Total possible marks: 100. Homeworks must be solved individually. Explain all your answers concisely. This set covers chapters 10–18 of the textbook *Introduction to Machine Learning*, 3rd. ed., by E. Alpaydin.

Exercise 1: linear classifier (10 points). Consider a binary linear classifier $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ with $\mathbf{w} = \binom{-3}{4}$ and $w_0 = -12$, where $\mathbf{x} \in \mathbb{R}^2$. Let class 1 be its positive side ($g(\mathbf{x}) > 0$) and class 2 its negative side ($g(\mathbf{x}) < 0$).

1. (4 points) Sketch the decision boundary in $\mathbb{R}^2$. Compute the points at which it intersects the coordinate axes. Indicate which is the positive side of the boundary (class 1).
2. (4 points) Compute the signed distance of the following points to the decision boundary: the origin; $\binom{-1}{3}$; $\binom{4}{6}$. Classify those points.
3. (2 points) Give a vector $\mathbf{u} \in \mathbb{R}^2$ that is parallel to the decision boundary and has norm 1.

Exercise 2: multilayer perceptrons (8 points). Construct manually a perceptron that calculates the NAND of its two inputs. That is, given a training set

$$\{(\mathbf{x}_n, y_n)\}_{n=1}^N = \left\{ \left(\tbinom{0}{0}, 1\right),\ \left(\tbinom{0}{1}, 1\right),\ \left(\tbinom{1}{0}, 1\right),\ \left(\tbinom{1}{1}, 0\right) \right\}$$

of 2D points in two classes $\{0, 1\}$, give numerical values of the perceptron's parameters that solve this classification problem.

Exercise 3: properties of the logistic and tanh **functions (10 points).** Consider the logistic function $\sigma(x) = \frac{1}{1+e^{-x}} \in (0, 1)$ for $x \in \mathbb{R}$. Prove the following properties:

1. (2 points) Inverse of logistic: $\sigma^{-1}(y) = \text{logit}(y) = \log\left(\frac{y}{1-y}\right) \in (-\infty, \infty)$ for $y \in (0, 1)$.
2. (2 points) Derivative of logistic: $\frac{d\sigma(x)}{dx} = \sigma'(x) = \sigma(x)(1 - \sigma(x))$.
3. (1 points) $\sigma(x) + \sigma(-x) = 1$.

Consider now the hyperbolic tangent $\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}} \in (-1, 1)$ for $x \in \mathbb{R}$. Work out the expression for:

1. (2 points) The inverse of tanh.
2. (2 points) The derivative of tanh, using the value of tanh itself.
3. (1 points) $\tanh(x) + \tanh(-x)$.

Exercise 4: RBF networks (20 points). Consider a Gaussian radial basis function (RBF) network $\mathbf{f}\colon \mathbb{R}^D \rightarrow \mathbb{R}^{D'}$ that maps input vectors $\mathbf{x} \in \mathbb{R}^D$ to output vectors $\mathbf{y} \in \mathbb{R}^{D'}$:

$$\mathbf{f}(\mathbf{x}) = \sum_{h=1}^{H} \mathbf{w}_h e^{-\frac{1}{2}\left\|\frac{\mathbf{x}-\boldsymbol{\mu}_h}{\sigma}\right\|^2} \qquad \text{or, elementwise:} \quad f_e(\mathbf{x}) = \sum_{h=1}^{H} w_{he} e^{-\frac{1}{2\sigma^2}\sum_{d=1}^{D}(x_d-\mu_{hd})^2} \quad e = 1,\dots,D'$$

where the RBF network parameters are the weight vectors $\{\mathbf{w}_h\}_{h=1}^H \subset \mathbb{R}^{D'}$, the centroids $\{\boldsymbol{\mu}_h\}_{h=1}^H \subset \mathbb{R}^D$ and the bandwidth $\sigma > 0$. We want to train $\mathbf{f}$ in a regression setting by minimizing the least-squares error with a fixed regularization parameter $\lambda \geq 0$, given a training set $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$:

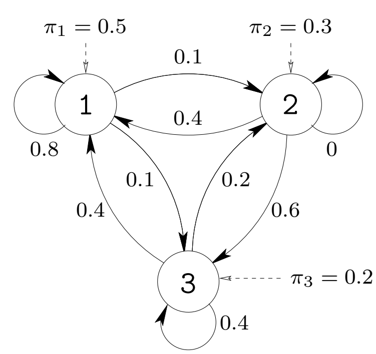
$$E\left(\{\mathbf{w}_h, \boldsymbol{\mu}_h\}_{h=1}^H, \sigma\right) = \sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n)\|^2 + \lambda \sum_{h=1}^{H} \|\mathbf{w}_h\|^2 = \sum_{n=1}^{N}\sum_{e=1}^{D'} (y_{ne} - f_e(\mathbf{x}_n))^2 + \lambda \sum_{h,e=1}^{H,D'} w_{he}^2. \tag{1}$$

A simple but approximate way to train the RBF network is by fixing the value of its bandwidth $\sigma > 0$ (this value is eventually cross-validated) and its centroids $\{\boldsymbol{\mu}_h\}_{h=1}^H$ (e.g. to a random subset of training points, or to the result of running k-means on the training set), and then optimizing eq. (1) over the weights (which results in a linear system).

Instead, we wish to train the RBF network parameters by gradient descent, as with multilayer perceptrons.

1. (15 points) Using the chain rule, compute the gradients of E in eq. (1) wrt the parameters:
   (a) The weights $\{\mathbf{w}_h\}_{h=1}^H$: $\frac{\partial E}{\partial w_{he}} = \dots$ for $h = 1,\dots,H$ and $e = 1,\dots,D'$.
   (b) The centroids $\{\boldsymbol{\mu}_h\}_{h=1}^H$: $\frac{\partial E}{\partial \mu_{hd}} = \dots$ for $h = 1,\dots,H$ and $d = 1,\dots,D$.
   (c) The bandwidth σ: $\frac{\partial E}{\partial \sigma} = \dots$
2. (5 points) What would be a good initialization for these parameters (to start gradient descent)?

Exercise 5: 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`.

Exercise 6: 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.

Exercise 7: graphical models (6 points). Consider the following two graphical models defined on binary random variables, given by their joint distributions:

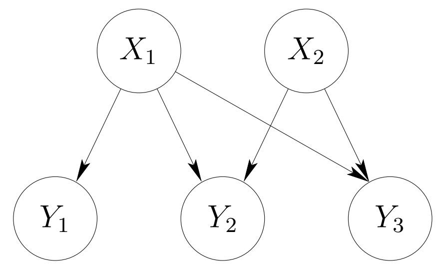
$$p(X,Y,Z) = p(Z|X,Y)\,p(Y|X)\,p(X) \qquad \text{and} \qquad p(X,Y,Z) = p(Z)\,p(Y|Z)\,p(X)$$

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 8: 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) $p(X_2|\overline{Y}_1, Y_2, \overline{Y}_3)$.

Exercise 9: ensemble learning (9 points). Consider the setting of regression from input vectors $\mathbf{x} \in \mathbb{R}^D$ to a single real output $y \in \mathbb{R}$. Imagine we have trained L learners $f_1, \dots, f_L$: $\mathbb{R}^D \to \mathbb{R}$ in some way (e.g. each on a bootstrapped sample from a training set). We combine them using their average: $f(\mathbf{x}) = \frac{1}{L}\sum_{l=1}^{L} f_l(\mathbf{x})$. What kind of model is the resulting f in each of the following cases? Be as specific as possible. *Hint*: we give the answer to the first case below.

1. (0 points) If $f_1, \dots, f_L$ are polynomials of degree q.
   Answer: f is another polynomial of degree q, whose coefficients are equal to the average of the corresponding coefficients in $f_1, \dots, f_L$.
2. (3 points) If $f_1, \dots, f_L$ are Gaussian RBF networks each with H centroids.
3. (3 points) If $f_1, \dots, f_L$ are linear regressors.
4. (3 points) If $f_1, \dots, f_L$ are MLPs each with a single hidden layer of H sigmoidal units and an output linear unit.