1 Solution: Multilayer Perceptron

(a)

$$a_k(\boldsymbol{x}) = \sum_{j=1}^{4} w_{kj}^{(1)} x_j + b_k, \quad z_k(\boldsymbol{x}) = g(a_k(\boldsymbol{x})), \quad k = 1, \dots, 5,$$
$$y(\boldsymbol{x}) = \sum_{k=1}^{5} w_k^{(2)} z_k(\boldsymbol{x}) + c.$$

It is OK to not write the “$(\boldsymbol{x})$” arguments.

(b) Each hidden unit has 5 parameters, and the output unit has 6 parameters, so $5 \cdot 5 + 6 = 31$ parameters in total.

(c) Problem: Overfitting. Guard: Split data into training and validation set. Monitor validation set error. Stop (early) once this error starts to increase or levels off.
One point for overfitting. Two points only for a complete description of early stopping.
If somebody correctly describes regularization: full points as well.

(d) This error function is not bounded below, since $z^5 \to -\infty$ as $z \to -\infty$. It is driven to $-\infty$ by making the $y(\boldsymbol{x}_i)$ ever larger negative.

(e) The gradient component is computed by the chain rule:

$$\frac{\partial E_i}{\partial w_{54}^{(1)}} = \frac{\partial E_i}{\partial y} \frac{\partial y}{\partial z_5} \frac{\partial z_5}{\partial a_5} \frac{\partial a_5}{\partial w_{54}^{(1)}} = (y(\boldsymbol{x}_i) - t_i)^5 w_{15}^{(2)} g'(a_5(\boldsymbol{x}_i))(\boldsymbol{x}_i)_4.$$

Since $g'(a) = 1 - g(a)^2$, we can replace $g'(a_5(\boldsymbol{x}_i))$ by $1 - z_5(\boldsymbol{x}_i)^2$. Last point not required for full points. Deductions for wrong notation (say, dropping the i index).

2 Solution: Decision Theory

(a) $p(x_1, x_2) = p(x_1)p(x_2) = \mathrm{I}_{\{0 \leq x_1 \leq 1\}}\mathrm{I}_{\{0 \leq x_2 \leq 1\}}$ (independent and uniform). The conditional probability is

$$P(t = 1|x_2) = \int P(t = 1|x_1, x_2)\, dx_1 = \begin{cases} \frac{2}{3} & x_2 \leq 1/3 \\ 1 & 1/3 < x_2 \leq 2/3 \\ \frac{1}{3} & x_2 > 2/3 \end{cases}$$

$f^*(x_2) = 0$ iff $P(t = 1|x_2) < 1/2$, so

$$f^*(x_2) = 0 \quad \Leftrightarrow \quad x_2 > 2/3.$$

One point for $P(t = 1|x_2)$, one point for $f^*(x_2)$. If $P(t = 1|x_2)$ wrong, but $f^*(x_2)$ correct given wrong distribution: one point.

(b) The shaded area $\mathcal{A}$ consists of the two squares $[0, 1/3] \times [2/3, 1]$ and $[0, 1/3] \times [0, 1/3]$. The Bayes error is

$$\int \mathrm{I}_{\{(x_1,x_2)\in\mathcal{A}\}} p(x_1)p(x_2)\, dx_1 dx_2 = (1/3)^2 + (1/3)^2 = 2/9.$$

One point for correct regions, one point for number. If (a) wrong, but answer here correct given (a): One point (instead of two).

3 Solution: Perceptron

(a) First gap: $t_i y(x_{i1}, x_{i2}) \leq 0$.
Second gap: $\boldsymbol{w} \leftarrow \boldsymbol{w} + t_i \boldsymbol{x}_i$.
Third gap: $t_j y(x_{j1}, x_{j2}) > 0$ for all $j = 1, \dots, n$.
Deductions (half points) for small things ("$<$" instead of "$\leq$"). No points for answer which does not lead to a working algorithm.

(b) The dataset must be linearly separable. There must exist some weight vector $\boldsymbol{w}$ so that $t_i \boldsymbol{w}^T \boldsymbol{x}_i > 0$ for all $i = 1, \dots, n$.
Term "linearly separable" enough to get the point.

(c) $\boldsymbol{x}_1$ is wrong, so $\boldsymbol{w}_1 = [2, 2, 1]^T$. The separating line passes through $[0, -1/2]$ (hint) and has slope -1. Given $\boldsymbol{w}_1$, the point $\boldsymbol{x}_2$ is correct, so $\boldsymbol{w}_2 = [2, 2, 1]^T$. But $\boldsymbol{x}_3$ is wrong, so $\boldsymbol{w}_3 = [1, 3, 0]^T$. The separating line passes through the origin now and has slope $-1/3$. It separates the dataset.
One point for correct solution (line and weight vector) for every step. In general, half point for each correct element. Do not make them pay for an earlier error, if correct from there on.

