University of Salzburg — Lecturer: Roland Kwitt

Machine Learning (911.236)

Exercise sheet **D**

In the following question, we take the first step towards answering the question of *what is the best prediction function (/hypothesis) h* regardless of the training data?

Exercise 1. 5 P.

First, remember that we defined the generalization error of a hypothesis $h \in H$ for *general loss functions* $l : H \times X \times Y \rightarrow \mathbb{R}_+$ as

$$L_D(h) = \mathbb{E}_{z \sim D}[l(h, (x, y))] \ ,$$

where D is a distribution over $Z = X \times Y$ and $z = (x, y) \in Z$. In fact (disregarding the explicit dependency on h for better readability), we could also simply write

$$L_D(h) = \mathbb{E}[l(y, h(x))] \ .$$

Now, using the *total law of expectation*, we write

$$L_D(h) = \mathbb{E}[\mathbb{E}[l(y, h(x))|x]] \ .$$

For a fixed $x = x' \in X$, this changes to

$$L_D(h) = \mathbb{E}_{x' \sim D|_x}[\mathbb{E}[l(y, h(x))|x = x']] \ . \tag{1}$$

To understand why we have rewritten $L_D(h)$ this way, your task is, **first**, to write Eq. (1) as an integral over X. Then, **second** assume X is finite and write the integral as a sum over $x' \in X$. Upon defining the *conditional risk* for any $u \in Y$ as

$$r(u|x') = \mathbb{E}[l(y, u)|x = x'] \ ,$$

please (**third**) argue why we can say that the minimizer of $L_D(h)$ is equal to a minimizer $u \in Y$ of $r(u|y)$ for any $x' \in X$?

Note that this will generalize to non-finite X as well, and we will see (in the lecture) that we will obtain what is called the *Bayes predictor*.

Total #points: 5 P.