Mixture-of-ITEs

François G

May 2023

1 The nonparametric model

Let X denote a random vector of pretreatment covariates, and Y the random variable denoting the individualized treatment effect (ITE). Let the random vector $(C_1, \dots, C_K)^T$ denote a one-hot encoded identifier for the cluster $k = 1, \dots, K$ an observation belongs to. We assume the following nonparametric data generating process.

Denoting $\rho_1(X) \stackrel{\text{def}}{=} \mathbb{E}[C_1|X], \quad \dots, \quad \rho_K(X) \stackrel{\text{def}}{=} \mathbb{E}[C_K|X]$, the probability of an observation belonging to cluster $1, \dots, K$ respectively, we have

$$(C_1, \dots, C_K)^T|X \sim \mathit{Multinomial}\Big(n = 1, k = K, p = \big(\rho_1(X), \dots, \rho_K(X)\big)^T\Big).$$

Denoting $q_k(X) \stackrel{\text{def}}{=} \mathbb{E}[Y|X, C_k = 1]$, and assuming these functions exist for all $k = 1, \dots, K$, we have

$$\begin{aligned}\mathbb{E}[Y|X] &= \sum_{k=1}^{K} \mathbb{P}(C_k = 1|X)\,\mathbb{E}[Y|X, C_k = 1] \\ &= \sum_{k=1}^{K} \rho_k(X) q_k(X).\end{aligned}$$

We assume that given X, the random variable Y is sampled from a probability density $f_{Y|X}(y|x)$ with expected value $\mathbb{E}[Y|X = x] = \sum_{k=1}^{K} \rho_k(x) q_k(x)$. Consistent with the terminology of Jacobs et al. [1] and Jordan and Jacobs [2] we call "expert networks" the functions $q_k(\cdot),\ k = 1, \dots, K$ and "gating network" the function $\{\rho(\cdot)\}_{k=1}^{K}$.

2 Fitting algorithm

To estimate $\{\rho_k(\cdot)\}_{k=1}^K$ we need posit a density for $f_{Y|X,K}(y|x,k)$. In the fitting algorithm below we posit that $f_{Y|X,K}(y|x,k)$ is a normal distribution.

Algorithm 1 The nonparametric EM-like procedure for estimating $\{\rho_k(\cdot)\}_{k=1}^K$.

Input: Data $(X_i, Y_i)_{1\leq i\leq n}$, and $K \in \mathbb{N}$ the numbers of clusters.
Initialize the prior probabilities associated with the nodes of the tree as

$$g_{k,i} \leftarrow 1/K \quad \text{for} \quad k = 1, \dots, K,$$

and use the shorthand

$$G = \begin{bmatrix} g_{1,1} & \cdots & g_{1,K} \\ \cdots & \cdots & \cdots \\ g_{n,1} & \cdots & g_{n,K} \end{bmatrix}.$$

Initialize the individual predictions from the expert networks e.g.,

$$\mu_{k,i} \sim \mathcal{U}_{[-1,1]} \quad \text{for} \quad k = 1, \dots, K.$$

Iterate until convergence on G:

Compute individual contributions to each expert's likelihood as

$$L_{k,i} \leftarrow \mathcal{N}_{\mathcal{L}}(Y_i|\mu = \mu_{k,i}, \sigma^2 = 1) \quad \text{for} \quad k = 1, \dots, K.$$

Compute the posterior probabilities associated with the nodes of the tree as ▷ E-step

$$h_{k,i} \leftarrow \frac{g_{k,i}L_{k,i}}{\sum_{l=1}^K g_{l,i}L_{l,i}} \quad \text{for} \quad k = 1, \dots, K.$$

For each expert network fit $\hat{q}_k(\cdot), k = 1, \dots, K$ separately ▷ M-step
with a weighted nonparametric classifier with features X_i, labels Y_i and weights $h_{k,i}$.
For the gating network jointly fit $\{\hat{\rho}_k(\cdot)\}_{k=1}^K$ as a multiclass classification problem with features X_i, and labels $(h_{1,i}, \dots, h_{K,i})$.
Update the predictions from the expert networks as

$$\mu_{k,i} \leftarrow \hat{q}_k(X_i) \quad \text{for} \quad k = 1, \dots, K.$$

Update the prior probabilities associated with the nodes of the tree as

$$g_{k,i} \leftarrow \hat{\rho}_k(X_i) \quad \text{for} \quad k = 1, \dots, K.$$

Return: $\{\hat{\rho}_k(\cdot)\}_{k=1}^K$

References

[1] Robert A Jacobs et al. “Adaptive mixtures of local experts”. In: *Neural computation* 3.1 (1991), pp. 79–87.

[2] Michael I Jordan and Robert A Jacobs. “Hierarchical mixtures of experts and the EM algorithm”. In: *Neural computation* 6.2 (1994), pp. 181–214.