L173 Pre-reading response for session 8

D.J. Wischik

Skim-read *A unifying view on dataset shift in classification*, Moreno-Torres et al. 2012, §§1–4, and read *Invariant risk minimization*, Arjovsky et al. 2019, §§1–3 as far as the optimization problem labelled (IRM).

Consider the ground-truth data generating process, a variant of Example 1 in Arjovsky et al.:

$$X_1 \leftarrow N(\mu_e, 1), \qquad Y \leftarrow \alpha X_1 + N(0,1), \qquad X_2 \leftarrow Y + N(\nu_e, 1),$$

where the noise terms are independent. We are interested in learning to predict the response Y given predictors $X = (X_1, X_2)$, across several environments $e \in \mathcal{E}$, which differ only in μ_e and ν_e; the coefficient $\alpha \neq 0$ is the same in every environment. Write $(X, Y) \sim q_e$ for the joint distribution in environment e.

1. Suppose there are two environments, $\mathcal{E} = \{\text{train}, \text{test}\}$, and they differ in both μ_e and ν_e. How would Moreno-Torres et al. describe this shift?

2. Now there are three training environments, $\mathcal{E}_{\text{tr}} = \{1, 2, 3\}$, with (μ_e, ν_e) taking values $(0,0)$, $(1,0)$, $(0,1)$ respectively. We wish to find an invariant predictor. We will choose $\mathcal{H} = \mathbb{R}$ and $\Phi(x_1, x_2) = \phi_1 x_1 + \phi_2 x_2$, allow any $w : \mathbb{R} \to \mathbb{R}$, and use the squared-error risk

   $$R^e(w \circ \Phi) = \mathbb{E}_{(X,Y)\sim q_e}\big(Y - w(\Phi(X))\big)^2.$$

   For which (ϕ_1, ϕ_2) is there a w such that (Φ, w) satisfies the constraint in (IRM)? Give the resulting predictor $w \circ \Phi$. *[Hint. In this problem we can find* $\arg\min_{\bar{w}} R^e(\bar{w} \circ \Phi)$ *exactly.]*

You can answer question 1 in two or three lines; your answer should refer to specific passages from Moreno-Torres et al. For question 2, just give a few key steps of your working, not the full detailed algebra of the derivation. AI usage is permitted.