Artificial neural networks

- Feedforward NN with one hidden layer:

$$\hat{\mathbf{y}} = \varphi(\mathbf{W}^{(2)}\sigma^{(1)}(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) + \mathbf{b}^{(2)}) = \varphi\left(\sum_{k=1}^{K} W_{kl}^{(2)}\, \sigma^{(1)}\left(\sum_{j=1}^{J} W_{jk}^{(1)} x_j + b_j^{(1)}\right) + b_k^{(2)}\right)$$

$\mathbf{x}$ = input layer

$\hat{\mathbf{y}}$ = output prediction layer

$\boldsymbol{\theta}$ = parameters to estimate = $\{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(2)}\}$

$$\sigma(\mathbf{z}) = \begin{cases} \max(\mathbf{0}, \mathbf{z}) & \text{relu, defacto standard} \\ \left(1 + e^{-\mathbf{z}}\right)^{-1} & \text{sigmoid, old school} \\ \text{many} & \text{variations on these and others} \end{cases}$$

$$\varphi(\mathbf{z}) = \begin{cases} h\tan \mathbf{z} & \textit{regression} \\ \frac{e^{\mathbf{z}}}{\sum_{\mathbf{z}} e^{\mathbf{z}}} & \textit{classification} \end{cases}$$

NN parameter estimation for regression

- Model: $\mathbf{Y} = \hat{\mathbf{Y}} + \boldsymbol{\epsilon} = f(\mathbf{X};\boldsymbol{\theta}) + \boldsymbol{\epsilon}$ *where* $f(\mathbf{X};\boldsymbol{\theta})$ expresses our neural network

$$\boldsymbol{\epsilon} \sim N(\mathbf{Y} - f(\mathbf{X};\boldsymbol{\theta}), \boldsymbol{\Sigma})$$

$$= \sqrt{\frac{1}{(2\pi)^N \det \boldsymbol{\Sigma}}}\, e^{-\frac{1}{2}(\mathbf{X} - f(\mathbf{X};\boldsymbol{\theta}))^T \boldsymbol{\Sigma}^{-1}(\mathbf{X} - f(\mathbf{X};\boldsymbol{\theta}))}$$

$$:= P(\mathbf{Y} \mid \mathbf{X};\boldsymbol{\theta}) \leftarrow \mathscr{L}(\mathbf{D};\boldsymbol{\theta})$$

- Optimization:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \operatorname*{argmin}_{\boldsymbol{\theta}} \; -\log P(\mathbf{Y} \mid \mathbf{X};\boldsymbol{\theta}) \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} \; -\log\left(\sqrt{\frac{1}{(2\pi)^N \det \boldsymbol{\Sigma}}}\exp\left[-\frac{1}{2}\left(\mathbf{X} - f(\mathbf{X};\boldsymbol{\theta})\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X} - f(\mathbf{X};\boldsymbol{\theta})\right)\right]\right) \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} \; -\frac{1}{2}\log((2\pi)^N \det \boldsymbol{\Sigma}) - \frac{1}{2}\left(\mathbf{X} - f(\mathbf{X};\boldsymbol{\theta})\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X} - f(\mathbf{X};\boldsymbol{\theta})\right) \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} \; -\left(\mathbf{X} - f(\mathbf{X};\boldsymbol{\theta})\right)^T\left(\mathbf{X} - f(\mathbf{X};\boldsymbol{\theta})\right) \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} \; -\sum_{i=1}^{M}\left(\mathbf{y}^{(i)} - f(\mathbf{x}^{(i)};\boldsymbol{\theta})\right)^T\left(\mathbf{y}^{(i)} - f(\mathbf{x}^{(i)};\boldsymbol{\theta})\right) \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} \; -\sum_{i=1}^{M}\sum_{j=1}^{N}\left(\mathbf{y}_j^{(i)} - f(\mathbf{x}^{(i)};\boldsymbol{\theta})_j\right)^2
\end{aligned}$$

$\leftarrow$ least squares

$\boldsymbol{\Sigma}$ is diagonal, strictly positive, independent of $\mathbf{x}$; it doesn't affect $\hat{\boldsymbol{\theta}}$