Solution Sheet 13

Tutorial Exercises

T13.1 Combinatorics, Cross-Validation and Disorder

Suppose we have a set of $N = 10$ samples and want to run cross-validation (CV) on that sample.

(a) Should we call the set learning, training, test, or validation set?

(b) How many iterations does the CV have, if we use 5-fold CV?

(c) Assume that we have a binary class label and that the labels in the 5 folds are given by

$$(0, 1), (0, 0), (0, 0), (1, 0), (0, 0).$$

Calculate the average disorder in the validation sets.

(d) Instead, we let the CV iterate over *all* possible partitions of the proportion 8 : 2 samples. What is the number of iterations in that case?

(e) Suppose that two of the samples are absolutely identical. Determine the probability that these fall into the same validation set in at least one of the CV iterations for both flavors of CV.

Solution:

(a) It is the learning set (if considered, the test set is already kept apart from the learning set).

(b) 5-fold CV has 5 iterations. In each iteration, one of the folds plays the role of the validation set, the other four are used for training.

(c)

$$D(1, 1) = 1, D(2, 0) = 0.$$

Average disorder test sets:

$$\frac{1}{5}\left(D(0, 1) + D(0, 1) + D(2, 0) + D(2, 0) + D(2, 0)\right) = \frac{2}{5}$$

(d) If all samples are distinguishable, the number of iterations is given by the number of unique validation sets, being

$$\binom{10}{2} = \frac{10!}{2!8!} = 45.$$

(This type of CV is called *exhaustive*.)

(e) 5-fold CV: first, consider the randomness in the ordering of the 10 samples. This equals the number of permutations of a set of size $n = 10$, which is $n! = 10!$. Second, the fold are just formed by the indexes $(1, 2)$, $(3, 4)$, $(5, 6)$, $(7, 8)$, $(9, 10)$ in ascending order. But in how many of the permutations do the samples fall into the same fold? This happens exactly if the first of the two samples has an odd index and the second the subsequent even index, or if the second has an odd index and the first the subsequent even index. In each of these $2 \cdot 5$ cases, the remaining eight samples can be arranged arbitrarily in 8! ways. In total, we get that the probability that the two identical samples are in the same fold is

$$Pr(\text{two copies in a validation set}|\text{5-fold CV}) = \frac{2 \cdot 5 \cdot 8!}{10!} = \frac{1}{9}.$$

It is clear that it is impossible for the samples can fall only in one fold together in a partition, so "having both samples in at least one of the validation sets" is equivalent to "having both in exactly one validation set". Thus, the answer is 1/9. Exhaustive: we run over all possible $8 : 2$ partitions, which guarantees that we will put the two samples to the same validation set in one case. So,

$$Pr(\text{two copies in a validation set}|\text{exhaustive CV}) = 1.$$

T13.2 Independent Events

A Laplacian dice is thrown $n > 1$ times. Define the event A that both an even and an odd number have been thrown, and the event B that among all thrown numbers, there is at most one odd number.

(a) Determine $Pr(A \cap B)$.

(b) Write down the joint probability mass function of A and B for the case $n = 2$.

Solution:

(a) Clearly, $A \cap B$ is equivalent to having exactly one odd number, so

$$Pr(A \cap B) = \binom{n}{1} \frac{1}{2} \cdot \frac{1}{2^{n-1}}.$$

(This is the probability of having one success in n trials where the success is throwing an odd number.)

(b)

		$A = 0$	$A = 1$	
		only odd or even	mixed	
$B = 0$	two odd	$\frac{1}{2} \cdot \frac{1}{2}$	0	$\frac{1}{2} \cdot \frac{1}{2}$
$B = 1$	no or one odd	$\frac{1}{2} \cdot \frac{1}{2}$	$2 \cdot \frac{1}{2} \cdot \frac{1}{2}$	$\frac{1}{2} \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} \cdot \frac{1}{2}$
		$\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}$	$2 \cdot \frac{1}{2} \cdot \frac{1}{2}$	

The values for $f_{A,B}(a, b)$ are in the inner 2-by-2 matrix.

T13.3 Continuous Random Variables

Let Z be a continuous random variable satisfying

$$f_Z(z) = \sqrt{\frac{2}{\pi}} c e^{-cz^2}$$

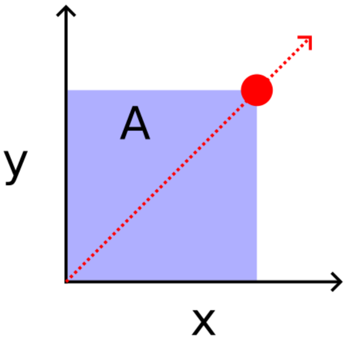
where $V_Z = \mathbb{R}_0^+$ and $c > 0$.

