CSCI567TA2

Youzhi Qu USCID: 5045807938

October 2019

1 Problem 1 Convergence of Perceptron Algorithm

1.1

$$\begin{aligned} \boldsymbol{w}_{k+1} &= \boldsymbol{w}_k + y_i \boldsymbol{x}_i \\ \boldsymbol{w}_{k+1}^T &= \boldsymbol{w}_k^T + y_i \boldsymbol{x}_i^T \\ \boldsymbol{w}_{k+1}^T \boldsymbol{w}_{opt} &= \boldsymbol{w}_k^T \boldsymbol{w}_{opt} + y_i \boldsymbol{x}_i^T \boldsymbol{w}_{opt} \end{aligned}$$

because $\boldsymbol{w}_{opt}$ is the optimal hyperplane which linearly separates the classes with maximum margin and data is linearly separable.

$$y_i \boldsymbol{x}_i^T \boldsymbol{w}_{opt} = |\boldsymbol{x}_i^T \boldsymbol{w}_{opt}| \geq \gamma \|\boldsymbol{w}_{opt}\|$$

$$\boldsymbol{w}_{k+1}^T \boldsymbol{w}_{opt} \geq \boldsymbol{w}_k^T \boldsymbol{w}_{opt} + \gamma \|\boldsymbol{w}_{opt}\|$$

1.2

$$\begin{aligned} \|\boldsymbol{w}_{k+1}\|^2 &= \boldsymbol{w}_{k+1}^T \boldsymbol{w}_{k+1} \\ &= (\boldsymbol{w}_k + yi\boldsymbol{x}_i)^T (\boldsymbol{w}_k + yi\boldsymbol{x}_i) \\ &= \|\boldsymbol{w}_k\|^2 + 2y_i \boldsymbol{w}_k^T \boldsymbol{x}_i + \|y_i \boldsymbol{x}_i\|^2 \end{aligned}$$

because $\|y_i \boldsymbol{x}_i\|^2 = 1$ and $y_i \boldsymbol{w}_k^T \boldsymbol{x}_i \leq 0$,

$\|\boldsymbol{w}_{k+1}\|^2 \leq \|\boldsymbol{w}_k\|^2 + 1$

1.3

1.3.1

according to 1.1 result and M is the total number of mistakes,

$\boldsymbol{w}_{k+1}^T \boldsymbol{w}_{opt} \geq \boldsymbol{w}_0^T \boldsymbol{w}_{opt} + M\gamma \|\boldsymbol{w}_{opt}\|$

because $\boldsymbol{w}_{\boldsymbol{0}} = 0$,

$\boldsymbol{w}_{k+1}^T \boldsymbol{w}_{opt} \geq M\gamma \|\boldsymbol{w}_{opt}\|$

according to Cauchy-Schwartz inequality,

$\|\boldsymbol{w}_{k+1}\| \|\boldsymbol{w}_{opt}\| \geq \boldsymbol{w}_{k+1}^T \boldsymbol{w}_{opt} \geq M\gamma \|\boldsymbol{w}_{opt}\|$

then

$\|\boldsymbol{w}_{k+1}\| \geq \gamma M$

1.3.2

according to the result of 1.2 and M is the total number of mistakes,

$\|\boldsymbol{w}_{k+1}\|^2 \leq \|\boldsymbol{w}_0\|^2 + M$

because $\boldsymbol{w_0} = 0$,

$\|\boldsymbol{w}_{k+1}\|^2 \leq M$

$\|\boldsymbol{w}_{k+1}\| \leq \sqrt{M}$

all in all, $\gamma M \leq \|\boldsymbol{w}_{k+1}\| \leq \sqrt{M}$

1.4

according to the result of 1.3

$$\begin{aligned} \gamma M &\leq \sqrt{M} \\ \gamma^2 M^2 &\leq M \\ M &\leq \gamma^{-2} \end{aligned}$$

2 Problem 2 Logistic Regression

2.1

because $\nabla_z \sigma(z) = \sigma(z)(1-\sigma(z))$ and $1-\sigma(z) = \sigma(-z)$

$$\begin{aligned} \nabla_{\boldsymbol{w}} L(\boldsymbol{w}, b) &= -\textstyle\sum_n \nabla_{\boldsymbol{w}} \{y_n \ln[\sigma(\boldsymbol{w}^T \boldsymbol{x}_n + b)] + (1-y_n) \ln[1-\sigma(\boldsymbol{w}^T \boldsymbol{x}_n + b)]\} \\ &= -\textstyle\sum_n \{y_n \nabla_{\boldsymbol{w}} \ln[\sigma(\boldsymbol{w}^T \boldsymbol{x}_n + b)] + (1-y_n) \nabla_{\boldsymbol{w}} \ln[\sigma(-\boldsymbol{w}^T \boldsymbol{x}_n - b)]\} \\ &= -\textstyle\sum_n \{y_n \nabla_{\boldsymbol{w}} \ln[\sigma(\boldsymbol{w}^T \boldsymbol{x}_n + b)] + (1-y_n) \nabla_{\boldsymbol{w}} \ln[\sigma(\boldsymbol{w}^T \boldsymbol{x}_n + b)]\} \\ &= -\textstyle\sum_n [y_n \boldsymbol{x_n} - y_n \boldsymbol{x_n} \sigma(\boldsymbol{w}^T \boldsymbol{x}_n + b) + y_n \boldsymbol{x_n} \sigma(\boldsymbol{w}^T \boldsymbol{x}_n + b) - \boldsymbol{x}_n \sigma(\boldsymbol{w}^T \boldsymbol{x}_n + b)] \\ &= -\textstyle\sum_n (y_n - \sigma(\boldsymbol{w}^T \boldsymbol{x}_n + b)) \boldsymbol{x}_n \\ &= \textstyle\sum_n (\sigma(\boldsymbol{w}^T \boldsymbol{x}_n + b) - y_n) \boldsymbol{x}_n \end{aligned}$$

