F1 Introduktion, vetenskaplig modellering, regressionsanalys

TIF385
31/10-23

Tre delar:

1. Maskininlärning:
   - grundläggande begrepp.
   - ingredienser (optimering, validering).
   - neurala nätverk.
   - etiskt perspektiv.
2. Bayesiansk inferens:
   - formellt perspektiv på inlärning.
   - tolkning av sannolikhet.
   - Bayesiansk linjär regression.
3. Stokastiska processer:
   - Markovkedjor.
   - MCMC (Markov chain Monte Carlo) för Bayesiansk inferens.

Vetenskaplig modellering:

$y = y(x)$ $\begin{cases} y & \text{beroende (respons-) variabel/-ler.} \\ x & \text{oberoende (prediktor-) variabel/-ler.} \end{cases}$

Introducera modell M (approximation).

$y \approx M(\theta; x)$, θ modellparameter/-rar. (reella)

y, x, θ kan vara olika typer av tal. Oftast betraktas $\cdot \in \mathbb{R}^p$.

Kategorisering:

- Deterministisk modell: (input + fixa parametrar → output)
  - Explicit: Väldefinierade matematiska funktioner.
    - Linjär.
    - Icke-linjär.
  - Implicit: T.ex. lösning av diffekv. eller egenvärdesproblem.
- Stokastisk modell: innehåller fluktuationer etc.

Vetenskapliga modeller är oftast parametriska.

T.ex. $x(t) = A\sin(\omega t + \phi)$.

- parametrarna har betydelse i sig själva.
- bygger oftast på en fysikalisk insikt/hypotes.

Alt. konstruera en flexibel klass av modeller och en inlärningsalgoritm för att hitta samband.
Kräver oftast en stor mängd observationer (träningsdata).

⇒ Maskininlärning (ML) (icke-parametriska modeller).

Regressionsanalys:

Att hitta samband, oftast med statistiska metoder.
Mest vanligt förekommande i ML är optimering.

Antag modell $M(\Theta; x)$, N_d observationer $\{y_1, \dots, y_{N_d}\}$,

motsv. obercende variabel $\{x_1, \dots, x_{N_d}\}$.

Tillsammans utgör detta en datamängd ①.

Definiera kostnadsfunktion $C(\Theta)$ som "mäter" hur väl modellen beskriver data.

T.ex. $C(\Theta) = \frac{1}{N_d}\sum_{i=1}^{N_d} \frac{[y_i - M(\Theta_i; x_i)]^2}{\sigma_i^2}$, σ_i skalfaktor s.a C dim.löst.

Hitta $\Theta^* = \underset{\Theta \in \mathbb{R}^p}{\operatorname{argmin}}\, C(\Theta)$.

Detta är ett optimeringsproblem.

De vanligaste optimeringsmetoderna bygger på gradientstegsoptimering (gradient descent).

Betrakta $C(\Theta)$ som en flervariabelfunktion.

Grundalgoritm:

1. Starta från slumpmässigt vald punkt Θ_0.
2. Vid Θ_n: • evaluera gradienten $\nabla_\Theta C(\Theta)\big|_{\Theta=\Theta_n} \equiv \nabla C_n$.
   • välj "inlärningshastighet" η_n
3. $\Theta_{n+1} = \Theta_n - \eta_n \nabla C_n$ för fixt antal iterationer eller tills gradienten i Θ_k är liten.

F2 Linjära modeller

TIF385
31/10-23

Linjära modeller:

$y \in \mathbb{R}$, $x \in \mathbb{R}$, $\Theta = \mathbb{R}^p$.

Linjära modeller innebär linjärt parameterberoende.

$M(\Theta; x) = \sum_{j=0}^{p-1} \Theta_j f_j(x)$, $f_j(x)$ beror ej på Θ.

Ex.

- polynommodell: $M(\Theta; x) = \sum_{j=0}^{p-1} \Theta_j x^j = \Theta_0 + \Theta_1 x + \Theta_2 x^2 + \dots + \Theta_{p-1} x^{p-1}$.
- trigonometrisk modell: $M(\Theta; x) = A_0 + \sum_{n=1}^{p} (A_n \cos nx + B_n \sin nx)$,

$$\Theta = \{A_0, \dots, A_p, B_0, \dots, B_p\}.$$

Θ_0 kallas intercept (skärningspunkt) eller bias.

Regressionsanalys med linjära modeller:

Data $\mathbb{D} = [y_1, \dots, y_{N_d}]^T$ (kolumnvektor).

Mot varje y_i ämns en oberoende variabel x_i.

$y_i \approx M(\Theta; x_i) = \sum_{j=0}^{p-1} \Theta_j f_j(x_i)$.

Introducera designmatrisen $\mathbb{X} = \begin{bmatrix} f_0(x_1) & \cdots & f_{p-1}(x_1) \\ \vdots & & \vdots \\ f_0(x_{N_d}) & \cdots & f_{p-1}(x_{N_d}) \end{bmatrix}$ $(N_d \times p)$

med parametervektorn $\Theta = [\Theta_0, \dots, \Theta_{p-1}]^T$

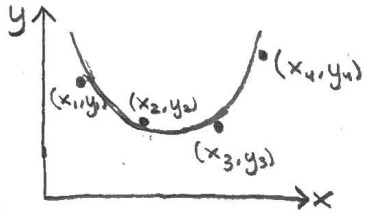
$$\Rightarrow \mathbb{D} = \begin{bmatrix} y_0 \\ \vdots \\ y_{N_d} \end{bmatrix} \approx \mathbb{X}\Theta.$$

Introducera residualer $\varepsilon_i = y_i - M_i$ $(M_i = M(\Theta; x_i))$

$\Rightarrow \mathbb{D} = \mathbb{X}\Theta + \mathbb{E}$, $\mathbb{E} = [\varepsilon_1, \dots, \varepsilon_{N_d}]^T$.

Ex.

$M(\Theta; x) = \Theta_0 + \Theta_1 x + \Theta x^2$, $\mathbb{D} = [y_1, y_2, y_3, y_4]^T$.



$$\mathbb{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{bmatrix} \Rightarrow \mathbb{D} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{bmatrix} \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \Theta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

Hitta Θ som minimerar $|\mathbb{E}|^2$. (vanligaste sättet att minimera $\mathbb{E}$)

Formulera som kostnadsfunktion $C(\Theta) = \underbrace{\sum_{i=1}^{N_d} \varepsilon_i^2}_{=|\varepsilon|^2} = \sum_{i=1}^{N_d} (y_i - M_i)^2 =$

$= \{M_i = (\mathbb{X}\Theta)_i\} = (\mathbb{D} - \mathbb{X}\Theta)^T(\mathbb{D} - \mathbb{X}\Theta)$ $(= \varepsilon^T\varepsilon)$.

Söker $\Theta^* = \underset{\Theta \in \mathbb{R}^p}{\operatorname{argmin}} (\mathbb{D} - \mathbb{X}\Theta)^T(\mathbb{D} - \mathbb{X}\Theta)$ (optimeringsproblem).

$C(\Theta)$ konvex funktion av $\Theta \Rightarrow$ en extrempunkt (minimipunkt).

$$\nabla_\Theta C(\Theta) = \begin{bmatrix} \frac{\partial C}{\partial \theta_0} \\ \vdots \\ \frac{\partial C}{\partial \theta_{p-1}} \end{bmatrix} = \left\{ \frac{\partial C}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \left(\sum_{i=1}^{N_d} (y_i - \theta_0 f_0(x_i) - \ldots - \theta_{p-1} f_{p-1}(x_i))^2 \right) = \right.$$

$$\left. = -2 \sum_{i=1}^{N_d} f_j(x_i)(y_i - \theta_0 f_0(x_i) - \ldots - \theta_{p-1} f_{p-1}(x_i)) \right\} = -2\mathbb{X}^T(\mathbb{D} - \mathbb{X}\Theta).$$

Söker Θ^* s.a. $\nabla_\Theta C(\Theta)\big|_{\Theta=\Theta^*} = 0$.

$\Rightarrow -2\mathbb{X}^T\mathbb{D} + 2\mathbb{X}^T\mathbb{X}\Theta^* = 0 \Rightarrow \mathbb{X}^T\mathbb{D} = \mathbb{X}^T\mathbb{X}\Theta^*$ (normalekvationen).

Hitta $(\mathbb{X}^T\mathbb{X})^{-1}$ s.a. $\underline{\Theta^* = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{D}}$.

$\mathbb{X}^T (p \times N_d)$, $\mathbb{X} (N_d \times p)$, $\mathbb{D} (N_d \times 1) \Rightarrow \Theta^* (p \times 1)$ (dim. ok).

Regressionsresidualerna $\varepsilon^* = \mathbb{D} - \mathbb{X}\Theta^* = 0$ vid "perfekt" anpassad regression (osannolikt).

Linjär regression inom fysik:

- Basfunktionerna $f_j(x)$ är fysikaliskt motiverade.
- y och x är fysikaliska storheter (med enheter).
- Designar och utför experiment för att samla in data $\mathbb{D}$.
- Modellparametrarna är intressanta i sig.
- Kan ha kunskaper om parametrarna från andra fysikaliska samband.
- Kan ofta estimera storlek på fel.

Linjär regression inom ML:

- En form av övervakad (supervised) maskininlärning.
- Ofta kombineras oberoende variabler och basfunktion till prediktor. T.ex. $x, x^2, \frac{1}{\sqrt{x}}, 1 \Rightarrow \mathbb{X} = [1, x, x^2, x^{-1/2}]$.

- Oftast konstateras en lämplig kostnadsfunktion och man prövar sig fram med olika prediktorer.
- Parametrarna är oftast inte intressanta i sig själva. Ibland skrivs modellen $\hat{y}(x)$ (beror ej på parametrar θ).
- Begränsad möjlighet att påverka mängden träningsdata.
- Svårt att estimera "generaliseringsförmåga" (d.v.s. då modellen används för förutsägelser).

F3 Maskininlärning, klassificering

TIF385
01/11-23

Maskininlärning:

Fokus på övervakad (supervised) inlärning.

Tre ingredienser:

1. Data.
2. Matematisk modell.
3. Inlärningsalgoritm.

Kategorisering av ML efter typ av uppgift:

i) Regressionsanalys: - hitta samband (funktionslikande) mellan prediktor- och responsvariabler.

ii) Klassificering: - kategorisera.
- tränas med kategoriserad data.

iii) Klusteralgoritmer: - hitta mönster i data.
- dela in i olika kluster.
- kan ibland utföras utan fördefinierade kategorier.

iv) Dimensionsreduktion: - identifiera de viktigaste prediktorerna.

