TOTEUTUSDOKUMENTTI

Ohjelman nimi: KNN, k-lähimmän naapurin algoritmi

Päivämäärä: 28.10.2025

YLEISKUVAUS

KNN-luokka, k-lähimmän naapurin algoritmi käsinkirjoitettujen numeroiden MNIST tietokanta luokitteluun. Käyttäjä voi valita etäisyysmitan (4 kpl), naapurien lukumäärän, koulutuskuvien määrän sekä vaikuttaa tarkastelualueen kokoon.

LUOKAT JA RAKENNE

KNN-luokan ominaisuuksia:

Etäisyysmitat: Euklidinen sekä mitat D22, D23, ja D23 ilman keskiarvoistamista viitteen (Dubuisson & Jain, 1994) mukaisesti. Koulutusdatana käytetään MNIST tietokannan kuvia.

Ennen tunnistusta kuvat muutetaan mustavalkoisiksi, pikselillä joko arvo 0 tai 1 (valkoinen) sekä koordinaattilistoiksi.

Etäisyyden laskeminen tehdään kahdessa vaiheessa, ensin tutkitaan kuvien pikseleiden lähialue boolean-listasta, mikäli vastaavuutta ei löydy toisesta kuvasta, seuraavaksi käydään läpi toisen kuvan koordinaattilista. Mikäli vastaavuus löytyy boolean-listasta on etäisyyden laskeminen nopeaa. Boolean listalla tarkistettavan alueen koko määritetään ympyrän halkaisijalla.

Ohjelma laskee k lähintä naapuria ja palauttaa sen numeron, jota esiintyy useimmin naapurien joukossa. Ohjelma on toteutettu ilman valmiiden kirjastojen funktioita. Sen tehokkuutta verrataan k-lähimmän naapurin algoritmiin, joka perustuu numpy kirjasto funktioihin.

AIKA- JA TILAVAATIVUUDET

Oman analyysin perusteella, katsomalla algoritmin rakennetta, algoritmin aikavaatimus riippuu siitä mitä etäisyysmittaa käytetään. Jos käytetään Euklidista etäisyysmittaa aikavaatimus on O(np), missä n on koulutuskuvien lukumäärä ja p on valkoisten pikselien lukumäärä. Mikäli käytetään viitteen (Dubuisson & Jain, 1994) etäisyysmittoja: D22, D23, ja D23 ilman keskiarvoistamista, on aikavaatimus $O(np^2)$.

Tilavaativuutta arvioidaan koulutuskuvien viemän muistin perusteella. Tilavaativuus on O(nd), missä n on koulutuskuvien lukumäärä ja d on yhden kuvan tarvitsema tila.

SUORITUSKYKY- JA O-ANALYYSIVERTAILU

Suorituskykyvertailua on esitetty testausdokumentin kohdassa ”Ohjelman toiminnan mahdollinen empiirinen testaus”, jossa on verrattu eri etäisyysmittoja, koulutuskuvien lukumäärää ja naapurien lukumäärää.

O-analyysivertailu on osittain esitetty kohdassa ”aika- ja tilavaativuudet”, mainittakoon että viitteen (Dubuisson & Jain, 1994) etäisyysmitat ovat tarkempia kuin Euklidinen etäisyysmitta.

TYÖN MAHDOLLISET PUUTTEET JA PARANNUSEHDOTUKSET

Ohjelmaa pitäisi nopeuttaa, 60000 kuvan käyttäminen koulutuksessa vie ehkä hieman liikaa aikaa.

Tarkkuudesta sen verran että en itse tunnista jokaista MNIST tietokannan kuvaa, jos piirrän kuvan. Jos en tietäisi että kyseessä on kuva numerosta, voisin luulla useita kuvia joksikin muuksi.

Tarkkuuteen alle 4% virheellisiä tunnistuksia pitäisi päästä noin 12 000 koulutuskuvalla.

LAAJOJEN KIELIMALLIEN (CHATGPT YMS.) KÄYTTÖ

Tässä työssä hyödynnettiin OpenAI:n ChatGPT-kielimallia (versio GPT-4, syyskuu 2025):

Haettu ohjeita, miten käytetään poetry ja pylint. Testien kirjoittamiseen EI OLE hyödynnetty.

Haettu ohjeita aiheista: aika- ja tilavaativuudet sekä suorituskyky- ja O-analyysivertailu.

LÄHTEET

Dubuisson, M.-P., & Jain, A. K. (1994). A modified Hausdorff distance for object matching. Proceedings of 12th International Conference on Pattern Recognition, 566–568. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=576361

k-nearest neighbors algorithm. (2025). viewed 2 September 2025, https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Harjoitustyön testaaminen. (2025). viewed 2 September 2025, https://algolabra-hy.github.io/testing-fi