Lecture 14

Final exam: 12/28, close book

Softmax regression and deep learning

The setup: Data X, m dimensional vector, label $Y \in \{1, \cdots, M\}$

samples (x_i, y_i), $i = 1, \cdots, n$, denote $P_{X,Y}(x, y)$: empirical distribution

Goal: Train a model to predict Y from X, by the training data.

- The linear classifier is often implemented by the softmax regression.

The architecture of deep learning:

$$X \to \boxed{\text{Neural Network (NN)}} \to f(x) \longrightarrow \boxed{\text{softmax regression}} \to \hat{Y}$$

$f(x)$: k-dim feature, in general $k \ll m$

Softmax regression: given $f(x) \in \mathbb{R}^k$, train weights $g(Y) \in \mathbb{R}^k$ and bias $b(Y) \in \mathbb{R}$ to model the conditional distribution $P_{Y|X}(y|x)$ by the softmax function

$$\tilde{P}_{Y|X}^{(g,b)}(y|x) = \frac{e^{f^T(x) g(y) + b(y)}}{\sum_{y'} e^{f^T(x) g(y') + b(y')}}$$

Train $g(Y), b(Y)$ to minimize K-L divergence $D(P_{Y|X} \| \tilde{P}_{Y|X}^{(g,b)})$:

$$(g^*(Y), b^*(Y)) = \arg\min_{g,b} D(P_{X,Y} \| \tilde{P}_{Y|X}^{(g,b)} \cdot P_X)$$

$$\Rightarrow \text{get } \tilde{P}_{Y|X}^{*}(y|x) = \tilde{P}_{Y|X}^{(g^*, b^*)}(y|x)$$

Prediction: For given data X', $\hat{Y} = \arg\max_y \tilde{P}_{Y|X}^{*}(y|x')$ (MAP)

- Implemented by the gradient decent algorithm, by taking $\frac{1}{n}\sum_{i=1}^{n}\log \widetilde{P}^{(g,b)}_{Y|X}(y_i|x_i)$ as the loss function

Log-Loss.

We focus on the questions:

(i) What kind of feature $f(x)$ is good for predicting the label Y in softmax regression? What feature should be generated by the neural network?

(ii) How the weight and bias should be designed for predicting the label? Can we get theoretical interpretation for optimal weight and bias from the view of information theory / statistics?

- We will show that $f(x)$ and $g(Y)$ should follow the principal of maximal correlation.

Assumptions: Let's assume X, Y are discrete and $k=1$, for simplicity. Let's further assume X, Y are weakly dependent, i.e.,

$P_{XY}(x,y) = P_X(x)\cdot P_Y(y) + \varepsilon \cdot Q_{XY}(x,y)$, $\forall x,y$, ε is some small quantity

↑ local assumption

The local approximation of the softmax function: $\widetilde{P}^{(g,b)}_{Y|X}$ denote by $\widetilde{P}_{Y|X}$

$$\widetilde{P}_{Y|X}(y|x) = \frac{e^{f(x)g(y)+b(y)}}{\sum_{y'} e^{f(x)g(y')+b(y')}} = \frac{P_Y(y)\,e^{f(x)g(y)+b'(y)}}{\sum_{y'} P_Y(y')\,e^{f(x)g(y')+b'(y')}}$$

$b'(y) = b(y) - \log P_Y(y)$

since X, Y are weakly dependent, $\widetilde{P}_{Y|X}$ should be close to P_Y, i.e.,

$\widetilde{P}_{Y|X}(y|x) = P_Y(y) + \varepsilon \cdot ?$

$\Rightarrow e^{f(x)g(y)+b'(y)} = 1 + \varepsilon \cdot ? \Rightarrow f(x)\cdot g(y) + b'(y) = \varepsilon \cdot ?$

$\Rightarrow e^{f(x)g(y)+b'(y)} = 1 + f(x)g(y) + b'(y) + \varepsilon^2 \cdot ?$

$$\sum_{y'} P_Y(y')\, e^{f(x)g(y')+b'(y)} = \sum_{y'} P_Y(y')\left(1+f(x)g(y')+b'(y')+\varepsilon^2\cdot ?\right)$$

$$= 1 + f(x)\cdot \mathbb{E}_{P_Y}[g(Y)] + \mathbb{E}_{P_Y}[b'(Y)] + \varepsilon^2\cdot ?$$

$$\Rightarrow \widetilde{P}_{Y|X}(y|x) = \frac{P_Y(y)\left(1+f(x)g(y)+b'(y)+\varepsilon^2\cdot ?\right)}{1+\underbrace{f(x)\cdot \mathbb{E}_{P_Y}[g(Y)]+\mathbb{E}_{P_Y}[b'(Y)]}_{\varepsilon\cdot ?}+\varepsilon^2\cdot ?}$$

