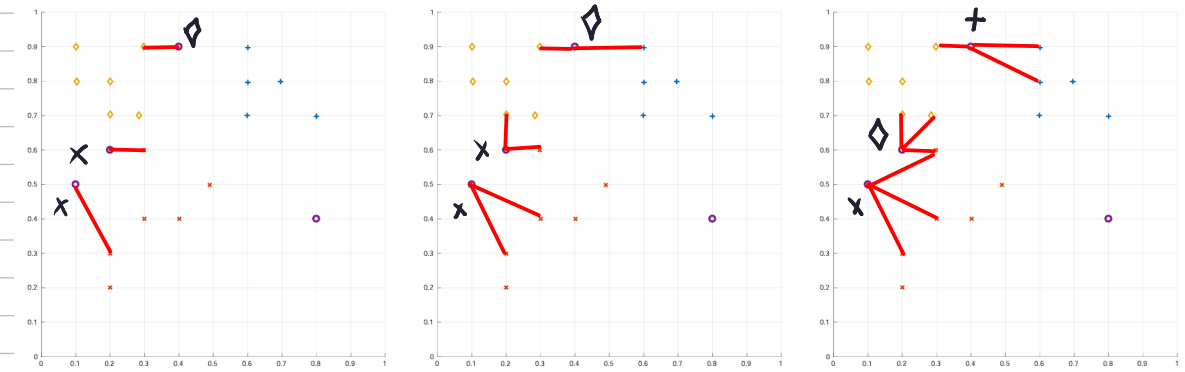
Question 1. (5 points) Graphically show how classification of the unknown instances, depicted as circles, is performed based on the K-nearest neighbors algorithm for $K = 1, 3$ and 5. Textually describe how K-NN works, just designing the result is not sufficient.

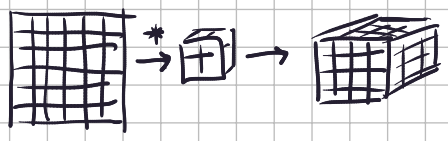


the KNN consider the k nearest samples to a point x and assigns to it the most popular class between them.

the red line connects the points to the closest.

Question 2. (5 points) Briefly describe the main parts that form a single convolutional layer and their function.

Part 1: Convolution) We perform a 2D cross-correlation with a kernel that generates a less wide but more deep tensor:



Part 2: Gating) We apply an activation function on the output tensor.

Part 3: Pooling) We reduce the size of the output tensor, for example, by averaging the components:

1	2	7	9
2	5	1	4
10	2	1	1
3	2	1	1

→

2	5
4	1

Question 3. (5 points) Describe how Principle Components are identified based on the principle of variance maximization.

Consider a data set $D = \{(x_i, t_i)\}_{i=1}^{N}$. Let u a vector. We have $x_i, u \in \mathbb{R}^d$. let S to be the variance matrix of D, we have that $u^T S u$ is the variance of the samples projected on u. IF we want to maximize $u^T S u$ we solve $\underset{u}{\text{argmax}}\ u^T S u$ with Lagrangian multipliers and we find $Su - \lambda u = 0 \Rightarrow Su = \lambda u \Rightarrow u$ is an eigenvector of S with λ eigenvalue. So, the space of dim $M < d$ where we maximize the variance of the projected sample is: $\text{Span}\{u_1 \cdots u_m\}$ where $u_1 \cdots u_m$ are m eigenvector of S associated with the biggest eigenvalues.

Question 4. (5 points)

A car driver in Rome has to move from one side of the Tiber river to the other every morning. There are three possible alternative paths passing to three different bridges and the paths are known. The driver wants to minimize the time to reach the target location, and due to traffic conditions, it is not guaranteed that the shortest path is also the quickest way. Moreover, traffic conditions are unpredictable and non-deterministic, but stationary.

1. Describe a complete model for this problem based on MDP, specifying all its elements.
2. Describe how to solve the problem based on Reinforcement Learning and determine the exact training rule to use to learn the best behavior.
3. Discuss the strategy for balancing exploration and exploitation in this case.

There is an MDP with a single state and 3 actions (3 paths). The reward function is $\frac{1}{T}$ with T the time. Let p_1, p_2 and p_3 to be the paths and x the single state:

$\delta(x, p_i) = x \quad \forall i$

$\mathbb{P}(r_i = \ell) = \mathcal{D}_i(\ell)$ ← unknown distribution.

$\| \; r(x, p_i, x)$

We want to estimate the expected value of r_i and choose p_i that maximize r_i. We consider:

```
Estimate (x, p_i, r_i, K) {
    E_1, E_2, E_3 = 0        // measure the reward
    For t = 1 ... K {
        ε = 1/t
        with probability 1-ε {
            choose i with max E_i
            E_i += r(x, p_i)
        }
        else {
            choose RANDOM i
            E_i += r(x, p_i)
        }
    }
    return p_i that max E_i.
}
```

ε-greedy explore then exploit

Question 5. (5 points)

Confusion matrices are used to summarize the result of a classification algorithm.

1. Provide a formal definition of a confusion matrix.
2. Provide an example (with invented numbers) of a confusion matrix for a classification problem for the target function $f : \{0,1\}^4 \to \{Low, Medium, High\}$ and a data set containing 150 samples (of which 70 for Low, 30 for Medium and 50 for High).
3. Describe how to compute classification accuracy, given a confusion matrix.

The confusion matrix is a metric for an hypothesis in classification problems. Is the matrix C where C_{ij} = number of samples of class i classified as j.

	Low	Medium	High
Low	40	2	4
Medium	20	25	5
High	10	3	41

The accuracy can be computed as:

$$\text{accuracy} = \left(\sum_i \sum_j C_{ij}\right)^{-1} \cdot \text{trace}(C)$$

Question 6. (5 points)

Consider a data set D for scoring different schools with the following real-valued attributes: staff salaries per pupil (x_1), teacher's test score (x_2), parents' education (x_3), school grade (y).

For this problem, an expert of the domain proposes to use the following model.

$$y = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_1 x_2 + \theta_5 x_3^2$$

1. Define an error function for this model.
2. Discuss if a linear model for regression can be used in this case.
3. Describe an iterative approach to solve the problem.

We consider $E = \frac{1}{2}\sum_{i=1}^{N}(t_i - y(x_i))^2$. We consider

$\phi(x_1, x_2, x_3) = (x_1, x_2, x_3, x_1 x_2, x_3^2)$ with a linear model

$y(x) = w^T \phi(x) + w_0$, $w \in \mathbb{R}^5$, $w_0 \in \mathbb{R}$

We consider ∇E and update: $w \leftarrow w - \eta \nabla E$.