Total possible marks: 100. Homeworks must be solved individually. Explain all your answers concisely. This set covers chapters 13–15 and 18 of the textbook *Introduction to Machine Learning*, 3rd. ed., by E. Alpaydin.

Exercise 1: kernel machines (18 points). Consider the XOR binary classification problem, with a training set $\{(\mathbf{x}_n, y_n)\}_{n=1}^N \subset \mathbb{R}^2 \times \{-1,+1\}$ given by $\{((\begin{smallmatrix}-1\\+1\end{smallmatrix}),+1), ((\begin{smallmatrix}+1\\-1\end{smallmatrix}),+1), ((\begin{smallmatrix}-1\\-1\end{smallmatrix}),-1), ((\begin{smallmatrix}+1\\+1\end{smallmatrix}),-1)\}$, which is not linearly separable. Construct a SVM to learn a nonlinear discriminant function as follows.

1. (2 points) Define a feature function $\mathbf{z} = \boldsymbol{\phi}(\mathbf{x}) = (\begin{smallmatrix}x_1\\x_1x_2\end{smallmatrix}) \in \mathbb{R}^2$. Evaluate it at each training point to obtain $\mathbf{z}_n = \boldsymbol{\phi}(\mathbf{x}_n)$ and plot them in the new space $\mathbf{z}$. Verify the points are now linearly separable.
2. (8 points) Find a linear discriminant $G(\mathbf{z}) = \mathbf{w}^T\mathbf{z} + w_0$ such that $\operatorname{sgn}(G(\boldsymbol{\phi}(\mathbf{x}_n))) = y_n\ \forall n$. Make sure this discriminant has the maximum margin, give the value of the margin, find the support vectors, and plot the separating hyperplane and the SVs. *Hint*: there is no need to solve any QP, just use geometric intuition. Explain your answer.
3. (6 points) Write the nonlinear discriminant $g(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + w_0$ in the original $\mathbf{x}$-space. Draw in the plane: the boundary $g(\mathbf{x}) = 0$, and the curves corresponding to each of the training points (i.e., the sets $\{\mathbf{x} \in \mathbb{R}^2\colon\ g(\mathbf{x}) = g(\mathbf{x}_n)\}$ for each $n = 1, \dots, 4$).
4. (2 points) Relate this result to the solution of the primal and dual QP for a kernel SVM. Specifically, write the kernel function $K(\mathbf{x}, \mathbf{y})$ that results from the choice of $\boldsymbol{\phi}$ above, the general form of the kernel SVM discriminant function $g(\mathbf{x})$, and guess what the values of the Lagrange multipliers α_n should be if we solved the dual QP. *Hint*: again, use geometric intuition and symmetry.

Exercise 2: kernel machines (18 points). Consider binary classification for $\mathbf{x} \in \mathbb{R}^2$ using the nonlinear discriminant function $g(\mathbf{x}) = x_1^2 + x_2^2 - 1$ and assigning a label $y = \operatorname{sgn}(g(\mathbf{x})) \in \{-1,+1\}$ to an instance $\mathbf{x}$.

1. (4 points) g splits the plane into two class regions. Draw them (or explain them in plain English). Are they linearly separable?

Consider a new feature function $\boldsymbol{\phi}\colon\ \mathbf{x} \in \mathbb{R}^2 \rightarrow \mathbb{R}^D$, which creates D features made up (in some way) of the original features x_1, x_2.

2. (10 points) Define a function $\boldsymbol{\phi}$ such that, using the new features $\mathbf{z} = \boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^D$, we can define a linear discriminant function $G(\mathbf{z}) = \mathbf{w}^T\mathbf{z}$ that is equivalent to g, i.e., $G(\boldsymbol{\phi}(\mathbf{x})) = g(\mathbf{x})$. Relate this to kernel functions and support vector machines.
3. (4 points) Define what the kernel function $K(\cdot,\cdot)$ would be that corresponds to your feature function $\boldsymbol{\phi}$.

Exercise 3: graphical models (6 points). Consider the following two graphical models defined on binary random variables, given by their joint distributions:

$$p(X,Y,Z) = p(Z|X,Y)\,p(Y|X)\,p(X) \qquad \text{and} \qquad p(X,Y,Z) = p(Z)\,p(Y|Z)\,p(X)$$

For each of them:

1. (4 points) Prove that $\sum_{X,Y,Z} p(X,Y,Z) = 1$.
2. (2 points) Draw the graphical model.

Exercise 4: graphical models (21 points).

