University of Salzburg/Dept. of Computer Science — Lecturer: Roland Kwitt

Machine Learning (911.236)

Exercise sheet **B**

Concentration inequalities & Bayes classifier

Exercise 1. 10 P.

Assume that someone gives you a coin and he/she tells you that the coin is *biased*. We try to answer the question of how many flips of the coin do you need to decide the direction of the bias (so, towards "heads" or "tails")? Consider the outcome of n coin flips as a sequence of independent Bernoulli random variables, $X_1, \ldots, X_n$, with success probability p (let "heads" denote 1 and "tails" denote 0). Assume you estimate p with the empirical average $\hat{p}_n = 1/n \sum_i X_i$. If $\hat{p}_n \geq 1/2$ you decide a bias towards "heads", and "tails" otherwise.

Part A (4 points): Show that if

$$n > \frac{1}{2\epsilon^2} \log\left(\frac{1}{\delta}\right) \tag{1}$$

then you decide correctly with probability $1 - \delta$ for $\delta \in (0, 1)$ fixed. Remember that if $p = 1/2 - \epsilon, \epsilon > 0$, you make an error if $\hat{p}_n \geq 1/2$ (and vice versa).

Part B (2 points): What is the problem with the sample complexity in Eq. (1)? Can you construct an interval estimate of p, i.e., something of the form

$$\mathbb{P}\left[p \in (\hat{p}_n - T, \hat{p}_n + T)\right] \geq 1 - \delta \quad .$$

Part C (4 points): The interval of *Part B* will only be valid for a fixed n. Now consider a sequence of sample sizes and construct a sequence of intervals such that for *all* of them the true parameter p is contained with probability of at least $1 - \delta$. Hint: use the union bound and let δ be an appropriate function of n, i.e., $\delta(n)$.

Exercise 2. 10 P.

For this exercise, consider the **Bayes classifer**. Let $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$. The Bayes classifier, $f^* : \mathcal{X} \to \{0, 1\}$ can be written as

$$f^*(x) = \begin{cases} 1, & \text{if } \eta(x) := \mathbb{P}[y = 1|x] \geq 1/2 \\ 0, & \text{else} \quad . \end{cases} \tag{2}$$

Show that for any classifier $g : \mathcal{X} \to \{0, 1\}$

$$L_{\mathcal{D}}(f*) \leq L_{\mathcal{D}}(g) \tag{3}$$

holds, implying

$$\mathbb{P}[f^*(x) \neq y|x] \leq \mathbb{P}[g(x) \neq y|x] \quad .$$

Your task is to show that this is actually the case. Start with:

$$\begin{aligned} \mathbb{P}[g(x) \neq y|x] &= 1 - \mathbb{P}[g(x) = y|x] \\ &= 1 - (\mathbb{P}[y = 1, g(x) = 1|x] + \mathbb{P}[y = 0, g(x) = 0|x]) \\ &\ldots \end{aligned} \tag{4}$$

Sketch: (1) Re-write in terms of indicator functions and switch to expectations, then (2) use properties of expectation and (3) appropriately simplify using the definition of $\eta(x)$. Once this is done, consider the difference

$$\mathbb{P}[g(x) \neq y|x] - \mathbb{P}[f^*(x) \neq y|x] = \ldots \tag{5}$$

and make a case distinction, i.e., (1) x s.t. $\eta(x) \geq 1/2$ and (2) x s.t. $\eta(x) < 1/2$, to show

$$\mathbb{P}[g(x) \neq y|x] - \mathbb{P}[f^*(x) \neq y|x] \geq 0 \tag{6}$$

which will imply Eq. (3).

Remark(s). In order to be able to write

$$\mathbb{P}[y = 1|x]$$

we have to make some restrictions, as the atomic event $\mathcal{D}_x(\{x\}) = 0$, where $\mathcal{D}_x$ denotes the distribution over unlabeled domain points. What we will do for now is to assume that $\mathcal{D}$ (i.e., our distribution over $\mathcal{X} \times \mathcal{Y}$) admits a *joint density function* $f(x, y)$ and assume $f_x(x) > 0$ (otherwise, we define the term below to be 0). In that case, the *conditional density* is given as

$$f_{y|x}(y) = \frac{f_{x,y}(x, y)}{f_x(x)} \ ,$$

and the probability of $y = 1$, conditioned on x, can be written as

$$\mathbb{P}[y = 1|x] = \frac{f(x, 1)}{f_x(x)} \ .$$

As another remark, if you can show Eq. (6) holds, then you could just integrate out x and actually obtain $L_{\mathcal{D}}(f^*) \leq L_{\mathcal{D}}(g)$.

I will make some remarks on this during the lab.

Exercise 3.

This is *Exercise 3* from exercise sheet A.

Exercise 4. 2 P.

(Exercise 2.10 from the Mohri book). For important questions, President Mouth relies on expert advice. He selects an appropriate advisor from a collection of 2,800 experts (our $\mathcal{H}$).

Assume that laws are proposed in a random fashion independently and identically according to some distribution $\mathcal{D}$, determined by an unknown group of senators. Assume that President Mouth can find and select an expert senator out of $\mathcal{H}$ who has consistently voted with the majority for the last $m = 200$ laws. Give a bound on the probability that such a senator incorrectly predicts the global vote for a future law. What is the value of the bound with 95% confidence?

Total #points: 22 P.