CH2 Stochastic gradient descent example

- The featrue $\phi(x)$ is represented as follows:

$$\phi(x) = \begin{bmatrix} \#\text{ of pretty} \\ \#\text{ of good} \\ \#\text{ of bad} \\ \#\text{ of plot} \\ \#\text{ of not} \\ \#\text{ of scenery} \end{bmatrix} = (\#\text{ of pretty}, ..., \#\text{ of scenery})$$

- The gradient of $\text{Loss}_{\text{hinge}}(x, y, \mathbf{w}) = \max\{0, 1 - (\mathbf{w} \cdot \phi(x))y\}$ is

$$\nabla_{\mathbf{w}} \text{Loss}_{\text{hinge}}(x, y, \mathbf{w}) = \begin{cases} -\phi(x)y, & \text{if } (\mathbf{w} \cdot \phi(x))y < 1 \\ \mathbf{0}, & \text{otherwise} \end{cases}$$

- For each of exmaples, (x_i, y_i) where $i = 1, .., 4$, the gradient $\nabla_{\mathbf{w}} \text{Loss}_{\text{hinge}}(x_i, y_i, \mathbf{w})$ when $(\mathbf{w} \cdot \phi(x))y < 1$ is as follows:

$$\nabla_{\mathbf{w}} \text{Loss}_{\text{hinge}}(x_1, y_1, \mathbf{w}) = -(1, 1, 0, 0, 0, 0), \text{where } \phi(x_1) = (1, 1, 0, 0, 0, 0) \text{ and } y_1 = 1$$
$$\nabla_{\mathbf{w}} \text{Loss}_{\text{hinge}}(x_2, y_2, \mathbf{w}) = (0, 0, 1, 1, 0, 0), \text{where } \phi(x_2) = (0, 0, 1, 1, 0, 0) \text{ and } y_2 = -1$$
$$\nabla_{\mathbf{w}} \text{Loss}_{\text{hinge}}(x_3, y_3, \mathbf{w}) = (0, 1, 0, 0, 1, 0), \text{where } \phi(x_3) = (0, 1, 0, 0, 1, 0) \text{ and } y_3 = -1$$
$$\nabla_{\mathbf{w}} \text{Loss}_{\text{hinge}}(x_4, y_4, \mathbf{w}) = -(1, 0, 0, 0, 0, 1), \text{where } \phi(x_4) = (1, 0, 0, 0, 0, 1) \text{ and } y_4 = 1$$

- Steps of SGD
  - Initialize: $\mathbf{w} \leftarrow \mathbf{0}$
  - $i = 1 \Rightarrow (\mathbf{w} \cdot \phi(x_i))y_i = ((0, 0, 0, 0, 0, 0) \cdot (1, 1, 0, 0, 0, 0))(1) = 0 < 1$
    Then, $\mathbf{w} \leftarrow \mathbf{w} - \nabla_{\mathbf{w}} \text{Loss}_{\text{hinge}}(x_i, y_i, \mathbf{w})$
    $\therefore \mathbf{w} = (0, 0, 0, 0, 0, 0) + (1, 1, 0, 0, 0, 0) = (1, 1, 0, 0, 0, 0)$
  - $i = 2 \Rightarrow (\mathbf{w} \cdot \phi(x_i))y_i = ((1, 1, 0, 0, 0, 0) \cdot (0, 0, 1, 1, 0, 0))(-1) = 0 < 1$
    Then, $\mathbf{w} \leftarrow \mathbf{w} - \nabla_{\mathbf{w}} \text{Loss}_{\text{hinge}}(x_i, y_i, \mathbf{w})$
    $\therefore \mathbf{w} = (1, 1, 0, 0, 0, 0) - (0, 0, 1, 1, 0, 0) = (1, 1, -1, -1, 0, 0)$
  - $i = 3 \Rightarrow (\mathbf{w} \cdot \phi(x_i))y_i = ((1, 1, -1, -1, 0, 0) \cdot (0, 1, 0, 0, 1, 0))(-1) = -1 < 1$
    Then, $\mathbf{w} \leftarrow \mathbf{w} - \nabla_{\mathbf{w}} \text{Loss}_{\text{hinge}}(x_i, y_i, \mathbf{w})$
    $\therefore \mathbf{w} = (1, 1, -1, -1, 0, 0) - (0, 1, 0, 0, 1, 0) = (1, 0, -1, -1, -1, 0)$
  - $i = 4 \Rightarrow (\mathbf{w} \cdot \phi(x_i))y_i = ((1, 0, -1, -1, -1, 0) \cdot (1, 0, 0, 0, 0, 1))(1) = 1$
    Weights are not updated becasue of zero gradient.
  - $i = 1 \Rightarrow (\mathbf{w} \cdot \phi(x_i))y_i = ((1, 0, -1, -1, -1, 0) \cdot (1, 1, 0, 0, 0, 0))(1) = 1$
    Weights are not updated becasue of zero gradient.
  - $i = 2 \Rightarrow (\mathbf{w} \cdot \phi(x_i))y_i = ((1, 0, -1, -1, -1, 0) \cdot (0, 0, 1, 1, 0, 0))(-1) = 2$
    Weights are not updated becasue of zero gradient.

- $i = 3 \Rightarrow (\mathbf{w} \cdot \phi(x_i))y_i = ((1, 0, -1, -1, -1, 0) \cdot (0, 1, 0, 0, 1, 0))(-1) = 1$
  Weights are not updated becasue of zero gradient.
- Weights are converged.

- $$\therefore \mathbf{w} = \begin{bmatrix} \text{weight of pretty} \\ \text{weight of good} \\ \text{weight of bad} \\ \text{weight of plot} \\ \text{weight of not} \\ \text{weight of scenery} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ -1 \\ -1 \\ 0 \end{bmatrix}$$