Week 04 - Part 02

So far: We talked about Logistic Regression and linear classification for binary classes.

Today: Multi-class Logistic Regression

Multi-class Logistic Regression

- Label: ...
- Hypothesis Set: Let $\Omega = \{ \underline{w}_{(1)}, \underline{w}_{(2)}, \ldots, \underline{w}_{(c)} \}$ be the weight vectors for c classes.
    - Hypothesize that

$$\mathbb{P}[y_n = i \mid \underline{x}_n] = \underline{\hspace{4cm}}$$

- Error Criterion:

- To minimize $E_{in}(\Omega)$, we need to use GD. Hence, we must find the gradient.

$$\nabla_{\Omega} e_n(\Omega) = \begin{bmatrix} \quad \end{bmatrix} = \begin{bmatrix} \quad \end{bmatrix}$$

SGD update:

In each iteration t :

Computing $\nabla_{\underline{w}(i)} \, e_n(\Omega_t)$

- $e_n(\Omega_t) = -\log \hat{P}_{\Omega_t}(y_n \mid \underline{x}_n) = -\log \dfrac{e^{\underline{w}^T(y_n)\,\underline{x}_n}}{\sum_{j=1}^{C} e^{\underline{w}^T(j)\,\underline{x}_n}}$

  $= -\underline{w}^T(y_n)\,\underline{x}_n + \log\left(\sum_{j=1}^{C} e^{\underline{w}^T(j)\,\underline{x}_n}\right)$

- For $i = y_n$,

For $i \neq y_n$,

$$\nabla_{\underline{w}(i)} e_n(\Omega_t) = \nabla_{\underline{w}(i)} \left(-\underline{w}^T(y_n)\, \underline{x}_n + \log\left(\sum_{j=1}^{c} e^{\underline{w}^T(j)\, \underline{x}_n} \right) \right)$$

$$=$$

Softmax Logistic Regression for Binary Classification (C=2)

- What is the relation between Softmax Logistic regression for C=2 and binary logistic regression that we studied last week.

- $$\hat{P}_\Omega(1|\underline{x}_n) = \frac{e^{\underline{w}^T(1)\underline{x}}}{e^{\underline{w}^T(1)\underline{x}} + e^{\underline{w}^T(2)\underline{x}_n}} = \frac{e^{(\underline{w}^T(1)-\underline{w}^T(2))\underline{x}}}{1 + e^{(\underline{w}^T(1)-\underline{w}^T(2))\underline{x}}}$$

- $$\hat{P}_\Omega(2|\underline{x}_n) = \frac{e^{\underline{w}^T(2)\underline{x}_n}}{e^{\underline{w}^T(1)\underline{x}_n} + e^{\underline{w}^T(2)\underline{x}_n}} = 1 - \hat{P}_\Omega(1|\underline{x}_n)$$

- It is Logistic Regression with $\underline{w} = \underline{w}(1) - \underline{w}(2)$

Can we use GD/SGD for Linear Regression? Yes, you can

- $E_{in}(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N} e_n(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N}(\underline{w}^T\underline{x}_n - y_n)^2.$

- $\underline{w}_{ls}$, the optimal Solution.
  - $\underline{w}_{ls} =$
  - Complexity :

- $E_{in}(\underline{w})$ is Convex. So, with GD/SGD, as $t\to\infty$, $\underline{w}_t$ Converges to the optimal Solution, i.e., $\underline{w}_{ls}$.
  - $\nabla_{\underline{w}} e_n(\underline{w}) =$
  - Time Complexity of SGD, Full-GD, Min_batch GD:
    - SGD:
    - Full GD:
    - Mini GD:

■ Benefits of GD over Least-squares method:

1-

2-