$\frac{1}{1+\varepsilon\cdot ?} = 1-\varepsilon\cdot ?$

$$= P_Y(y)\left(1+\underbrace{f(x)g(y)+b'(y)}_{\varepsilon\cdot ?}+\varepsilon^2\cdot ?\right)\left(1-\underbrace{f(x)\cdot \mathbb{E}_{P_Y}[g(Y)]-\mathbb{E}_{P_Y}[b'(Y)]}_{\varepsilon\cdot ?}-\varepsilon^2\cdot ?\right)$$

$$= P_Y(y)\left(1+f(x)\underbrace{\left(g(y)-\mathbb{E}_{P_Y}[g(Y)]\right)}_{\tilde{g}(y)}+\underbrace{\left(b'(y)-\mathbb{E}_{P_Y}[b'(Y)]\right)}_{\tilde{b}(y)}+\varepsilon^2\cdot ?\right)$$

$$= P_Y(y)\left(1+f(x)\cdot\tilde{g}(y)+\tilde{b}(y)+\varepsilon^2\cdot ?\right)$$

⊛ $\tilde{g}(y) = g(y)-\mathbb{E}_{P_Y}[g(Y)]$, $\tilde{b}(y) = b'(y)-\mathbb{E}_{P_Y}[b'(Y)]$

The local approximation for K-L divergence:

If $Q_1(x)$ and $Q_2(x)$ are close, i.e., $Q_2(x) = Q_1(x)+\varepsilon\cdot Q'(x)$

Then $D(Q_1\|Q_2) = \sum_x Q_1(x)\log\frac{Q_1(x)}{Q_2(x)} = -\sum_x Q_1(x)\log\frac{Q_2(x)}{Q_1(x)} = -\sum_x Q_1(x)\log\left(1+\frac{Q_2(x)-Q_1(x)}{Q_1(x)}\right)$

$\log(1+x) = x-\frac{1}{2}x^2+\cdots$ ↑ Taylor's expansion.

$$= -\sum_x Q_1(x)\left(\frac{Q_2(x)-Q_1(x)}{Q_1(x)}-\frac{1}{2}\left(\frac{Q_2(x)-Q_1(x)}{Q_1(x)}\right)^2+\varepsilon^3\cdot ?\right)$$

$$= \frac{1}{2}\underbrace{\sum_x \frac{(Q_2(x)-Q_1(x))^2}{Q_1(x)}}_{\chi^2(Q_1\|Q_2):\ \chi^2\text{-distance}} + \varepsilon^3\cdot ?$$

- Without loss of generality, we assume $f(x)$ to be zero-mean ($\mathbb{E}_{P_X}[f(x)]=0$), otherwise replace $f(x)$ by $f(x)-\mathbb{E}_{P_X}[f(x)]$, and $b(y)$ by $b'(y)+\mathbb{E}_{P_X}[f(x)]\cdot g(y)$

The local approximation of Log-Loss:

both close to $P_X \cdot P_Y$

$$D(P_{XY} \| \tilde{P}_{Y|X} \cdot P_X) = \frac{1}{2} \sum_{x,y} \frac{(\tilde{P}_{Y|X}(y|x) \cdot P_X(x) - P_{XY}(x,y))^2}{P_{XY}(x,y)} + \varepsilon^3 \cdot ?$$

($\varepsilon^2 \cdot ?$; $P_{XY}(x,y) \approx P_X(x) \cdot P_Y(y) + \varepsilon \cdot ?$)

$$= \frac{1}{2} \sum_{x,y} \frac{(\tilde{P}_{Y|X}(y|x) \cdot P_X(x) - P_{XY}(x,y))^2}{P_X(x) \cdot P_Y(y)} + \varepsilon^3 \cdot ?$$

$$= \frac{1}{2} \left[\sum_{x,y} \frac{P_{XY}^2(x,y)}{P_X(x) \cdot P_Y(y)} - 2 \cdot \sum_{x,y} \frac{\tilde{P}_{Y|X}(y|x) \cdot P_X(x) \cdot P_{XY}(x,y)}{P_X(x) \cdot P_Y(y)} + \sum_{x,y} \frac{\tilde{P}_{Y|X}^2(y|x) \cdot P_X^2(x)}{P_X(x) \cdot P_Y(y)} \right] + \varepsilon^3 \cdot ?$$

const.

