Exercise 1: Bayes' rule (6 points). Suppose that 5% of competitive athletes use performance-enhancing drugs and that a particular drug test has a 2% false positive rate and a 1.5% false negative rate.

1. (3 points) Athlete A tests positive for drug use. What is the probability that Athlete A is using drugs?
2. (3 points) Athlete B tests negative for drug use. What is the probability that Athlete B is not using drugs?

Exercise 2: Bayesian decision theory: losses and risks (11 points). Consider a classification problem with K classes, using a loss $\lambda_{ik} \geq 0$ if we choose class i when the input actually belongs to class k, for $i, k \in \{1, \dots, K\}$.

1. (2 points) Write the expression for the expected risk $R_i(\mathbf{x})$ for choosing class i as the class for a pattern $\mathbf{x}$, and the rule for choosing the class for $\mathbf{x}$.

Consider a two-class problem with losses given by the matrix $\lambda_{ik} = \left(\begin{smallmatrix} 0 & 1 \\ \lambda_{21} & 0 \end{smallmatrix}\right)$.

2. (3 points) Give the optimal decision rule in the form "$p(C_1|\mathbf{x}) > \dots$" as a function of λ_{21}.
3. (3 points) Imagine we consider both misclassification errors as equally costly. When is class 1 chosen (for what values of $p(C_1|\mathbf{x})$)?
4. (3 points) Imagine we want to be very conservative when choosing *class 2* and we seek a rule of the form "$p(C_2|\mathbf{x}) > 0.99$" (i.e., choose class 2 when its posterior probability exceeds 99%). What should λ_{21} be?

Exercise 3: association rules (6 points). Given the following data of transactions at a supermarket, calculate the support and confidence values of the following association rules: meat → avocado, avocado → meat, yogurt → avocado, avocado → yogurt, meat → yogurt, yogurt → meat. What is the best rule to use in practice?

transaction #	items in basket
1	meat, avocado
2	yogurt, avocado
3	meat
4	yogurt, meat
5	avocado, meat, yogurt
6	meat, avocado

Exercise 4: true- and false-positive rates (10 points). Consider the following table, where $\mathbf{x}_n$ is a pattern, y_n its ground-truth label (1 = positive class, 2 = negative class) and $p(C_1|\mathbf{x}_n)$ the posterior probability produced by some probabilistic classification algorithm:

n	1	2	3	4	5
y_n	1	2	2	1	2
$p(C_1\|\mathbf{x}_n)$	0.6	0.7	0.5	0.9	0.2

We use a classification rule of the form "$p(C_1|\mathbf{x}) > \theta$" where $\theta \in [0, 1]$ is a threshold.

1. (8 points) Give, *for all possible values of* $\theta \in [0, 1]$, the predicted labels and the corresponding confusion matrix and classification error.
2. (2 points) Plot the corresponding pairs $(\mathrm{fp}, \mathrm{tp})$ as an ROC curve.

Exercise 5: ROC curves (8 points). Imagine we have a classifier A that has false-positive and true-positive rates $\mathrm{fp_A}, \mathrm{tp_A} \in [0, 1]$ such that $\mathrm{fp_A} > \mathrm{tp_A}$ (that is, this classifier is below the diagonal on the ROC space). Now consider a classifier B that negates the decision of A, that is, whenever A predicts the positive class then B predicts the negative class and vice versa. Compute the false-positive and true-positive rates $\mathrm{fp_B}, \mathrm{tp_B}$ for classifier B. Where is this point in the ROC space?

Exercise 6: least-squares regression (14 points). Consider the following model, with parameters $\mathbf{\Theta} = \{\theta_1, \theta_2, \theta_3\} \subset \mathbb{R}$ and an input $x \in \mathbb{R}$:

$$h(x; \mathbf{\Theta}) = \theta_1 + \theta_2 \sin 2x + \theta_3 \sin 4x \in \mathbb{R}.$$

1. (2 points) Write the general expression of the least-squares error function of a model $h(x; \mathbf{\Theta})$ with parameters $\mathbf{\Theta}$ given a sample $\{(x_n, y_n)\}_{n=1}^N$.
2. (2 points) Apply it to the above model, simplifying it as much as possible.
3. (6 points) Find the least-squares estimate for the parameters.
4. (4 points) Assume the values $\{x_n\}_{n=1}^N$ are uniformly distributed in the interval $[0, 2\pi]$. Can you find a simpler, approximate way to find the least-squares estimate ? *Hint*: approximate $\frac{1}{N}\sum_{n=1}^N f(x_n)$ by an integral.