Fokus på i) och ii).

(Numeriskt exempel på linjär regression i kap. 8.)

Klassificering:

Binär klassificering: (0 eller 1)

Ex. Två prediktorer $\mathbf{x} = [x_1, x_2]$, respons $y \in \{0, 1\}$,

träningsdata $\mathcal{T} = \{\mathbf{x}_i, y_i\}_{i=1}^{N_d}$

Skapa en "hård" klassificerare (hard classifier) genom att göra en linjär modell

$z = w_0 + w_1 x_1 + w_2 x_2$, $\mathbf{w} = [w_0, w_1, w_2]^T$ parametrar, $z \in \mathbb{R}$.

Projicera på $\{0, 1\}$ genom "perceptron"

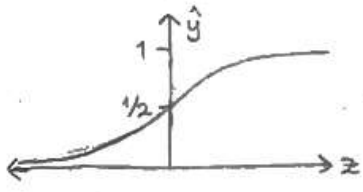
$$\hat{y} = \hat{y}(z) = \frac{\text{sign}(z)+1}{2} = \begin{cases} 1 \text{ om } \text{sign}(z)=1 \ (z>0) \\ 0 \text{ om } \text{sign}(z)=-1 \ (z<0) \end{cases}$$

D.v.s. $\hat{y}_i = \hat{y}(z_i)$, $z_i = w_0 + w_1 x_{1,i} + w_2 x_{2,i}$, eller

$$\hat{\mathbf{y}} = \hat{y}(\mathbf{z}), \quad \mathbf{z} = \mathbb{X}\mathbf{w}, \quad \mathbb{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ \vdots & \vdots & \vdots \\ 1 & x_{1,N_0} & x_{2,N_0} \end{bmatrix}.$$

Skapa en "mjuk" klassificerare genom sigmoidfunktionen

$$\hat{y} = \hat{y}(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}.$$



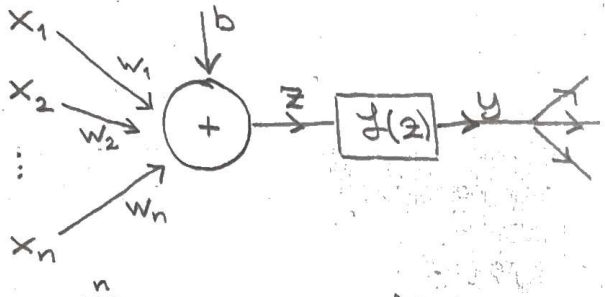
Tolka $\hat{y}$ som sannolikhet:

$$\hat{y} = \mathbb{P}(y(\mathbf{x}) = 1), \quad 1 - \hat{y} = \mathbb{P}(y(\mathbf{x}) = 0).$$

Artificiella neurala nätverk:

Byggstenarna för neurala nätverk kallas neuroner.

Matematisk modell av neuron:



$z = \sum_{j=1}^{n} w_j x_j + b$, $y = f(z)$ aktiveringsfunktion.

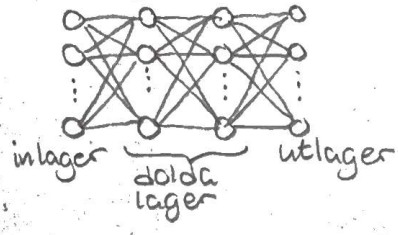
Ex.

$f(z) = \frac{\text{sign}(z)+1}{2}$, $f(z) = \frac{1}{1+e^{-z}}$ (sigmoid).

Ett neuralt nätverk (NN) består av flera ihopkopplade neuroner (noder) som kan delas in i olika typer (beroende på uppgift).

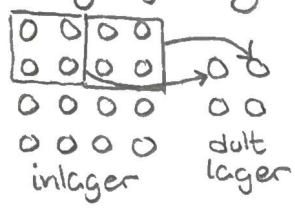
1. Framåtriktade NN (FFNN):

   - Många olika användningsområden inkl. djupa NN (djupa → med dolda lager mellan in- och utsignal).
   - Oftast övervakad inlärning.
   - Oftast fullt anslutna → alla noder i ett lager skickar utsignaler till alla noder i nästa lager.



2. Neurala konvolutionsnätverk (CNN):

- Ej fullt anslutna.
- Utnyttjar geometrisk struktur.



- Används framförallt vid bildigenkänning.

3. Återkopplade NN (RNN):

- Ej enbart framåtriktade.
- Sekventiell information (ordning spelar roll).
- Används framförallt vid text och tal.

4. Autoencoders:

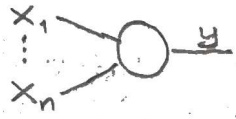
- För övervakad inlärning (eller Boltzmann machines).

FFNN definieras av sin:

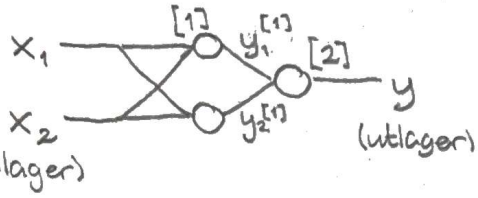
- Arkitektur.
- Aktiveringsfunktion(er).
- Inlärningsalgoritm.

Arkitektur:

- Antal lager (L) (inkl. de med aktiveringsfunktioner → inkl. utlagret, exkl. inlagret.
- Antal noder per lager N_l, $l = 1, \ldots, L$.
- Signalpropageringen beskrivs bäst med matris-vektor-operationer.
- Grafiskt förenklad bild av noder:



Litet nätverk:



$$y_1^{[1]} = f(z_1^{[1]}), \quad z_1^{[1]} = w_{11}^{[1]} x_1 + w_{12}^{[1]} x_2 + b_1^{[1]}.$$

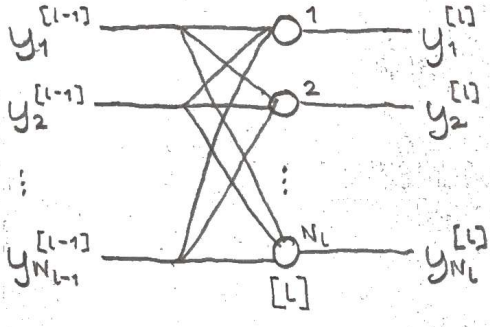
Med matris-vektor-operationer:

$$\left.\begin{aligned} &\mathbf{x} = [x_1, x_2]. \\ &\mathbf{y}^{[1]} = [y_1^{[1]}, y_2^{[1]}]. \\ &W^{[1]} = \begin{bmatrix} w_{11}^{[1]} & w_{21}^{[1]} \\ w_{12}^{[1]} & w_{22}^{[1]} \end{bmatrix}. \\ &b^{[1]} = [b_1^{[1]}, b_2^{[1]}]. \end{aligned}\right\} \Rightarrow \begin{cases} \mathbf{y}^{[1]} = f(\mathbf{z}^{[1]}), \\ \mathbf{z}^{[1]} = \mathbf{x} W^{[1]} + b^{[1]}. \\ \text{Utsignal } y = y^{[2]} = f(z^{[2]}), \\ z^{[2]} = \mathbf{y}^{[1]} W^{[2]} + b^{[2]}, \\ W^{[2]} = \begin{bmatrix} w_1^{[2]} \\ w_2^{[2]} \end{bmatrix}. \end{cases}$$

Beteckning: $w_{12}^{[1]}$ ← lager; 1 ↑ nod; 2 ↑ insignal

Djupt nätverk:

Betraktat lager l:



$$\mathbf{y}^{[l]} = f(\mathbf{z}^{[l]}),$$

$$\underset{[1\times N_l]}{\mathbf{z}^{[l]}} = \underset{[1\times N_{l-1}]}{\mathbf{y}^{[l-1]}} \cdot \underset{[N_{l-1}\times N_l]}{W^{[l]}} + \underset{[1\times N_l]}{b^{[l]}}.$$

$\Rightarrow$ # parametrar i lager l är $N_l(N_{l-1}+1)$.

Ex. Hur många parametrar har ett FFNN som ska klassificera bilder (8×8 pixlar) av handskrivna siffror 0-9? Det har 10 dolda lager med 20 noder var. Utlagret ger 10 utsignaler y_j, $j=0,\dots,9$, där $y_j = \mathbb{P}(t=j) \in [0,1]$ (mjuk klassificerare).

$$\left.\begin{aligned} \text{Dolt lager 1: } & 20(64+1) = 1300 \\ \text{2-10: } & 20(20+1) = 420 \\ \text{Utlager: } & 10(20+1) = 210 \end{aligned}\right\} \overset{\Sigma}{\Rightarrow} \underline{5290 \text{ parametrar.}}$$

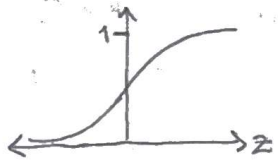
F5 Aktiveringsfunktioner, optimering

TIF385
07/11-23

Aktiveringsfunktioner:

a) Sigmoid: $f(z)=\frac{1}{1+e^{-z}} \in [0,1]$, $f(z)=\tanh(z) \in [-1,1]$.

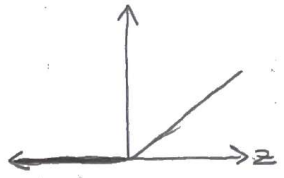
- Nollskild $\forall z$.
- Små gradienter $\Rightarrow$ Långsam inlärning.



b) Modifierad linjär enhet (ReLU):

$f(z)=\max(z,0)$

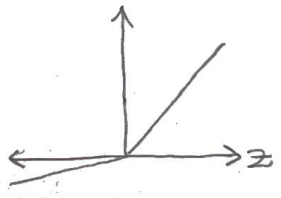
- Oviktiga noder är "tysta".
- $f'(z)$ är enkel, $f'(z)=0$ då $z<0$.



c) Läckande ReLU:

$f(z)=\max(0,z)+\alpha\min(0,z)$, $\alpha \ll 1$.

- Inga "tysta" noder.

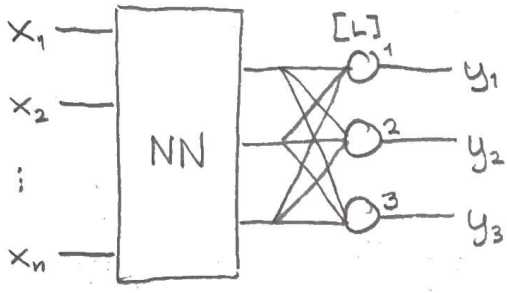


d) ELU: (exp.)