$$\sum_{x,y} \frac{P_{XY}(x,y)(1+f(x)\tilde{g}(y)+\tilde{b}(y)) \cdot P_X(x) \cdot P_{XY}(x,y)}{P_X(x) \cdot P_Y(y)}$$

$$\sum_{x,y} P_{XY}(x,y)\left(1+f(x)\tilde{g}(y)+\tilde{b}(y)\right)$$

$$1 + \mathbb{E}_{P_{XY}}[f(x)\tilde{g}(y)] + \mathbb{E}_{P_Y}[\tilde{b}(y)]$$

$$\sum_{x,y} \frac{P_{XY}(x,y)(1+f(x)\tilde{g}(y)+\tilde{b}(y))^2 \cdot P_X^2(x)}{P_X(x) P_Y(y)}$$

$$\sum_{x,y} P_X(x) \cdot P_Y(y)(1+f(x)\tilde{g}(y)+\tilde{b}(y))^2$$

$$\sum_{x,y} P_X(x) P_Y(y)(1+f^2(x)\tilde{g}^2(y)+\tilde{b}^2(y))$$

$$+ 2 \sum_{x,y} P_X(x) P_Y(y)\left(f(x)\tilde{g}(y)+\tilde{b}(y) + f(x)\tilde{g}(y)\tilde{b}(y)\right)$$

$$1 + \underbrace{\mathbb{E}_{P_X}[f(x)] \cdot \mathbb{E}_{P_Y}[\tilde{g}(y)]}_{\text{var}(f(x)) \cdot \text{var}(\tilde{g}(y))} + \mathbb{E}_{P_Y}[\tilde{b}^2(y)]$$

Note that $f(x), \tilde{g}(y), \tilde{b}(y)$ are zero-mean functions

$$\Rightarrow D(P_{XY} \| \tilde{P}_{Y|X} \cdot P_X) = \text{const.} - \mathbb{E}_{P_{XY}}[f(x) \cdot \tilde{g}(y)] + \frac{1}{2}\text{var}(f(x))\text{var}(\tilde{g}(y)) + \frac{1}{2}\mathbb{E}_{P_Y}[\tilde{b}^2(y)]$$

In order to minimize the Log-loss, we need to design $f, \tilde{g}, \tilde{b}$ to maximize.

$$\mathbb{E}_{P_{XY}}[f(x) \cdot \tilde{g}(y)] - \frac{1}{2}\text{var}(f(x)) \cdot \text{var}(\tilde{g}(y)) - \frac{1}{2}\mathbb{E}_{P_Y}[\tilde{b}^2(y)]$$

$\tilde{b}(y) = 0$

$\Rightarrow b(y) = \log P_Y(y) + \text{const.}$

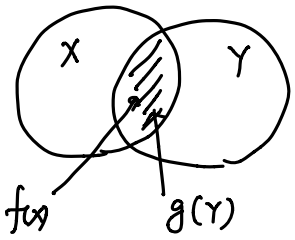
$$\Rightarrow \begin{cases} f(x) = \text{var}^{-1}(\tilde{g}(y)) \cdot \mathbb{E}_{P_{Y|X}}[\tilde{g}(y)|X=x] \\ \tilde{g}(y) = \text{var}^{-1}(f(x)) \cdot \mathbb{E}_{P_{X|Y}}[f(x)|Y=y] \end{cases}$$

$\Rightarrow f(x)$ and $\tilde{g}(y)$ are the HGR maximal correlation functions !

$$\sum_{x,y} P_{XY}(x,y) f(x) \tilde{g}(y) - \frac{1}{2}\left(\sum_x P_X(x) f^2(x)\right) \cdot \left(\sum_y P_Y(y) \tilde{g}^2(y)\right) \quad \text{— (*)}$$

$$\frac{\partial}{\partial f(x)}(*) \Rightarrow \sum_y P_{XY}(x,y) \cdot \tilde{g}(y) - P_X(x) \cdot f(x) \cdot \underbrace{\sum_y P_Y(y) \cdot \tilde{g}^2(y)}_{\text{var}(\tilde{g}(y))} = 0 \Rightarrow f(x) = \text{var}^{-1}(\tilde{g}(y)) \cdot \mathbb{E}_{P_{Y|X}}[\tilde{g}(y)|X=x]$$

- Deep neural networks and softmax regression are trying to extract the most correlated / informative features of the data and label.



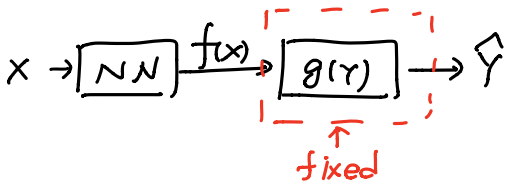
- Softmax regression : Fix input feature $f(x)$

$$f(x) \rightarrow \boxed{g(Y)} \rightarrow \hat{Y}$$

optimal weight $g^*(y) = \mathbb{E}[f(x) \mid Y=y]$

  - The forward updating

- The backward updating :



  - The optimal feature the NN should generate : $f^*(x) = \mathbb{E}[g(Y) \mid X=x]$

- Forward + Backward = Gradient decent.