Problem 1. (Naive Bayes)
We aim to predict the probability of success for a data point $x = (x_1, \ldots, x_d)$ with features $x_i \in \mathbb{R}$ for $i = 1, \ldots, d$. Success corresponds to class 1 and failure corresponds to class 0. Let $P(1)$ denote the probability of success determined based on a training set.

1. Suppose you are given a test point $x^t = (x_1^t, \ldots, x_d^t)$ and $f_{x_i|1}$ is the conditional probability density of feature x_i given success. Give the expression of the probability that the test point is a success, by filling in the blanks below and using feature independence with the Naive Bayes assumption.

$$P(1 \mid x^t) = \frac{\ldots\ldots\ldots\ldots\ldots\ldots\ldots P(1)}{\ldots\ldots\ldots\ldots}.$$

Solution:

$$P(1 \mid x^t) = \frac{\prod_{i=1}^d f_{x_i|1}(x_i^t)P(1)}{\sum_{y\in\{0,1\}} \prod_{i=1}^d f_{x_i|y}(x_i^t)P(y)} \propto \prod_{i=1}^d f_{x_i|1}(x_i^t)P(1).$$

2. Suppose the number of features is $d = 1$ and the test data point is $x^t = 4$. Assume the following is known to you: $f_{x|1}(4) = 0.17$, $f_{x|0}(4) = 0.24$ and $P(1) = 0.6$. Using the Naive Bayes approach to determine whether the test point $x^t = 4$ is classified as a success or not.

Solution: The probability $P\left(\text{success} = 1 \mid x^t\right)$ that the test point is a success can be computed as follows:

$$\begin{aligned} P\left(\text{success} = 1 \mid x^t\right) &= \frac{f_{x|1}(4)P(1)}{f_{x|0}(4)P(0) + f_{x|1}(4)P(1)} \\ &= \frac{0.17 \times 0.6}{0.24 \times 0.4 + 0.17 \times 0.6} \\ &= 17/33. \end{aligned}$$

Since we know $P\left(\text{success} = 1 \mid x^t\right) + P\left(\text{success} = 0 \mid x^t\right) = 1$ and $17/33 > 1/2$, we conclude that the test point x^t is more likely to lead to a success.

Exercise 2. (Naive Bayes)
In this problem we aim to predict the likelihood of failure during manufacturing of an item based on the pressure and temperature applied to the object. Consider the following hypothetical dataset consisting of 10 measurements:

Failure (F)	1	0	0	0	0	1	0	0	0	0
Temperature (T) $°C$	170	100	80	140	60	230	100	60	140	120
Pressure (P) MPa	430	220	200	160	240	370	200	180	240	160

1. Based on this dataset, what is the empirical probability for an item to be defective ($F = 1$)?

Solution:

$$P(F=1) = \frac{2}{10} = 0.2.$$

2. Given each of the two classes, 1 (failure) and 0 (no failure), for each feature vector, determine the empirical means $\mu_{1,T}, \mu_{1,P}, \mu_{0,T}, \mu_{0,P}$ and variances $\sigma^2_{1,T}, \sigma^2_{1,P}, \sigma^2_{0,T}, \sigma^2_{0,P}$. Based on this, determine the corresponding two Gaussian distributions for each of the two features.

   Solution:

$$\begin{aligned}
\mu_{1,T} &= \frac{170+230}{2} = 200 \\
\mu_{1,P} &= \frac{430+370}{2} = 400 \\
\mu_{0,T} &= \frac{100+80+140+60+100+60+140+120}{8} = 100 \\
\mu_{0,P} &= \frac{220+200+160+240+200+180+240+160}{8} = 200 \\
\sigma^2_{1,T} &= \frac{(200-170)^2+(200-230)^2}{2} = 30^2 = 900 \\
\sigma^2_{1,P} &= \frac{(430-400)^2+(370-400)^2}{2} = 30^2 = 900 \\
\sigma^2_{0,T} &= \frac{(100-100)^2+(80-100)^2+(140-100)^2+(60-100)^2}{8} \\
&\quad + \frac{(100-100)^2+(60-100)^2+(140-100)^2+(120-100)^2}{8} \\
&= \frac{20^2 \times (0+1+4+4+0+4+4+1)}{8} = 900 \\
\sigma^2_{0,P} &= \frac{(220-200)^2+(200-200)^2+(160-200)^2+(240-200)^2}{8} \\
&\quad + \frac{(200-200)^2+(180-200)^2+(240-200)^2+(160-200)^2}{8} \\
&= \frac{20^2 \times (1+0+4+4+0+1+4+4)}{8} = 900
\end{aligned}$$

3. Suppose you have the choice between two classification models that use the means and variances derived previously to predict a new datapoint $x_{test} = (T', P')$:

   - Model A: Naive Bayes Classifier.
   - Model B: Classify (T', P') according to its Euclidean (L_2) distance to the class means. For example, classify (T', P') as failure if

$$(T' - \mu_{1,T})^2 + (P' - \mu_{1,P})^2 < (T' - \mu_{0,T})^2 + (P' - \mu_{0,P})^2.$$

   Now suppose you are given the datapoint $x_{test} = (T', P') = (151, 302)$.

   (a) Predict the output of Model A.

(b) Predict the output of Model B.

(c) Are the above two outputs different? If yes, why?

Solution:

(a)

$$\begin{aligned} P(F=1 \mid x_{test}) &= \frac{f_{x_{test}}(x_{test} \mid F=1)P(F=1)}{f_{x_{test}}(x_{test})} \\ &\propto \frac{1}{2\pi\sigma_{1,T}\sigma_{1,P}} e^{-\frac{(T'-\mu_{1,T})^2}{2\sigma_{1,T}^2}} e^{-\frac{(P'-\mu_{1,P})^2}{2\sigma_{1,P}^2}} \times 0.2 \\ &\propto \frac{1}{2\pi \times 30 \times 30} e^{-\frac{49^2}{2\times 30^2}} e^{-\frac{98^2}{2\times 30^2}} \times 0.2 \\ &\propto 4.48 \times 10^{-8} \end{aligned}$$

$$\begin{aligned} P(F=0 \mid x_{test}) &= \frac{f_{x_{test}}(x_{test} \mid F=0)P(F=0)}{f_{x_{test}}(x_{test})} \\ &\propto \frac{1}{2\pi\sigma_{0,T}\sigma_{0,P}} e^{-\frac{(T'-\mu_{0,T})^2}{2\sigma_{2,T}^2}} e^{-\frac{(P'-\mu_{0,P})^2}{2\sigma_{0,P}^2}} \times 0.8 \\ &\propto \frac{1}{2\pi \times 30 \times 30} e^{-\frac{51^2}{2\times 30^2}} e^{-\frac{102^2}{2\times 30^2}} \times 0.8 \\ &\propto 1.03 \times 10^{-7} \end{aligned}$$

Since $P(F=0 \mid x_{test}) > P(F=1 \mid x_{test})$, the Naive Bayes classifer will predict it as a success.

(b) For Model B, we compare the L_2 distance of x_{test} to the respective class means

$$\begin{aligned} D_1^2 &= (T'-\mu_{1,T})^2 + (P'-\mu_{1,P})^2 = (151-200)^2 + (302-400)^2 = 5 \times 49^2 \\ D_0^2 &= (T'-\mu_{0,T})^2 + (P'-\mu_{0,P})^2 = (151-100)^2 + (302-200)^2 = 5 \times 51^2 \end{aligned}$$

Since $(T'-\mu_{1,T})^2 + (P'-\mu_{1,P})^2 < (T'-\mu_{0,T})^2 + (P'-\mu_{0,P})^2$, Model B will predict it as a failure.

