Exercise 1: Bayes' rule (6 points). Let $D \in \{0,1\}$ indicate the use of drugs and $T \in \{0,1\}$ the test result. We are told that 5% of athletes use drugs and that the test has 2% false positive rate and 1.5% false negative rate. Hence $P(D=1) = 0.05$, $P(T=1|D=0) = 0.02$ and $P(T=0|D=1) = 0.015$.

1. Positive test. Athlete A tests positive for drug use. The probability that A is using drugs is:

$$
\begin{aligned}
P(D=1|T=1) &= \frac{P(T=1|D=1)\,P(D=1)}{P(T=1)} && \text{Bayes rule} \\
&= \frac{P(T=1|D=1)\,P(D=1)}{\sum_d P(T=1,D=d)} && \text{marginalization} \\
&= \frac{P(T=1|D=1)\,P(D=1)}{\sum_d P(T=1|D=d)\,P(D=d)} && \text{product rule} \\
&= \frac{(1-0.015)(0.05)}{(0.02)(1-0.05)+(1-0.015)(0.05)} && \text{substitution} \\
&= 0.72161.
\end{aligned}
$$

2. Negative test. Athlete B tests negative for drug use. The probability that B is not using drugs is:

$$
\begin{aligned}
P(D=0|T=0) &= \frac{P(T=0|D=0)\,P(D=0)}{P(T=0)} && \text{Bayes rule} \\
&= \frac{P(T=0|D=0)\,P(D=0)}{\sum_d P(T=0,D=d)} && \text{marginalization} \\
&= \frac{P(T=0|D=0)\,P(D=0)}{\sum_d P(T=0|D=d)\,P(D=d)} && \text{product rule} \\
&= \frac{(1-0.02)(1-0.05)}{(1-0.02)(1-0.05)+(0.015)(0.05)} && \text{substitution} \\
&\approx 0.99919.
\end{aligned}
$$

Exercise 2: Bayesian decision theory: losses and risks (11 points).

1. The risk of choosing class i is $R_i(\mathbf{x}) = \sum_{k=1}^{K} \lambda_{ik}\, p(C_k|\mathbf{x})$.
   We choose the class with minimum risk: $\arg\min_{i=1,\dots,K} R_i(\mathbf{x})$.

2. $R_1(\mathbf{x}) = \lambda_{11}\, p(C_1|\mathbf{x}) + \lambda_{12}\, p(C_2|\mathbf{x}) = p(C_2|\mathbf{x}) = (1 - p(C_1|\mathbf{x}))$.
   $R_2(\mathbf{x}) = \lambda_{21}\, p(C_1|\mathbf{x}) + \lambda_{22}\, p(C_2|\mathbf{x}) = \lambda_{21} p(C_1|\mathbf{x})$.
   Hence, we pick class 1 if $R_1(\mathbf{x}) < R_2(\mathbf{x}) \Leftrightarrow p(C_1|\mathbf{x}) > 1/(1+\lambda_{21})$.

3. If both misclassification errors are equally costly then $\lambda_{21} = 1$, so we pick class 1 if $p(C_1|\mathbf{x}) > 1/2$.

4. We know that $p(C_2|\mathbf{x}) = (1 - P(C_1|\mathbf{x}) = \lambda_{21}/(1+\lambda_{21})$. We want to pick class 2 when $p(C_2|\mathbf{x}) > 0.99 \Rightarrow \lambda_{21} = 0.99/(1-0.99) = 99$, i.e., the cost of predicting class 1 when the true class is 2 (λ_{12}) is much smaller than the cost of predicting class 2 when the true class is 1 (λ_{21}).

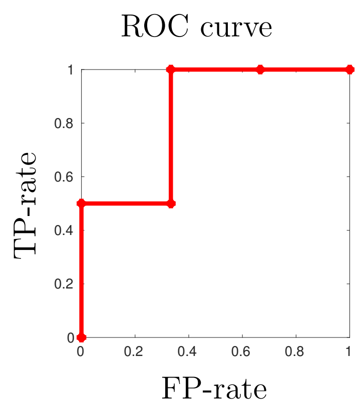
Exercise 3: association rules (6 points).

association rule	support	confidence
meat → avocado	3/6	3/5
avocado → meat	3/6	3/4
yogurt → avocado	2/6	2/3
avocado → yogurt	2/6	2/4
meat → yogurt	2/6	2/3
yogurt → meat	2/6	2/5

Rule "avocado → meat" has significant support (50%) and high confidence (75%), so whenever a customer buys avocado, the system should suggest buying meat.

