Laskentaintensiiviset tilastolliset menetelmät (kuulustelija: Petri Koistinen)
Tentti 20.1. 2004.
Merkitse vastauspaperiin oman nimen lisäksi joko sosiaaliturvatunnus tai opiskelijanumero.

Tiedoksi: Voit antaa kurssista nimettömästi palautetta vastaamalla matematiikan laitoksen järjestämään kurssikyselyyn. Linkit löytyvät sekä kurssin kotisivulta että matematiikan laitoksen kotisivulta.

1. Selitä lyhyesti (esim. kaavan avulla) seuraavat käsitteet.

a) Jensenin epäyhtälö (2 p),

b) jatkuvan jakauman kvantiilifunktio (2 p),

c) aineiston empiirinen kvantiilifunktio (2 p).

2. Tahdomme arvioida integraalia $\int g(x)f(x)\,\mathrm{d}x$, jossa f on tiheysfunktio, josta osataan generoida satunnaisarvoja $X_1, X_2, \ldots$, ja g on reaaliarvoinen funktio.

a) Selosta, miten integraalia arvioidaan Monte Carlo -menetelmällä.

b) Alustavien tarkastelujen perusteella var $g(X_1)$ on välillä $0.8 \ldots 1.0$. Miten suuri otoskoko tarvitaan, jotta integraalin arvion keskihajonta olisi korkeintaan 10^{-3}?

3. Johda Gibbsin otantaan perustuva simulointialgoritmi kohdetiheydelle f, jossa

$$f(x_1, x_2, x_3) = cx_1^4 x_2^3 x_3^2 (1 - x_1 - x_2 - x_3), \quad \text{kun}$$
$$x_1 > 0, x_2 > 0, x_3 > 0 \text{ ja } x_1 + x_2 + x_3 < 1,$$

ja f on nolla muualla.

Ohjeita: betajakauman $B(a, b), a > 0, b > 0$ tiheysfunktio on normalisointivakiota vaille $u^{a-1}(1-u)^{b-1}$ välillä $0 < u < 1$ ja nolla muualla. Selvitä itsellesi, mikä on muuttujan kU tiheysfunktio, jos $U \sim B(a, b)$ ja k on vakio.

4. Kirjoita pseudokoodilla EM-algoritmit (a) suurimman uskottavuuden estimaatin hakuun ja (b) Bayes-mallin posteriorijakauman moodin hakuun. Selitä käyttämäsi merkinnät.

5. Selosta valintasi mukaan toinen seuraavista parametrittomista regressiofunktion estimointimenetelmistä, k-lähinaapuriestimaatti tai Nadarayan–Watsonin estimaatti. Selosta, miten menetelmän silotusparametri voidaan valita v-kertaisella ristiinvalidoinnilla. Mitän varten mentelmää pidetään parametrittomana, vaikka sen tulos riippuu silotusparametrin arvosta?

Matematiikan ja tilastotieteen laitos
Laskentaintensiiviset tilastolliset menetelmät
Loppukoe 10.8.2005 (kuulustelija: Petri Koistinen)

Ohje tehtäviin 1 ja 3: liitteessä on lueteltu jakaumia, joista osataan generoida satunnaisarvoja.

1. Esitä algoritmi, jolla voidaan simuloida arvoja satunnaismuuttujalle, jolla on normalisoimaton tiheysfunktio

$$h(x) = \frac{x^2}{1+x}\,\mathrm{e}^{-x}\,1_{[0,\infty)}(x).$$

2. Tahdomme arvioida integraalia $\int g(x)f(x)\,\mathrm{d}x$, jossa f on tiheysfunktio, josta osataan generoida satunnaisarvoja $X_1, X_2, \ldots$, ja g on reaaliarvoinen funktio.

a) Selosta, miten integraalia arvioidaan Monte Carlo -menetelmällä.

b) Alustavien tarkastelujen perusteella $\mathrm{var}\, g(X_1)$ on välillä $0.8 \ldots 1.0$. Miten suuri otoskoko tarvitaan, jotta integraalin arvion keskihajonta olisi korkeintaan 10^{-3}?