(c) The two classifer outputs are different. The reason that Model A predicts $F=0$ instead of $F=1$, despite x_{test} being closer to $(\mu_{1,T}, \mu_{1,P})$ is because of class imbalance, that is, the probability of a defective product is much lower than that of a successful product. This results in Model A assigning more weightage to the class $F=0$ compared to $F=1$ when making classification predictions. This is explained as follows:

Since we have that $\sigma_{1,T} = \sigma_{0,T} = \sigma_{1,P} = \sigma_{0,P} = \sigma$, the class conditional likelihoods can be written as

$$f_{x_{test}}(x_{test} \mid F=i) \propto e^{-\frac{D_i^2}{2\sigma^2}}, D_i^2 = (T'-\mu_{i,T})^2 + (P'-\mu_{i,P})^2$$

Because $D_1 < D_0$, we have $f(x_{test} \mid F=0) > f(x_{test} \mid F=1)$, that is, the class conditional likelihood for x_{test} is higher for $F=0$.

However, since $P(F = i \mid x_{test}) \propto f(x_{test} \mid F = i)P(F = i)$, and the probability of a defective product is lesser than of a non-defective product ($P(F = 0) = 0.2, P(F = 1) = 0.8$), the posterior probability is lesser for $F = 0$.

The above example shows the importance of class imbalance in machine learning. A familiar example that you have already seen in class is spam filtering, where spam emails are much less frequent than non-spam emails.

Problem 3. (Dynamical systems, discretization, RNN)

Consider the following dynamical system: $\ddot{x}(t) = -x(t)$.

1. Verify that $x(t) = a\sin(t) + b\cos(t)$ is a solution to this system for any $a, b \in \mathbb{R}$.

   Solution: We have $x(t) = a\sin(t) + b\cos(t)$. Its first derivative is $\dot{x}(t) = a\cos(t) - b\sin(t)$, and its second derivative is $\ddot{x}(t) = -a\sin(t) - b\cos(t) = -x(t)$.

2. Given $x(0) = c_1$, $\dot{x}(0) = c_2$, determine $a, b \in \mathbb{R}$ in terms of c_1, c_2.

   Solution: $a\sin(0) + b\cos(0) = b = c_1, \quad a\cos(0) - b\sin(0) = a = c_2$.

3. By defining $s_1 = x(t)$ and $s_2 = \dot{x}(t)$ verify that the system can be put in state-space form as follows.

$$\begin{bmatrix} \dot{s}_1(t) \\ \dot{s}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \end{bmatrix} \tag{0.1}$$

   What are the eigenvalues of the system matrix $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$?

   Solution: We have $\dot{s}_1(t) = \dot{x}(t) = s_2(t)$ and $\dot{s}_2(t) = \ddot{x}(t) = -x(t) = -s_1(t)$, so

$$\begin{bmatrix} \dot{s}_1(t) \\ \dot{s}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \end{bmatrix}.$$

   For $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, the characteristic polynomial is

$$\det(A - \lambda I) = \begin{vmatrix} -\lambda & 1 \\ -1 & -\lambda \end{vmatrix} = \lambda^2 + 1 = 0.$$

   The eigenvalues are $\lambda_{1,2} = \pm i$.

4. Consider the phase plot of the system, that is, plotting $s_1(t)$ on the x-axis and $s_2(t)$ on the y-axis. For a given initial condition $s_1(0) = c_1$, $s_2(0) = c_2$, verify that phase-plot will be a circle centered at $(0, 0)$ with radius $\sqrt{c_1^2 + c_2^2}$.

