1 Linear Algebra and Probability Theory

For the next two exercises, let X and Y be two random variables with the *joint cumulative distribution function* (cdf)

$$F_{X,Y} = \begin{cases} 1 - e^{-x} - e^{-y} + e^{-x-y} & x, y \geq 0 \\ 0 & \text{else} \end{cases}.$$

Problem 1 [3 points] Determine the *marginal probability density functions* (pdfs) f_X and f_Y.

Identify the marginal distributions.

Problem 2 [2 points] Are X and Y independent? Prove your claim.

2 kNN

Problem 3 [1 point] You are testing three new sensors. You would like to classify them using the nearest neighbour approach with Euclidean distance ($d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$).

Sensor	Output	Nearest Neighbour (current)	Nearest Neighbour (expected)
sensor1	$[1, 150]$	sensor2	sensor3
sensor2	$[2, 110]$	sensor3	sensor1 or sensor3
sensor3	$[1, 100]$	sensor2	sensor1

Your sensors output two parameters each. This output is in the above table. Since this is a controlled test of the sensors, you know that sensor1 and sensor3 belong to the same group and that sensor2 belongs to a different group. We can see that sensor2 is causing trouble for the classification of the other sensors. We would expect the output to look like the fourth column in the above table. How can you fix this problem without changing your distance measure from Euclidean?

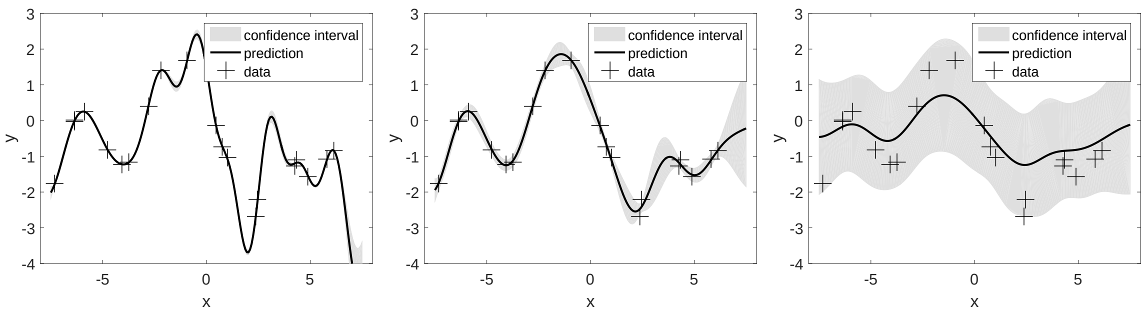
Problem 4 [1 point] In general, especially with data which might have many uninformative parameters, why can the use of Euclidean distance as the distance measure be problematic? (Even after the problem from above has been fixed?)

3 Gaussian Processes

We have a data set $x \in \mathbb{R}^1$. You are given Gaussian processes $f \sim GP$ with mean function $m(x) = 0$, covariance function $k(x, x')$, and a noisy observation $\epsilon \sim \mathcal{N}(0, \sigma_y^2)$.

Problem 5 [2 points] Assume the covariance function is $k(x, x') = (xx' + 1)^2$, and the observation data $x_1 = -\frac{1}{2}, x_2 = 2$ is given, write down the distribution of $p(f(x_1), f(x_2))$. What is the relationship of $f(x_1)$ and $f(x_2)$? Describe your reasoning.

Problem 6 [2 points] We have a squared exponential (SE) kernel. With different values of σ_y^2, the GP models are shown in Fig. 1. Which model is best? What causes the other two to be not good? Explain your answer.



A B C

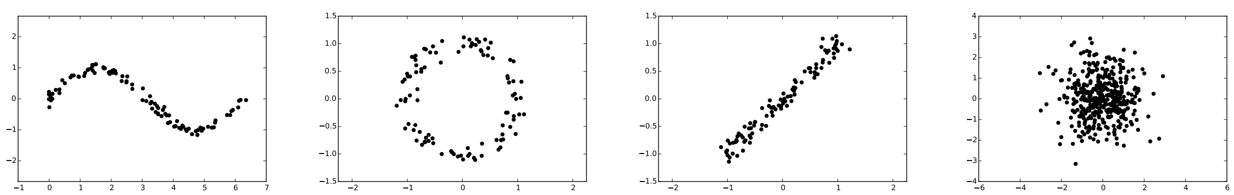
Figure 1: Gaussian Processes

4 Neural networks

Problem 7 [3 points] A neural network with activation functions $\tanh(\cdot)$ in the hidden units is initialised with all parameters (weights, biases) set to 0. Can it learn? Explain your answer.

Problem 8 [1 point] A neural network with activation functions $\tanh(\cdot)$ in the hidden units is initialised with all parameters (weights, biases) set to 1. Can it learn? Explain your answer.

5 Unsupervised learning



A B C D

Figure 2: Four datasets

Problem 9 [3 points] PCA was performed on dataset A, B and C (Fig.: 2). The three results are:

	Result 1	Result 2	Result 3
normalised eigenvalues =	$[0.5, 0.5]$	$[0.95, 0.05]$	$[0.99, 0.01]$
normalised eigenvectors =	$\left[\begin{pmatrix}0.78\\0.63\end{pmatrix}, \begin{pmatrix}0.63\\-0.78\end{pmatrix}\right]$	$\left[\begin{pmatrix}0.96\\-0.27\end{pmatrix}, \begin{pmatrix}0.27\\0.96\end{pmatrix}\right]$	$\left[\begin{pmatrix}0.71\\0.71\end{pmatrix}, \begin{pmatrix}0.71\\-0.71\end{pmatrix}\right]$

Unfortunately the results got mixed up. Which result corresponds to which dataset (A, B, C)? Explain your answer.

Problem 10 [1 point] What would the result on dataset D look like (Fig.: 2)? Use the same notation as in Problem 9.

6 Kernels

Consider the following algorithm.

Algorithm 1: Counting something

input : Character string x of length m (one based indexing)
input : Character string y of length n (one based indexing)
output: A number $s \in \mathbb{R}$

```
s ← 0;
for i ← 1 to m do
    for j ← 1 to n do
        if x[i] == x[j] then
            s ← s + 1;
```

Problem 11 [1 point] Explain, in no more than two sentences, what the above algorithm is doing.

Problem 12 [4 points] Let $\mathcal{S}$ denote the set of strings over a finite alphabet of size v. Define a function $K : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ as the output of running algorithm 1 on a pair of strings x, y. Show that $K(x, y)$ is a valid kernel.