3. Johda Gibbsin otantaan perustuva simulointialgoritmi kohdetiheydelle f, jossa

$$f(x_1,x_2,x_3) = cx_1^4x_2^3x_3^2(1-x_1-x_2-x_3), \quad \text{kun}$$
$$x_1>0, x_2>0, x_3>0 \text{ ja } x_1+x_2+x_3<1,$$

ja f on nolla muualla.

Vihje: Selvitä itsellesi, mikä on muuttujan kU tiheysfunktio, jos U:lla on tiheysfunktio f ja k on vakio.

4. Meillä on havaintoja $y_1, \ldots, y_n$ satunnaismuuttujista $Y_1, \ldots, Y_n$, joille

$$Y_i = \begin{cases} X_i, & \text{jos } X_i < c \\ c, & \text{jos } X_i \geq c, \end{cases} \qquad i = 1, \ldots, n,$$

ja jossa satunnaismuuttujat $X_i, i = 1, \ldots, n$ ovat riippumattomia, ja kullakin niillä on eksponenttijakauma $\mathrm{Exp}(\theta)$. Tässä c on tunnettu vakio, mutta $\theta > 0$ on tuntematon parametri.

Johda EM-algoritmi parametrin θ suurimman uskottavuuden estimaatin laskemiseksi.

5. Selosta valintasi mukaan toinen seuraavista parametrittomista regressiofunktion estimointimenetelmistä: k-lähinaapuriestimaatti tai Nadarayan–Watsonin estimaatti. Mitä varten selostamaasi mentelmää pidetään parametrittomana, vaikka sen tuottama tulos riippuu silotusparametrin arvosta?

Selosta lisäksi, miten menetelmän silotusparametri voidaan valita V-kertaisella ristiinvalidoinnilla.

Liite: jakaumia

Betajakauma. Betajakaumalla $\mathrm{Be}(a, b)$ parametreilla $a > 0$ ja $b > 0$ on tiheysfunktio

$$\frac{1}{B(a,b)}\, x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1.$$

Eksponenttijakauma. Eksponenttijakaumalla $\mathrm{Exp}(\lambda)$ parametrilla $\lambda > 0$ on tiheysfunktio

$$\lambda \mathrm{e}^{-\lambda x}, \quad x \geq 0.$$

Gammajakauma. Gammajakaumalla $\mathrm{Gamma}(a, b)$ parametreillä $a > 0$ ja $b > 0$ on tiheysfunktio

$$\frac{b^a}{\Gamma(a)}\, x^{a-1}\mathrm{e}^{-bx}, \quad x > 0.$$

Normaalijakauma. Normaalijakaumalla $N(\mu, \sigma^2)$ on tiheysfunktio

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right).$$

Poissonin jakauma. Poissonin jakaumalla parametrilla $\theta > 0$ on ptnf

$$p(k) = \frac{1}{k!}\, \theta^k \mathrm{e}^{-\theta}, \quad k = 0, 1, 2, \ldots$$

Binomijakauma. Binomijakaumalla $\mathrm{Bin}(n, p)$ on ptnf

$$\binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Multinomijakauma. Multinomijakaumalla $\mathrm{Mult}(n, p_1, \ldots, p_m)$, jossa $0 \leq p_i \leq 1$ ja $\sum_i p_i = 1$, on ptnf

$$p(y_1, \ldots, y_m) = \begin{cases} \dfrac{n!}{y_1! \cdots y_m!} p_1^{y_1} \cdots p_m^{y_m} & \text{kun } y_1 + \cdots + y_m = n, \\ 0 & \text{muuten.} \end{cases}$$

Matematiikan ja tilastotieteen laitos
Laskentaintensiiviset tilastolliset menetelmät
Loppukoe 26.1.2006 (Kuulustelija: Petri Koistinen)

Laskutehtävissä annetaan pisteitä myös osittaisista vastauksista, joissa selitetään, millä periaatteella täydelliseen vastaukseen tarvittavat suureet saataisiin laskettua.