Exercise 4: true- and false-positive rates (10 points). In the following table, $\hat{y}_n$ is the predicted label for pattern $\mathbf{x}_n$, "ground truth" contains the true labels y_n, $\mathbf{C}$ is the 2×2 confusion matrix (with counts, not rates) and E the classification error (in %).

θ	$[0, 0.2)$	$[0.2, 0.5)$	$[0.5, 0.6)$	$[0.6, 0.7)$	$[0.7, 0.9)$	$[0.9, 1]$	ground truth
$\hat{y}_1$	1	1	1	2	2	2	1
$\hat{y}_2$	1	1	1	1	2	2	2
$\hat{y}_3$	1	1	2	2	2	2	2
$\hat{y}_4$	1	1	1	1	1	2	1
$\hat{y}_5$	1	2	2	2	2	2	2
$\mathbf{C}$	$\begin{array}{c\|c} 2 & 0 \\ \hline 3 & 0 \end{array}$	$\begin{array}{c\|c} 2 & 0 \\ \hline 2 & 1 \end{array}$	$\begin{array}{c\|c} 2 & 0 \\ \hline 1 & 2 \end{array}$	$\begin{array}{c\|c} 1 & 1 \\ \hline 1 & 2 \end{array}$	$\begin{array}{c\|c} 1 & 1 \\ \hline 0 & 3 \end{array}$	$\begin{array}{c\|c} 0 & 2 \\ \hline 0 & 3 \end{array}$	
E	60%	40%	20%	40%	20%	40%	



Exercise 5: ROC curves (8 points). Imagine the true class is negative. Then, A predicts positive with a rate $\mathrm{fp_A}$ (i.e., $\mathrm{fp_A}$% of the times), and so B (which reverses A's decision) predicts positive with a rate $1 - \mathrm{fp_A}$. So $\mathrm{fp_B} = 1 - \mathrm{fp_A}$. Now imagine the true class is positive. Then, A predicts positive with a rate $\mathrm{tp_A}$ (i.e., $\mathrm{tp_A}$% of the times), and so B (which reverses A's decision) predicts positive with a rate $1 - \mathrm{tp_A}$. So $\mathrm{tp_B} = 1 - \mathrm{tp_A}$.

If the ROC point $(\mathrm{fp_A}, \mathrm{tp_A})$ for A is below the diagonal, then the ROC point for B $(\mathrm{fp_B}, \mathrm{tp_B}) = (1 - \mathrm{fp_A}, 1 - \mathrm{tp_A})$ is above the diagonal, symmetrically opposite to $(0.5, 0.5)$ (the center of the ROC space).

Exercise 6: least-squares regression (14 points).

1. Least-squares error of $\boldsymbol{\Theta}$ given sample: $E(\boldsymbol{\Theta}) = \sum_{n=1}^{N} (y_n - h(x_n; \boldsymbol{\Theta}))^2$.

2. $E(\theta_1, \theta_2, \theta_3) = \sum_{n=1}^{N} (y_n - \theta_1 - \theta_2 \sin 2x_n - \theta_3 \sin 4x_n)^2$.

3. Taking partial derivatives wrt the parameters and simplifying:

$$\frac{\partial E}{\partial \theta_1} = -2\sum_{n=1}^{N}(y_n - \theta_1 - \theta_2 \sin 2x_n - \theta_3 \sin 4x_n) = -2N(\overline{y} - \theta_1 - \theta_2\,\overline{\sin 2x} - \theta_3\,\overline{\sin 4x})$$

$$\frac{\partial E}{\partial \theta_2} = -2\sum_{n=1}^{N}(y_n - \theta_1 - \theta_2 \sin 2x_n - \theta_3 \sin 4x_n)\sin 2x_n = -2N(\overline{y\,\sin 2x} - \theta_1\,\overline{\sin 2x} - \theta_2\,\overline{\sin^2 2x} - \theta_3\,\overline{\sin 4x\,\sin 2x})$$

$$\frac{\partial E}{\partial \theta_3} = -2\sum_{n=1}^{N}(y_n - \theta_1 - \theta_2 \sin 2x_n - \theta_3 \sin 4x_n)\sin 4x_n = -2N(\overline{y\,\sin 4x} - \theta_1\,\overline{\sin 4x} - \theta_2\,\overline{\sin 2x\,\sin 4x} - \theta_3\,\overline{\sin^2 4x})$$

where we use the notation $\overline{y\,\sin 2x} = \frac{1}{N}\sum_{n=1}^{N} y_n \sin 2x_n$, $\overline{\sin^2 2x} = \frac{1}{N}\sum_{n=1}^{N}\sin^2 2x_n$, etc. Equating this to zero yields the least-squares estimate, which is the solution to the following linear system:

$$\begin{pmatrix} 1 & \overline{\sin 2x} & \overline{\sin 4x} \\ \overline{\sin 2x} & \overline{\sin^2 2x} & \overline{\sin 2x\,\sin 4x} \\ \overline{\sin 4x} & \overline{\sin 2x\,\sin 4x} & \overline{\sin^2 4x} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = \begin{pmatrix} \overline{y} \\ \overline{y\,\sin 2x} \\ \overline{y\,\sin 4x} \end{pmatrix}.$$

