

Introducere in invatarea automata

Invatare supervizata
Estimatori si Regresie

Ion Necoara & Daniela Lupu

2023

Learning paradigms in ML

- ▶ **Supervised learning:** Supervised learning occurs when an algorithm is trained using “*labeled data*,” or data that is tagged with a label so that an algorithm can successfully learn from it.
- ▶ **Unsupervised learning:** Unsupervised algorithms use *unlabeled* data to train an algorithm. In this process, the algorithm finds patterns in the data itself and creates its own data clusters.
- ▶ **Reinforcement learning:** Reinforcement learning is a machine learning technique in which positive and negative values are assigned to desired and undesired actions. The goal is to encourage programs to avoid the negative training examples and seek out the positive, learning how to maximize rewards through trial and error. Reinforcement learning can be used to direct unsupervised machine learning.
- ▶ **Semi-supervised learning:** Semi-supervised learning uses a *mix of labeled and unlabeled* data to train an algorithm. In this process, the algorithm is first trained with a small amount of labeled data before being trained with a much larger amount of unlabeled data.

ML types

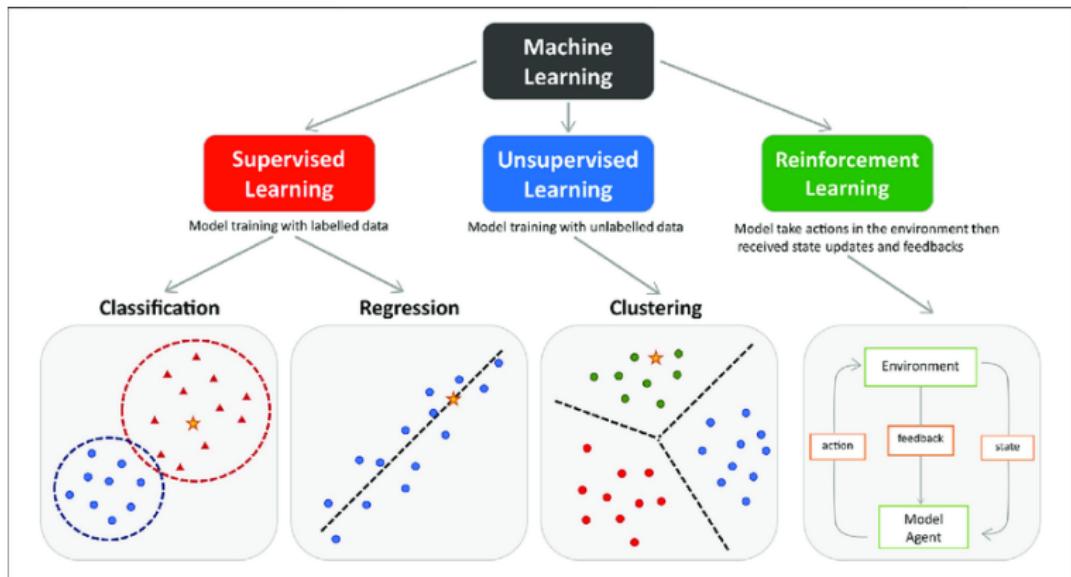
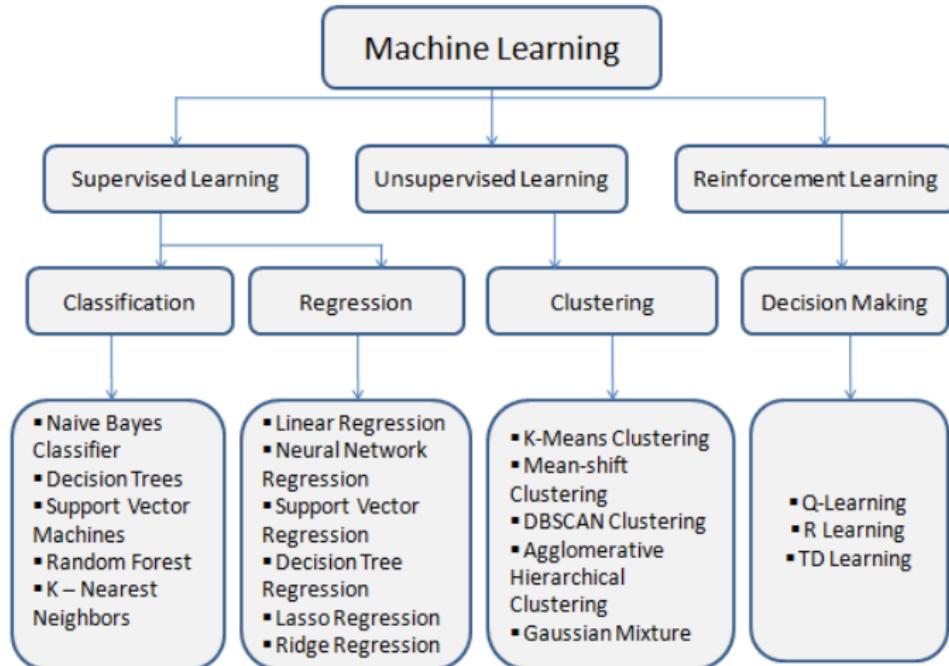


