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

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,

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

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,

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

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?

Solution:

a)

$$\mathrm{E}(X_i) = \sum_{j=1}^{6} j \cdot \Pr(X_i = j) = \tfrac{7}{2}$$

$$\mathrm{E}(Y_n) = \mathrm{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathrm{E}(X_i) = \left(\tfrac{7}{2}\right)^n$$

$$\mathrm{E}(X_i^2) = \sum_{j=1}^{6} j^2 \cdot \Pr(X_i = j) = \tfrac{91}{6}$$

$$\mathrm{Var}(Y_n) = \mathrm{E}(Y_n^2) - \mathrm{E}(Y_n)^2 = \mathrm{E}\left(\prod_{i=1}^{n} X_i^2\right) - \mathrm{E}(Y_n)^2 = \left(\tfrac{91}{6}\right)^n - \left(\tfrac{7}{2}\right)^{2n}.$$

b) First note that (using independence)

$$\begin{aligned}
\mathrm{E}(\widehat{X}_i) &= p\\
\mathrm{Var}(\widehat{X}_i) &= p(1-p)\\
\mathrm{E}(Z_n) &= \mathrm{E}\left(\sum_{i=1}^{n}\widehat{X}_i\right) = \sum_{i=1}^{n}\mathrm{E}(\widehat{X}_i) = n\cdot p\\
\mathrm{Var}(Z_n) &= \mathrm{Var}\left(\sum_{i=1}^{n}\widehat{X}_i\right) = \sum_{i=1}^{n}\mathrm{Var}(\widehat{X}_i) = n\cdot p(1-p).
\end{aligned}$$

Let $Z \coloneqq Z_{300}$. Then $\mathrm{E}(Z) = 50$ and $\mathrm{Var}(Z) = 125/3$.

Markov Let $\hat{Z}_{300}$ be the number of die rolls that are not a 6 in 300 rolls. Clearly $\Pr(Z \le 10) = \Pr(\hat{Z}_{300} \ge 290)$ and $\mathrm{E}(\hat{Z}_{300}) = 250$. Then

$\Pr(Z \le 10) = \Pr(\hat{Z}_{300} \ge 290) \le \frac{250}{290} \approx 0.862$

Chebychev $\Pr(Z \le 10) \le \Pr(|Z - \mathrm{E}(Z)| \ge 40) \le \frac{125/3}{40^2} = \frac{5}{192} \approx 0.026$

Chernoff $\Pr(Z \le (1-c)\,\mathrm{E}(Z)) \le \exp\big(-\mathrm{E}(Z)\cdot c^2/2\big)$. With $c = 4/5$ we get $\Pr(Z \le 10) \le \exp(-50\cdot(\frac{4}{5})^2/2) \approx 1.125\cdot 10^{-7}$

Hoeffding $\Pr(Z \le \mathrm{E}(Z) - dn) \le \exp^{-2nd^2}$. With $n = 300$ and $d = 40/300 = 2/15$ we get $\Pr(Z \le 10) \le \exp(-2\cdot 300\cdot (2/15)^2) \approx 2.33\cdot 10^{-5}$

Chernoff gives the best bound here.

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

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

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

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

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

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 $H(X \mid Y) = H(X)$.

Solution:

For convenience, we omit the ranges from the sum operators and pull the -1 out of the logarithms.

a) It holds that

$$\begin{aligned}
H(X,Y) &= -\sum_{x,y} \Pr(X = x, Y = y) \cdot \log \big(\Pr(X = x, Y = y) \big) \\
&= -\sum_{x,y} \Pr(X = x \mid Y = y) \Pr(Y = y) \cdot \log \big(\Pr(X = x \mid Y = y) \Pr(Y = y) \big) \\
&= -\sum_{x,y} \Pr(X = x \mid Y = y) \Pr(Y = y) \cdot \log \big(\Pr(X = x \mid Y = y) \big) \\
&\qquad -\sum_{x,y} \Pr(X = x \mid Y = y) \Pr(Y = y) \cdot \log \big(\Pr(Y = y) \big) \\
&= H(X \mid Y) - \sum_{y} \Pr(Y = y) \cdot \log(\Pr(Y = y)) \underbrace{\sum_{x} \Pr(X = x \mid Y = y)}_{=1} \\
&= H(X \mid Y) + H(Y).
\end{aligned}$$

b) If X and Y are independent, it holds that

$$\begin{aligned} H(X \mid Y) &= -\sum_y \Pr(Y=y) \sum_x \Pr(X=x \mid Y=y) \cdot \log\big(\Pr(X=x \mid Y=y)\big) \\ &= -\sum_y \Pr(Y=y) \sum_x \Pr(X=x) \cdot \log\big(\Pr(X=x)\big) \\ &= -\sum_x \Pr(X=x) \cdot \log\big(\Pr(X=x)\big) = H(X). \end{aligned}$$