1. Esitä simulointialgoritmi jakaumalle, jonka tiheysfunktio on normalisointia vaille $h(x) = (1 + \cos(x)) \exp(-x^2/2)$.

2. Arvioimme integraalia $\int g(x)f(x)\,\mathrm{d}x$, jossa f on tiheysfunktio, josta osaamme generoida riippumattomia satunnaisarvoja $X_1, X_2, \ldots$, ja g on reaaliarvoinen funktio, jonka arvoja osaamme laskea.

a) Selosta, miten integraalia arvioidaan Monte Carlo -menetelmällä.

b) Alustavien tarkastelujen perusteella $\mathrm{Var}\, g(X_1)$ on välillä $0.8 \ldots 1.0$. Miten suuri otoskoko tarvitaan, jotta integraalin arvion keskihajonta olisi korkeintaan 10^{-3}?

3. Tarkastellaan Bayes-päättelyä mallissa, jossa on kaksi parametria θ ja a. Parametri a on kokonaisluku joukosta $\{1, \ldots, 10\}$, ja sen priorijakauma on kyseisen diskreetin joukon tasajakauma (jonka ptnf on vakio $1/10$). Ehdolla a parametrilla θ on gammajakauma $\mathrm{Gamma}(a, 1)$. Ehdolla θ ja a satunnaismuuttujat Y_i, $i = 1, \ldots, n$ ovat riippumattomia, ja Y_i noudattaa Poissonin jakaumaa parametrilla θ. Satunnaismuuttujien Y_i arvot y_i on havaittu, $i = 1, \ldots, n$.

Esitä jokin Markovin ketjuihin perustuva Monte Carlo -menetelmä, jolla voidaan simuloida parametrien (yhteis)posteriorijakaumaa.

4. Olemme saaneet havaintoja $y_1, \ldots, y_n$ satunnaismuuttujista $Y_1, \ldots, Y_n$, jotka noudattavat mallia

$$Y_i = \begin{cases} 0, & \text{jos } X_i \le c \\ 1, & \text{jos } X_i > c, \end{cases}$$

jossa satunnaismuuttujat X_i, $i = 1, \ldots, n$ ovat riippumattomia, ja kukin niistä noudattaa eksponenttijakaumaa $\mathrm{Exp}(\theta)$. Tässä c on tunnettu vakio, ja θ on tuntematon parametri.

Johda EM-algoritmi suurimman uskottavuuden estimaatin laskemiseksi.

5. Vastaa, valintasi mukaan, **toiseen** seuraavista kysymyksistä.

a) Selosta, miten lasketaan studentisoitu saapasremmiluottamusväli.

b) Selosta, miten regressiofunktiota arvioidaan lokaalin lineaarisen regression avulla.

Liite: Jakaumia

Betajakauma. Betajakaumalla $\mathrm{Be}(a, b)$ parametreilla $a > 0$ ja $b > 0$ on tf

$$\frac{1}{B(a,b)}\, x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1.$$

Eksponenttijakauma. Eksponenttijakaumalla $\mathrm{Exp}(\lambda)$ parametrilla $\lambda > 0$ on tf

$$\lambda \mathrm{e}^{-\lambda x}, \quad x \geq 0,$$

ja sen odotusarvo on $1/\lambda$ ja varianssi $1/\lambda^2$.

Gammajakauma. Gammajakaumalla $\mathrm{Gamma}(a, b)$ parametreillä $a > 0$ ja $b > 0$ on tf

$$\frac{b^a}{\Gamma(a)}\, x^{a-1}\mathrm{e}^{-bx}, \quad x > 0,$$

ja sen odotusarvo on a/b ja varianssi a/b^2.

Normaalijakauma. Normaalijakaumalla $N(\mu, \sigma^2)$ on tf

$$\frac{1}{\sqrt{2\pi}\sigma}\, \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right),$$

ja sen odotusarvo on μ ja varianssi σ^2.

