SC201 HW2

tags: `SC201` `course`

Milestone 1 - Mini Reviews

Logistic Regression

- Setup
  - feature (x)
  - feature extractor:
    - [“pretty”, “good”, “bad”, “plot”, “not”, “scenery”]
  - feature vector $\Phi(x)$:
    - e.g., $[1, 0, 1, 0, 0, 0]$
  - weight (w)
- Loss function for Logistic regression

$$L(y, h) = -(y \log(h) + (1 - y) \log(1 - h))$$

- Sigmoid function (logistic function)

$$h = \sigma(w \cdot \Phi(x)) = \frac{1}{1 + e^{-w \cdot \Phi(x)}}$$

- Gradient Descent

$$w = w - \alpha \frac{\partial L(y, h)}{\partial w}, \text{where}$$
$$\frac{\partial \mathrm{L(y, h)}}{\partial \mathrm{w}} = (\mathrm{h} - \mathrm{y})\Phi(\mathrm{x})$$

Step 0. Setup

- Learning rate: $\alpha = 0.5$
- Training Data:

```
1. (0) pretty bad
2. (1) good plot
3. (0) not good
4. (1) pretty scenery
```

- Initial weights: $w_{t_0} = [0, 0, 0, 0, 0, 0]$

Review 1: pretty bad

- $y_1 = 0$ (negative review)
- $\Phi(x_1) = [1, 0, 1, 0, 0, 0]$
- Gradient Descent, calculate w_{t_1}
  - $w_{t_0} \cdot \Phi(x_1)$
    $= [0, 0, 0, 0, 0, 0] \cdot [1, 0, 1, 0, 0, 0] = 0$

- $h_1 = \sigma(w_{t_0} \cdot \Phi(x_1))$
  $= \sigma(0) = \frac{1}{1+e^0} = 0.5$
- $\alpha(h_1 - y_1)$
  $= 0.5(0.5 - 0) = 0.25$
- $w_{t_1} = w_{t_0} - \alpha(h_1 - y_1)\Phi(x_1)$
  $= [0, 0, 0, 0, 0, 0] - 0.25[1, 0, 1, 0, 0, 0] = [-0.25, 0, -0.25, 0, 0, 0]$

Review 2: good plot

- $y_2 = 1$ (positive review)
- $\Phi(x_2) = [0, 1, 0, 1, 0, 0]$
- Gradient Descent, calculate w_{t_2}
  - $w_{t_1} \cdot \Phi(x_2)$
    $= [-0.25, 0, -0.25, 0, 0, 0] \cdot [0, 1, 0, 1, 0, 0] = 0$
  - $h_2 = \sigma(w_{t_1} \cdot \Phi(x_2))$
    $= \sigma(0) = \frac{1}{1+e^0} = 0.5$
  - $\alpha(h_2 - y_2)$
    $= 0.5(0.5 - 1) = -0.25$
  - $w_{t_2} = w_{t_1} - \alpha(h_2 - y_2)\Phi(x_2)$
    $= [-0.25, 0, -0.25, 0, 0, 0] + 0.25[0, 1, 0, 1, 0, 0] = [-0.25, 0.25, -0.25, 0.25, 0, 0]$

Review 3: not good

- $y_3 = 0$ (negative review)
- $\Phi(x_3) = [0, 1, 0, 0, 1, 0]$
- Gradient Descent, calculate w_{t_3}
  - $w_{t_2} \cdot \Phi(x_3) =$
    $[-0.25, 0.25, -0.25, 0.25, 0, 0] \cdot [0, 1, 0, 0, 1, 0] = 0.25$
  - $h_3 = \sigma(w_{t_2} \cdot \Phi(x_3))$
    $= \sigma(0.25) = \frac{1}{1+e^{-0.25}} = 0.562176$
  - $\alpha(h_3 - y_3)$
    $= 0.5(0.562176 - 0) = -0.281088$
  - $w_{t_3} = w_{t_2} - \alpha(h_3 - y_3)\Phi(x_3)$
    $= [-0.25, 0.25, -0.25, 0.25, 0, 0] - 0.281088[0, 1, 0, 0, 1, 0]$
    $= [-0.25, -0.031088, -0.25, 0.25, -0.281088, 0]$

Review 4: pretty scenery

- $y_4 = 1$ (positive review)
- $\Phi(x_4) =$ [1, 0, 0, 0, 0, 1]
- Gradient Descent, calculate w_{t_3}
  - $w_{t_3} \cdot \Phi(x_4)$
    $= [-0.25, -0.031088, -0.25, 0.25, -0.281088, 0] \cdot [1, 0, 0, 0, 0, 1] = -0.25$
  - $h_4 = \sigma(w_{t_3} \cdot \Phi(x_4))$
    $= \sigma(-0.25) = \frac{1}{1+e^{0.25}} = 0.437823$
  - $\alpha(h_4 - y_4)$
    $= 0.5(0.437823 - 1) = -0.281088$

  - $w_{t_4} = w_{t_3} - \alpha(h_4 - y_4)\Phi(x_4)$
    $= [-0.25, -0.031088, -0.25, 0.25, -0.281088, 0] + 0.281088[1, 0, 0, 0, 0, 1]$
    $= [0.031088, -0.031088, -0.25, 0.25, -0.281088, 0.281088]$

Milestone 2 - Derivatives

Derive

$$\begin{gathered}\frac{\partial L(y,h)}{\partial w} = (h-y)\Phi(x),\ \text{where} \\ L(y,h) = -(y\log(h) + (1-y)\log(1-h)), \\ h = \sigma(w\cdot\Phi(x)) = \frac{1}{1+e^{-w\cdot\Phi(x)}}\end{gathered}$$

