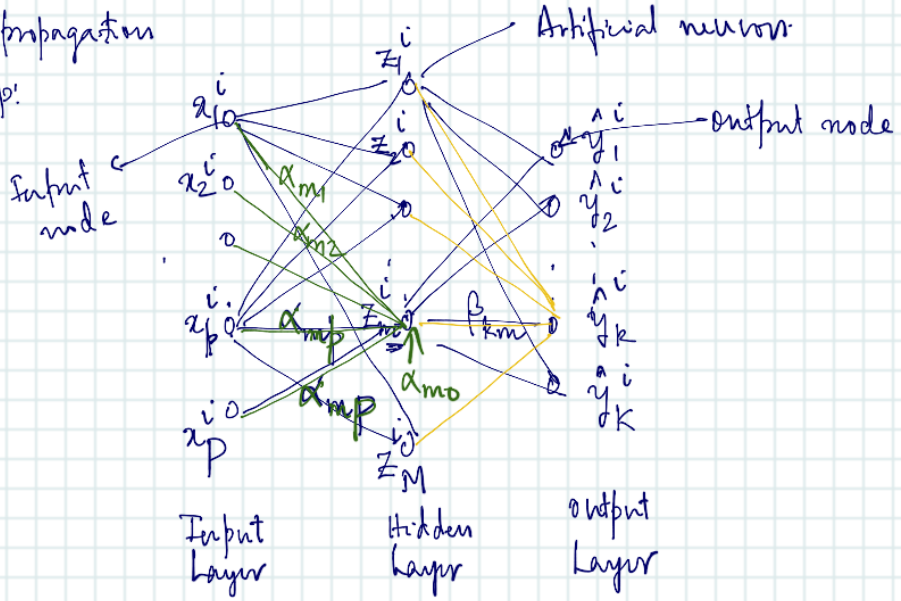
8/9/23

Deep Learning

- Recap
- ANN
- Gradient descent
- Back propagation
- Recap!



Multilayer Perceptron

Goals:
- Find the input output relation ✓
- In the supervised learning setting, describe & define loss $\mathcal{D} = \{(\underline{x}_1, y_1), (\underline{x}_2, y_2) \cdots (\underline{x}_n, y_n)\}$
- Find a method to minimize loss
- Let $\underline{x}^i = [1, x_1^i, \cdots x_p^i \cdots x_p^i]^T$,

  $\underline{z}^i = [1, z_1^i, \cdots z_m^i \cdots z_M^i]^T$

  $\hat{y}^i = [\hat{y}_1^i, \cdots \hat{y}_k^i, \cdots \hat{y}_K^i]^T$

- Input output relation:

Input-hidden layer relation: → non-linearity

$$z_m^i = \sigma\left(\langle \underline{x}^i, \underline{\alpha}_m \rangle\right)$$

z_m^i: data index, node index; $\underline{x}^i$: ith data point; $\underline{\alpha}_m$: weight vector

$$\underline{\alpha}_m = [\alpha_{m0}, \alpha_{m1}, \cdots \alpha_{mp}]^T$$

α_{m0}: bias

$$z_m^i = \sigma\left(\sum_{j=0}^{p} x_j^i \cdot \alpha_{mj}\right)$$

$$z_m^i = \sigma\left(\underbrace{1 \cdot \alpha_{m0}}_{\text{bias term}} + x_1^i \alpha_{m1} + \cdots + x_p^i \cdot \alpha_{mp}\right) \quad -(1)$$

$$\underline{z}^i = [1, z_1^i, z_2^i \cdots \underbrace{z_m^i}, \cdots z_M^i]^T$$

$$\underline{z}^i = [1, \sigma(\langle \underline{x}^i, \underline{\alpha}_1 \rangle), \sigma(\langle \underline{x}^i, \underline{\alpha}_2 \rangle), \cdots, \sigma(\langle \ , \ \rangle) \cdots \sigma(\langle \underline{x}^i, \underline{\alpha}_M \rangle)]^T \quad -(2)$$

parameters in the input - hidden layer node connections $= M \cdot (P+1)$

- Hidden layer - output layer relation:

$$\hat{y}_k^i = \text{softmax}\left(\langle \underline{z}^i, \underline{\beta}_k \rangle\right) \qquad \underline{\beta}_k = [1, \beta_{k1}, \cdots, \beta_{kM}]^T$$

$$\hat{y}_k^i = \frac{e^{\langle \underline{z}^i, \underline{\beta}_k \rangle}}{\sum_{j=1}^{K} e^{\langle \underline{z}^i, \underline{\beta}_j \rangle}} \quad -(3)$$

$$\hat{y}_k^i = \text{softmax}\left(\langle [1, \sigma(\langle \underline{x}^i, \underline{\alpha}_1 \rangle) \cdots, \sigma(\langle \underline{x}^i, \underline{\alpha}_M \rangle)]^T, \underline{\beta}_k \rangle\right) \quad -(4)$$

$\Rightarrow \hat{y}_k^i = f(\underline{x}^i; \underline{\theta})$ where $f(\ )$ is defined in (4),

$$\theta = \{\underline{\alpha}_1, \underline{\alpha}_2, \cdots, \underline{\alpha}_M, \underline{\beta}_1, \underline{\beta}_2 \cdots \underline{\beta}_K\}$$

- Risk or cost : $R(\theta) = \sum_{i=1}^{n} d(\underline{y}^i, \hat{\underline{y}}^i)$