2.2

after one iteration:

$$\begin{aligned} \boldsymbol{w_1} &= \boldsymbol{w_0} - \lambda \nabla_{\boldsymbol{w}} L(\boldsymbol{w}, b) \\ &= 0 - 0.001 \times (0.5 \times 1 - 0.5 \times 1 - 0.5 \times 1 - 0.5 \times 1] \\ &= 0.001 \end{aligned}$$

$$(x_1, y_1) = (1, 0) \quad p(y=1|x) = \sigma(wx) > 0.5 \quad \textit{prediction}: 1$$

$$(x_2, y_2) = (1, 1) \quad p(y=1|x) = \sigma(wx) > 0.5 \quad \textit{prediction}: 1$$

$$(x_3, y_3) = (1, 1) \quad p(y=1|x) = \sigma(wx) > 0.5 \quad \textit{prediction}: 1$$

$$(x_4, y_4) = (1, 1) \quad p(y=1|x) = \sigma(wx) > 0.5 \quad \textit{prediction}: 1$$

training accuracy: $\frac{3}{4} = 0.75$

2.3

because $\boldsymbol{w} = (0.001)$,

$$(x_1, y_1) = (-1, 0) \quad p(y=1|x) = \sigma(wx) < 0.5 \quad prediction: 0$$

$$(x_2, y_2) = (1, 1) \quad p(y=1|x) = \sigma(wx) > 0.5 \quad prediction: 1$$

$$(x_3, y_3) = (1, 0) \quad p(y=1|x) = \sigma(wx) > 0.5 \quad prediction: 1$$

test accuracy: $\frac{2}{3} \approx 0.67$

3 Problem 3 Backpropagation

3.1

$$\begin{aligned} \frac{\partial L}{\partial v_{jk}} &= \frac{\partial L}{\partial \hat{y_1}} * \frac{\partial \hat{y_1}}{\partial v_{jk}} + \frac{\partial L}{\partial \hat{y_2}} * \frac{\partial \hat{y_2}}{\partial v_{jk}} + \frac{\partial L}{\partial \hat{y_3}} * \frac{\partial \hat{y_3}}{\partial v_{jk}} \\ &= \sum_{i=1}^{3} \frac{\partial L}{\partial \hat{y_i}} * \frac{\partial \hat{y_i}}{\partial v_{jk}} \end{aligned}$$

$\frac{\partial L}{\partial \hat{y_i}} = -\frac{y_i}{\hat{y_i}} = -y_i \hat{y_i}^{-1}$

$\frac{\partial \hat{y_i}}{\partial v_{jk}} = \frac{\partial \hat{y_i}}{\partial o_j} * \frac{\partial \hat{o_j}}{\partial v_{jk}}$

$$\frac{\partial \hat{y_i}}{\partial o_j} = \begin{cases} \frac{exp(o_i)(\sum_{i=1}^{3} exp(o_i) - exp(o_j))}{(\sum_{i=1}^{3} exp(o_i))^2} & i = j \\ \\ -\frac{exp(o_i)exp(o_j)}{(\sum_{i=1}^{3} exp(o_i))^2} & i \neq j \end{cases}$$

$$\frac{\partial \hat{y_i}}{\partial o_j} = \begin{cases} \hat{y_i}(1 - \hat{y_i}) & i = j \\ \\ -\hat{y_i}\hat{y_j} & i \neq j \end{cases}$$

$\frac{\partial \hat{o_j}}{\partial v_{jk}} = z_k$

$$\frac{\partial L}{\partial v_{jk}} = \sum_{i=1}^{3} -y_i \hat{y_i}^{-1} * \begin{cases} \hat{y_i}(1 - \hat{y_i}) z_k & i = j \\ \\ -\hat{y_i}\hat{y_j} z_k & i \neq j \end{cases}$$

3.2

$\frac{\partial L}{\partial w_{ki}} = \sum_{j=1}^{3} \frac{\partial L}{\partial \hat{y_j}} * \frac{\partial \hat{y_j}}{\partial w_{ki}}$

$\frac{\partial L}{\partial \hat{y_j}} = -\frac{y_j}{\hat{y_j}} = -y_j \hat{y_j}^{-1}$

$\frac{\partial \hat{y_j}}{\partial w_{ki}} = \frac{\partial \hat{y_j}}{\partial z_k} * \frac{\partial z_k}{\partial w_{ki}}$

$$\frac{\partial \hat{y_j}}{\partial z_k} = \frac{\partial \hat{y_j}}{\partial o_j} * \frac{\partial \hat{o_j}}{\partial z_k} = \hat{y_j}(1 - \hat{y_j})v_{jk}$$

because $\frac{\partial tanh(x)}{\partial x} = 1 - tanh^2(x)$

$\frac{\partial z_k}{\partial w_{ki}} = (1 - z_k^2)x_i$

$$\frac{\partial L}{\partial w_{ki}} = \sum_{j=1}^{3} -y_j(1 - \hat{y_j})(1 - z_k^2)x_i v_{jk}$$