambda}{2}(\theta_j + \Delta\theta_j)^2$$

It is easy to compute the **closed-form** of the optimal $\Delta\theta_j$ by taking derivative on $\Delta\theta_j$ and setting the derivative as 0.

$$-\frac{1}{N} \sum_i w_i \left(z_i - \Delta\theta_j \cdot x_{ij}\right) x_{ij} + \lambda(\theta_j + \Delta\theta_j) = 0$$

Therefore we get

$$\Delta\theta_j^* = \frac{-\lambda\theta_j + \frac{1}{N}\sum_i w_i z_i x_{ij}}{\lambda + \frac{1}{N}\sum_i w_i x_{ij}^2}$$