Ans

- $\frac{\partial L}{\partial h}$

$$\begin{gathered}\frac{\partial L}{\partial h} = \frac{\partial}{\partial h}(-(y\log h + (1-y)\log(1-h))) \\ = -(y(\frac{1}{h}) + (1-y)(\frac{1}{1-h})(-1)) \\ = -(\frac{y}{h} - \frac{1-y}{1-h})\end{gathered}$$

- $\frac{\partial h}{\partial k}$

$$\begin{gathered}\frac{\partial h}{\partial k} = \frac{\partial}{\partial k}(\sigma(k)) = \frac{\partial}{\partial k}\left(\frac{1}{1+e^{-k}}\right) \\ = \frac{\partial}{\partial k}\left(\frac{1}{1+e^{-k}}\cdot\frac{e^k}{e^k}\right) = \frac{\partial}{\partial k}\left(\frac{e^k}{e^k+1}\right) \\ = \frac{(e^k)(e^k+1) - (e^k)(e^k)}{(1+e^{-k})^2} = \frac{e^k}{(1+e^{-k})^2} \\ = h(1-h) = \frac{e^k}{e^k+1}\left(1 - \frac{e^k}{e^k+1}\right)\end{gathered}$$

- $\frac{\partial k}{\partial w}$

$$\frac{\partial k}{\partial w} = \frac{\partial}{\partial w}(w\cdot\Phi(x)) = \Phi(x)$$

- $\frac{\partial L}{\partial w} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial k}\frac{\partial k}{\partial w}$

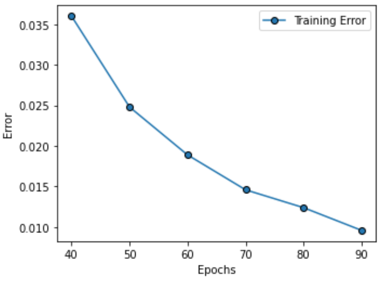
$$\begin{gathered}\frac{\partial L}{\partial w} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial k}\frac{\partial k}{\partial w} \\ = -(\frac{y}{h} - \frac{1-y}{1-h})\cdot h(1-h)\cdot\Phi(x) \\ = -y(1-h)\Phi(x) + (1-y)h\cdot\Phi(x) \\ = (h-y)\Phi(x)\end{gathered}$$

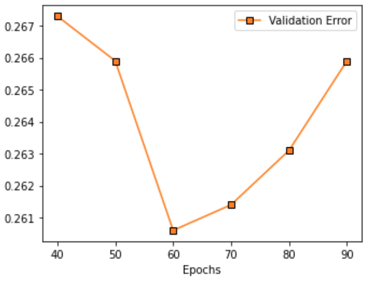
Milestone 4 - Sentiment Classification

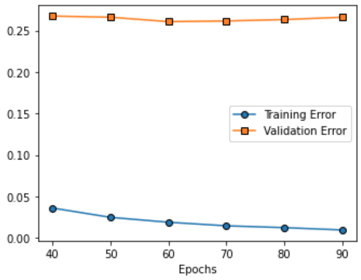
Q: Is there any difference when changing the epoch from 40 to 70/80/90?

When the epoch in the range from 40 to 90:

1. Training error is lower than validation error.
2. The training error decreases when the number of epochs are increased.
3. The validation error decreases first, hint a minumum, and then increases when the number of epochs are increased. It suggests that the model starts to overfit when the epoch higher than 60.





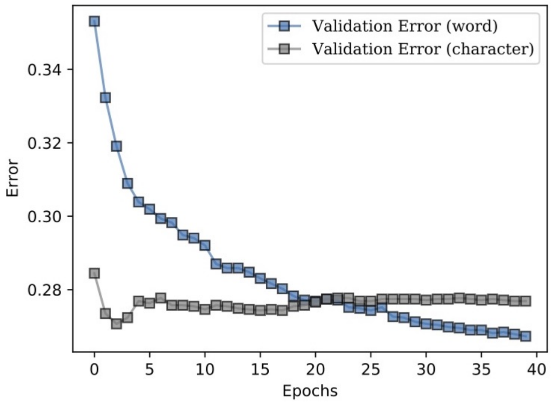


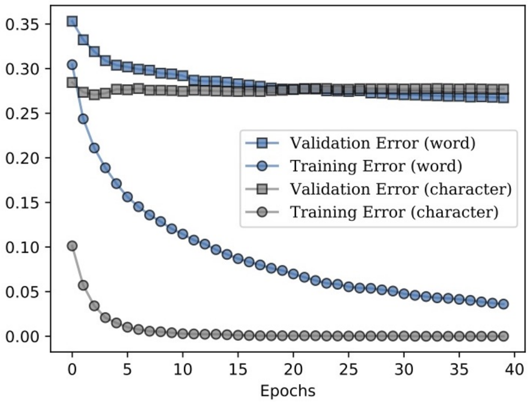
Milestone 5 - Finishing up

Q: Is there any difference between using extractCharacterFeatures and extractWordFeatures? Why?

When the epoch in the range from 0 to 40:

1. The training error of using character features is generally lower than that of using word features.
2. The training error decreases when the number of epochs are increased for both cases
3. The validation error of using word features decreases when the number of epochs are increased. However, the error of using character features decreases first, hint a minumum, and then slightly increases when the number of epochs are increased. The difference suggests that using character features will more likely lead to overfit.


Validation Error (word)
Validation Error (character)
Error
Epochs


Validation Error (word)
Training Error (word)
Validation Error (character)
Training Error (character)
Epochs