Solution: We compute

$$\frac{d}{dt}(s_1(t)^2 + s_2(t)^2) = 2(s_1(t)\dot{s_1}(t) + s_2(t)\dot{s_2}(t)) = 2(s_1(t)(s_2(t)) + s_2(t)(-s_1(t))) = 0.$$

Given that $\frac{d}{dt}(s_1(t)^2+s_2(t)^2) = 0$, it follows that $s_1(t)^2+s_2(t)^2 = C$ for some constant C. This along with the initial condition being $(s_1(0), s_2(0)) = (c_1, c_2)$ implies that $s_1(t)^2 + s_2(t)^2 = s_1(0)^2 + s_2(0)^2 = c_1^2 + c_2^2$ for all t.

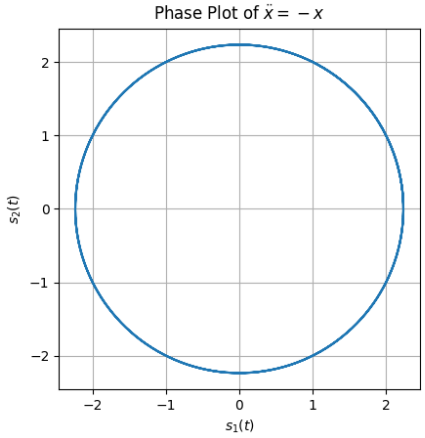

Figure 1: Phase plot for the continuous-time system.

5. Derive the Euler discretizations of the continuous-time system in differential equation (0.1) with discretizations parameter $\delta > 0$. Denote the discretized state by $(s_{1,k}, s_{2,k})$. What are the eigenvalues of the resulting system matrix?

$$\begin{bmatrix} s_{1,k+1} \\ s_{2,k+1} \end{bmatrix} = \begin{bmatrix} 1 & \delta \\ -\delta & 1 \end{bmatrix} \begin{bmatrix} s_{1,k} \\ s_{2,k} \end{bmatrix}$$

Solution: For $\dot{s}(t) = As(t)$ with $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, the forward Euler update is $s_{k+1} = s_k + \delta A s_k = (I + \delta A)s_k$, giving

$$I + \delta A = \begin{bmatrix} 1 & \delta \\ -\delta & 1 \end{bmatrix}.$$

For the eigenvalues of $A_d = \begin{bmatrix} 1 & \delta \\ -\delta & 1 \end{bmatrix}$, we set $\det(\tilde{A} - \lambda I) = (1-\lambda)^2 + \delta^2 = 0$, so the eigenvalues are $\lambda_{1,2} = 1 \pm i\delta$.

6. Provide the phase plot of the discretized system, namely, plot $s_2(k)$ against $s_1(k)$. You may use $\delta = 0.05$, $c_1 = 1, c_2 = 2$. Observe that in this phase plot, we see spirals growing out in contrast to the purely rotational dynamics in the continuous-time system. Thus, the trajectories of the system grow unbounded.

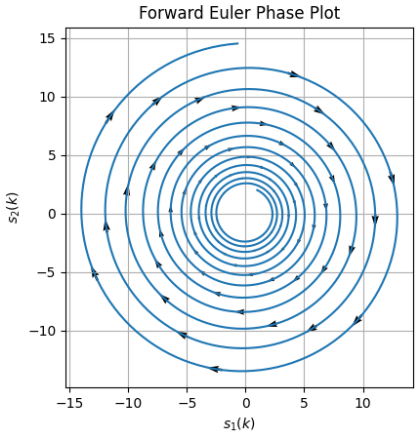

Figure 2: Phase plot with forward Euler.

7. Let us consider the so-called *backward* Euler discretization, where $\dot{s}(t) \approx \frac{s(t)-s(t-\delta)}{\delta}$. Verify that with the above discretization, the system dynamics can be written as

$$\begin{bmatrix} s_{1,k+1} \\ s_{2,k+1} \end{bmatrix} = \frac{1}{1+\delta^2} \begin{bmatrix} 1 & \delta \\ -\delta & 1 \end{bmatrix} \begin{bmatrix} s_{1,k} \\ s_{2,k} \end{bmatrix}$$

Hint: you will have to solve for $s_{1,k}, s_{2,k}$ in terms of $s_{1,k-1}, s_{2,k-1}$. After solving for s_k in terms of s_{k-1}, we shift the index $k \mapsto k+1$ so that the dynamics are expressed in the usual form $s_{k+1} = A_b\, s_k$.