(a) Determine the constant c.

(b) Suppose an object that moves with constant speed v in the positive direction on the diagonal line $x = y$ in the 2D plane until it stops at some random time t. The probability density function of the stopping time is given by

$$f_T(t) = e^{-t}.$$

Which area (see sketch below) is enclosed at the expected stopping time? Is this equal to the expected value of the enclosed area?



Solution:

(a)

$$\int_0^\infty f_Z(z)dz = \frac{1}{2} \cdot \int_{-\infty}^\infty \sqrt{\frac{2}{\pi}} c e^{-cz^2} dz \overset{!}{=} 1.$$

$$\frac{1}{2} \cdot \int_{-\infty}^\infty \sqrt{\frac{2}{\pi}} e^{-u^2/2} \frac{\sqrt{2}}{\sqrt{c}} du$$

$$= \int_{-\infty}^\infty \sqrt{\frac{1}{2\pi}} e^{-u^2/2} sqrt\frac{2}{c} du$$

$$= 1 \cdot \frac{\sqrt{2}}{\sqrt{c}} \overset{!}{=} 1. \qquad \Rightarrow c = 2.$$

(substitution: $u = \sqrt{\frac{c}{2}}$ and comparison with standard normal distribution)

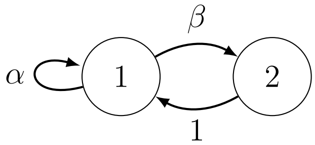
(b) Since t is exponentially distributed with parameter $\lambda = 1$, the expected stopping time equals $\frac{1}{\lambda} = 1$. The area enclosed at this time is $x(t) \cdot y(t) = v \cdot 1 \cdot v \cdot 1 = v^2$. This is not the same as the expected area, which would be calculated by

$$\mathbb{E}(A) = \int_0^\infty A(t) f_T(t) dt = \int_0^\infty v^2 t^2 e^{-t} dt$$

where we have t^2 in the integral (compared to t in the integral for $\mathbb{E}(T)$).

T13.4 Markov Chains, Estimation and Hypothesis

We work with a two-state Markov chain as shown below. Suppose $0 < \alpha, \beta < 1$.



(a) Does the Markov Chain always run into the same state distribution?

(b) Suppose we have sample $\boldsymbol{s} = (s_1, s_2, s_2, \ldots, s_n)$ of size n, denoting the sequence of states ($s_i \in \{0, 1\}$). Find the maximum likelihood estimator for α and β.

(c) We measured $\boldsymbol{s} = (1, 1, 1, 2, 1, 2, 1, 1, 2, 1)$,

$$\boldsymbol{s} = \begin{array}{llllllllll} (1, & 1, & 1, & 2, & 1, & 2, & 1, & 1, & 2, & 1, \\ 1, & 1, & 1, & 2, & 1, & 2, & 1, & 1, & 2, & 1, \\ 1, & 1, & 1, & 2, & 1, & 2, & 1, & 1, & 2, & 1, \\ 1, & 1, & 1, & 2, & 1, & 2, & 1, & 1, & 2, & 1, \\ 1, & 1, & 1, & 2, & 1, & 2, & 1, & 1, & 2, & 1) \end{array}$$

Does this data support the hypothesis $H_0 : \alpha = 2/5$ on the confidence level of 95%?

Solution:

(a) The Markov chain is irreducible and aperiodic, hence ergodic, and thus has a unique stationary distribution, independent of the starting state distribution.

(b) Define the Bernoulli random variable X for the transitions from state 1 with $X = 1$ if $1 \to 1$ and $X = 0$ if $1 \to 0$. Clearly, the success probability of X equals α. The maximum likelihood estimator for the success probability of a Bernoulli random variable is just the sample mean. So, we count the relative frequency of transitions $1 \to 1$ among all transitions starting at 1, and this is the MLE for α. The MLE for β is $1 - \alpha_{mle}$.

(c) We count 34 transitions starting at 1, out of which 19 lead back to 1. We do an approximate binomial test:

$$Z = \frac{X - n \cdot \alpha}{\sqrt{n\alpha\beta}} = \frac{19 - 34 \cdot 2/5}{\sqrt{34 \cdot 2/5 \cdot 3/5}} \approx 1.89.$$

The critical value of the corresponding two-sided test is $z_{1-\alpha/2} = z_{0.975} = 1.96$. Since $z < z_{0.975}$, we can keep the null hypothesis that $\alpha = 2/5$.

