HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos
PL 68 (Gustaf Hällströmin katu 2b)
00014 HELSINGIN YLIOPISTO

582631 Johdatus koneoppimiseen
Laskuharjoitus 1
ke 29.10.2008 klo 10-12
sali B221

Näissä harjoituksissa tutustutaan `Matlab`iin ja palautellaan mieliin joitain todennäköisyyslaskennan perustuloksia. Tehtäviä kannattaa yrittää tehdä itsenäisesti etukäteen, mutta niiden tekemistä voi jatkaa myös ohjatusti itse laskuharjoitustilaisuudessa. Harjoitusten tekemisessä saattaa olla hyötyä Jarmo Hurrin `Matlab`-materiaalista ja `Matlab`in online-dokumentaatiosta. Alkupään tehtävistä on suoraan linkit dokumentaatioon, mutta loppua kohden sopivat funktiot pitää löytää itse. . .

Ensimmäisiin laskuharjoituksiin osallistuminen on pakollista! Jos et millään pääse paikalle, asiasta voi neuvotella Mika Urtelan tai Matti Kääriäisen kanssa.

1. Generoi `Matlab`illa 100 näytettä normaalijakautuneesta satunnaismuuttujasta, jonka keskiarvo on 0 ja varianssi 1 (kts. funktio randn). Visualisoi generoimaasi otosta (kts. esim. hist). Järjestä otoksen luvut siten suuruusjärjestykseen (kts. funktio sort). Piirrä kuvaaja (kts. funktio plot), jossa järjestetty otos on x-akselilla, ja järjestyksessä i:nteen x-arvoon liittyvä y on $i/100$ (kts. operaattori :).

   Osaatko sanoa mitä tuli piirrettyä?

2. Kirjoita `Matlab`-funktio, joka

   - ottaa argumenttinaan tiedoston nimen
   - lukee tiedostosta datamatriisin `X` (kts. funktio load)
   - tulostaa matriisin `X` rivi- ja sarakesummat pylväskuvaajina (kts. funktiot sum, bar, ja figure tai subplot)
   - palauttaa kaksi paluuarvoa: matriisin `X` koon (vektori; kts. funktio size) ja matriisin `X` alkioiden summan.

   Testaa funktiosi toimintaa generoimallasi ja tallentamallasi (kts. save) tiedostolla.

3. Tässä todennäköisyyslaskennan tärkeimpiin tuloksiin kuuluvaa *suurten lukujen lakia* empiirisesti.

   Muistin virkistykseksi: Suurten lukujen laki sanoo, että jos satunnaismuuttujan X odotusarvo on äärellinen ja $X_1, \ldots, X_n$ ovat riippumattomia ja samoin jakautuneita kuin X, niin

   $$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mathbb{E}[X]$$

   todennäköisyydellä 1. Intuitiivisemmin (ja epätarkemmin) tämän voi ilmaista esimerkiksi seuraavasti: kun n on “riittävän suuri”, satunnaismuuttujien X_i otoskeskiarvo on “suurella todennäköisyydellä” “lähellä” vakiota $\mathbb{E}[X]$.

Havainnollista suurten lukujen lakia empiirisesti: Valitse satunnaismuuttujalle X jakauma, generoi jakaumasta riittävän suuri otos, ja piirrä kuvaaja siitä miten otoskeskiarvo käyttäytyy otoksen koon kasvaessa. Toista koe varmistaaksesi, että empiirinen havaintosi pitää paikkaansa "suurella todennäköisyydellä".

4. Palautetaan mieliin, että *odotusarvo on lineaarinen operaattori*: kaikille satunnaismuuttujille X ja Y pätee, että

$$\mathbb{E}[\alpha X] = \alpha\mathbb{E}[X]$$

ja

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Olkoon F^n satunnaismuuttuja, joka kuvaa satunnaispermutaation (kts. funktio randperm) *kiintopisteiden* (eli pisteiden jotka jäävät permutaatiossa paikalleen) lukumäärää.

Seuraavassa selvitetään satunnaismuuttujan F^n odotusarvoa ensin empiirisesti ja sitten analyyttisesti odotusarvon lineaarisuusominaisuutta hyödyntäen.

(a) Kirjoita `Matlab`-funktio, joka saa parametrinaan luvun n, ja palauttaa näytteen satunnaismuuttujan F^n jakaumasta.

(b) Generoi muutamilla n:n arvoilla sopivan kokoinen otos F^n:n jakaumasta, ja piirrä otoksista histogrammit (kts. funktio hist). Laske otoksien keskiarvot, ja arvaa niiden perusteella mikä $\mathbb{E}[F^n]$ voisi olla.

(c) Sovella odotusarvon lineaarisuutta $\mathbb{E}[F^n]$:n laskemiseen. (Vihje: Kirjoita $F^n = \sum_{i=1}^{n} F_i^n$, missä satunnaismuuttuja F_i^n saa arvon 1 jos i pysyy permutaatiossa paikallaan ja muuten arvon 0. Laske $\mathbb{E}[F_i^n]$, ja sovella odotusarvon lineaarisuutta.)

5. (a) Kirjoita `Matlab`-funktio, joka saa parametreinaan todennäköisyyden p, kahden 1-ulotteisen normaalijakauman odotusarvo- ja varianssiparametrit, ja otoskoon n. Tuloksenaan funktio palauttaa kaksi n-alkioista vektoria `x` ja `y`, missä parit (`x(i)`, `y(i)`) on generoitu toisistaan riippumattomasti seuraavasti: Todennäköisyydellä p `y(i)=0` ja `x(i)` valitaan ensimmäisestä normaalijakaumasta; muuten `y(i)=1` ja `x(i)` valitaan toisesta normaalijakaumasta.

(b) Visualisoi kirjoittamasi funktion generoimaa dataa eri parametrien arvoilla.

(c) Lisätehtävä: Yleistä funktiosi siten, että datan `x`-komponentit noudattavat parametrien kuvaamaa 2-ulotteista normaalijakaumaa. Visualisoi.