(d) It can help to normalize the datapoints $\boldsymbol{x}_i$ to unit norm. $\|\boldsymbol{x}_4\| = (9+16+1)^{1/2} = \sqrt{26}$. After preprocessing:

$$\frac{1}{\sqrt{26}}[3, 4, 1]^T.$$

4 Solution: Kernel Methods

(a) We need to show that

$$K_\varepsilon(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\phi}_\varepsilon(\boldsymbol{x})^T \boldsymbol{\phi}_\varepsilon(\boldsymbol{y}).$$

The feature mapping is

$$\boldsymbol{\phi}_\varepsilon(\boldsymbol{x}) = \left[\varepsilon^2, x_1^2, x_2^2, \sqrt{2}\varepsilon x_1, \sqrt{2}\varepsilon x_2, \sqrt{2}x_1 x_2\right]^T .$$

Another ordering is OK as well, of course.
It is enough to state the feature map.

(b) Using the ordering above, $\tilde{w}_j = w_j$ for $j = 2, 3, 6$, while

$$\tilde{w}_1 = \varepsilon^2 w_1,\ \tilde{w}_4 = \varepsilon w_4,\ \tilde{w}_5 = \varepsilon w_5.$$

Half point if solution is correct, based on wrong (a).

(c) This statement is wrong. It is true that the function spaces are the same for all kernels K_ε, but the SVM criterion includes a regularization term $\|\boldsymbol{w}\|^2$. If we fix a function $y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}_\varepsilon(\boldsymbol{x}) = \tilde{\boldsymbol{w}}^T \boldsymbol{\phi}_1(\boldsymbol{x})$, say, then $\|\boldsymbol{w}\|^2 \neq \|\tilde{\boldsymbol{w}}\|^2$, so the same function is regularized differently, which means that the minimizing functions will in general be different for the same data, but different kernels.
One point for mentioning "regularization", but not giving a proper argument.

5 Solution: Naive Bayes Classification

(a) Just count the number of ones per column, divide by number of rows.

m	1	2	3	4	5	6	7
$\hat{p}_m^{(0)}$	2/7	2/7	2/7	5/7	6/7	2/7	6/7
$\hat{p}_m^{(1)}$	6/8	5/8	2/8	3/8	2/8	3/8	5/8

And $\hat{P}(t=0) = 7/15$, $\hat{P}(t=1) = 8/15$. Point only for correct solution.

(b)

$$P(t=0)P(\boldsymbol{x}_*|t=0) = (7/15)\cdot 7^{-7} 5\cdot 5\cdot 2\cdot 5\cdot 6\cdot 5\cdot 6 = 7^{-6}5^4 2^3 3^2/15,$$
$$P(t=1)P(\boldsymbol{x}_*|t=1) = (8/15)\cdot 8^{-7} 2\cdot 3\cdot 2\cdot 3\cdot 2\cdot 5\cdot 5 = 8^{-6}5^2 2^3 3^2/15.$$

Only the ratio matters: $7^{-6}5^2$ for $t=0$ vs. 8^{-6} for $t=1$.

$$\hat{P}(t=0|\boldsymbol{x}_*) = \frac{7^{-6}5^2}{7^{-6}5^2 + 8^{-6}} = \frac{1}{1+5^{-2}(7/8)^6},$$

so that $a = 1/25$, $b = 7/8$.
One point each for each joint distribution (if stated), final point for correct a, b. Two points for correct a, b without any derivation. Award points if errors are only due to errors in (a).

(c) The probabilities $P(\boldsymbol{x}_*, t)$ are products of independent parts, one for each feature, so just have to fix the contributions from $m = 2, 3$. $\hat{p}_m^{(0)}$ from 2/7 to 5/7 ($m = 2, 3$), $\hat{p}_2^{(1)}$ from 5/8 to 3/8, $\hat{p}_3^{(1)}$ from 2/8 to 6/8. In $P(\boldsymbol{x}_*, t=0)$, $5\cdot 2$ becomes $2\cdot 5$, so nothing changes. In $P(\boldsymbol{x}_*, t=1)$, $3\cdot 2$ becomes $5\cdot 6$, so $P(\boldsymbol{x}_*, t=1)$ is multiplied by 5. The part ab^6 in (c) is equal to $P(\boldsymbol{x}_*, t=1)/P(\boldsymbol{x}_*, t=0)$, so that $d = b = 7/8$, $c = 5a = 1/5$.
No points for wrong solution, unless error due only to errors in (a).