Notera att aktiveringsfunktionen i utlagret väljs efter syfte.

Ex. NN ska förutsäga position i $\mathbb{R}^3$.

$y_j=(y_1,y_2,y_3)$, $y_i \in \mathbb{R}$, ex.v. $y_j=y_j^{[L]}=z_j^{[L]}$



Ex. NN ska förutsäga fyra diskreta sannolikheter.

$y_j=\mathbb{P}(t=j)$, $\sum_{j=1}^{4} y_j=1$.

Låt $\tilde{y}_j=y_j^{[L]}=\sigma(z_j^{[L]})$ (sigmoid) och transformera enligt

$y_j=\frac{\tilde{y}_j}{\sum_{j=1}^{4}\tilde{y}_j}$.

Inlärningsalgoritm:

a) Val av kostnadsfunktion:

Jämför modellens utdata med träningsdata (ex.v. medelkvadratfel (MSE)).

$C(y_j, \hat{y}_j)$, y_j data, $\hat{y}_j$ modell utdata.

b) Optimeringsalgoritm:

Optimera vikterna för att minimera kostnadsfunktion genom någon form av gradientstegsoptimering.

$\frac{\partial C}{\partial w_{ij}^{[l]}}$ $\forall l, i, j$ (bakåtpropagering).

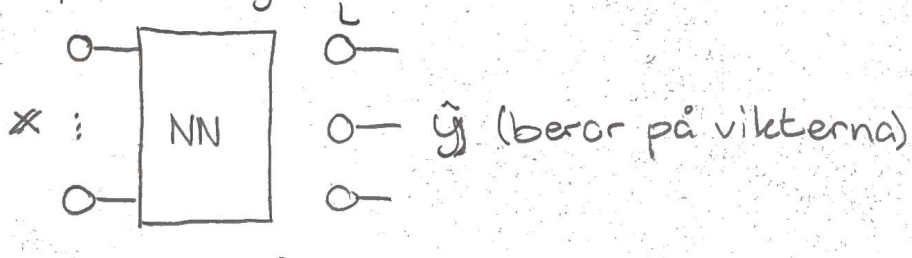
c) Uppdelning av data:

i) träningsdata.
ii) valideringsdata. (se modellvalidering)
iii) testdata.

d) Inlärningsprocedur:

Hur många iterationer med gradientstegsoptimering? Variera arkitektur, aktiveringsfunktion?

Optimering:



Kostnadsfunktion $C(y_j, \hat{y}_j)$.

Justera vikterna iterativt mha grad.stegsopt:

$w_{ij}^{[l]}{}_{(\text{iteration } n+1)} \leftarrow w_{ij}^{[l]}{}_{(\text{iteration } n)} - \eta \delta_{ij}^{[l]}$, $\delta_{ij}^{[l]} = \frac{\partial C}{\partial w_{ij}^{[l]}}$.

Pss för bias:

$b_i^{[l]} \leftarrow b_i^{[l]} - \eta \delta_i^{[l]}$

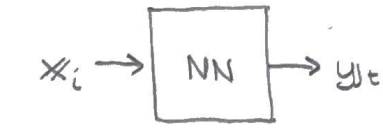
Notera att C beror på $\hat{y}_j$.

$\hat{y}_j = y_j^{[L]} = \mathcal{J}(z^{[L]}) = \mathcal{J}(y_j^{[L-1]} w^{[L]} + b^{[L]}) = \mathcal{J}(\mathcal{J}(z^{[L-1]} w^{[L]} + b^{[L]})) = \ldots,$

dvs alla gradienter kan beräknas med bakåtpropagering.

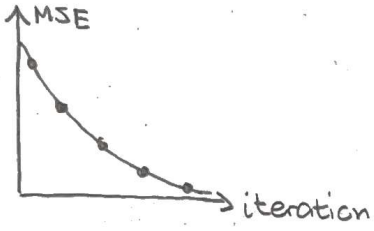
Inlärning:

Träningsdata $x_t = \begin{bmatrix} x_1 \\ \vdots \\ x_{N_t} \end{bmatrix}$, $y_{jt} = \begin{bmatrix} y_{j1} \\ \vdots \\ y_{jN_t} \end{bmatrix}$.



Vill optimera nätverket s.a. MSE(y_j, $\hat{y}_j$) blir så liten som möjligt.

Vikterna justeras iterativt.



(iterationer mäts ofta i epoker (1 epok = all träningsdata har använts en gång))

F6 Modellvalidering

TIF385
13/11-23

Modellvalidering:

ML-modeller har ofta stor flexibilitet ⇒ risk för överanpassning mot träningsdata och opålitliga prediktioner.

Använd linjär regression med polynommodell som prototypex.

$\hat{y}(x) = \sum_{i=0}^{p} \theta_i x^i$ (ökande p ⇒ ökande komplexitet).

(se ex. på Canvas: Files/demo/demo-ModelValidation-skeleton.ipynb)

Överanpassad modell:

Har anpassats till brus lika mycket som verklig respons.

Underanpassad modell:

Ignorerar variabilitet i data som faktiskt motsvarar verklig respons.

Detta kan vara svårt att upptäcka visuellt. Behöver andra metoder.

Kännetecken:

1. Stora värden på modellens parametrar vid överanpassning.
2. Valideringsdata kan användas för att detektera både över- och underanpassning.

Stora parametervärden kan undvikas genom regularisering.

Istället för att optimera $\Theta^* = \underset{\Theta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} C(\Theta)$, optimeras en modifierad kostnadsfunk., ex.v.

$$\Theta^*_{\lambda,2} = \underset{\Theta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} C_{\lambda,2}(\Theta), \quad C_{\lambda,2}(\Theta) = C(\Theta) + \lambda \|\Theta\|_2^2.$$

(2-norm av parametervektorn)

Detta kallas Ridge-regularisering.

Lasso-regularisering: $C_{\lambda,1}(\Theta) = C(\Theta) + \lambda \|\Theta\|_1^2$.

Notera: $\|\Theta\|_2^2 = \sum_{i=1}^{p+1} \theta_i^2$, $\|\Theta\|_1^2 = \left(\sum_{i=1}^{p+1} |\theta_i|\right)^2$.

λ är en hyperparameter.

Validering:

Bygger på att dela upp tillgänglig data.

Data → Valideringsdata (D_{val})
Data → Träningsdata (D_{train})

Optimera modellen mot D_{train} ⇒ minimera $C(\underbrace{y_{j|train}}_{\text{respons i } D_{train}}, \underbrace{\hat{y}_j(x_{train})}_{\text{modell-prediktioner}})$ ⇒

⇒ θ^* tränad modell ⇒ Beräkna $\hat{y}_j(x_{val})$, jämför med $y_{j|val}$,

där $D_{val} = (x_{val}, y_{j|val})$.

Jämförelsen sker mha en förlustfunktion. Ex. för regression

$$E(y_j, \hat{y}_j) = \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} (y_i - \hat{y}(x_i))^2$$

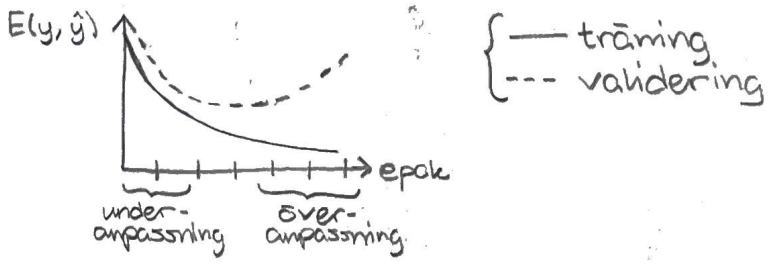
Kan vara (men måste inte) identisk med kostnadsfunk.
Ex. för klassificering

$$E(y_j, \hat{y}_j) = \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} I(y_i \neq \hat{y}_i), \quad I(y_i \neq \hat{y}_i) = \begin{cases} 0 \text{ om } y_i = \hat{y}_i \\ 1 \text{ om } y_i \neq \hat{y}_i \end{cases}.$$

Två exempel på användning av valideringsdata:

a) Inlärningskurva:

Vid iterativ träning av ett NN ex.v.



b) Optimering av hyperparametrar:

Betrakta ex.v. ordningen p hos en polynommodell som hyperparameter.

Testa flera val $\{p_i\}_i$. För varje val optimeras modellen mot träningsdata och jämförs mot valideringsdata. Valet av p som minimerar felet mot valideringsdata är då det "bästa" valet.

F7 Rättvis AI

TIF385
14/11-23

Rättviseaspekten vid träning och validering av ML-modeller:

- Kostnadsfunktion används vid träning (optimering) av en ML-modell mot given träningsdata.
- Felfunktion används för att validera modellen mot valideringsdata.

Huvudsakligen betraktas tekniska aspekter, men val av fel-/kostnadsfunktion tillsammans med vilken data som är tillgänglig kan påverka hur rättvis modellen är.

Betraktа binär klassificering, vilken kan utvärderas med felklassificeringsfrekvensen.

$$E_{val} = \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} (1 - \delta_{y_i, \hat{y}(x_i)}),$$

(y_i: måldata; $\hat{y}(x_i)$: ML-modellens förutsägelse)

Notera $y_i \in \{0, 1\}$, $\hat{y}(x_i) \in \{0, 1\}$.

Sanningsmatrisen:

	$y=0$	$y=1$	Σ
$\hat{y}=0$	TN	FN	N*
$\hat{y}=1$	FP	TP	P*
Σ	N	P	N_{val}

T - true
F - false
P - positive
N - negative

FP kallas ibland för typ-1 fel.
FN -"- typ-2 fel.

Beroende på sammanhang kan endera av dessa två feltyper ge allvarliga konsekvenser.

Modellens träffsäkerhet:

$$\frac{TN+TP}{N_{val}} = 1 - \text{fel.klass.frekv.} = 1 - \frac{FN+FP}{N_{val}}.$$

Andel falskt positiva: $\frac{FP}{N}$.

-"- negativa: $\frac{FN}{P}$.

Modellens precision: $\frac{TP}{(TP+FP)} = \frac{TP}{P^*}$.

Notera att det finns en risk att något/flera av dessa mått blir olika för olika grupper inom populationen.