Solution: For $\dot{s}(t) = As(t)$, backward Euler gives $s_k - \delta A s_k = s_{k-1} \quad \Rightarrow \quad (I - \delta A)s_k = s_{k-1}$.

With $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, we get $\quad I - \delta A = \begin{bmatrix} 1 & -\delta \\ \delta & 1 \end{bmatrix}, \quad (I - \delta A)^{-1} = \frac{1}{1+\delta^2} \begin{bmatrix} 1 & \delta \\ -\delta & 1 \end{bmatrix}$.

Hence

$$s_k = (I - \delta A)^{-1} s_{k-1} = \frac{1}{1+\delta^2} \begin{bmatrix} 1 & \delta \\ -\delta & 1 \end{bmatrix} s_{k-1}.$$

Shifting the index $k \leftarrow k+1$ yields

$$\begin{bmatrix} s_{1,k+1} \\ s_{2,k+1} \end{bmatrix} = \frac{1}{1+\delta^2} \begin{bmatrix} 1 & \delta \\ -\delta & 1 \end{bmatrix} \begin{bmatrix} s_{1,k} \\ s_{2,k} \end{bmatrix}.$$

8. What are the eigenvalues of the discretized system dynamics $A_b = \frac{1}{1+\delta^2} \begin{bmatrix} 1 & \delta \\ -\delta & 1 \end{bmatrix}$, corresponding to backward Euler discretization approach?

   Provide the phase plot of the discretized system above, for $\delta = 0.05$, $c_1 = 1, c_2 = 2$. Observe that in this phase plot, we see spirals growing in, in contrast to the purely rotational dynamics

in the continuous time system. Thus, the trajectories of the system are approaching the equilibrium at $(0, 0)$.

Solution: Similarly to the forward Euler's case, the eigenvalues of A_b are $\lambda_{1,2} = \frac{1 \pm i\delta}{1+\delta^2}$.

We have $|\lambda_{1,2}| = \frac{1}{\sqrt{1+\delta^2}} < 1$, so the discrete-time system is a stable spiral.

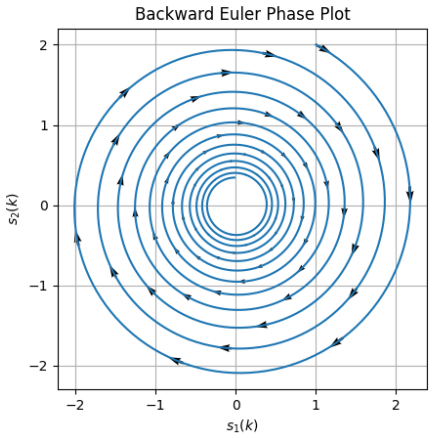

Figure 3: Phase plot with backward Euler.

9. As you can observe from the above two subparts, neither forward, nor backward Euler discretization is effective in capturing the perfect oscillations of the continuous-time signal $y(t) = a\sin(t) + b\cos(t)$. The issue is that the eigenvalues of the continuous-time system matrix A are purely imaginary and the forward Euler discretization results in eigenvalues with magnitude slightly larger than one (and thus, growing oscillations and unbounded growth of the solution), whereas the backward Euler discretization results in eigenvalues with magnitude slightly smaller than 1 (thus, damping oscillations and convergence to zero of the solutions). Another discretization approach is using central difference discretization $\dot{s}(t) \approx \frac{s(t+\delta)-s(t-\delta)}{2\delta}$. Verify that with the central difference discretization, the discrete-time system can be written as

$$\begin{bmatrix} s_{1,k+1} \\ s_{2,k+1} \end{bmatrix} = \begin{bmatrix} 1-\delta^2 & \delta \\ -\delta & 1 \end{bmatrix} \begin{bmatrix} s_{1,k} \\ s_{2,k} \end{bmatrix}$$

What are the eigenvalues of $A_c = \begin{bmatrix} 1-\delta^2 & \delta \\ -\delta & 1 \end{bmatrix}$? Provide the phase plot of the above discretized system. As you can see the phase plot now matches that of the continuous-time system much better.