4. If $x_1, \dots, x_N$ are uniformly distributed in $[0, 2\pi]$, we can approximate integrals as finite sums, e.g. $\frac{1}{N}\sum_{n=1}^{N}\sin x_n \approx \frac{1}{2\pi}\int_0^{2\pi}\sin 2x\,dx = 0$, $\frac{1}{N}\sum_{n=1}^{N}\sin^2 x_n \approx \frac{1}{2\pi}\int_0^{2\pi}\sin^2 2x\,dx = \frac{1}{2}$, etc., where the approximation error tends to zero as $N \to \infty$. Hence we approximate the matrix of the previous linear system as follows:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} \approx \begin{pmatrix} \overline{y} \\ \overline{y\,\sin 2x} \\ \overline{y\,\sin 4x} \end{pmatrix} \Rightarrow \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} \approx \begin{pmatrix} \overline{y} \\ 2\,\overline{y\,\sin 2x} \\ 2\,\overline{y\,\sin 4x} \end{pmatrix}$$

so the approximately optimal model is $h(x) = \overline{y} + 2\,\overline{y\,\sin 2x}\,\sin 2x + 2\,\overline{y\,\sin 4x}\,\sin 4x$.
Try it in Matlab!

Exercise 7: maximum likelihood estimate (15 points).

1. After substitution, $\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty}\frac{e^{-\theta}\theta^x}{x!} = e^{-\theta}\left(\sum_{x=0}^{\infty}\frac{\theta^x}{x!}\right)$, we recognize the term in parenthesis as Taylor's expansion of e^{θ} around 0, therefore $\sum_{x=0}^{\infty} p(x) = e^{-\theta}e^{\theta} = 1$.

2. Log-likelihood of $\mathbf{\Theta}$ given iid sample: $\mathcal{L}(\mathbf{\Theta}) = \sum_{n=1}^{N}\log p(x_n; \mathbf{\Theta})$.

3. Defining $\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$:

$$\mathcal{L}(\theta) = \sum_{n=1}^{N}\log p(x_n;\theta) = \sum_{n=1}^{N}(-\theta + x_n \log\theta - \log(x_n!)) = -N\left(\theta - \overline{x}\log\theta + \frac{1}{N}\sum_{n=1}^{N}\log(x_n!)\right).$$

4. Taking the derivative wrt θ and equating it to zero:

$$\frac{\partial \mathcal{L}}{\partial \theta} = -N\left(1 - \frac{\overline{x}}{\theta}\right) = 0 \Rightarrow \theta = \overline{x},$$

that is, the MLE for θ is the sample average.

Exercise 8: multivariate Bernoulli distribution (20 points).

1. The log-likelihood for the Bernoulli parameter of class C_k is:

$$\mathcal{L}(\boldsymbol{\theta}_k) = \sum_{n \in C_k} \log p(\mathbf{x}_n; \boldsymbol{\theta}_k) = \sum_{n \in C_k} \Big(x_{nd} \log \theta_{kd} + (1 - x_{nd}) \log (1 - \theta_{kd}) \Big)$$
$$= \left(\sum_{n \in C_k} x_{nd} \right) \log \left(\frac{\theta_{kd}}{1 - \theta_{nd}} \right) + N_k \log (1 - \theta_{kd}),$$

where class C_k has N_k points out of the N points in the sample. To maximize the log-likelihood, we take derivatives wrt the parameters and equate to zero:

$$\frac{\partial \mathcal{L}}{\partial \theta_{kd}} = \left(\sum_{n \in C_k} x_{nd} \right) \frac{1}{\theta_{kd}(1 - \theta_{kd})} - \frac{N_k}{1 - \theta_{kd}} = 0 \Rightarrow \theta_{kd} = \frac{1}{N_k} \sum_{n \in C_k} x_{nd},$$

that is, the MLE for θ_{kd} is the average of the points in class C_k over dimension d.

2. Discriminant function: $g_k(\mathbf{x}) = \log p(\mathbf{x}|C_k) + \log p(C_k)$, for $k = 1, \dots, K$.
Classification rule: choose $\arg\max_{k=1,\dots,K} g_k(\mathbf{x})$.

