Statistics 221 Final Project: C++ team

December 5, 2013

Here we derive the SGD and implicit updates for three commonly-used loss functions. The optimization problem is:

$$\min_{\theta} \left(\sum_t L(y_t, \hat{y}_t) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2 \right) \tag{1}$$

where $\hat{y}_t = \text{sign}(\boldsymbol{x_t^T} \boldsymbol{\theta_t})$

1 Log-loss

The log-loss function is given by:

$$L(y, \hat{y}) = \log(1 + \exp(-y\hat{y})) \tag{2}$$

1.1 SGD update

The SGD update is given by:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \alpha_t \nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) \tag{3}$$

where

$$Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) = \log\left(1 + \exp(-y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t})\right) + \frac{\lambda}{2} \|\boldsymbol{\theta_t}\|^2 \tag{4}$$

The gradient $\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t})$ can be calculated as follows:

$$\begin{aligned}
\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) &= \left(\frac{1}{1 + \exp(-y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t})} \right) \left(\exp(-y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t}) \right) (-y_t \cdot \boldsymbol{x_t}) + \lambda \boldsymbol{\theta_t} \\
&= \frac{-y_t \exp(-y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t})}{1 + \exp(-y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t})} \cdot \boldsymbol{x_t} + \lambda \boldsymbol{\theta_t} \\
&= \frac{-y_t}{\exp(y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t}) + 1} \cdot \boldsymbol{x_t} + \lambda \boldsymbol{\theta_t}
\end{aligned}$$

So, the SGD update for the log-loss function is:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \alpha_t \left(\frac{-y_t}{\exp(y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t}) + 1} \cdot \boldsymbol{x_t} + \lambda \boldsymbol{\theta_t} \right) \tag{5}$$

1.2 Implicit update

The implicit update can be derived in much the same way as above. We have the following:

$$\begin{aligned}\boldsymbol{\theta_{t+1}} &= \boldsymbol{\theta_t} - \alpha_t \nabla Q(\boldsymbol{\theta_{t+1}}, y_t, \boldsymbol{x_t}) \\ &= \boldsymbol{\theta_t} - \alpha_t \left(\frac{-y_t}{\exp(y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_{t+1}}) + 1} \cdot \boldsymbol{x_t} + \lambda \boldsymbol{\theta_{t+1}} \right)\end{aligned}$$

2 Hinge loss

The hinge loss function is given by:

$$L(y, \hat{y}) = \max(0, 1 - y\hat{y}) \tag{6}$$

2.1 SGD update

The SGD update is given by:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \alpha_t \nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) \tag{7}$$

where

$$Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) = \max(0, 1 - y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t}) + \frac{\lambda}{2} \|\boldsymbol{\theta_t}\|^2 \tag{8}$$

The gradient $\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t})$ can be calculated as follows, where we consider two cases depending on the sign of $1 - y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t}$:

1. If $1 - y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t} < 0$, then $Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) = \frac{\lambda}{2} \|\boldsymbol{\theta_t}\|^2$. Then,

$$\begin{aligned}\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) &= \nabla \left(\frac{\lambda}{2} \|\boldsymbol{\theta_t}\|^2 \right) \\ &= \lambda \boldsymbol{\theta_t}\end{aligned}$$

2. If $1 - y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t} \geq 0$, then $Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) = 1 - y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t} + \frac{\lambda}{2} \|\boldsymbol{\theta_t}\|^2$. Then,

$$\begin{aligned}\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) &= \nabla \left(1 - y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t} + \frac{\lambda}{2} \|\boldsymbol{\theta_t}\|^2 \right) \\ &= \nabla \left(-y_t \cdot \boldsymbol{x_t^T} \boldsymbol{\theta_t} \right) + \lambda \boldsymbol{\theta_t} \\ &= -y_t \cdot \boldsymbol{x_t} + \lambda \boldsymbol{\theta_t}\end{aligned}$$

where the last step is accomplished by noting that:

$$\begin{aligned}\nabla_\theta \left(\boldsymbol{x^T\theta}\right) &= \nabla_\theta \left(\sum_{i=1}^{|\boldsymbol{x}|} x_i\theta_i\right) \\ &= \left(\frac{\partial}{\partial\theta_1}\left(\sum_{i=1}^{|\boldsymbol{x}|} x_i\theta_i\right), \frac{\partial}{\partial\theta_2}\left(\sum_{i=1}^{|\boldsymbol{x}|} x_i\theta_i\right), \ldots\right) \\ &= (x_1, x_2, \ldots) \\ &= \boldsymbol{x}\end{aligned}$$

