Exercise 1:

In supervised learning, we typically assume that the data set $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$ originates from a data generating process $\mathbb{P}_{xy}$ in an i.i.d manner, i.e., $\mathcal{D} \sim (\mathbb{P}_{xy})^n$. One could split data set $\mathcal{D}$ with n observations into subsets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ of sizes n_{train} and n_{test} with $n_{\text{train}} + n_{\text{test}} = n$. Both subsets can be represented with index vectors $J_{\text{train}} \in \{1, \ldots, n\}^{n_{\text{train}}}$ and $J_{\text{test}} \in \{1, \ldots, n\}^{n_{\text{test}}}$, respectively. For such an index vector J of length m, one can define a corresponding vector of labels $\mathbf{y}_J = \left(y^{(J^{(1)})}, \ldots, y^{(J^{(m)})}\right) \in \mathcal{Y}^m$ and a corresponding matrix of prediction scores $\boldsymbol{F}_{J,f} = \left(f(\mathbf{x}^{(J^{(1)})}), \ldots, f(\mathbf{x}^{(J^{(m)})})\right) \in \mathbb{R}^{m \times g}$ for a model f. For regression tasks, $g = 1$ and $\boldsymbol{F}_{J,f}$ is a vector.

For a learner $\mathcal{I}$, n_{train} training observations and a performance measure ρ, the **generalization error** can be formally expressed as:

$$\mathrm{GE}(\mathcal{I}, n_{\text{train}}, \rho) = \lim_{n_{\text{test}} \to \infty} \mathbb{E}_{\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}} \sim \mathbb{P}_{xy}} \left[\rho\left(\mathbf{y}_{J_{\text{test}}}, \boldsymbol{F}_{J_{\text{test}}, \mathcal{I}(\mathcal{D}_{\text{train}})}\right)\right], \tag{1}$$

where $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ are independently sampled from $\mathbb{P}_{xy}$.

1) **What is the generalization error? Describe the formula above in your own words.**

In practice, the data generating process $\mathbb{P}_{xy}$ is usually unknown. However, assume we can sample as many times as we like from $\mathbb{P}_{xy}$.

2) **Explain how you could empirically estimate the generalization error $\mathrm{GE}(\mathcal{I}, n_{\text{train}} = 100, \rho)$ of a learner $\mathcal{I}$ trained on $n_{\text{train}} = 100$ observations and evaluated on performance measure ρ, given that you can sample from $\mathbb{P}_{xy}$ as often as you like.**

In addition to an unknown data-generating process $\mathbb{P}_{xy}$, supervised learning is often restricted to a data set $\mathcal{D}$ of fixed size n. Therefore, the true generalization error $\mathrm{GE}(\mathcal{I}, n, \rho)$ remains unknown. In this case, hold-out splitting is a simple procedure that can be used to estimate the generalization error:

$$\widehat{\mathrm{GE}}_{J_{\text{train}}, J_{\text{test}}}(\mathcal{I}, |J_{\text{train}}|, \rho) = \rho\left(\mathbf{y}_{J_{\text{test}}}, \boldsymbol{F}_{J_{\text{test}}, \mathcal{I}(\mathcal{D}_{\text{train}})}\right), \tag{2}$$

where $J_{\text{train}} \in \{1, \ldots, n\}^{n_{\text{train}}}$ specifies the subset of $\mathcal{D}$ the learner $\mathcal{I}$ is trained on, with $|J_{\text{train}}| = n_{\text{train}} < n$.

3) **Explain how the choice of $|J_{\text{train}}|$ may influence the bias of $\widehat{\mathrm{GE}}_{J_{\text{train}}, J_{\text{test}}}(\mathcal{I}, |J_{\text{train}}|, \rho)$ wrt $\mathrm{GE}(\mathcal{I}, n, \rho)$.**

4) **Explain how the choice of $|J_{\text{train}}|$ may influence the variance of $\widehat{\mathrm{GE}}_{J_{\text{train}}, J_{\text{test}}}(\mathcal{I}, |J_{\text{train}}|, \rho)$.**