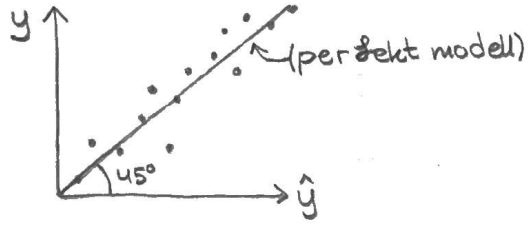
Detta kan ske pga:

- snedvriden data.
- val av kostnadsfunktion.

Svårt att upptäcka. Medvetenhet är viktigt. Presentera och utvärdera modellen utifrån olika rättviseaspekter.

Ex. Compasalgoritmen.

Ett sätt att undersöka hel för en regressionsmodell (motsv, sannings matris) är paritetsplot.



Begränsningar i träningsdata:

Målet är att hitta $y \approx \underbrace{\hat{y}(x)}_{\text{(ML-modell som tränas på träningsdata)}}$.

Bristfällig data ⇒ bristfällig modell.

Fokuserar på snedvriden data (biased data). Ex.v.

- Systematisk snedvridning:
  - En viss typ av data favoriseras (är vanligare) medan annan ignoreras.
- Automatisk snedvridning:
  - Data samlas in genom automatiserad process, ex.v. cookies.
- Urvalsnedvriden data:
  - Urvalet representerar inte populationen.
  - Generellt vanligaste typen.
- Språklig snedvridning:

Det finns etiska principer och riktlinjer för pålitlig AI. Det finns 7 krav:

1. Mänskligt agentskap och tillsyn.
2. Teknisk robusthet och säkerhet.
⋮
4. Transparens.
⋮
7. Ansvarsskyldighet.

F8 Sannolikhet repetition & introduktion

TIF385
21/11-23

Repetition:

Sannolikhetsmått: $\mathbb{P}: A \to [0,1]$. (A: händelse i utfallsrum S)

Betingad sannolikhet: $\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$. (*) ($\cap$: snitt; $B|A$: B givet A)

A, B oberoende händelser $\Rightarrow \mathbb{P}(B|A) = \mathbb{P}(B) \overset{(*)}{\Rightarrow} \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Marginal sannolikhet:

$B_1, \dots, B_n$ unika och spänner upp utfallsrummet, $\mathbb{P}(B_i) > 0 \ \forall i \in [1,n]$

$\Rightarrow \mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i) \overset{(*)}{=} \sum_{i=1}^{n} \mathbb{P}(A \cap B_i)$.

Slumpvariabel: $X: S \to \mathbb{R}$.

Fokuserar på utfallen $x \in \mathbb{R}$.

Fördelningsfunktion: $P: \mathbb{R} \to [0,1]$ som uppfyller $\mathbb{P}(X \le x) = P(x)$.

(Notera att $\mathbb{P}(X = x) = 0$ då X kontinuerlig.)

Ibland skrivs $P_X(x)$.

Täthetsfunktion: $p: \mathbb{R} \to [0, \infty)$ s.a. $P(x) = \int_{-\infty}^{x} p(\tilde{x})\,d\tilde{x}$.

Notera att

- $p(x) \ge 0$.
- Enheter $[P] = 1$, $[p] = [x^{-1}]$.
- $\int_{-\infty}^{\infty} p(\tilde{x})\,d\tilde{x} = 1$.
- $\mathbb{P}(a \le X \le b) = \mathbb{P}(b) - \mathbb{P}(a) = \int_{a}^{b} p(\tilde{x})\,d\tilde{x}$.

($\mathbb{P}(x \le X \le x + dx) = p(x)\,dx$ för infinitesimalt intervall.)

Marginal täthetsfunktion:

Betrakta den bivariata fördelningsfunktionen

$P(x,y) \equiv \mathbb{P}(X \le x, Y \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} p(\tilde{x},\tilde{y})\,d\tilde{y}\,d\tilde{x}$.

$p(\tilde{x}) = \int_{-\infty}^{\infty} p(\tilde{x},\tilde{y})\,d\tilde{y}$.

Vad är sannolikhet?

Kan man mäta $\mathbb{P}(A)$? Ja, om man kan göra ett stort antal observationer N och tolka den relativa frekvensen

$$\frac{N_A}{N} \approx \mathbb{P}(A), \quad N_A = \#\text{gånger som } A \text{ observerades.}$$

Detta kan endast tillämpas på slumpvariabler.

Bayesiansk sannolikhet:

Tolkar sannolikhet som en grad av tro. Kan tillämpas på alla intressanta vetenskapliga frågeställningar.

Produktregeln: $\mathbb{P}(A \cap B \mid I) \overset{(*)}{=} \mathbb{P}(A \mid B, I)\mathbb{P}(B \mid I) = \mathbb{P}(B \mid A, I)\mathbb{P}(A \mid I)$.

(I: tillgänglig information)

$$\Rightarrow \mathbb{P}(A \mid B, I) = \frac{\mathbb{P}(B \mid A, I)\mathbb{P}(A \mid I)}{\mathbb{P}(B \mid I)}. \quad \text{(Bayes sats)}$$

- $\mathbb{P}(A \mid I)$ - A priorisannolikhet.
- $\mathbb{P}(A \mid B, I)$ - A posteriorisannolikhet.

Ex.

A = Higgsbosonens massa är mellan $[m_H, m_H + dm_H]$.

$$\Rightarrow \mathbb{P}(A \mid I) = p(m_H \mid I)\, dm_H.$$

B = Observerar datapunkt i $[D, D + dD]$.

$$\Rightarrow \mathbb{P}(B \mid I) = p(D \mid I)\, dD.$$

$$\mathbb{P}(A \mid B, I) = p(m_H \mid D \cap I)\, dm_H.$$

$$\mathbb{P}(B \mid A, I) = p(D \mid m_H \cap I)\, dD.$$

$$\text{Bayes sats} \Rightarrow p(m_H \mid D, I)\, dm_H = \frac{p(D \mid m_H, I)\, dD\, p(m_H \mid I)\, dm_H}{p(D \mid I)\, dD} =$$

$$= \frac{p(D \mid m_H \cap I)\, p(m_H \mid I)}{p(D \mid I)},$$

där $p(D \mid m_H, I)$ är trolighetsfunktion (likelihood),

$p(D \mid I)$ är marginell trolighet.

Vi är egentligen intresserade av posteriori-fördelningen.

$$p(p_H \mid D, I) = \frac{p(D \mid p_H, I)\, p(p_H, I)}{p(D, I)} = \frac{1}{N} p(D \mid p_H, I) p(p_H \mid I).$$

($p(D, I)$: beror ej på p_H; $\frac{1}{N}$: normalisering)

Likformig apriori-fördelning:

$$p(p_H \mid I_L) = \begin{cases} 1 & \text{för } p_H \in [0,1]. \\ 0 & \text{annars.} \end{cases}$$

Godtrogen apriorifördelning:

$p(p_H \mid I_G) = N(0{,}5;\ 0{,}01^2)$ för $p_H \in [0,1]$, 0 annars.

(N: normalfördelning, (μ), (σ))

$$\propto \exp\left(-\frac{(p_H - 0{,}5)^2}{2 \cdot 0{,}01^2}\right) \text{ för } p_H \in [0,1].$$

Likformig inferens: $p(p_H \mid D, I) = \frac{1}{N} p_H^{N_H} (1 - p_H)^{N - N_H} \cdot 1$ för $p_H \in [0,1]$.

F9 Inferens, data och modeller, Bayesiansk linjär regression

TIF385
22/11-23

Inferens:

Slutsats som dras utifrån kontext.

Deduktiv inferens:

$A, B, \ldots$ = premisser som betraktas som sanna.

H = slutsats.

$\mathbb{P}(H|A,B,\ldots) = \{0, 1\}$ ($0 = H$ falskt, $1 = H$ sant).

Induktiv inferens:

Premisserna påverkar slutsatsen, men tillåter inte att avgöra sannolikhetshalten definitivt. (vanligaste inferensen)

Statistisk inferens används för att kvantifiera styrkan hos induktiv inferens via sannolikheter.

Data och modeller:

Betrakta en observation D av ett fysikaliskt fenomen. En modell $M(\Theta)$ beskriver detta fenomen.

$$D = M(\Theta) + \delta D + \delta M. \quad (*)$$

δD: exp. osäkerhet (stumvariabel); δM: modellosäkerhet (motverkar överanpassning)

Datan lär oss om modellens parametrar. Formaliseras med Bayes teorem enligt

$$p(\Theta|D,I) = \frac{p(D|\Theta,I)\,p(\Theta|I)}{p(D|I)}.$$

$$(*) \Rightarrow p(D|\Theta,I) \equiv \mathcal{L}(\Theta).$$

Att inferera posteriorifördelningen $p(\Theta|D,I)$ kallas för <u>parameterestimering</u>.

Förutsägelser:

Antag att modellen $M(\Theta)$ förutsäger ett annat fysikaliskt fenomen, för vilket det ej ännu finns data. Kallar detta framtida data F.

$D \Rightarrow$ lärdom om $M \Rightarrow$ förutsäg F, d.v.s.

$$p(F|D,I) = \int_\Omega p(F,\Theta|D,I)\,d\Theta = \int_\Omega \underbrace{p(F|\Theta,D,I)}_{\text{modellens förutsägelse}}\,\underbrace{p(\Theta|D,I)}_{\text{a posteriori-fördelning}}\,d\Theta.$$

I: modell ingår här; integralen: marginalisering

Detta kallas för en a-posteriori-förutsägelse.

Dessa integraler utförs i praktiken av Monte Carlo-tekniker

Bayesiansk linjär regression:

Betrakta modellen $M(\Theta; x) = \sum_{j=0}^{p-1} \Theta_j f_j(x)$.

Anpassas till data $D = \begin{pmatrix} y_1 \\ \vdots \\ y_{N_d} \end{pmatrix}$ med designmatris $X = \begin{pmatrix} f_0(x_1) & \cdots & f_{p-1}(x_1) \\ \vdots & & \vdots \\ f_0(x_{N_d}) & \cdots & f_{p-1}(x_{N_d}) \end{pmatrix}$

$D = X\Theta + \varepsilon$,

där $\Theta = \begin{pmatrix} \Theta_0 \\ \vdots \\ \Theta_{p-1} \end{pmatrix}$ (parametrar), $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{N_d} \end{pmatrix}$ ← residualer.

Minimering av $\|\varepsilon\|_2^2$ leder till normalekvationen med lösningen

$\Theta^* = (X^T X)^{-1} X^T D$.

