MULTIVARIATE LOSS FUNCTIONS

- In MTP: For a feature vector **x**, predict a tuple of scores
 f(**x**) = (*f*(*x*)₁, *f*(*x*)₂, ..., *f*(*x*)_{*l*})[⊤] for *l* targets with a function
 (hypothesis) *f* : *X* → ℝ^{g₁} × ··· × ℝ^{g_l}.
- Following loss minimization in machine learning, we need a *multivariate loss function*

$$L: (\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_l) \times (\mathbb{R}^{g_1} \times \cdots \times \mathbb{R}^{g_l}) \to \mathbb{R}.$$

- In multi-target regression: $\mathcal{Y}_1 = \ldots = \mathcal{Y}_l = \mathbb{R}$, and $g_1 = \ldots = g_l = 1$.
- In multi-label binary classification: $\mathcal{Y}_1 = \ldots = \mathcal{Y}_l = \{0, 1\}$, and $g_1 = \ldots = g_l = 1$.



MULTIVARIATE LOSS FUNCTIONS

• We treat two categories: Decomposable and instance-wise





• L is decomposable over targets if

$$L(\mathbf{y}, f) = \frac{1}{l} \sum_{m=1}^{l} L_m(y_m, f(\mathbf{x})_m)$$

with single-target losses L_m .

• Example: *Squared error loss* (in multivariate regression):

$$L_{\text{MSE}}(\mathbf{y}, f) = \frac{1}{l} \sum_{m=1}^{l} (y_m - f(\mathbf{x})_m)^2.$$

• Can also be used for cases with missing entries.

INSTANCE-WISE LOSSES

• Hamming loss averages over mistakes in single targets:

$$L_{H}(\mathbf{y},\mathbf{h}) = \frac{1}{l} \sum_{m=1}^{l} \mathbb{1}_{[y_m \neq h_m(\mathbf{x})]},$$

where $h_m(\mathbf{x}) := [f(\mathbf{x})_m \ge c_m]$ is the threshold function for target *m* with threshold c_m .

- Hamming loss is identical to the average *0/1 loss* and is decomposable.
- The *subset 0/1 loss* checks for entire correctness and is not decomposable:

$$L_{0/1}(\mathbf{y},\mathbf{h}) = \mathbb{1}_{[\mathbf{y}\neq\mathbf{h}]} = \max_{m} \mathbb{1}_{[y_m\neq h_m(\mathbf{x})]}$$



HAMMING VS. SUBSET 0/1 LOSS

• The risk minimizer for the Hamming loss is the marginal mode:

$$f^*(\mathbf{x})_m = \arg \max_{y_m \in \{0,1\}} \Pr(y_m \mid \mathbf{x}), \quad m = 1, \dots, l,$$

while for the subset 0/1 loss it is the *joint mode*:

$$f^*(\mathbf{x}) = \arg \max_{\mathbf{y}} \Pr(\mathbf{y} \mid \mathbf{x}).$$

• Marginal mode vs. joint mode:

У	Pr(y)		
0000	0.30		
0111	0.17	Marginal mode: Joint mode:	1111
1011	0.18		0000
1101	0.17		
1110	0.18		