Poissonin jakauma. Poissonin jakaumalla parametrilla $\theta > 0$ on ptnf

$$p(k) = \frac{1}{k!}\, \theta^k \mathrm{e}^{-\theta}, \quad k = 0, 1, 2, \ldots,$$

ja sen odotusarvo on θ ja varianssi on θ.

Binomijakauma. Binomijakaumalla $\mathrm{Bin}(n, p)$ on ptnf

$$\binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n,$$

ja sen odotusarvo on np ja varianssi $np(1-p)$.

Multinomijakauma. Multinomijakaumalla $\mathrm{Mult}(n, (p_1, \ldots, p_m))$, jossa $0 \leq p_i \leq 1$ ja $\sum_i p_i = 1$, on ptnf

$$p(y_1, \ldots, y_m) = \begin{cases} \dfrac{n!}{y_1! \cdots y_m!} p_1^{y_1} \cdots p_m^{y_m} & \text{kun } y_1 + \cdots + y_m = n, \\ 0 & \text{muuten.} \end{cases}$$

Matematiikan ja tilastotieteen laitos
Laskentaintensiiviset tilastolliset menetelmät
Loppukoe 9.8.2006 (Kuulustelija: Petri Koistinen)

Huom. Laskutehtävissä annetaan pisteitä myös epätäydellisistä vastauksista, joissa selitetään, millä periaatteella täydelliseen vastaukseen tarvittavat suureet saataisiin laskettua.

1. Esitä simulointialgoritmi jakaumalle, jonka tiheysfunktio on normalisointia vaille $h(x) = (2 + \sin(x)) \exp(-x^2/2)$.

2. Arvioimme integraalia $\int g(x)f(x)\,\mathrm{d}x$, jossa f on tiheysfunktio, josta osaamme generoida riippumattomia satunnaisarvoja $X_1, X_2, \ldots$, ja g on reaaliarvoinen funktio, jonka arvoja osaamme laskea.

a) Selosta, miten integraalia arvioidaan Monte Carlo -menetelmällä.

b) Alustavien tarkastelujen perusteella $\mathrm{Var}\, g(X_1)$ on välillä $0.8 \ldots 1.0$. Miten suuri otoskoko tarvitaan, jotta integraalin arvion keskihajonta olisi korkeintaan 10^{-3}?

3. Tarkastellaan Bayes-päättelyä mallissa, jossa on kaksi parametria θ ja a. Parametri a on kokonaisluku joukosta $\{1, \ldots, 10\}$, ja sen priorijakauma on kyseisen diskreetin joukon tasajakauma (jonka ptnf on vakio $1/10$). Ehdolla a parametrilla θ on gammajakauma $\mathrm{Gamma}(a, 1)$. Ehdolla θ ja a satunnaismuuttujat Y_i, $i = 1, \ldots, n$ ovat riippumattomia, ja Y_i noudattaa Poissonin jakaumaa parametrilla θ. Satunnaismuuttujien Y_i arvot y_i on havaittu, $i = 1, \ldots, n$.

Esitä jokin Markovin ketjuihin perustuva Monte Carlo -menetelmä, jolla voidaan simuloida parametrien (yhteis)posteriorijakaumaa.

4. Olemme saaneet havaintoja $y_1, \ldots, y_n$ satunnaismuuttujista $Y_1, \ldots, Y_n$, jotka noudattavat mallia

$$Y_i = \begin{cases} 0, & \text{jos } X_i \le c \\ 1, & \text{jos } X_i > c, \end{cases}$$

jossa satunnaismuuttujat X_i, $i = 1, \ldots, n$ ovat riippumattomia, ja kukin niistä noudattaa eksponenttijakaumaa $\mathrm{Exp}(\theta)$. Tässä c on tunnettu vakio, ja θ on tuntematon parametri.

Johda EM-algoritmi suurimman uskottavuuden estimaatin laskemiseksi.

5. Vastaa, valintasi mukaan, **toiseen** seuraavista kysymyksistä.

a) Selosta, miten lasketaan studentisoitu saapasremmiluottamusväli.

b) Selosta, miten regressiofunktiota arvioidaan lokaalin lineaarisen regression avulla.