Gör istället Bayesiansk inferens.

$$p(\Theta \mid D, I) = \frac{p(D \mid \Theta, I)\, p(\Theta \mid I)}{p(D \mid I)}.$$

Trolighetsfunktion:

Under antagandet att residualerna är oberoende, är

$$p(\varepsilon \mid I) = \prod_{i=1}^{N_d} p(\varepsilon_i \mid I).$$

Under antagandet att $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}[\varepsilon_i] = \sigma_i^2$, samt att trolighetsfunktionen beskrivs av en normalfördelning, är

$$p(\varepsilon_i \mid I) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\frac{\varepsilon_i^2}{\sigma_i^2}\right).$$

$D_i = \underbrace{(X\Theta)_i}_{\equiv M_i} + \varepsilon_i$. (**)

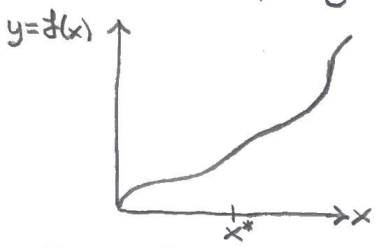
Osäkerhet i ε_i leder till osäkerhet i D_i, men hur?

Felpropagering, variabelbyte:

Betrakta slumpvariabel X, $x \sim p_X(x \mid I)$ ("fördelad enligt").

Betrakta $Y = f(X)$ (unik avbildning).

Vad blir $p_I(y|I)$?



Betrakta godtycklig punkt x^* och litet intervall δx.

$\Rightarrow \mathbb{P}(x^* - \frac{\delta x}{2} \leq x \leq x^* + \frac{\delta x}{2}) = p_x(x^*|I)\,\delta x.$

Avbildningen ger $y^* = f(x^*)$.

Konservering av sannolikhet innebär att det måste finnas ett intervall δy s.a.

$$p_x(x^*|I)\,\delta x = p_I(y^*|I)\,\delta y.$$

Detta ska gälla för alla punkter. Låt $\delta x, \delta y \to 0$.

$\Rightarrow$ $p_x(x|I) = p_I(y = y(x)|I)\left|\frac{dy}{dx}\right|$ eller $p_I(y|I) = p_x(x = x(y)|I)\left|\frac{dx}{dy}\right|$.

Jacobian

Betrakta (**). Ovanstående ger

$$p(D_i|\Theta, I) = p(\varepsilon_i(D_i)|\Theta, I)\cdot 1 = \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{1}{2}\frac{(D_i - M_i)^2}{\sigma_i^2}\right).$$

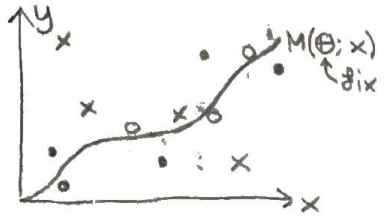
F10 Bayesiansk linjär regression TIF385 28/11-23

Bayesiansk linjär regression: (forts.)

$M(\Theta; x) = \sum_{j=0}^{p-1} \Theta_j g_j(x)$ anpassas till data $D_{obs} = \begin{pmatrix} y_1 \\ \vdots \\ y_{N_d} \end{pmatrix}$.

Bayes teorem ger $p(\Theta | D_{obs}, I) \propto p(D_{obs} | \Theta, I)\, p(\Theta | I)$.

Trolighetsfunktion - beskriver sannolikheten att observera en uppsättning data givet en specifik datagenererande process (i detta fall $M(\Theta; x)$).



$\begin{cases} \bullet\ D_1 \\ \circ\ D_2 \\ \times\ D_3 \end{cases}$ (observerade datapunkter).

$p(D_2 | \Theta, I) > p(D_1 | \Theta, I) \gg p(D_3 | \Theta, I)$.

Med den observerade datan vill man hitta den modell som maximerar trolighetsfunktionen.

$p(D_{obs} | \Theta, I) \equiv \mathcal{L}(\Theta)$.

Betrakta felmodell för residualerna ε som oberoende slumpvariabler.

$p(\varepsilon_i | I) = N(0, \sigma_i^2)$ (normalförd.)

$\Rightarrow p(D_i | \Theta, I) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(D_i - M_i)^2}{2\sigma_i^2}\right)$, $M_i = M(\Theta; x_i)$.

Oberoende fel ger

$$p(D | \Theta, I) = \prod_{i=1}^{N_d} p(D_i | \Theta, I) = \frac{1}{(2\pi)^{N_d/2}} \frac{1}{|\Sigma_\varepsilon|^{1/2}} \exp\left(-\frac{1}{2}(D - \mathbb{X}\Theta)^T \Sigma_\varepsilon^{-1} (D - \mathbb{X}\Theta)\right),$$

$\Sigma_\varepsilon = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{N_d}^2 \end{pmatrix}$ kovariansmatris, $\mathbb{X}$ designmatris.

I specialfallet då $\sigma_i^2 = \sigma_\varepsilon^2\ \forall i$, gäller

$$p(D | \Theta, I) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{N_d/2}} \exp\left(-\frac{1}{2} \frac{(D - \mathbb{X}\Theta)^T (D - \mathbb{X}\Theta)}{\sigma_\varepsilon^2}\right).$$

För en linjär modell kan man skriva om den normalfördelade troligheten till något som är proportionellt mot

$$\exp\left(-\frac{1}{2}(\Theta - \Theta^*)^T \Sigma_\Theta^{-1} (\Theta - \Theta^*)\right), \quad \Theta^* = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T D, \quad \Sigma_\Theta^{-1} = \frac{\mathbb{X}^T\mathbb{X}}{\sigma_\varepsilon^2}.$$

- Det ser ut som en normalfördelning för Θ, men är ej normaliserad.

- Moden (maximum) är lika med Θ^*, dvs lösningen till normalekvationen. ⇒ linjär regression motsvarar "Maximum likelihood estimate" (MLE).

A priori-fördelning:

Betrakta två möjligheter:

a) $p(\Theta \mid I_a) = \begin{cases} \left(\frac{1}{\Delta\Theta}\right)^{N_p} & \text{om } \Theta_i \in [-\frac{\Delta\Theta}{2}, \frac{\Delta\Theta}{2}] \ \forall i. \\ 0 & \text{annars.} \end{cases}$

dvs likformig fördelning med bredd $\Delta\Theta$.

b) $p(\Theta \mid I_b) = \frac{1}{(2\pi\sigma_\Theta^2)^{N_p/2}} \exp\left(-\frac{\Theta^T\Theta}{\sigma_\Theta^2}\right)$.

dvs en i.i.d normalfördelning för p parametrar.

A posteriori-fördelning:

Betrakta analys av a):

$$p(\Theta \mid D_{obs}, I_a) \propto \begin{cases} \exp\left(-\frac{1}{2}(\Theta-\Theta^*)^T\Sigma_\Theta^{-1}(\Theta-\Theta^*)\right) & \text{om } \Theta_i \in [-\frac{\Delta\Theta}{2}, \frac{\Delta\Theta}{2}] \ \forall i. \\ 0 & \text{annars.} \end{cases}$$

- Givet $\Theta_i^* \in [-\frac{\Delta\Theta}{2}, \frac{\Delta\Theta}{2}]$, har denna a-posteriori-fördelning samma mod som trolighetsfunktionen.
- Kan tolkas som en pdf för Θ.

Betrakta analys av b):

$$p(\Theta \mid D_{obs}, I_b) = \left\{\begin{matrix}\text{produkt av två} \\ \text{normalförd.}\end{matrix}\right\} \propto \exp\left(-\frac{1}{2}(\Theta-\tilde{\Theta})^T\tilde{\Sigma}_\Theta^{-1}(\Theta-\tilde{\Theta})\right),$$

$$\tilde{\Theta} = \tilde{\Sigma}_\Theta\Sigma_\Theta^{-1}\Theta^*, \quad \tilde{\Sigma}^{-1} = \Sigma_\Theta^{-1} + \frac{1}{\sigma_\Theta^2}\mathbb{I}_{p\times p}.$$

- Inferensen ger en normalfördelning
- Jämfört med a priori:

$\mathbb{E}[\Theta]$: $0 \xrightarrow{\text{data}} \tilde{\Theta}$.

$\text{Var}[\Theta]$: $\sigma_\Theta^2\mathbb{I}_{p\times p} \xrightarrow{\text{data}} \tilde{\Sigma}_\Theta$.

Posteriorifördelningar och kostnadsfunktioner:

Betrakta modell $M(\Theta; x)$ och data $D = \begin{pmatrix} y_1 \\ \vdots \\ y_{N_d} \end{pmatrix}$.

$D_i = y_i$, $M_i = M(\Theta; x_i)$.

Förutsätt (oändligt stor) likformig apriorifördelning och normalfördelad trolighetsfunktion.

A posteriori:

$$p(\Theta | D, I) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N_d}\frac{(D_i - M_i)^2}{\sigma_\varepsilon^2}\right).$$

Moden kallas MAP (maximum a posteriori).

$$\Theta^* = \underset{\Theta \in \mathbb{R}^p}{\operatorname{argmax}}\, p(\Theta | D, I) = \underset{\Theta \in \mathbb{R}^p}{\operatorname{argmax}}\, \log p = \underset{\Theta \in \mathbb{R}^p}{\operatorname{argmin}} \overbrace{\sum_{i=1}^{N_d}(D_i - M_i)^2}^{\text{(vanlig kostnadsfunktion)}}.$$

Antag att apriorifördelningen är:

$$p(\Theta | I) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{p}\frac{\Theta_i^2}{\sigma_\Theta^2}\right).$$

Vilket ger a posteriorifördelningen:

$$p(\Theta | D, I) \propto \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{N_d}\frac{(D_i - M_i)^2}{\sigma_\varepsilon^2} + \sum_{i=1}^{p}\frac{\Theta_i^2}{\sigma_\Theta^2}\right)\right).$$

MAP-punkten för denna är:

$$\Theta^* = \underset{\Theta \in \mathbb{R}^p}{\operatorname{argmin}}\left(\sum_{i=1}^{N_d}(D_i - M_i)^2 + \underbrace{\frac{\sigma_\varepsilon^2}{\sigma_\Theta^2}}\sum_{i=1}^{p}\Theta_i^2\right).$$

(motsvarar Ridgeregularisering)

Modellering med osäkerhet:

Felpropagering:

Betrakta parametrisk modell. Dataanalys för att estimera parametrarna har genomförts.