Exercise 7: maximum likelihood estimate (15 points). A discrete random variable $x \in \{0, 1, 2 \dots\}$ follows a Poisson distribution if it has the following probability mass function:

$$p(x; \theta) = \frac{e^{-\theta}\theta^x}{x!}$$

where the parameter is $\theta > 0$.

1. (2 points) Verify that $\sum_{x=0}^{\infty} p(x) = 1$.
2. (2 points) Write the general expression of the log-likelihood of a probability mass function $p(x; \mathbf{\Theta})$ with parameters $\mathbf{\Theta}$ for an iid sample $x_1, \dots, x_N$.
3. (5 points) Apply it to the above distribution, simplifying it as much as possible.
4. (6 points) Find the maximum likelihood estimate for the parameter θ.

Exercise 8: multivariate Bernoulli distribution (20 points). Consider a multivariate Bernoulli distribution where $\boldsymbol{\theta} \in [0, 1]^D$ are the parameters and $\mathbf{x} \in \{0, 1\}^D$ the binary random vector:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{d=1}^{D} \theta_d^{x_d} (1 - \theta_d)^{1 - x_d}.$$

1. (5 points) Compute the maximum likelihood estimate for $\boldsymbol{\theta}$ given a sample $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$.

Let us do document classification using a D-word dictionary (element d in $\mathbf{x}_n$ is 1 if word d is in document n and 0 otherwise) using a multivariate Bernoulli model for each class. Assume we have K document classes for which we have already obtained the values of the optimal parameters $\boldsymbol{\theta}_k = (\theta_{k1}, \dots, \theta_{kD})^T$ and prior distribution $p(C_k) = \pi_k$, for $k = 1, \dots, K$, by maximum likelihood.

2. (2 points) Write the discriminant function $g_k(\mathbf{x})$ for a probabilistic classifier in general (not necessarily Bernoulli), and the rule to make a decision.
3. (5 points) Apply it to the multivariate Bernoulli case with K classes. Show that $g_k(\mathbf{x})$ is linear on $\mathbf{x}$, i.e., it can be written as $g_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0}$ and give the expression for $\mathbf{w}_k$ and w_{k0}.
4. (3 points) Consider $K = 2$ classes. Show the decision rule can be written as "if $\mathbf{w}^T\mathbf{x} + w_0 > 0$ then choose class 1", and give the expression for $\mathbf{w}$ and w_0.
5. (5 points) Compute the numerical values of $\mathbf{w}$ and w_0 for a two-word dictionary where $\pi_1 = 0.7$, $\boldsymbol{\theta}_1 = \left(\begin{smallmatrix}0.2\\0.8\end{smallmatrix}\right)$ and $\boldsymbol{\theta}_2 = \left(\begin{smallmatrix}0.3\\0.6\end{smallmatrix}\right)$. Plot in 2D all the possible values of $\mathbf{x} \in \{0, 1\}^D$ and the boundary corresponding to this classifier.

Exercise 9: Gaussian classifiers (10 points). Consider a binary classification problem for $\mathbf{x} \in \mathbb{R}^D$ where we use Gaussian class-conditional probabilities $p(\mathbf{x}|C_1) \sim \mathcal{N}(\boldsymbol{\mu}, \sigma_1^2\mathbf{I})$ and $p(\mathbf{x}|C_2) \sim \mathcal{N}(\boldsymbol{\mu}, \sigma_2^2\mathbf{I})$. That is, they have the same mean and the covariance matrices are isotropic but different. Compute the expression for the class boundary. What shape is it?