6 Solution: Maximum Margin Perceptron and SVM

(a) Support vectors are 1, 4 (class $+1$) and 8 (class -1). The maximum margin separating line is given by the equation $x_2 = x_1$, it passes through the origin.

(b) $\boldsymbol{w}_0$ is proportional to $[1, -1]^T$, so

$$\boldsymbol{w}_0 = \frac{1}{\sqrt{2}}[1, -1]^T, \quad b = 0, \quad \kappa = \frac{3}{2}\sqrt{2}.$$

$b = 0$ because the separating line passes through the origin. The margin κ is the length of the vector $(3/2)[1, -1]^T$ for the orthogonal projection of point 8 onto the line.
One point for each correct subanswer. If students swap x_1, x_2: No deduction (axes are not labeled).

(c) Three support vectors: 1, 4, 8.

(d) Denote the datapoints by $\boldsymbol{x}_1$, $\boldsymbol{x}_4$, $\boldsymbol{x}_8$. The mirror image of $\boldsymbol{x}_8$ w.r.t. the separating line is the midpoint $(\boldsymbol{x}_1 + \boldsymbol{x}_4)/2$, so we have that

$$\frac{1}{2}(\boldsymbol{x}_1 + \boldsymbol{x}_4) = \boldsymbol{x}_8 + 2\kappa\boldsymbol{w}_0,$$

so that

$$\boldsymbol{w}_0 = \frac{1}{3\sqrt{2}}\left(\boldsymbol{x}_1/2 + \boldsymbol{x}_4/2 - \boldsymbol{x}_8\right).$$

7 Solution: Least Squares Regression

(a)

$\langle x \rangle = \frac{1}{n}\sum_{i=1}^{n} x_i \quad = 0$

$\langle x^2 \rangle = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \quad = 2$

$\langle xt \rangle = \frac{1}{n}\sum_{i=1}^{n} x_i t_i \quad = 2$

$\langle t \rangle = \frac{1}{n}\sum_{i=1}^{n} t_i \quad = -2/5$

(b)

$$\frac{\partial E}{\partial a} = \sum_i (ax_i + b - t_i)x_i = n\left(a\langle x^2\rangle + b\langle x\rangle - \langle tx\rangle\right),$$

$$\frac{\partial E}{\partial b} = \sum_i (ax_i + b - t_i) = n\left(a\langle x\rangle + b - \langle t\rangle\right).$$

Correct solution, but no derivation: One point only.

(c) Subtract $\partial E/\partial b$ times $\langle x \rangle$ from $\partial E/\partial a$ gives

$$a_* = \frac{\langle tx\rangle - \langle t\rangle\langle x\rangle}{\langle x^2\rangle - \langle x\rangle^2}, \quad b_* = a_*\langle x\rangle + \langle t\rangle.$$

This gives $a_* = 2/2 = 1$ and $b_* = -2/5$. The line has equation $t = x - 2/5$, it has slope 1 and passes the t axis at $t = -2/5$.
One point for a_*, b_*, one point for correct line (slope should be one, offset in correct ballpark).

(d) The answer is "decreases or stays the same". The new model contains the old model (set $c = 0$), so the error cannot grow and will typically decrease.
No point if explanation is missing, or if wrong box is ticked. A slightly incomplete answer like "model has more parameters" is OK.

8 Solution: Principal Components Analysis

(a) We compute the sample covariance matrix

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T, \quad \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i.$$

The first principal components direction $\boldsymbol{u}$ is the unit norm eigenvector of $\boldsymbol{S}$ corresponding to the largest eigenvalue of $\boldsymbol{S}$.

Alternative, also correct: Compute $\boldsymbol{S}$, then

$$\boldsymbol{u} = \operatorname*{argmax}_{\boldsymbol{v}:\,\|\boldsymbol{v}\|=1} \boldsymbol{v}^T\boldsymbol{S}\boldsymbol{v}.$$

No point if mean is not subtracted, or if “largest” or “leading” is not stated.

(b) In the following solution, the horizontal is the first, the vertical the second axis. But it is fine to use the other ordering as well.

