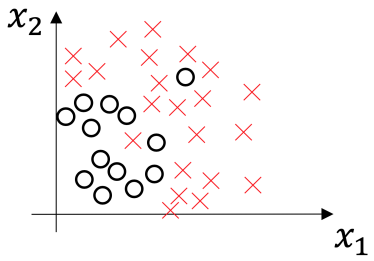
Resit Advanced Machine Learning
8 January 2019, 18.30–21.15

This exam consists of 5 problems, each consisting of several questions. All answers should be motivated, including calculations, formulas used, etc. The use of a calculator is not allowed.

Question 1: Bias-variance trade-off
Consider the following training data for solving the classification problem illustrated in the following figure. The circles (○) represent one class and the crosses (×) represent the other class.



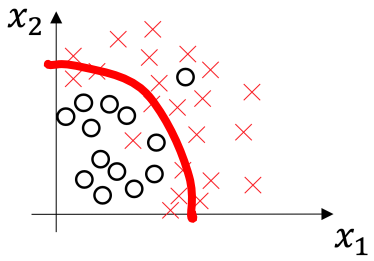
We attempt to solve this binary classification task with a machine learning model. In doing so, there is a trade-off between the bias and the variance of the model.

(a) Please explain what the bias-variance trade-off is.

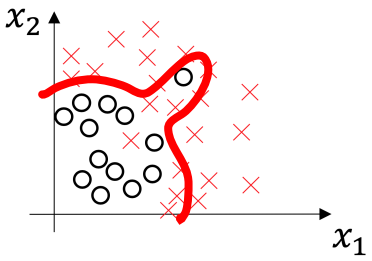
Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Variance is the variability of model prediction for a given data point or a value that tells us the spread of our data. If our model is too simple, then it may have high bias and low variance. As the model increases in parameters, it is going to have a high variance and low bias. Therefore, one needs to find a good balance.

The outcome of the machine learning model is a decision boundary that aims to separate the two classes. Draw a potential decision boundary for the following cases having the stated characteristics, and explain your answer.

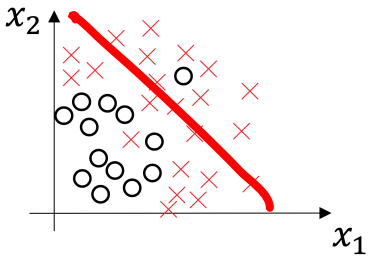
(b) low bias and low variance.



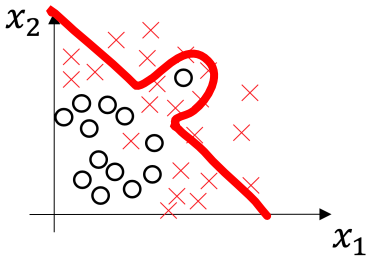
(c) low bias and high variance.



(d) high bias and low variance.



(e) high bias and high variance.



Question 2: Decision trees

For this question, we are going to try to determine whether a particular type of food is appealing based on the food's temperature, taste, and size. For this purpose, we are given data represented in the following table.

The impurity will be measured by the entropy. Note that for a given set S, the entropy is given by

$$E(S) = -p\log_2(p) - q\log_2(q),$$

with p the fraction of positive samples, and $q = 1 - p$ the fraction of negative samples.

Appealing	Temperature	Taste	Size
No	Hot	Salty	Small
No	Cold	Sweet	Large
No	Cold	Sweet	Large
Yes	Cold	Sour	Small
Yes	Hot	Sour	Small
No	Hot	Salty	Large
Yes	Hot	Sour	Large
Yes	Cold	Sweet	Small
Yes	Cold	Sweet	Small
No	Hot	Salty	Large

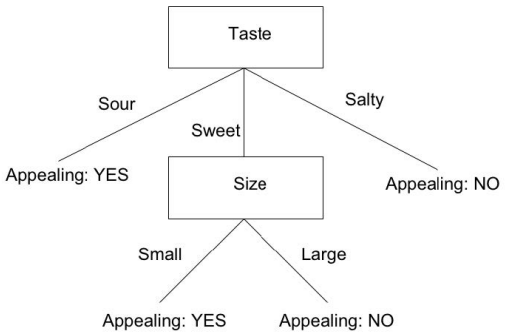
(a) What is the initial entropy of *Appealing*?

$-5/10 \cdot \log(5/10) - 5/10 \cdot \log(5/10)$.

(b) Assume that *Taste* is chosen for the root of the decision tree. What is the information gain associated with this attribute, i.e., how much decrease in initial entropy do you get upon splitting on this variable?