Solution: For $A_c = \begin{bmatrix} 1-\delta^2 & \delta \\ -\delta & 1 \end{bmatrix}$,

$$\det(A_c - \lambda I) = (1-\delta^2-\lambda)(1-\lambda) + \delta^2 = \lambda^2 + (\delta^2-2)\lambda + 1.$$

$$\lambda_{1,2} = \frac{2-\delta^2 \pm \sqrt{\delta^2(\delta^2-4)}}{2} = 1 - \frac{\delta^2}{2} \pm i\,\frac{\delta}{2}\sqrt{4-\delta^2}.$$

Observe that

$$|\lambda_{1,2}|^2 = \left(1 - \frac{\delta^2}{2}\right)^2 + \left(\frac{\delta}{2}\sqrt{4-\delta^2}\right)^2 = 1.$$

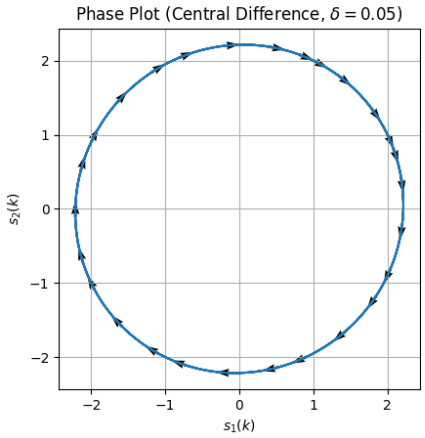

Figure 4: Phase plot with central difference.

10. Now, letting $y_k = s_{1,k}$ verify that any of the discretized dynamics above can be written as a recurrent neural network without any input, with two hidden states and the identity function as the activation.

Solution: A vanilla RNN without input has the form

$$h_{k+1} = g_s(W_{ss}s_k), \qquad y_k = g_o(W_{os}s_k),$$

where s_k is the hidden state, g_s and g_o are the activations, W_{ss} is the recurrent weight matrix, and W_{os} is the output matrix. Since there is no input, we set $W_{sa} = 0$.

Here we take the hidden state $s_k = \begin{bmatrix} s_{1,k} \\ s_{2,k} \end{bmatrix}$, choose the activation as the identity $g_s(z) = z$ and $g_o(z) = z$, and set

$$W_{ss} \in \left\{ \begin{bmatrix} 1 & \delta \\ -\delta & 1 \end{bmatrix}, \frac{1}{1+\delta^2}\begin{bmatrix} 1 & \delta \\ -\delta & 1 \end{bmatrix}, \begin{bmatrix} 1-\delta^2 & \delta \\ -\delta & 1 \end{bmatrix} \right\},$$

corresponding to forward Euler, backward Euler, or central difference, respectively. Then the state update

$$s_{k+1} = W_{ss}s_k$$

exactly matches the discretized dynamics. For the output we want $y_k = s_{1,k}$, so we choose

$$W_{os} = \begin{bmatrix} 1 & 0 \end{bmatrix},$$

which gives $y_k = W_{os}s_k = s_{1,k}$.

Remark: The system in differential equation 0.1 is a Hamiltonian, and the energy is preserved (see Part 4 of the question). For such systems, Euler discretization is not effective as it does not preserve the energy. The so-called Symplectic discretization ensures energy preservation and will result in a discretized dynamics that are identical (up to a linear transformation) to that of the central difference approach.