$(x_i, y_i)_{i=1,2,\dots,n}$

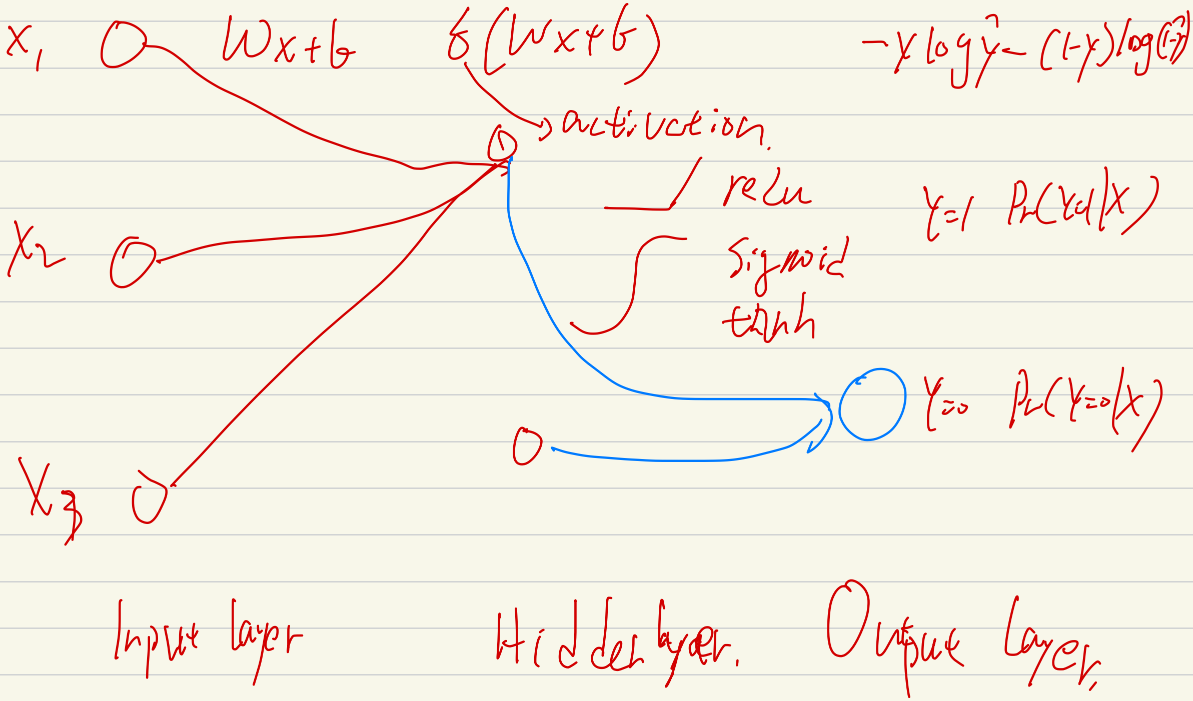
$$\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i;\theta))$$

$f(x, \theta)$ is a DNN

$x \rightarrow$ feature, $\theta \rightarrow$ parameter.

L: loss

- regression: $(y-\hat{y})^2$
- classification: $-y\log\hat{y}-(1-y)\log(1-\hat{y})$



$$f(X;\theta)=\sigma_2(W\sigma_1(WX+b)+b)$$

① Loss?: ✓.

② How to find $\hat{\theta} \in \arg\min \mathcal{L}(\theta)$

Gradient Descent.

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \alpha \nabla_\theta \mathcal{L}(\hat{\theta}_t)$$

③ How to estimate $\nabla_\theta \mathcal{L}(\hat{\theta})$?

(i) BP / Chain-Rule.

$$y = f(z) \quad z = g(x)$$

$$\frac{dy}{dx} = \frac{df}{dz}\Big|_{z=g(x)} \cdot \frac{dg}{dx}\Big|_{x}.$$

(ii) SGD: $B \subseteq D$

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \alpha \cdot \frac{1}{B} \sum_{i=1}^{B} \nabla_\theta \mathcal{L}(\hat{\theta}_t)$$

(iii) Adam