Figure: Coloured dots and triangles represent the training data. Yellow stars represent the new data which can be predicted by the trained model.(Courtesy of [1])



In acest curs ne concentrăm pe regresie (liniara, ridge, lasso, logistic,...) și estimatori folosind date cu "labels"!

INVATARE SUPERVIZATA!

Scop statisticii este de a face inferente/deductii pe baza datelor.

Etapele statisticii:



Proiectarea experimentului

- Colectarea de date utile si de calitate



Statistica descriptivă

- Indicatori: medie, varianta, etc
- Grafice: histogramme, grafice cu linii



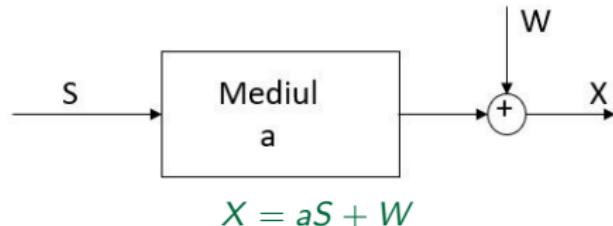
Statistica inferentială

- Analizeaza datele pentru a deduce proprietatile unei distributii asunse de probabilitati
- testarea ipotezelor si derivarea estimarilor



"To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of" (Ronald A. Fisher)

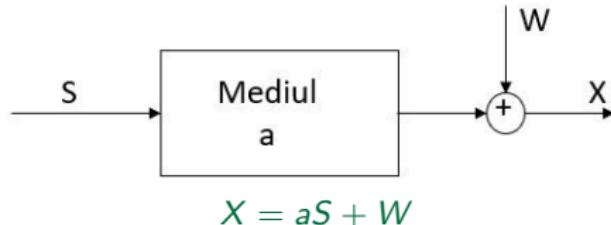
Tipuri de inferenta



► Abordarea "black box"

- Se cunoaste: S
- Se observa: X
- Se deduce: a

Tipuri de inferenta



► Abordarea "black box"

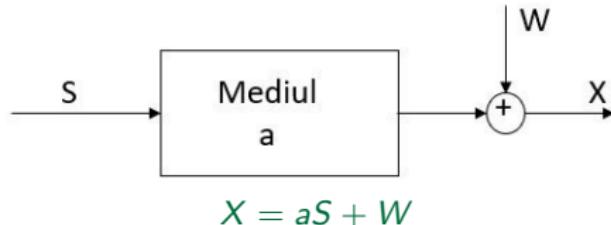
- Se cunoaste: S
- Se observa: X
- Se deduce: a

► Estimare variabile

- Se cunoaste: a
- Se observa: X
- Se deduce: S

Desi interpretarea celor doua metode este diferita, structura matematica este aceeasi.