Propagera osäkerheter till prediktioner:

Tre metoder:

1. Marginalisering.
2. Variabeltransformation.
3. Approximativ felpropagering.

Marginalisering:

Ex. 1 Modell med två parametrar Θ, Φ, varav bara Θ är av fysikaliskt intresse.

Inferens ger $p(\Theta, \Phi \mid D, I)$. ($\Phi$: nuissance parameter)

$$p(\Theta \mid D, I) = \int p(\Theta, \Phi \mid D, I)\, d\Phi \quad \text{(marginalisering)}.$$

Ex. 2 $z = f(x,y)$. Givet $p_{x,y}(x, y \mid D, I)$, vad är $p_z(z \mid D, I)$?

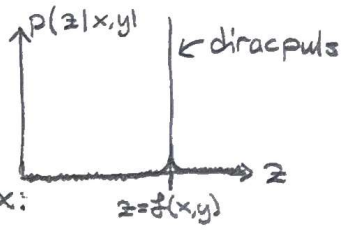
Använd marginalisering och produktregeln.

$$p(z \mid D, I) = \int p(x, y, z \mid D, I)\, dy dx = \int p(z \mid x, y, D, I) \cdot p(x, y \mid D, I)\, dy dx.$$

Eftersom x, y ger z, är

$$p(z \mid x, y, D, I) = \delta(z - f(x,y))$$

$$\Rightarrow p(z \mid D, I) = \int \delta(z - f(x,y))\, p(x, y \mid D, I)\, dy dx.$$



Kan evaluera integral mha deltafunktioner.

(se t.ex. 17.1 ($z = X + Y$))

Variabeltransformation:

Betrakta $\mathbf{x} = \{x_j\}_{j=1}^{M}$, $\mathbf{y} = \{y_j\}_{j=1}^{M}$, $x_j = x_j(\mathbf{y})$, $y_j = y_j(\mathbf{x})$.

Om $p_Y(\mathbf{y} \mid I)$ är känd, vad är $p_x(\mathbf{x} \mid I)$?

Konservering av sannolikhet ger

$$p_x(\mathbf{x} \mid I) = p_Y(\mathbf{y}(\mathbf{x}) \mid I) \left| \frac{\partial(y_1, \ldots, y_M)}{\partial(x_1, \ldots, x_M)} \right|. \quad \leftarrow \text{(Jacobian)}$$

T.ex. för $y = y(x)$:

$$p_x(x \mid I) = p_Y(y(x) \mid I) \left| \frac{dy}{dx} \right|.$$

Approximativ felpropagering:

Betrakta $Z = f(X, Y, \ldots)$,

$\mathbb{E}[X] = x_0$, $\mathbb{E}[Y] = y_0$, ...

$\mathrm{Var}[X] = \sigma_x^2$, $\mathrm{Var}[Y] = \sigma_y^2$, ...

Vad är $\mathbb{E}[z]$, $\mathrm{Var}[z]$?

- Oftа gäller $\mathbb{E}[z] = z_0 = f(x_0, y_0, \dots)$ åtminstone approximativt.

För $\mathrm{Var}[z]$:

- Differentiera sambandet runt $(x_0, y_0, \dots)$, dvs Taylorutveckla runt $(x_0, y_0, \dots)$ och behåll upp t.om. linjära termer i $(x-x_0)$, $(y-y_0)$, ...

$$\begin{cases} \delta X = X - x_0 \\ \delta Y = Y - y_0 \\ \delta z = z - z_0 \dots \end{cases}$$

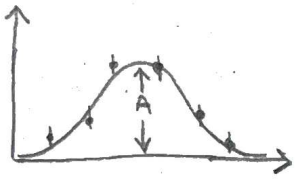
- Kvadrera detta uttryck och beräkna väntevärdet. Notera att

$$\begin{aligned} \mathbb{E}[\delta z^2] &= \mathbb{E}[(z - \mathbb{E}[z])^2] = \sigma_z^2 \\ \mathbb{E}[\delta X^2] &= \dots = \sigma_x^2 \\ \mathbb{E}[\delta Y^2] &= \dots = \sigma_y^2 \dots \end{aligned}$$

Var försiktig! Antagligen fås termer $\mathbb{E}[\delta X \delta Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathrm{Cov}[X, Y]$, alltså påverkar det om $X, Y, \dots$ är beroende av varandra.

Ex. 17.4

Mät ett spektrum och estimera amplitud A.



$\mathbb{E}[A] = A_0$.
$\mathrm{Var}[A] = \sigma_A^2$.

Givet att $A = |f|^2$, f komplex strukturfaktor.

Uppskatta $|f| = f_0 + \sigma_f^2$.

$A_0 = f_0^2 \Rightarrow f_0 = \sqrt{A_0}$.

Differentiera uttrycket.

$$\delta A = 2 f \delta f$$

Taylorutveckla runt $f = f_0$.

$$A = f_0^2 + \left.\frac{dA}{df}\right|_{f=f_0} (f - f_0) + O((f-f_0)^2) = f_0^2 + 2f_0(f-f_0) + O((f-f_0)^2).$$

Definiera $\delta A = A - A_0$, $\delta f = f - f_0$. $\Rightarrow \delta A = 2 f_0 \delta f$.

$\Rightarrow \delta A^2 = 4\mathcal{J}_0^2 \delta\mathcal{J}^2$

$\mathbb{E}[\delta A^2] = \mathbb{E}[4\mathcal{J}_0^2 \delta\mathcal{J}^2] \Rightarrow \sigma_A^2 = 4\mathcal{J}_0^2 \mathbb{E}[\delta\mathcal{J}^2] = 4\mathcal{J}_0^2 \sigma_{\mathcal{J}}^2$.

$\sigma_{\mathcal{J}} = \frac{\sigma_A}{2\sqrt{A_0}}$, dvs $\mathcal{J} = \sqrt{A_0} \pm \frac{\sigma_A}{2\sqrt{A_0}}$.

Använd även pdf:er och variabeltransformation:

$$p_{\mathcal{J}}(\mathcal{J} \mid D, I) = p_A(A(\mathcal{J}) \mid D, I) \left| \frac{dA}{d\mathcal{J}} \right| = 2\mathcal{J}\, p_A(A(\mathcal{J}) \mid D, I).$$

Konkret om

$$p_A(A \mid D, I) = \frac{1}{\sqrt{2\pi}\sigma_A} \exp\left(-\frac{1}{2}\frac{(A-A_0)^2}{\sigma_A^2}\right) \Rightarrow p_{\mathcal{J}}(\mathcal{J} \mid D, I) =$$

$$= \frac{2\mathcal{J}}{\sqrt{2\pi}\sigma_A} \exp\left(-\frac{1}{2}\frac{(\mathcal{J}^2-\mathcal{J}_0^2)^2}{\sigma_A^2}\right).$$

(Se figurer i kompendiet med vad som kan gå fel med approximationen)

F12 Stokastiska processer, Markovkedjor

TIF385
05/12-23

Stokastiska processer:

Deterministiska processer - givet resultat för given input.

Stokastiska processer - beskriver fenomen som kännetecknas av osäkerheter och slumpmässiga förlopp.

Ex. Tillväxtmodell.

Antalet bakterier $y(t)$ växer deterministiskt med tillväxtfaktor λ

$$\Rightarrow \frac{dy}{dt} = \lambda y(t) \Rightarrow \{y(0)=y_0\} \Rightarrow y(t) = y_0 e^{\lambda t}.$$

Detta beskriver inte det stokastiska förloppet.

$\lambda \Delta t$ = sannolikheten att en bakterie delar på sig under ett kort tidsintervall.

Def. Stokastisk process:

[En uppsättning indexerade slumpvariabler X_t med gemensamt utfallsrum S.]

Diskreta processer: $t \in \{0, 1, 2, \ldots\}$.

Kontinuerliga processer: $t \in \mathbb{R}$, $t \in [0, \infty)$, ...

Olika sekvenser av en given process kommer ge olika utfall. Processen beskrivs med (betingade) sannolikheter

$$P_{X_0}(x_0),\ P_{X_1|X_0}(x_1|x_0),\ \ldots,\ P_{X_n|X_0,\ldots,X_{n-1}}(x_n|x_0,\ldots,x_{n-1}).$$

Ex. Slumpvandring (diskret, endimensionell).

$$S_n = a + \sum_{i=1}^{n} X_i \leftarrow \text{slumpvariabel, } S=\{-1, 1\}.$$

S_n: position efter n steg; a: startposition

$\{S_t : t \in \mathbb{N}\}$ är den intressanta stokastiska processen.

Notera att $P_{S_n|S_0,\ldots,S_{n-1}}(s_n|s_0,\ldots,s_{n-1}) = P_{S_n|S_{n-1}}(s_n|s_{n-1}) =$

$$= \begin{cases} 0{,}5 & \text{för } s_n = s_{n-1}+1. \\ 0{,}5 & \text{för } s_n = s_{n-1}-1. \\ 0 & \text{för alla andra } s_n. \end{cases}$$

Dvs att ett utfall beror endast av det senasta utfallet. Detta kallas Markovkedja.

För kontinuerliga utfallsrum:

$$\mathbb{P}_{X_n|X_0,\dots,X_{n-1}}(x_n|x_0,\dots,x_{n-1}) = \mathbb{P}_{X_n|X_{n-1}}(x_n|x_{n-1}).$$

Stationära Markovkedjor:

En Markovkedja är stationär om de betingade sannolikheterna ej beror på tidsindexet, dvs för diskreta utfallsrum

$$\mathbb{P}_{X_{n+1}|X_n}(j|i) = \mathbb{P}_{X_1|X_0}(j|i) = T(i,j) \quad \forall n,i,j.$$

T kallas övergångsmatris $(|S|\times|S|)$.

För kontinuerliga utfallsrum:

$$\mathbb{P}_{X_{n+1}|X_n}(x_j|x_i) = \mathbb{P}_{X_1|X_0}(x_j|x_i) = T(x_i,x_j),$$

där T är en (kontinuerlig) övergångsdensitet.

Ibland skrivs T som $T(x_j \leftarrow x_i)$, T_{ij}.

Gränsdistributioner:

Betrakta diskret utfallsrum.

~~Frekvens~~sfunktionen blir en vektor π med element

$$\pi_j = \mathbb{P}_X(j), \quad j \in S.$$

En vektor π är gränsdistributionen till Markovkedjan som beskrivs av övergångsmatrisen T om

1) $\pi_j \geq 0 \quad \forall j, \ j \in S.$

2) $\sum_j \pi_j = 1.$

3) $\pi_j = \lim_{n\to\infty} T^n(i,j) \quad \forall i,j \in S.$

dvs $\pi = \lim_{n\to\infty} \alpha T^n$, α godtycklig startdistribution.