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

Solution:

Let S be the set of examples. Let Y be the random variable for the outcome and use letter A, I, S, C for the features. We have $\Pr(Y = 1) = \frac{2}{3}$ and $\Pr(Y = 0) = \frac{1}{3}$.

$$H(Y) = \frac{1}{3}\log 3 + \frac{2}{3}\log\frac{3}{2} \approx 0.918$$

a) First consider feature A:

$$H(Y_{A,<30}) = \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = 1$$

$$H(Y_{A,>40}) = \frac{3}{5}\log\frac{5}{3} + \frac{2}{5}\log\frac{5}{2} \approx 0.971$$

$$H(Y_{A,\text{30-40}}) = 1\log 1 = 0$$

So the information gain is

$$G(S, A) = H(Y) - \left(\frac{6}{15}H(Y_{A,<30}) + \frac{5}{15}H(Y_{A,>40}) + \frac{4}{15}H(Y_{A,\text{30-40}})\right) \approx 0.195$$

Next concsider feature I:

$$H(Y_{I,L}) = \frac{1}{4}\log 4 + \frac{3}{4}\log\frac{4}{3} \approx 0.811$$

$$H(Y_{I,M}) = \frac{1}{3}\log 3 + \frac{2}{3}\log\frac{3}{2} \approx 0.918$$

$$H(Y_{I,H}) = \frac{3}{5}\log\frac{5}{3} + \frac{2}{5}\log\frac{5}{2} \approx 0.971$$

So the information gain is

$$G(S,I) = H(Y) - \left(\frac{4}{15}H(Y_{I,L}) + \frac{6}{15}H(Y_{I,M}) + \frac{5}{15}H(Y_{I,H})\right) \approx 0.011$$

Next concsider feature S:

$$H(Y_{S,Y}) = \frac{1}{7}\log 7 + \frac{6}{7}\log\frac{7}{6} \approx 0.592$$

$$H(Y_{S,N}) = \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = 1$$

So the information gain is

$$G(S,S) = H(Y) - \left(\frac{7}{15}H(Y_{S,Y}) + \frac{8}{15}H(Y_{S,N})\right) \approx 0.109$$

Finally concsider feature C:

$$H(Y_{C,E}) = \frac{3}{7}\log\frac{7}{3} + \frac{4}{7}\log\frac{7}{4} \approx 0.985$$

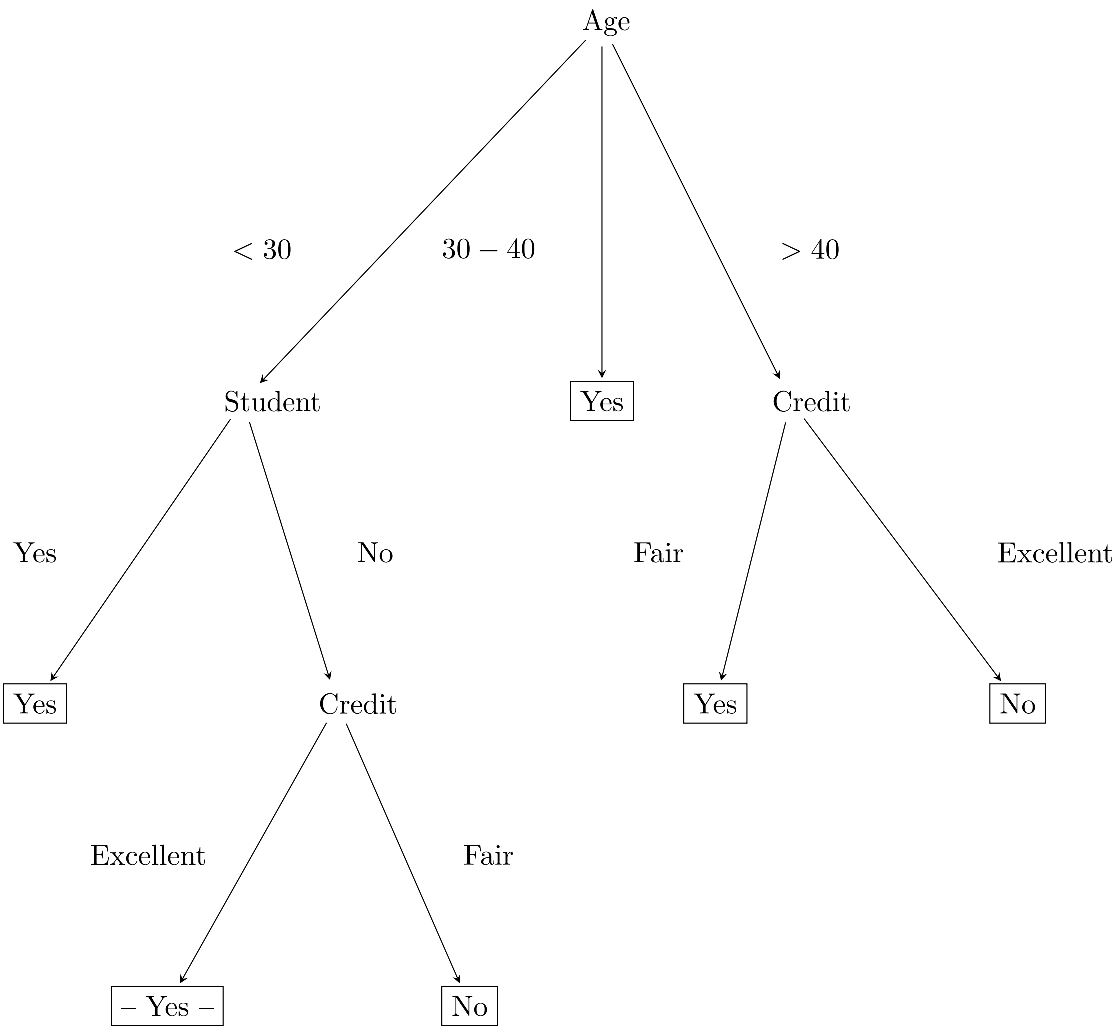
$$H(Y_{C,F}) = \frac{1}{4}\log 4 + \frac{3}{4}\log\frac{4}{3} \approx 0.811$$

