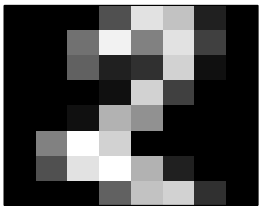
Kötelező feladat **12p. (+5p. opc)**

Felügyelt gépi tanulás.



Egy kettes számjegy.

Az **OCR** tanulási halmazban kézzel írott számjegyek képei találhatók. Az tanulási és teszt adatokat egyenként $8 \times 8 + 1 = 65$ hosszúságú vektorok alakjában tároljuk, ahol az első 64 szám a 8×8-as bittérkép szürke-árnyalatának a kódja, az utolsó érték pedig az osztály kódja 0 és 9 között.

Az adathalmaz a következő állományokból áll:

- `optdigits.tra` – tanulási adatok (3823 darab);
- `optdigits.tes` – teszt adatok (1797 darab);
- `optdigits.tra` – az adatok leírása.

Feladat:

1. Írjunk egy távolságszámoló függvényt, amely két, egyenként $64 = 8 \times 8$ hosszú adatra megadja azok euklideszi távolságát. **2pt.**

2. Az adatok terében definiáljunk egy általánosított skaláris szorzatot – kernel függvényt: $\langle \mathbf{x}, \mathbf{y} \rangle \stackrel{\text{def}}{=} k(\mathbf{x}, \mathbf{y})$ – majd definiáljuk a skaláris szorzat függvényeként a távolságot:

$$\|\mathbf{x} - \mathbf{y}\|^2 = \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle = k(\mathbf{x}, \mathbf{x}) + k(\mathbf{y}, \mathbf{y}) - 2\, k(\mathbf{x}, \mathbf{y}).$$

   Implementáljuk a lineáris, polinomiális és Gauss-féle kerneleket. **2pt.**

3. Implementáljuk a kNN (k-Nearest Neighbor – k legközelebbi szomszéd) algoritmust úgy, hogy az tetszőleges kernelt használhasson. **3pt.**

4. Implementáljuk a centroid módszert úgy, hogy az tetszőleges kernelt használhasson. **3pt.**

5. Számítsuk ki a tanulási és a teszt hibát minden osztályra. **2pt.**

6. 5-szörös kereszt-megerősítést (*cross-validation*) használva határozzuk meg a kernelek optimális paramétereit a kNN és centroid módszerekhez. **opc *5pt.**

Skaláris szorzatok:

$$k_{\text{lin}}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} x_i y_i, \quad k_{\text{pol}}(\mathbf{x}, \mathbf{y};\, p) = \left(1 + \sum_{i=1}^{d} x_i y_i\right)^p, \quad k_{\text{gauss}} = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right)$$

Adatbázis: *Optical Recognition of Handwritten Digits*

http://archive.ics.uci.edu/ml/machine-learning-databases/optdigits