F13 Markovkedjor

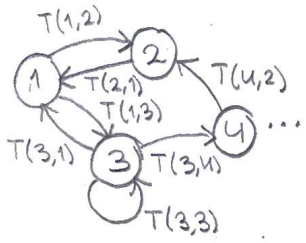
TIF385
06/12-23

Markovkedjor (forts.):

Betrakta diskret stationär Markovkedja med diskret utfallsrum. Detta beskrivs entydigt med

1) $p_{X_0}(i) = \mathbb{P}(X_0 = i)$ (frekvensfunktion för X_0).

2) $T(i,j) = \mathbb{P}_{X_{n+1}|X_n}(j|i)$ (övergångsmatris).



Initiär $\pi^{(0)} = (\pi_1^{(0)}, \dots, \pi_{|S|}^{(0)})$, $\pi_i^{(0)} = p_{X_0}(i)$.

($|S|$ ← # möjliga utfall)

Vad blir π^1?

$$\pi_i^{(1)} = p_{X_1}(i) = \sum_j p_{X_0,X_1}(j,i) = \sum_j p_{X_1|X_0}(i|j)\, p_{X_0}(j) = \sum_j T(j,i)\pi_j^{(0)} \Rightarrow$$

$$\Rightarrow \pi^{(1)} = \pi^{(0)}T \Rightarrow \pi^{(n)} = \pi^{(0)}T^n.$$

Under vissa förutsättningar existerar en gränsfördelning π så att

$$\pi = \lim_{n\to\infty} \alpha T^n, \quad \alpha = \pi^0 \text{ godtycklig startfördelning.}$$

Stationärfördelning:

Vektorn π är en stationärfördelning till Markovkedjan (som beskrivs av T) om

1) $\pi_j \geq 0 \ \forall j, \ j \in S$.

2) $\sum_j \pi_j = 1$.

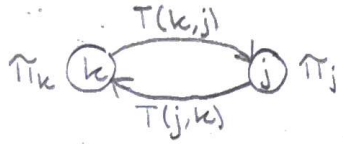
3) $\pi_j = \sum_i \pi_i T(i,j) \ \forall j$ $(\pi = \pi T)$.

En gränsfördelning är alltid en stationärfördelning, men det motsatta gäller inte alltid.

Reversibla kedjor:

En stationär Markovkedja (övergångsmatris T) med stationärfördelning π är reversibel om

$$\pi_k T(k,j) = \pi_j T(j,k) \quad \forall j,k \in S.$$



Detta beskriver en detaljerad balans av förflyttad sannolikhetsvekt. Sambandet innebär också att

$$P_{X_{n-1},X_n}(k,j) = P_{X_{n-1}|X_n}(j,k),$$

dvs invariant under tidsreversion.

Denna egenskap garanterar att processen är stationär.

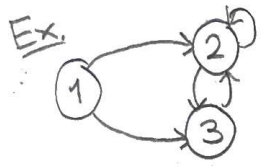
Existensvillkor för gränsdistribution:

1) Icke-reducerbar: Alla tillstånd kan nås oavsett startpunkt.

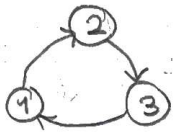
2) Positivt återkommande: Sannolikheten att återkomma till ett tillstånd någon gång är 1 $\forall i$.

3) Icke-periodisk: Sekvenser av tillstånd upprepas inte genom hela kedjan.

Ex.



reducerbar
ej positivt återkommande
ej periodisk



(*)

ej reducerbar
positivt återkommande
periodisk

Studera (*):

- Har en stationär fördelning $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.
- Men detta är ej en gränsfördelning, ty exv

$$\pi^{(0)} = (1,0,0), \quad \pi^{(1)} = (0,1,0), \ldots$$

Repetition: (Markovkedjor)

Betrakta stationär Markovkedja med diskret utfallsrum.

$\mathbb{P}_{X_{n+1}|X_n}(j \mid i) = T(i,j) \quad \forall i,j \in S$ ober. av n.

$\pi_j^{(n)} = \mathbb{P}_{X_n}(j)$.

Vid jämvikt:

$\pi = \pi T$, π kedjans stationärfördelning.

Ett starkare villkor som garanterar jämvikt är detaljerad balans

$\pi_i T(i,j) = \pi_j T(j,i)$.

I vissa fall är stationärfördelningen även kedjans gränsfördelning, dvs kedjan konvergerar mot π oavsett startfördelning α.

$\pi = \lim_{n\to\infty} \alpha T^n$.

Tillräckliga (men ej nödvändiga) villkor för existens av gränsfördelning:

- Irreducibel.
- Positivt återkommande.
- Aperiodisk.

Metropolisdesign:

Man kan konstruera en Markovkedja (diskret utfallsrum) så att den har en speciell gränsfördelning π.

Bevis:

Introducera en stegförslagsfunktion:

$S(i,j) = \mathbb{P}_{X_{n+1}|X_n}(\text{föreslå position } j \mid i)$.

- Måste vara en stokastisk matris, annars valfri.

$\sum_j S(i,j) = 1 \quad \forall i, \quad S(i,j) \in [0,1]$.

Introducera acceptansfunktion:

$A(i,j) = \mathbb{P}_{X_{n+1}|X_n}(\text{acceptera } j \mid i), \quad A(i,j) \in [0,1]$.

A och S är oberoende för $i \neq j$, vilket ger

$$T(i,j) = \mathbb{P}_{\xi_{n+1}|\xi_n}(j|i) = S(i,j)A(i,j), \quad i \neq j.$$

Detaljerad balans

$$\pi_i S(i,j)A(i,j) = \pi_j S(j,i)A(j,i) \Rightarrow \frac{A(i,j)}{A(j,i)} = \frac{\pi_j}{\pi_i}\frac{S(j,i)}{S(i,j)},$$

vilket uppfylls om

$$A(i,j) = \min\left(1, \frac{\pi_j S(j,i)}{\pi_i S(i,j)}\right).$$

Diagonalelementen fås genom normalisering

$$T(i,i) = 1 - \sum_{j\neq i} T(i,j) = \sum_j S(i,j) - \sum_{j\neq i} S(i,j)A(i,j) =$$

$$= S(i,i) + \sum_{j\neq i} S(i,j)[1 - A(i,j)].$$

$1 - A(i,j)$: sannolikheten att inte acceptera ett steg från $i \to j$.

Algoritm:

1. Välj $S(i,j)$.
2. Beräkna $A(i,j)\ \forall i \neq j \Rightarrow T(i,j)\ \forall i \neq j$.
3. Beräkna $T(i,i)\ \forall i$.

$\Rightarrow$ Markovkedja med stationärfördelning π.

För att vara säker på att π också är en gränsfördelning kan man bekräfta att T har tillräckliga egenskaper, ex.v aperiodisk, irreducibel, positivt återkommande.

Kontinuerliga, multivariata utfallsrum. Skillnader från tidigare:

Utfall: $i \to \Theta \in \mathbb{R}^p$.

Fördelningar: $\pi_i = p(i) \to p(\Theta)$.

Sannolikheter: $\mathbb{P}_{\xi_n}(i) = \pi_i \to \mathbb{P}(\xi_n \leq \Delta\Theta) = \int_{\Delta\Theta} p(\Theta')\,d\Theta'$.

För stationära Markovkedjor:

$T(i,j) \to T(\Theta,\Theta') = p_{\xi_{n+1}|\xi_n}(\Theta'|\Theta)$ (ober. av n).

uppfyller $\int_{\mathbb{R}^p} T(\Theta,\Theta')d\Theta' = 1\ \forall\Theta \in \mathbb{R}^p$.

Stationär Markovkedja i jämvikt:

$$p(\Theta) = \int_{\mathbb{R}^p} p(\Theta') T(\Theta', \Theta)\, d\Theta'.$$

Starkare villkor med detaljerad balans

$$p(\Theta) T(\Theta, \Theta') = p(\Theta') T(\Theta', \Theta').$$

Metropolis(-Hastings)algoritmen:

$$T(\Theta, \Theta') = J(\Theta, \Theta') A(\Theta, \Theta').$$

Stegförslagsfunktionen måste vara en pdf:

$$p_{z_{n+1}|z_n}(\text{föreslå } \Theta' | \Theta).$$

T.ex. likformig i en liten volym nära Θ,

$$J(\Theta, \Theta') = \begin{cases} \frac{1}{\Delta\Theta^p} & \text{då } |\Theta_i' - \Theta_i| \leq \frac{\Delta\Theta}{2} \ \forall i \in \{1, \ldots, p\}. \\ 0 & \text{annars.} \end{cases}$$

eller normalfördelad

$$J(\Theta, \Theta') = N(\Theta' | \Theta, \Sigma).$$

Acceptansfunktionen

$$A(\Theta, \Theta') = \min\left(1, \underbrace{\frac{p(\Theta')}{p(\Theta)} \frac{J(\Theta', \Theta)}{J(\Theta, \Theta')}}_{\equiv r \text{ (Metropoliskvoten)}}\right).$$

För symmetriska J fås

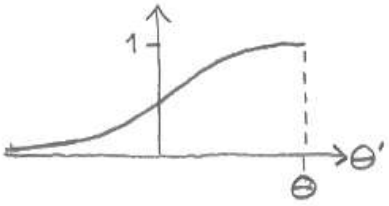
$$r = \frac{p(\Theta')}{p(\Theta)}.$$

Stickprov från en pdf

En (pseudo-)slumptalsgenerator kan producera

$\{u_i\}_{i=1}^N$, $u_i \sim U([0,1])$.

Hur produceras slumptal $\{\theta_i\}_{i=1}^N$, $\theta_i \sim N(0,1)$?



$$P(\theta) = \int_{-\infty}^{\theta} p(\theta')d\theta'.$$

$$P(\theta_i) = u_i \Rightarrow \theta_i = P^{-1}(u_i) \sim p(\theta).$$

Fungerar för analytiska, icke-inverterbara fördelningar.

Man kan använda MCMC-algoritmer för att dra stickprov.