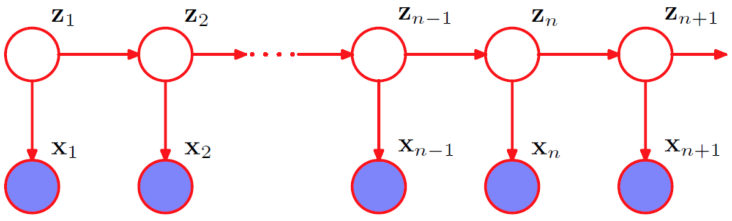
$[1 - 4/10 \cdot (2/4 \cdot \log(2/4) + 2/4 \cdot \log(2/4)) - 6/10 \cdot 0]$.

(c) Draw the full decision tree learned for this data (assuming that *Taste* is the root) without any pruning.



Question 3: Hidden Markov Models

Assume we are working with a Hidden Markov Model (HMM) given as in the picture below.



The latent variables $\mathbf{z}_i$ assume values in the set $\{1, 2, 3\}$, and the observations $\mathbf{x}_i$ also assume values in the set $\{1, 2, 3\}$. Recall that the model is fully defined by (π, A, ϕ). Let the initial distribution be given by $\pi = (1, 0, 0)$, the transition probability matrix A for the latent variables and the emission probability matrix ϕ by

$$A = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}, \qquad \phi = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

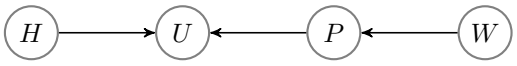
Remember that the probability that the latent variable moves from state i to state j is given by $A_{ij} = \mathbb{P}(\mathbf{z}_{n+1} = j \,|\, \mathbf{z}_n = i)$. Similarly, the probability that one observes j when the latent variable is in state i is given by $\phi_{ij} = \mathbb{P}(\mathbf{x}_n = j \,|\, \mathbf{z}_n = i)$.

(a) Suppose that we let HMM run with parameters (π, A, ϕ), and that we have observed the sequence $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_9) = (1, 3, 1, 2, 2, 3, 2, 3, 3)$. We are interested in the probabilities that explain the observations. For this, define the probability $\alpha_t(i) = \mathbb{P}(\mathbf{x}_1, \ldots, \mathbf{x}_t, \mathbf{z}_t = i)$. Thus, $\alpha_3(2) = \mathbb{P}(\mathbf{x}_1 = 1, \mathbf{x}_2 = 3, \mathbf{x}_3 = 1, \mathbf{z}_3 = 2)$. Fill in the following table.

t	$\alpha_t(1)$	$\alpha_t(2)$	$\alpha_t(3)$
1	1/2		
2		1/8	
3		1/32	
4			1/128
5			1/256
6			1/512
7			1/1024
8			1/2048
9			1/4096

Question 4: Graphical models

Consider the following graphical model.



(a) Which of the following independence statements follow from the above network structure? Please motivate your answer.

(a) H is independent of P
True

(b) W is independent of U given H
False

(c) H is independent of P given U
False

(b) Write the expression for the joint probability $\mathbb{P}(H, U, P, W)$ in its *reduced* factored form.

$\mathbb{P}(H, U, P, W) = \mathbb{P}(H)\mathbb{P}(W)\mathbb{P}(P \,|\, W)\mathbb{P}(U \,|\, H, P)$.

Question 5: Neural networks
The following questions pertain to the foundations of neural networks.

(a) Assume that you are not concerned with the training time of your neural network. When using a neural network it is best to include enough hidden units so that the training error can be reduced as much as possible. Can you explain why you agree or disagree with this statement?

False: minimizing training error will not necessarily minimize true test set error - overfitting may set in.

(b) A 1-layer neural network (i.e., there are no hidden layers) can only compute linear variations of AND, OR, and XOR. Can you explain why you agree or disagree with this statement?

False: A 1-layer neural network cannot compute an XOR.

(c) Please explain when it is possible to run a gradient descent algorithm. What is guaranteed by the algorithm, and what is not guaranteed?

The algorithm is guaranteed to converge to a local minimum of the error function. It is not guaranteed to converge to the global minimum (nor to a 'good' minimum, not always to the same minimum).

Question 6: Kaggle competition
These questions pertain to the Kaggle competition on flight delays.

(a) Which model(s) did you use in the competition? Elaborate on the choices on the model(s) and the corresponding (hyper)parameter(s).

(b) Which features (default and/or created) did you use in your model and why?

(c) Did overfitting play a role in the process of developing your solution? If so, how did it play a role and what did you do about it? If not, please explain why.

(d) How good do you think the performance of your model is? Please explain.

partial grade	1	2	3	4	5	6
(a)	6	6	9	6	6	4
(b)	3	6		3	6	4
(c)	3	6			6	4
(d)	3					3
(e)	6					

Final grade is: (sum of partial grades) / 10 + 1.0