T13.5 Decision Trees

We have a dataset of size 1000 × 3 where the columns are two features x_1 and x_2, a target y, all of binary nature. In Python, we get the following information about the data.

```
np.unique(X,return_counts=True)
>> (array([(0,0,0),(0,0,1),(0,1,0),(0,1,1),(1,0,0),(1,0,1),(1,1,0),(1,1,1)]),
    array([50,150,150,50,100,200,200,100]))
```

(a) Determine which feature should be used first in a univariate decision tree, predicting y.

(b) What can we say about the measures precision, recall, F_1 score, and accuracy of the obtained level-1 tree? (We view class 1 as the positive outcome.)

Solution:

(a) Baseline entropy (focus on last component in the triples):

$$H_b = D(500,500) = 1.$$

Split on x_1: $x_1 = 0$: $D(200,200)$ $x_1 = 1$: $D(300,300) \Rightarrow \bar{H} = D(1,1) = 1 \Rightarrow$ information gain 0.

Split on x_2: $x_2 = 0$: $D(150,350) = D(3,7)$ $x_2 = 1$: $D(350,150) = D(7,3) \Rightarrow \bar{H} = 1/2 \cdot D(3,7) + 1/2 \cdot D(7,3) = D(3,7) < 1$. $\Rightarrow$ information gain positive.

$\Rightarrow$ select x_2 for the first split.

(b) The decision rule will be $x_2 = 0 \Rightarrow y = 1$ vs. $x_2 = 1 \Rightarrow y = 0$ by majority vote in the two classes. Thus,

$$TP = 350, FP = 150, TN = 350, FN = 150.$$

It follows that

$$\begin{aligned}
\text{precision} &= \frac{TP}{TP+FP} = \frac{350}{500} = \frac{7}{10}, \\
\text{recall} &= \frac{TP}{TP+FN} = \frac{350}{500} = \frac{7}{10}, \\
F_1 = \text{precision} = \text{recall} &= \frac{7}{10} \quad \text{since the two measures are equal,} \\
\text{accuracy} &= \frac{TP+TN}{N} = \frac{700}{1000} = \frac{7}{10}.
\end{aligned}$$

T13.6 Neural Networks

We work with the same dataset as in the previous task, but now aim at finding a small-sized neural network to predict the probability $Pr(Y=1)$ depending on X_1 and X_2.

(a) Is the dataset linearly separable with respect to the class label Y?

(b) State the correct outputs, based on the data.

(c) Construct a neural network that replicates the structure of a level-2 decision tree with features x_1 or x_2 in the levels, and outputs the estimated probability of $Pr(Y=1)$. You are free to choose the architecture (how many neurons, how they are connected, which weights and biases are used, which activation function is used). Explicitly write down the weights achieving the correct output.

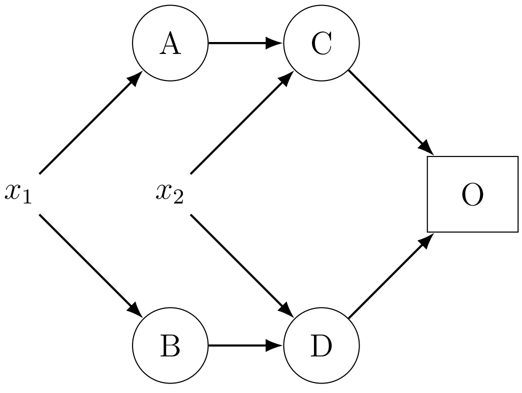
Solution:

(a) Clearly not, for each combination (x_1, x_2) in the input, we have samples of class 0 and of class 1. It is not even separable in a general sense if we use only these two features.

(b) The network should output the probability of class one, which we estimate from the relative frequencies.

input		output
x_1	x_2	$Pr(Y=1)$
0	0	$\frac{150}{200} = \frac{3}{4}$
0	1	$\frac{50}{200} = \frac{1}{4}$
1	0	$\frac{200}{300} = \frac{2}{3}$
1	1	$\frac{200}{300} = \frac{1}{3}$

(c)



$$w_{1,A} = -1, b_A = 0, w_{1,B} = 1, b_B = -1$$

$$w_{2,C} = 2/4, b_C = 1/4, w_{A,C} = 1, w_{2,D} = 1/3, b_D = 1/3, w_{B,D} = 1$$

$$w_{C,O} = 1, w_{D,O} = 1.$$

C and D use ReLU activation, A and B work with identity activation. C and D provide the output of the probabilities in the condition of $x_1 = 0$ and $x_1 = 1$ and depending on $x_2 = 0$ or $x_2 = 1$. A and B switch on or off C and D, depending on $x_1 = 0$ or $x_1 = 1$.

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4	0.99997	0.99997	0.99997	0.99997	0.99997	0.99997	0.99998	0.99998	0.99998	0.99998

Values of the cumulative distribution function Φ of the standard normal distribution. For example, $\Phi(1.55) \approx 0.93943$.