STATISTICAL MACHINE LEARNING (WS2024/25) HOMEWORK: EM ALGORITHM FOR PRIOR SHIFT ADJUSTMENT

Assignment (5 points) *Consider a training data $\mathcal{T}^m = ((x^i, y^i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \dots, m)$, where the samples are i.i.d. and drawn from the distribution $p_{\mathrm{tr}}(x, y) = p(x \mid y)\, p_{\mathrm{tr}}(y)$. From this training data, you have obtained estimates of:*

1. *The training prior: $\hat{p}_{\mathrm{tr}}(y)$,*
2. *The training posterior: $\hat{p}_{\mathrm{tr}}(y \mid x)$.*

At deployment, the data distribution changes due to a shift in the prior. The deployment data is drawn from the distribution $p_{\mathrm{de}}(x, y) = p(x \mid y)\, p_{\mathrm{de}}(y)$, where $p_{\mathrm{de}}(y) \neq p_{\mathrm{tr}}(y)$. You are provided with an unlabeled dataset $\mathcal{S}^n = (x^i \in \mathcal{X} \mid i = 1, \dots, n)$, where the samples are i.i.d. and drawn from the deployment marginal distribution $p_{\mathrm{de}}(x) = \sum_{y \in \mathcal{Y}} p_{\mathrm{de}}(x, y)$.

a) Estimating Deployment Prior with EM. *Implement the Expectation-Maximization (EM) algorithm to estimate the deployment prior $p_{\mathrm{de}}(y)$ using the unlabeled deployment samples $\mathcal{S}^n$. Use the training posterior $\hat{p}_{\mathrm{tr}}(y \mid x)$ in your implementation.*

b) Plugin Bayes Classifier for Deployment Data. *Given:*

- *The estimated deployment prior $\hat{p}_{\mathrm{de}}(y)$,*
- *The training posterior $\hat{p}_{\mathrm{tr}}(y \mid x)$,*
- *The training prior $\hat{p}_{\mathrm{tr}}(y)$,*
- *A loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$,*

implement a plugin Bayes classifier:

$$\hat{h}(x) = \operatorname*{arg\,min}_{y \in \mathcal{Y}} \sum_{y' \in \mathcal{Y}} \hat{p}_{\mathrm{de}}(y' \mid x) \ell(y, y') \,, \tag{1}$$

where the deployment posterior $\hat{p}_{\mathrm{de}}(y \mid x)$ accounts for the prior shift and is adapted from the training posterior. Ensure your implementation can efficiently handle deployment data.