Prof. Dr. M. Grohe E. Fluck, N. Runde

Exercise Sheet 2

Due date: Tuesday, May 02 until 13:00

- Please upload your solutions to RWTH Moodle.
- The due date at Tuesday, May 02 until 13:00.
- Hand in your solutions in groups of **two to three students**. If you need to change your group, contact algds@lics.rwth-aachen.de.
- Hand in the solutions of your group as a single PDF file.
- A discussion regarding this exercise sheet will take place on **Friday, May 05 14:30** in room AH II.

Exercise 1 (Dice and Tail Bounds) **2+4=6 points**

We roll a fair (six-sided) die $n \geq 1$ times independently.

a) Let X_i denote the outcome of the ith roll of the die, for $i = 1, \dots, n$. That is, X_i uniformly distributed random variable with values in $\{1, \dots, 6\}$.

We define another random variable Y_n to be the product of all die roll outcomes, that is,

$$Y_n \coloneqq \prod_{i=1}^{n} X_i.$$

Give expressions for $\mathrm{E}(Y_n)$ and $\mathrm{Var}(Y_n)$ (as functions of n).

b) Now let $\widehat{X}_i$ denote the $\{0,1\}$-valued random variable defined by

$$\widehat{X}_i = 1 \iff \text{the } i\text{th roll results in a 6.}$$

Again, the random variables $\widehat{X}_i$ are independent.

Let Z_n to be the number of 6's seen in n die rolls, that is,

$$Z_n \coloneqq \sum_{i=1}^{n} \widehat{X}_i.$$

Now let $n = 300$. We want to calculate upper bounds for the probability that at most 10 of the 300 die rolls result in a 6, i.e. the probability $\Pr(Z_{300} \leq 10)$.

To bound this probability, employ the

(i) Markov,

(ii) Chebyshev,

(iii) Chernoff, and

(iv) Hoeffding

inequalities / bounds from the lecture. Which inequality gives the best bound?

Exercise 2 (Joint and Conditional Entropy) **3+2=5 points**

Let X and Y be random variables with finite ranges $\mathrm{rg}(X)$ and $\mathrm{rg}(Y)$, defined over the same probability space $(\Omega, \mathcal{P})$. The *joint entropy* of X and Y is defined as

$$H(X,Y) = \sum_{\substack{x \in \mathrm{rg}(X) \\ y \in \mathrm{rg}(Y)}} \Pr(X = x, Y = y) \cdot \log\left(\frac{1}{\Pr(X = x, Y = y)}\right).$$

The *conditional entropy* of X given Y is defined as

$$H(X \mid Y) = \sum_{y \in \mathrm{rg}(Y)} \Pr(Y = y)\left(\sum_{x \in \mathrm{rg}(X)} \Pr(X = x \mid Y = y) \cdot \log\left(\frac{1}{\Pr(X = x \mid Y = y)}\right)\right),$$

where $\Pr(X = x \mid Y = y) = \frac{\Pr(X=x, Y=y)}{\Pr(Y=y)}$ is the conditional probability of $X = x$ given that $Y = y$.

a) Show that $H(X,Y) = H(X \mid Y) + H(Y)$.

b) Show that if X and Y are independent, then $H(X \mid Y) = H(X)$.

Exercise 3 (Information Gain) **3+2=5 points**

Consider the following data set regarding the question whether a person buys a new computer. Apply information gain to find a decision tree for this data set.

a) Calculate the information gain of each feature for the first split of the decision tree.

b) Explain your choice for the splits below the first level and draw the resulting decision tree.

Age	Income	Student	Credit Rating	Buys Computer
< 30	High	No	Fair	No
< 30	High	No	Excellent	No
30-40	High	No	Fair	Yes
> 40	Medium	No	Fair	Yes
> 40	Low	Yes	Fair	Yes
> 40	Low	Yes	Excellent	No
30-40	Low	Yes	Excellent	Yes
< 30	Medium	No	Fair	No
< 30	Low	Yes	Fair	Yes
> 40	Medium	Yes	Fair	Yes
< 30	Medium	Yes	Excellent	Yes
30-40	Medium	No	Excellent	Yes
30-40	High	Yes	Fair	Yes
> 40	Medium	No	Excellent	No
< 30	High	No	Excellent	Yes

Exercise 4 (Sample Size Bounds) **2+2=4 points**

We consider a Boolean classification problem in $\mathbb{R}^3$. Our hypothesis class $\mathcal{H}$ is the (finite) class of functions $\mathbf{x} \mapsto \operatorname{sgn}\left(\langle \mathbf{a}, \mathbf{x} \rangle\right)$ with $\mathbf{a} \in \{-1, 0, 1\}^3$.

Note: For the following tasks, try to make your bound as tight as possible using the suitable theorems from the lecture. Justify your answers. Please give your final results numeric, and round them to 2 decimal places.

a) Suppose we have 143 training examples available and suppose that for any such training sequence T, we are able to find a hypothesis h_T that achieves a training error of at most $3\,\%$. Give an upper bound for the generalisation error of h_T that holds with probability $> 90\%$.

b) Now suppose that the unknown target function is also of the shape $\mathbf{x} \mapsto \operatorname{sgn}\left(\langle \mathbf{a}, \mathbf{x} \rangle\right)$ with $\mathbf{a} \in \{-1, 0, 1\}^3$. Give a lower bound on the number of examples m that guarantees that the generalisation error of any consistent hypothesis is at most $1\,\%$ with probability greater than $90\,\%$.