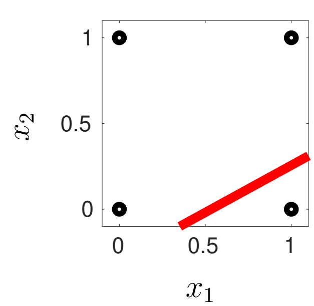
3. We have, for $k = 1, \dots, K$:

$$g_k(\mathbf{x}) = \log p(\mathbf{x}|C_k) + \log p(C_k) = \sum_{d=1}^{D} (x_d \log \theta_{kd} + (1 - x_d) \log (1 - \theta_{kd})) + \log \pi_k$$
$$= \sum_{d=1}^{D} \left(x_d \log \left(\frac{\theta_{kd}}{1 - \theta_{kd}} \right) \right) + \sum_{d=1}^{D} \log (1 - \theta_{kd}) + \log \pi_k = \mathbf{w}_k^T \mathbf{x} + w_{k0}$$

where $\mathbf{w}_k$ has elements $w_{kd} = \log \left(\frac{\theta_{kd}}{1-\theta_{kd}} \right)$ for $d = 1, \dots, D$ and $w_{k0} = \sum_{d=1}^{D} \log (1 - \theta_{kd}) + \log \pi_k$. Note: $\mathrm{logit}(\theta) = \log \left(\frac{\theta}{1-\theta} \right) \in (-\infty, \infty)$ for $\theta \in (0, 1)$ is called the *logit function* or *log odds* of θ. It is the inverse of the *logistic function* $\sigma(t) = \frac{1}{1+e^{-t}}$.

4. For $K = 2$ classes, we pick class 1 if $g_1(\mathbf{x}) > g_2(\mathbf{x}) \Leftrightarrow \mathbf{w}_1^T \mathbf{x} + w_{10} > \mathbf{w}_2^T \mathbf{x} + w_{20} \Leftrightarrow \mathbf{w}^T \mathbf{x} + w_0 > 0$ with $\mathbf{w} = \mathbf{w}_1 - \mathbf{w}_2$ and $w_0 = w_{10} - w_{20}$. This is true for any linear classifiers, not just those derived from a Bernoulli distribution.

5.
$$\mathbf{w}_1 = \begin{pmatrix} -1.3863 \\ 1.3863 \end{pmatrix} \quad w_{10} = -2.1893$$
$$\mathbf{w}_2 = \begin{pmatrix} -0.8473 \\ 0.4055 \end{pmatrix} \quad w_{20} = -2.4769$$
$$\mathbf{w} = \begin{pmatrix} -0.5390 \\ 0.9808 \end{pmatrix} \quad w_0 = 0.2877$$



Exercise 9: Gaussian classifiers (10 points). The boundary points $\mathbf{x} \in \mathbb{R}^D$ satisfy $p(C_1|\mathbf{x}) = p(C_2|\mathbf{x}) \Leftrightarrow \log p(\mathbf{x}|C_1) + \log \pi_1 = \log p(\mathbf{x}|C_2) + \log \pi_2$, where $\pi_i = p(C_i) \in (0,1)$ and $\sigma_1 > \sigma_2$ w.l.o.g. Substituting the Gaussian densities $p(\mathbf{x}|C_i) = (2\pi\sigma_i^2)^{-D/2} \exp\left(-\frac{1}{2}\|\mathbf{x} - \boldsymbol{\mu}\|^2/\sigma_i^2\right)$ and simplifying, we obtain $\|\mathbf{x} - \boldsymbol{\mu}\|^2 = r^2$ where

$$r^2 = 2\,\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 - \sigma_2^2}\,\log\left(\frac{\pi_2\,\sigma_1^D}{\pi_1\,\sigma_2^D}\right).$$

We have 3 cases, noting that $\pi_2 = 1 - \pi_1$:

- $\pi_1 < \frac{\sigma_1^D}{\sigma_1^D + \sigma_2^D} \Leftrightarrow r^2 > 0$: the class boundary is a circle in 2D or in general a hypersphere of radius r with center at $\boldsymbol{\mu}$. Its interior is C_2 and its exterior is C_1.
- $\pi_1 = \frac{\sigma_1^D}{\sigma_1^D + \sigma_2^D} \Leftrightarrow r^2 = 0$: the class boundary is the point $\mathbf{x} = \boldsymbol{\mu}$, so the entire $\mathbb{R}^D$ space except for $\boldsymbol{\mu}$ is classified as C_1.
- $\pi_1 > \frac{\sigma_1^D}{\sigma_1^D + \sigma_2^D} \Leftrightarrow r^2 < 0$: there is no boundary, the entire $\mathbb{R}^D$ space is classified as C_1.

Try it in Matlab!