Liite: Jakaumia

Betajakauma. Betajakaumalla $\mathrm{Be}(a, b)$ parametreilla $a > 0$ ja $b > 0$ on tf

$$\frac{1}{B(a,b)}\, x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1.$$

Eksponenttijakauma. Eksponenttijakaumalla $\mathrm{Exp}(\lambda)$ parametrilla $\lambda > 0$ on tf

$$\lambda \mathrm{e}^{-\lambda x}, \quad x \geq 0,$$

ja sen odotusarvo on $1/\lambda$ ja varianssi $1/\lambda^2$.

Gammajakauma. Gammajakaumalla $\mathrm{Gamma}(a, b)$ parametreillä $a > 0$ ja $b > 0$ on tf

$$\frac{b^a}{\Gamma(a)}\, x^{a-1}\mathrm{e}^{-bx}, \quad x > 0,$$

ja sen odotusarvo on a/b ja varianssi a/b^2.

Normaalijakauma. Normaalijakaumalla $N(\mu, \sigma^2)$ on tf

$$\frac{1}{\sqrt{2\pi}\sigma}\, \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right),$$

ja sen odotusarvo on μ ja varianssi σ^2.

Poissonin jakauma. Poissonin jakaumalla parametrilla $\theta > 0$ on ptnf

$$p(k) = \frac{1}{k!}\, \theta^k \mathrm{e}^{-\theta}, \quad k = 0, 1, 2, \ldots,$$

ja sen odotusarvo on θ ja varianssi on θ.

Binomijakauma. Binomijakaumalla $\mathrm{Bin}(n, p)$ on ptnf

$$\binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n,$$

ja sen odotusarvo on np ja varianssi $np(1-p)$.

Multinomijakauma. Multinomijakaumalla $\mathrm{Mult}(n, (p_1, \ldots, p_m))$, jossa $0 \leq p_i \leq 1$ ja $\sum_i p_i = 1$, on ptnf

$$p(y_1, \ldots, y_m) = \begin{cases} \dfrac{n!}{y_1! \cdots y_m!} p_1^{y_1} \cdots p_m^{y_m} & \text{kun } y_1 + \cdots + y_m = n, \\ 0 & \text{muuten.} \end{cases}$$

Matematiikan ja tilastotieteen laitos
Laskentaintensiiviset tilastolliset menetelmät
Loppukoe 12.4.2007 (Kuulustelija: Petri Koistinen)

Huom. Laskutehtävissä annetaan pisteitä myös epätäydellisistä vastauksista, joissa selitetään, millä periaatteella täydelliseen vastaukseen tarvittavat suureet saataisiin laskettua.

1. Esitä simulointialgoritmi jakaumalle, jonka tiheysfunktio on normalisointivakiota vaille $h(x) = (2 + \sin(x) + \sin^2(3x)) \exp(-x^2/2)$.

2. Tahdomme arvion odotusarvolle

$$I = E|X|^{1.6}, \quad \text{jossa} \quad X \sim N(0,1).$$

Selosta, miten odotusarvo saadaan arvioitua käyttämällä Monte Carlo -integrointia ja kontrollimuuttujaa X^2. Selosta myös, miten I:lle saadaan laskettua luottamusväli.

3. Seuraavassa hierarkkisessa Bayes-mallissa Y_i kuvaa utaretulehdusten tapausten lukumäärää i:nnessä karjalaumassa. Sillä ehdolla, että $\Theta_i = \theta_i, i = 1, \ldots, n$, ovat

$$Y_i \sim \text{Poisson}(\theta_i), \qquad i = 1, \ldots, n$$

riippumattomasti. Tässä ehdolla $B_i = b_i$, $i = 1, \ldots, n$, ovat

$$\Theta_i \sim \text{Gamma}(a, b_i), \qquad i = 1, \ldots, n$$

riippumattomasti ja lopulta

$$B_i \sim \text{Gamma}(c, d), \qquad i = 1, \ldots, n$$

riippumattomasti. Luvut a, c ja d ovat tunnettuja vakioita. Satunnaismuuttujien Y_i arvot y_i on havaittu, $i = 1, \ldots, n$. Suureet $\Theta_1, \ldots, \Theta_n$ ja $B_1, \ldots, B_n$ ovat parametreja.