One point for determining the covariance matrix $\boldsymbol{S}$. First, the empirical mean is zero by symmetry, so that $\boldsymbol{S} = (1/6)\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^T$. One way to compute $\boldsymbol{S}$ is through the data matrix. Here, only half of the points have to be used (faster):

$$\boldsymbol{S} = \frac{2}{6}\boldsymbol{X}^T\boldsymbol{X} = \frac{1}{3}\begin{bmatrix} 14 & 4 \\ 4 & 14 \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 3 & 3 \\ 1 & -1 \\ 2 & -2 \end{bmatrix}.$$

Good students may exploit the symmetry to directly determine the eigendecomposition. If $\boldsymbol{v} = [1,-1]^T$, $\mathbf{1} = [1,1]^T$, four points are $\pm\boldsymbol{v}$, $\pm 2\boldsymbol{v}$, two are $\pm 3\mathbf{1}$, so we have

$$\boldsymbol{S} = \frac{1}{6}\left(2(1+2^2)\boldsymbol{v}\boldsymbol{v}^T + 2\cdot 3^2\mathbf{1}\mathbf{1}^T\right) = \frac{5}{3}\boldsymbol{v}\boldsymbol{v}^T + 3\mathbf{1}\mathbf{1}^T.$$

The remaining two points are for demonstrating that $\boldsymbol{u} = [1,1]^T/\sqrt{2}$ is the first PC direction, with corresponding eigenvalue $\lambda = 6$. The brute force solution is to test both $\boldsymbol{u}$ and $[1,-1]^T/\sqrt{2}$ (hint). They are both eigenvectors, the eigenvalue for the latter is $10/3 < 6$, so $\boldsymbol{u}$ is the *first* PC direction. The demonstration that $\boldsymbol{u}$ is an eigenvector with $\lambda = 6$ *alone* results in 1 point only (of the two).

Smart students may guess that $\boldsymbol{u}$ is the first PC direction, compute $\lambda = 6$, then $\operatorname{tr}\boldsymbol{S} = 28/3 < 2\cdot 6$. Since the trace is the sum of eigenvalues, the second eigenvalue must be smaller than 6.

9 Solution: Maximum Likelihood

(a) The likelihood function is

$$\prod_{i=1}^{n} p(x_i|\gamma) = (\gamma^3/2)^n \prod_i x_i^2 \exp\left(-\gamma \sum_i x_i\right).$$

Providing the log likelihood function is also OK. It is also OK to drop multiplicative (or additive for log) constants (independent of γ).

(b) We maximize the log likelihood function w.r.t. γ. Here, we can immediate drop additive constants which do not depend on γ. If $\bar{x} = n^{-1}\sum_i x_i$, then

$$L(\gamma) = n \log \gamma^3 - \gamma n \bar{x} = n\left(3 \log \gamma - \gamma \bar{x}\right).$$

Then,

$$\frac{dL}{d\gamma} = n\left(3/\gamma - \bar{x}\right) = 0 \quad \Leftrightarrow \quad 1/\hat{\gamma} = \bar{x}/3.$$

Clearly, the derivative is a decreasing function of γ, so this must be the maximum point. Half a point deduction if only the stationary point is determined.

(c)

$$p_{\text{mixture}}(x) = \sum_{k=1}^{4} P(\omega_k)\frac{1}{2}\gamma_k^3 x^2 e^{-\gamma_k x}.$$

(d) There are $3 + 4 = 7$ independent parameters, only 3 for $P(\omega_k)$.

(e) First, we compute $p(x_i, \omega_k)$, dropping any common multiplicative constant independent of k:

$$p(x_i, \omega_1) = C8e^{-2x_i}, \quad p(x_i, \omega_2) = p(x_i, \omega_3) = p(x_i, \omega_4) = Ce^{-x_i}.$$

For $x_i = \log 2$: $e^{-x_i} = 1/2$, $e^{-2x_i} = (1/2)^2 = 1/4$, so that

$$P(\omega_1|x_i) = \frac{p(x_i, \omega_1)}{\sum_{k=1}^{4} p(x_i, \omega_k)} = \frac{2}{2 + 3 \cdot 1/2} = \frac{4}{7}.$$

(f) First,

$$\frac{\partial \log p(x_i)}{\partial \gamma_1} = \frac{1}{p(x_i)}\frac{\partial p(x_i|\omega_1)P(\omega_1)}{\partial \gamma_1} = \frac{p(x_i|\omega_1)P(\omega_1)}{p(x_i)}\frac{\partial \log p(x_i|\omega_1)}{\partial \gamma_1} = P(\omega_1|x_i)\frac{\partial \log p(x_i|\omega_1)}{\partial \gamma_1}.$$

The derivative on the right has already been determined in part (b). The log likelihood function is $L = \sum_i \log p(x_i)$. Plugging the derivative in:

$$\frac{\partial L}{\partial \gamma_1} = \sum_{i=1}^{n} P(\omega_1|x_i)\left(3/\gamma_1 - x_i\right) = 0 \quad \Leftrightarrow \quad \frac{3}{\hat{\gamma}_1}\sum_i P(\omega_1|x_i) = \sum_i P(\omega_1|x_i)x_i.$$

This gives the update equation with $C = 1/3$.

Full points only for a complete derivation. It is OK just to state the first line of the derivation above, but the rest must be derived.