Putting these results together, we have that:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \alpha_t \begin{cases} \lambda\boldsymbol{\theta_t} & y_t \cdot \boldsymbol{x_t^T\theta_t} > 1 \\ \lambda\boldsymbol{\theta_t} - y_t\boldsymbol{x_t} & \text{otherwise} \end{cases} \tag{9}$$

2.2 Implicit update

We now derive the implicit update for the hinge loss, which is given by:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \alpha_t \nabla Q(\boldsymbol{\theta_{t+1}}, y_t, \boldsymbol{x_t}) \tag{10}$$

The calculations are very similar to the SGD derivation above, and for the two cases depending on the sign of $1 - y_t \cdot \boldsymbol{x_t^T\theta_{t+1}}$ are:

1. If $1 - y_t \cdot \boldsymbol{x_t^T\theta_{t+1}} < 0$, then

$$\nabla Q(\boldsymbol{\theta_{t+1}}, y_t, \boldsymbol{x_t}) = \lambda\boldsymbol{\theta_{t+1}}$$

   Then, substituting this result into the implicit update equation above, we can solve to find:

$$\boldsymbol{\theta_{t+1}} = \frac{1}{1+\lambda\alpha_t}\boldsymbol{\theta_t} \tag{11}$$

2. If $1 - y_t \cdot \boldsymbol{x_t^T\theta_{t+1}} \geq 0$, then

$$\nabla Q(\boldsymbol{\theta_{t+1}}, y_t, \boldsymbol{x_t}) = -y_t \cdot \boldsymbol{x_t} + \lambda\boldsymbol{\theta_{t+1}}$$

Then, substituting this result into the implicit update equation above, we can solve to find:

$$\boldsymbol{\theta_{t+1}} = \frac{1}{1+\lambda\alpha_t}\left(\boldsymbol{\theta_t} + \alpha_t y_t \cdot \boldsymbol{x_t}\right) \tag{12}$$

Note that during implementation, we should be careful to make sure that we check the sign of $1-y_t\cdot\boldsymbol{x_t^T\theta_{t+1}}$ after the update, as the derivation of the implicit updates forced an assumption of the sign to begin with. If the assumption was wrong, then the other update function should be used.

3 Squared-loss

The squared-loss function is given by:

$$L(y, \hat{y}) = (y - \hat{y})^2 \tag{13}$$

3.1 SGD update

The SGD update is given by:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \alpha_t \nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) \tag{14}$$

where

$$Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) = \left(y_t - \boldsymbol{x_t^T \theta_t}\right)^2 + \frac{\lambda}{2}\|\boldsymbol{\theta_t}\|^2 \tag{15}$$

The gradient $\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t})$ can be calculated as follows:

$$\begin{aligned}
\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) &= 2\left(y_t - \boldsymbol{x_t^T \theta_t}\right)(-\boldsymbol{x_t}) \\
&= -2(y_t - \boldsymbol{x_t^T \theta_t})\boldsymbol{x_t}
\end{aligned}$$

So, the SGD update for the log-loss function is:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} + 2\alpha_t \left(y_t - \boldsymbol{x_t^T \theta_t}\right) \boldsymbol{x_t} \tag{16}$$

3.2 Implicit update

The implicit update can be derived in much the same way as above. We have the following:

$$\begin{aligned}
\boldsymbol{\theta_{t+1}} &= \boldsymbol{\theta_t} - \alpha_t \nabla Q(\boldsymbol{\theta_{t+1}}, y_t, \boldsymbol{x_t}) \\
&= \boldsymbol{\theta_t} + 2\alpha_t \left(y_t - \boldsymbol{x_t^T \theta_{t+1}}\right) \boldsymbol{x_t}
\end{aligned}$$