University of Salzburg | Lecturer: Roland Kwitt

Machine Learning (911.236)

Exercise sheet **B**

Exercise 1. 3 P.

Let (S, F, μ) be a measure space and $f : S \longrightarrow \mathbb{R}$ a real-valued measurable function. Further, let g be a real-valued function that is (1) *non-negative* on the range of f and (2) *non-decreasing*. Show that for any real number $t > 0$ and $g(t) \neq 0$,

$$\mu(\{x \in S : f(x) \geq t\}) \leq \frac{1}{g(t)} \int_{x \in S} g(f(x)) \,\mathrm{d}\mu(x) \tag{1}$$

holds. **Strategy** (suggestion): start by fixing $t > 0$ and looking at the set $B_t = \{x \in S : f(x) \geq t\}$. We can also define the indicator function $\mathbf{1}_{B_t}$ for the set B_t. Now, due to the non-decreasing and non-negative nature of g, we can claim

$$0 \leq g(t)\mathbf{1}_{B_t} \leq g(f(x))\mathbf{1}_{B_t} \enspace .$$

Go on from here by taking the Lebesgue integral with respect to μ and then complete the argument.

Exercise 2. 4 P.

We are going show the following theorem.

Theorem 1 (Chebychev inequality). *Let $X : \Omega \longrightarrow \mathbb{R}$ be a real-valued random variable on the probability space $(\Omega, \mathcal{F}, \mathcal{D})$. Assuming X has finite expected value m and finite non-zero variance[1] σ^2, it holds that for any real number $k > 0$,*

$$\mathbb{P}[|X - m| \geq k\sigma] \leq \frac{1}{k^2} \enspace .$$

Take the statement from **Exercise 1**, specifically Eq. (1), for granted. Further, we can take the fact that that continuous functions are measurable for granted (so you don't have to check $|X - m|$).

Remark 1. *Think about what the Chebychev inequality actually tells us: the probability that X deviates more than k standard deviations ($k\sigma$) from its expected value is less than $1/k^2$. This is always true!*

Exercise 3. 3 P.

Say you have a learning problem with $n < \infty$ input variables $v_1, \ldots, v_n$ that can only take on binary values, that is $v_i \in \{0, 1\}$; hence $\mathcal{X} = \{0, 1\}^n$. Also, our label set is $\mathcal{Y} = \{0, 1\}$. We look at the hypothesis class $\mathcal{H}$ of *all* boolean functions from $\mathcal{X} \longrightarrow \mathcal{Y}$. Use our results from the lecture (PAC learnability of finite hypothesis classes) and argue what the problem might be in this example, especially in terms of the number of required samples for any given choice of $\epsilon, \delta \in (0, 1)$.

Total #points: 10 P.

[1]We have $\sigma^2 = \mathbb{V}[X] = \int_{w \in \Omega} (X - m)^2 \,\mathrm{d}\mathcal{D}(w)$