Neurona = "afín + no linealidad"

$z = w^T x + b$, $a = \phi(z)$

- x: Entradas (características)
- w, b: Parámetros de aprendizaje.
- ϕ: Activación no lineal
- a: salida de la neurona.

Capas:

$h^{(1)} = \phi(W^{(1)} x + b^{(1)})$

$h^{(2)} = \phi(W^{(2)} h^{(1)} + b^{(2)})$

$\vdots$

$\hat{y} = g(W^{(L)} h^{(L-1)} + b^{(L)})$

Aprender $= \underbrace{\min_{\theta} \mathbb{E}_{(x,y)\sim \text{datos}}[\mathcal{L}(f_\theta(x), y)]}_{\text{minimizar una pérdida.}}$

Entrenamiento:

Descenso por gradiente

$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L} \longrightarrow$ Regla de la cadena con esteroides xD.

Red Neuronal = Modelo paramétrico enorme $f_\theta(x)$

Autoencoder $\begin{cases} \text{Encoder: } z = f_\phi(x) \\ \text{Decoder: } \hat{x} = g_\theta(z) = g_\theta(f_\phi(x)) \end{cases}$

Entrenamiento: $\min_{\theta\phi} \mathbb{E}_{x\sim D}[\mathcal{L}(x, \hat{x})]$

Autoencoder Undercomplete:

$x \in \mathbb{R}^D$, $z \in \mathbb{R}^d$: $d << D$

Si:

- f_ϕ y g_θ son lineales.
- $\mathcal{L}(\cdot)$ es MSE.
- No hay activaciones no lineales.

$\longrightarrow$ PCA.

VAE: $z \sim p(z)$ (prior, típicamente $\mathcal{N}(0, I)$)

$x \sim p_\theta(x|z)$ (decoder probabilístico).

El objetivo probabilístico ideal sería máxima verosimilitud:

$\max_\theta \sum_i \log p_\theta(x_i)$ donde $p_\theta(x) = \int p(z) p_\theta(x|z)\, dz$

Problema: $p_\theta(x)$ suele ser directamente intratable

Por lo tanto, el posterior:

$p(z|x) = \dfrac{p(z)\, p_\theta(x|z)}{p_\theta(x)}$ también es intratable.

Solución:

Aproximar posterior con una familia manejable

$q_\phi(z|x)$ (típicamente) $= \mathcal{N}(\mu_\phi(x), \operatorname{diag}(\sigma_\phi^2(x)))$

Entonces: VAE no da z sino su distribución.

ELBO (Evidence Lower BOund)

$p_\theta(x) = \int q_\phi(z|x) \cdot \dfrac{p_\theta(x|z)\, p(z)}{q_\phi(z|x)}\, dz$

Vía Jensen:

$\log p_\theta(x) \geq \underbrace{\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x) \| p_\theta(z))}_{\text{ELBO}}$

$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x) \| p_\theta(z))$

Maximizar ELBO $\Rightarrow$ Acercarse al verdadero $p_\theta(x)$ mientras se fuerza a $q_\phi(z|x)$ a ser cercano al prior real.