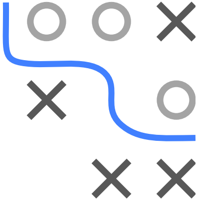
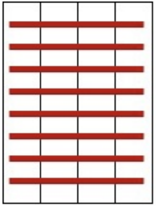
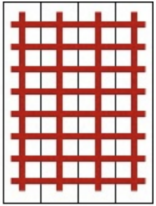
Advanced Machine Learning

Multi-Target Prediction: Loss Functions





instance-wise

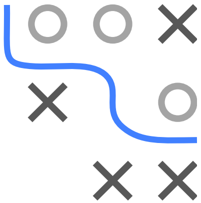


micro

Learning goals

- Get to know loss functions for multi-target prediction problems
- Know the Bayes predictor for Hamming loss and subset $0/1$ loss
- Understand the difference between macro-, micro-, and instance-wise-losses

MULTIVARIATE LOSS FUNCTIONS

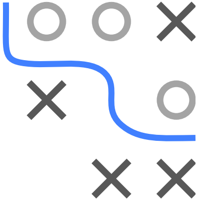


- In MTP: For a feature vector $\mathbf{x}$, predict a tuple of score vectors $f(\mathbf{x}) = (f(x)_1, f(x)_2, \dots, f(x)_l)^\top$ for l tasks with a function (hypothesis) $f : \mathcal{X} \to \mathbb{R}^{g_1} \times \dots \times \mathbb{R}^{g_l}$.
- Following loss minimization in machine learning , we need a *multivariate loss function*

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

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

MULTIVARIATE LOSS FUNCTIONS



- L is decomposable over instances if

$$L = \sum_{i=1}^{n} L(\mathbf{y}^{(i)}, f(\mathbf{x}^{(i)})) \, ,$$

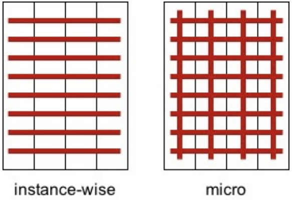
  i.e., as a sum of losses over all examples.

- L is decomposable over targets if

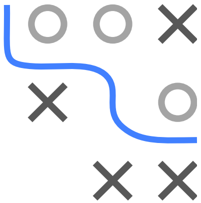
$$L(\mathbf{y}, f) = \sum_{m=1}^{l} L_m(y_m, f(\mathbf{x})_m)$$

  with single-target losses L_m.

- Two categories: micro- and instance-wise-losses.



MICRO LOSSES



- Micro-losses: L corresponds to aggregating the pointwise losses over the targets and instances.

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

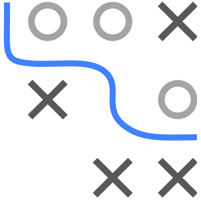
where $L_m : \mathcal{Y}_m \times \mathbb{R}^{g_m} \to \mathbb{R}$ in this case.

- Example *Squared error loss* (i.e. used in multivariate regression):

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

- Can also be used for cases with missing entries.

INSTANCE-WISE LOSSES



- Given a label vector $\mathbf{y}$, it computes the loss on single instance.
- *Hamming loss* averages over mistakes on individual scores:

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

  where $h_m(x) := [f(\mathbf{x})_m \geq c_m]$ is the threshold function for task m with threshold c_m.
- Hamming loss is identical to the average *0/1 loss* if computed on the entire dataset.
- The *subset 0/1 loss* simply checks for entire correctness:

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

HAMMING VS. SUBSET 0/1 LOSS

- The risk minimizer for the Hamming loss is the *marginal mode*:

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

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:

$\mathbf{y}$	$\Pr(\mathbf{y})$
0 0 0 0	0.30
0 1 1 1	0.17
1 0 1 1	0.18
1 1 0 1	0.17
1 1 1 0	0.18

Marginal mode: 1 1 1 1
Joint mode: 0 0 0 0