Datasta Tietoon, syksy 2011

TENTTI 11. 1. 2012 (note: problems in English on a separate paper)

1.

On annettuna d-alkioinen vektori $\mathbf{x} = (x_1, x_2, ..., x_d)^T$. Mitä tarkoittaa, kun sanotaan että se on normaalijakautunut? Montako parametriä tarvitaan, jotta sen normaalijakauma on yksikäsitteisesti määrätty?

2.

On annettu n mittausparia $(y(i), x(i))$, $i = 1, .., n$ joistakin muuttujista x, y joiden välillä arvellaan olevan seuraavanlainen lineaarinen yhteys: $y = \alpha + \theta x$. Mittauksiin sisältyy kuitenkin virhettä: $y(i) = \alpha + \theta x(i) + \epsilon(i)$ missä $\epsilon(i)$ on mittausvirhe ("kohina") i:nnessä pisteessä. Oletetaan että mittausvirhe $\epsilon(i)$ on normaalijakautunut keskiarvolla 0 ja keskihajonnalla σ.

Meillä on käytettävissä etukäteistietoa:

1. Arvellaan että kulmakerroin θ on suunnilleen 1. Mallitetaan tähän liittyvä epävarmuus olettamalla normaalinen priorijakauma jonka keskiarvo on 1 ja keskihajonta 0.5.

2. Arvellaan että regressiosuora kulkee suunilleen origon kautta, jolloin parametri α on osapuilleen nolla. Mallitetaan parametriin α liittyvä epävarmuus olettamalla että sillä on normaalinen priorijakauma jonka keskiarvo on 0 ja keskihajonta 0.1.

Laske Bayes-estimaatit parametreille α, θ.

3.

a) Selosta miten toimii c-means -ryhmittelyalgoritmi.

b) On annettuna seuraava datamatriisi:

$$X = \begin{bmatrix} 1 & 1 & 3 & 3 & 6 & 8 \\ 2 & 4 & 2 & 4 & 1 & 1 \end{bmatrix}$$

Ryhmittele sen sarakkeet 2 ryhmään (klusteriin) c-means-algoritmilla. Laita aluksi ryhmien keskipistevektorit arvoihin

$$\mathbf{m}_1 = \begin{bmatrix} 5 \\ 4 \end{bmatrix}; \ \mathbf{m}_2 = \begin{bmatrix} 8 \\ 4 \end{bmatrix}.$$

Piirrä kuva algoritmin toiminnasta.

4.

a) Mitä ovat kattavat joukot? Määrittele ne.

b) Luonnostele tasoittainen algoritmi kattavien joukkojen löytämiseksi (todistuksia ei tarvita).

c) Kaupassa on käynyt kymmenen asiakasta, jotka ovat ostaneet kassatietojärjestelmän mukaan seuraavat tuotteet, jossa p = pasta, t = tomaatti, o = sipuli ja b = basilika: $\{t, b\}$, $\{b\}$, $\{p\}$, $\{o, b\}$, $\{t, o, b\}$, $\{t, o\}$, $\{p, o, b\}$, $\{t, o, b\}$, $\{p\}$, $\{p, t, o, b\}$. Löydä kattavat joukot kuvailemalla tasoittaisen algoritmin toimintaa, kun kynnysarvo on 4.

(Jatkuu)

5.

a) kNN-luokitin (k:n lähimmän naapurin luokitin): kerro lyhyesti, mitä sillä tarkoitetaan ja mihin ja miten sitä käytetään yleisellä tasolla.

b) Käytössä on neljä opetuspistettä $\mathbf{d}_{cj} \in R^4$, joista kaksi ensimmäistä kuuluu luokkaan A ja kaksi jälkimmäistä luokkaan B:

$$\mathbf{d}_{A1} = \begin{bmatrix} 3 \\ -2 \\ -1 \\ -3 \end{bmatrix}, \quad \mathbf{d}_{A2} = \begin{bmatrix} -1 \\ 2 \\ -2 \\ -1 \end{bmatrix}, \quad \mathbf{d}_{B1} = \begin{bmatrix} 2 \\ 3 \\ -1 \\ -2 \end{bmatrix}, \quad \mathbf{d}_{B2} = \begin{bmatrix} 3 \\ 0 \\ 1 \\ 2 \end{bmatrix}$$

Luokittele uusi datapiste

$$\mathbf{x} = \begin{bmatrix} 2 & -2 & -2 & 2 \end{bmatrix}^T$$

kun käytät kNN-luokitinta (i) $k = 1$ naapurilla, (ii) $k = 3$ naapurilla. Käytä euklidista (neliöllistä) etäisyyttä. Mikä on datapisteen saama luokka kohdissa (i) ja (ii)?