So the information gain is

$$G(S,C) = H(Y) - \left(\frac{7}{15}H(Y_{C,E}) + \frac{8}{15}H(Y_{C,F})\right) \approx 0.026$$

So A is the best feature for the first split.

b) For the age group 30-40 we are done. For the age group > 40 the feature C finishes the job and thus, it also has the maximum information gain. So it remains the age group < 30 with its 6 examples. In this case the next best feature is S which correctly classifies all students. The remaining < 30 and not student can further be split, here Credit gives a higher information gain then Income. Lastly we have two candidates with different classifications and no more features to split on so we break the tie arbitrarily. If we choose No, we can also remove the last split on Credit.



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 $90\,\%$.

Solution:

a) Note with $m = 143$, $|\mathcal{H}| = 27$ and $\delta = 0.1$, from the Uniform Convergence theorem we know that for all ε that satisfy

$$143 \geq \tfrac{1}{2\varepsilon^2} \ln\left(\tfrac{2\cdot 27}{0.1}\right)$$

it holds that

$$\Pr(\forall h \in \mathcal{H}\colon\ |\mathrm{err}_T(h) - \mathrm{err}_{\mathcal{D}}(h)| \leq \varepsilon) > 0.9.$$

In particular, this holds for the smallest such ε, i. e. for

$$\varepsilon = \sqrt{\tfrac{1}{2\cdot 143} \cdot \ln\left(\tfrac{2\cdot 27}{0.1}\right)}.$$

This is roughly 0.1483, so in particular smaller than 0.15.

Then

$$\begin{aligned} 0.9 &< \Pr(\forall h \in \mathcal{H}\colon\ |\mathrm{err}_T(h) - \mathrm{err}_{\mathcal{D}}(h)| \leq 0.15) \\ &\leq \Pr(|\mathrm{err}_T(h_T) - \mathrm{err}_{\mathcal{D}}(h_T)| \leq 0.15) \\ &\leq \Pr(\mathrm{err}_{\mathcal{D}}(h_T) \leq \mathrm{err}_T(h_T) + 0.15) \\ &\leq \Pr(\mathrm{err}_{\mathcal{D}}(h_T) \leq 0.03 + 0.15). \end{aligned}$$

Thus, with probability greater than 90 %, it holds that $\mathrm{err}_{\mathcal{D}}(h_T) \leq 0.18$.

b) Here we can use the simple sample size bound with $\varepsilon = 0.01$ and $\delta = 0.1$. This gives the guarantee

$$\Pr(\forall h \in \mathcal{H}\colon \text{if } h \text{ consistent, then } \mathrm{err}_{\mathcal{D}}(h) \leq 0.01) > 0.9$$

if $m \geq \frac{1}{0.01} \cdot \ln\left(\frac{27}{0.1}\right) = 100 \cdot \ln(270) \approx 559.84$. Thus, 560 examples are sufficent.