Consider 3 binary random variables with joint distribution given by the table.

1. (14 points) Evaluate the following distributions: $p(X)$, $p(Y)$, $p(X,Y)$, $p(X|Z)$, $p(Y|Z)$, $p(Z|X)$, and $p(X,Y|Z)$.
2. (4 points) Show by direct evaluation that this distribution has the property that X and Y are marginally dependent, i.e., $P(X,Y) \neq p(X)\,p(Y)$ (for all values of X and Y); but that they become independent when conditioned on Z, i.e., $p(X,Y|Z) = p(X|Z)\,p(Y|Z)$ for all values of X, Y and Z.
3. (3 points) Show by direct evaluation that $p(X,Y,Z) = p(X)\,p(Z|X)\,p(Y|Z)$. Draw the corresponding directed graph for this graphical model.

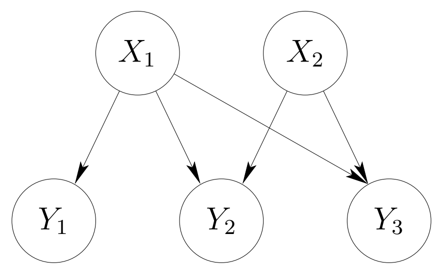
X	Y	Z	$p(X,Y,Z)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

Show your work in all cases.

Exercise 5: graphical models (21 points).

Consider a graphical model defined on binary random variables (where variables X_i correspond to diseases and variables Y_j to symptoms), given by the following diagram and conditional probability tables at each node.

Note: in the tables and the questions, the notation "$p(Y_3|\overline{X}_1, X_2)$" means "$p(Y_3 = 1|X_1 = 0, X_2 = 1)$", etc.

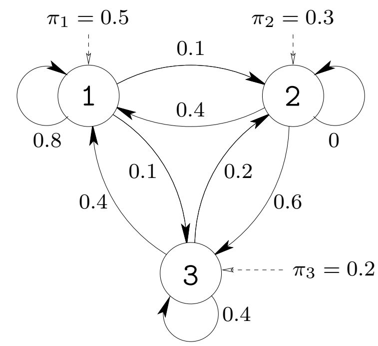


conditional probability tables at each node				
X_1 ("flu")	X_2 ("hayfever")	Y_1 ("fever")	Y_2 ("headache")	Y_3 ("fatigue")
$p(X_1) = 0.4$	$p(X_2) = 0.1$	$p(Y_1|X_1) = 0.8$	$p(Y_2|X_1, X_2) = 0.9$	$p(Y_3|X_1, X_2) = 0.7$
		$p(Y_1|\overline{X}_1) = 0.1$	$p(Y_2|X_1, \overline{X}_2) = 0.8$	$p(Y_3|X_1, \overline{X}_2) = 0.7$
			$p(Y_2|\overline{X}_1, X_2) = 0.7$	$p(Y_3|\overline{X}_1, X_2) = 0.3$
			$p(Y_2|\overline{X}_1, \overline{X}_2) = 0.1$	$p(Y_3|\overline{X}_1, \overline{X}_2) = 0.1$

1. (3 points) Give the expression of the joint distribution it defines over all the variables.
2. (18 points) Calculate the value of the following probabilities:
   (a) $p(\overline{Y}_2|X_1, \overline{X}_2)$.
   (b) $p(Y_1, Y_3|\overline{X}_1, \overline{X}_2)$.
   (c) $p(Y_1|X_2)$.
   (d) $p(Y_1)$.
   (e) $p(X_1|Y_1, \overline{Y}_2)$.
   (f) $p(X_2|\overline{Y}_1, Y_2, \overline{Y}_3)$.

Show your work in all cases.

Exercise 6: discrete Markov models (9 points).



Consider the discrete Markov model given by the diagram.

1. (3 points) Give the set of states of this discrete Markov model, its transition matrix $\mathbf{A}$ and its vector of initial state probabilities $\boldsymbol{\pi}$.

2. (6 points) Compute the probability of the following sequences: `12123`, `221`, `3`.

Show your work in all cases.

Exercise 7: discrete Markov models (7 points). Consider a discrete Markov model with two states `a`, `b`.

1. (5 points) We have a training set consisting of the following sequences: `bbbaa`, `baaaa`, `bbbbb`, `bbbba`. Give the maximum likelihood estimate of the parameters $(\mathbf{A}, \boldsymbol{\pi})$.

2. (2 points) Draw the corresponding discrete Markov model as in the previous exercise.

Show your work in all cases.