Esitä Gibbsin otanta-algoritmi parametrien posteriorijakaumalle.

4. Olemme saaneet havaintoja $y_1, \ldots, y_n$ satunnaismuuttujista $Y_1, \ldots, Y_n$, jotka noudattavat mallia

$$Y_i = \begin{cases} 0, & \text{jos } X_i \le c \\ 1, & \text{jos } X_i > c, \end{cases}$$

jossa satunnaismuuttujat X_i, $i = 1, \ldots, n$ ovat riippumattomia, ja kukin niistä noudattaa eksponenttijakaumaa $\text{Exp}(\theta)$. Tässä c on tunnettu vakio, ja θ on tuntematon parametri.

Johda EM-algoritmi suurimman uskottavuuden estimaatin laskemiseksi.

5. Vastaa, valintasi mukaan, **toiseen** seuraavista kysymyksistä.

a) Selosta, miten lasketaan studentisoitu saapasremmiluottamusväli.

b) Selosta, miten regressiofunktiota arvioidaan lokaalin lineaarisen regression avulla.

Liite: jakaumia

Betajakaumalla $\mathrm{Be}(a,b)$ parametreilla $a>0$ ja $b>0$ on tf

$$\frac{1}{B(a,b)}\,x^{a-1}(1-x)^{b-1},\quad 0<x<1.$$

Odotusarvo on $\frac{a}{a+b}$ ja varianssi on $\frac{ab}{(a+b)^2(a+b+1)}$. $B(a,b)$ on betafunktion arvo,

$$B(a,b)=\int_0^1 u^{a-1}(1-u)^{b-1}\,\mathrm{d}u=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Eksponenttijakaumalla $\mathrm{Exp}(\lambda)$ parametrilla $\lambda>0$ on tf

$$\lambda\,\mathrm{e}^{-\lambda x},\quad x\geq 0.$$

Odotusarvo on λ^{-1} ja varianssi on λ^{-2}.

Gammajakaumalla $\mathrm{Gamma}(a,b)$ parametreillä $a>0$ ja $b>0$ on tf

$$\frac{b^a}{\Gamma(a)}\,x^{a-1}\mathrm{e}^{-bx},\quad x>0.$$

Odotusarvo on a/b ja varianssi ab^{-2}. $\Gamma(a)$ on gammafunktion arvo,

$$\Gamma(a)=\int_0^\infty x^{a-1}\mathrm{e}^{-x}\,\mathrm{d}x,\qquad a>0.$$

Tunnetusti $\Gamma(a+1)=a\Gamma(a)$ kaikilla $a>0$, ja $\Gamma(1)=1$, joten $\Gamma(n)=(n-1)!$, kun $n=1,2,3,\ldots$

Normaalijakaumalla $N(\mu,\sigma^2)$ (jossa $\sigma^2>0$) on tf

$$\frac{1}{\sqrt{2\pi}\sigma}\,\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right).$$

Odotusarvo on μ ja varianssi on σ^2.

Binomijakaumalla $\mathrm{Bin}(n,p)$ on ptnf

$$\binom{n}{k}p^k(1-p)^{n-k},\quad k=0,1,\ldots,n.$$

Odotusarvo on np ja varianssi on $np(1-p)$.

Bernoullin jakauma $\mathrm{Bern}(p)$ on sama kuin $\mathrm{Bin}(1,p)$.

Poissonin jakaumalla $\mathrm{Poisson}(\theta)$ parametrilla $\theta>0$ on ptnf

$$\frac{1}{k!}\,\theta^k\mathrm{e}^{-\theta},\quad k=0,1,2,\ldots$$

Odotusarvo on θ ja varianssi on θ.