Metropolisalgoritmen:

1. Initialisera samplingen genom att dra θ_0 ur en startfördelning.
2. Dra stickprov från slumpvariabel X_{i+1} enligt följande:
   a) Givet utfallet θ_i för X_i, föreslå ϕ genom dragning ur $S(\theta_i, \phi)$.
   b) Beräkna $r = \frac{p(\phi)}{p(\theta_i)}$ (Metropolis - om S symmetrisk),
   $r = \frac{p(\phi)S(\phi,\theta_i)}{p(\theta_i)S(\theta_i,\phi)}$ (Metropolis-Hastings - om S ej symmetrisk)
   c) Om • $r \geq 1$: $\theta_{i+1} \leftarrow \phi$.
   • $r < 1$: Dra $u \sim U([0,1])$, om
   - $u \leq r$: $\theta_{i+1} \leftarrow \phi$.
   - $u > r$: $\theta_{i+1} \leftarrow \theta_i$.

Fortsätt tills kedjan har konvergerat (svårt) för då kommer alla nya utfall vara stickprov ur $p(\theta)$. Notera att man bara evaluerar $p(\theta)$ punktvis (i b). Algoritmen fungerar i höga dimensioner, men ineffektiv. (Mer sofistikerade algoritmer finns.

Bayesiansk inferens:

Utmaningar:

1. Utforska a posteriori-fördelningar $p(\Theta \mid D, I)$.

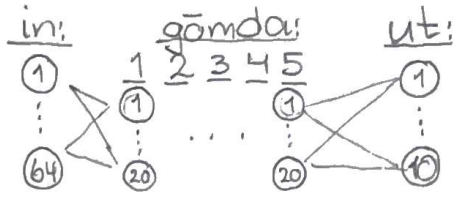
   - Var finns moden?
   - Finns det flera moder?
   - Vad är sannolikhetsmassan koncentrerad?
   - Parameterkorrelationer?
   - Hur ser marginalfördelningarna ut?

2. A posteriori-förutsägelser.

   - Evaluera modell $g(\Theta)$ för $\Theta \sim p(\Theta \mid D, I)$.
   - Förväntansvärde $\mathbb{E}_p[g(\Theta)] = \int_\Theta g(\Theta) p(\Theta \mid D, I)\, d\Theta$
   - Fördelning av modellprediktioner (PPD). $\{g(\Theta) : \Theta \sim p(\Theta \mid D, I)\}$.

Samtliga frågor kan besvaras med stickprov $\{\Theta_i\}_{i=1}^N$, $\Theta_i \sim p(\Theta \mid D, I)$.

2301-2 (Läs uppgiftsformulering själv)

a) 

Utsignal från lager 1:

$$y^1 = f(z^1), \quad y_i^1 = f(z_i^1) = \frac{1}{1+e^{-z_i^1}}.$$

y^1, z^1 har storleken (1×20).

Aktiveringarna ges av

$$z^1 = y^{in} W^1 + W_0^1.$$

(y^{in}: insignaler $(1\times N_{in})$; W^1: vikter $(N_{in}\times N_1)$; W_0^1: biasvikter $(1\times N_1)$)

b) Det finns vikter i alla gömda lager samt i utlagret. Varje nod (i) i ett lager (j) har $N_{j-1}+1$ vikter.
(vikter för varje nod i tidigare lagret + bias)

Antal parametrar per lager:

$$\underset{1}{20\cdot(64+1)} + \underset{2}{(20} + \underset{3}{20} + \underset{4}{20} + \underset{5}{20} + \underset{ut}{10)}\cdot(20+1) = \underline{3190 \text{ vikter}}.$$

c) Mjuk klassificerare; 10 klasser

$\Rightarrow$ utsignal $\hat{y}_i = P(t=i)$, $i \in \{0,\dots,9\}$, $\sum_{i=0}^{9} \hat{y}_i = 1$.

Använd t.ex. sigmoidfunktion i utlagret

$$y_i^{ut} = \frac{1}{1+e^{-z_i^{ut}}} \in [0,1].$$

Normalisera

$$\hat{y}_i = \frac{y_i^{ut}}{\sum_{i=0}^{9} y_i^{ut}},$$

vilket ger en diskret sannolikhetsfördelning $\hat{y}$.

Bevis 1 Prediktion $\hat{y}^\odot \equiv \hat{y}(x^\odot)$ jämförs med framtida observation $y^\odot \equiv y(x^\odot)$. $y(x) = f(x) + \varepsilon$, $f(x)$ underliggande funktionen, ε fel. $\mathbb{E}[\varepsilon]=0$, $\mathrm{Var}(\varepsilon)=\sigma_\varepsilon^2$.

Visa $\mathbb{E}[(y^0-\hat{y}^0)^2] = (f^0-\mathbb{E}[\hat{y}^0])^2 + \mathrm{Var}[\hat{y}^0] + \sigma_\varepsilon^2$,
där $f^0 = f(x^0)$.
(Bias-variance theorem)

$y^0 = f^0 + \varepsilon^0$, $\mathbb{E}[y^0] = f^0$, $\mathrm{Var}[y^0] = \sigma_\varepsilon^2$. $\quad +(I) \quad\quad (II)$

Söker $\mathbb{E}[(y^0-\hat{y}^0)^2] = \mathbb{E}[(f^0+\varepsilon^0-\hat{y}^0)^2] = \mathbb{E}[(f^0+\varepsilon^0-\mathbb{E}[\hat{y}^0]+\mathbb{E}[\hat{y}^0]-\hat{y}^0)^2]$.

$(I)^2$ • $\mathbb{E}[(f^0-\mathbb{E}[\hat{y}^0]+\varepsilon^0)^2] = \mathbb{E}[(f^0-\mathbb{E}[\hat{y}^0])^2] + \mathbb{E}[\varepsilon^{0^2}] + 2\mathbb{E}[(f^0-\mathbb{E}[\hat{y}^0])\varepsilon^0] =$

$= \left\{ \begin{array}{l} \text{Notera} \\ \bullet \mathrm{Var}[\varepsilon^0] = \mathbb{E}[(\varepsilon^0 - \underbrace{\mathbb{E}[\varepsilon^0]}_{=0})^2] = \mathbb{E}[\varepsilon^{0^2}] = \sigma_\varepsilon^2 \\ \bullet 2(f^0-\mathbb{E}[\hat{y}^0])\underbrace{\mathbb{E}[\varepsilon^0]}_{=0} = 0 \end{array} \right\} = \mathbb{E}[(f^0-\mathbb{E}[\hat{y}^0])^2] + \sigma_\varepsilon^2.$

$(II)^2$ • $\mathbb{E}[(\hat{y}^0-\mathbb{E}[\hat{y}^0])^2] = \mathrm{Var}[\hat{y}^0]$. (modellens prediktionsvarians)

$2(I)(II)$ • $2\mathbb{E}[(f^0-\mathbb{E}[\hat{y}^0]+\varepsilon^0)\underbrace{(\mathbb{E}[\hat{y}^0]-\hat{y}^0)}_{\mathbb{E}[(\bullet)]=\ldots=0}]$

Bevis 3| Betrakta $\mathbb{D} = \mathbb{X}\Theta + \mathbb{E}$, $p(\varepsilon_i|I) = \mathcal{N}(0,\sigma_\varepsilon^2) = \frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\exp\left[-\frac{1}{2}\frac{\varepsilon_i^2}{\sigma_\varepsilon^2}\right]$.

a) Betraktа observation D_i, jämför $M_i = (\mathbb{X}\Theta)_i$, $D_i = M_i + \varepsilon_i$.

$p(D_i|\Theta,I) = \{\text{var. byte}\} = p_\varepsilon(\varepsilon_i = D_i - M_i|\Theta,I)\underbrace{\left|\frac{d\varepsilon_i}{dD_i}\right|}_{=1} =$

$= \frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\exp\left[-\frac{1}{2}\frac{(D_i-M_i)^2}{\sigma_\varepsilon^2}\right]$.

Felen oberoende $\Rightarrow$

$\Rightarrow p(\mathbb{D}|\Theta,I) = \prod_{i=1}^{N_d} p(D_i|\Theta,I) = \left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{N_d/2}\exp\left[-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{N_d}(D_i-M_i)^2\right] =$

$= \left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{N_d/2}\exp\underbrace{\left[-\frac{1}{2}\frac{(\mathbb{D}-\mathbb{X}\Theta)^T(\mathbb{D}-\mathbb{X}\Theta)}{\sigma_\varepsilon^2}\right]}_{\equiv -L(\Theta)}$. $\square$

b) Taylorutvecka $L(\Theta)$ runt $\Theta^* = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{D}$.

$L(\Theta) = L(\Theta^*) + \sum_{j=0}^{N_p-1}\frac{\partial L}{\partial\theta_j}\Big|_{\Theta=\Theta^*}(\theta_j-\theta_j^*) + \frac{1}{2}\sum_{j=0}^{N_p-1}\Big(\sum_{i=0}^{N_p-1}\frac{\partial^2 L}{\partial\theta_i\partial\theta_j}\Big|_{\Theta=\Theta^*}$

$\cdot(\theta_i-\theta_i^*)(\theta_j-\theta_j^*) + \ldots = L(\Theta^*) + \nabla L\Big|_{\Theta=\Theta^*}(\Theta-\Theta^*) + \frac{1}{}$

$$+\frac{1}{2}(\Theta-\Theta^*)^T H|_{\Theta=\Theta^*}(\Theta-\Theta^*)+\dots$$

∇L och H ges i uppgiftsledtråden.

$$\nabla L|_{\Theta=\Theta^*} = \frac{\mathbb{X}^T D - \mathbb{X}^T \mathbb{X}\Theta^*}{\sigma_\varepsilon^2} = \{D=\mathbb{X}\Theta^*\} = 0.$$

$H = \frac{\mathbb{X}^T\mathbb{X}}{\sigma_\varepsilon^2}$ oberoende av $\Theta \Rightarrow \frac{\partial^{(n)}L}{\partial_{\dots}} = 0$ för $n \geq 3$.

Alltså, $L(\Theta) = L(\Theta^*) + \frac{1}{2}(\Theta-\Theta^*)^T \underbrace{\frac{\mathbb{X}^T\mathbb{X}}{\sigma_\varepsilon^2}}_{\equiv \Sigma_\theta^{-1}}(\Theta-\Theta^*)$.

$$p(D|\Theta,I) = \underbrace{\left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{N_0/2}\exp\left[-L(\Theta^*)\right]}_{=p(D|\Theta^*,I)}\exp\left[-\frac{1}{2}(\Theta-\Theta^*)^T\Sigma_\theta^{-1}(\Theta-\Theta^*)\right]. \quad \square$$

(Läs uppgiftsbeskrivningarna själv)