Tipuri de inferenta



► Abordarea "black box"

- Se cunoaste: S
- Se observa: X
- Se deduce: a

► Estimare variabile

- Se cunoaste: a
- Se observa: X
- Se deduce: S

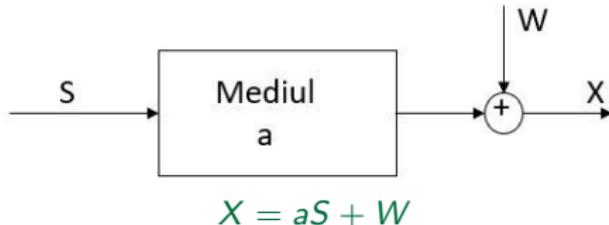
Desi interpretarea celor doua metode este diferita, structura matematica este aceeasi.

Testare ipoteze versus Estimare

► Problema testarea ipotezelor:

- Necunoscuta poate lua una sau mai multe valori → avem mai multe modele posibile. Care model este cel corect ?
- Scop: probabilitatea ca modelul ales sa fie incorect sa fie cat mai mica.

Tipuri de inferenta



► Abordarea "black box"

- Se cunoaste: S
- Se observa: X
- Se deduce: a

► Estimare variabile

- Se cunoaste: a
- Se observa: X
- Se deduce: S

Desi interpretarea celor doua metode este diferita, structura matematica este aceeasi.

Testare ipoteze versus Estimare

► Problema testarea ipotezelor:

- Necunoscuta poate lua una sau mai multe valori \rightarrow avem mai multe modele posibile. Care model este cel corect ?
- Scop: probabilitatea ca modelul ales sa fie incorect sa fie cat mai mica.

► Problema estimarii:

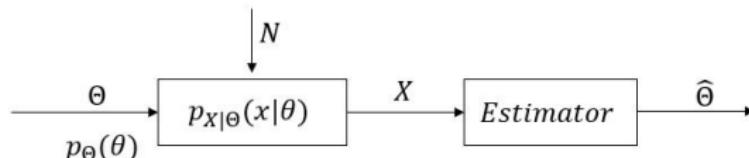
- necunoscuta ia valori numerice, poate fi chiar continua
- Scop: estimarea cea mai "apropiata" de adevar pentru o cantitate pe care nu o cunoastem

Statistica

Tipuri de inferenta

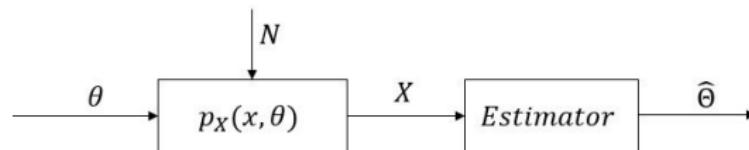
Bayesiana:

- ▶ **Abordare:** Tot ce este necunoscut se modeleaza ca o variabila aleatoare
- ▶ Se utilizeaza probabilitati apriori si teorema lui Bayes



Clasica: Fie θ parametrul necunoscut, de exemplu masa unui electron.

- ▶ **Abordare:** θ este un numar real constant, necunoscut (**nu este o variabila aleatoare**)



Statistica si inferenta sunt un exercitiu aplicat in probabilitate.

- ▶ In probabilitate exista un singur raspuns corect. Nu exista ambiguitate.
- ▶ In statistica nu exista un raspuns unic. Se pot aplica diferite metode ce vor produce un raspuns cu proprietati diferite.

Inferenta Bayesiana

Inferenta Bayes

Formularea problemei

Fie Θ o necunoscută și X un set de observații independente generate de aceasta.
Obiectiv: Aflați necunoscuta Θ .

Abordarea Bayes:

- ▶ Θ tratată ca o v.a ce are o distribuție de probabilitate p_Θ sau f_Θ .
- ▶ Probabilitatea apriorică este ceea ce se cunoaște despre Θ înainte de observații.

Observații:

- ▶ Nu este posibil să aflăm cu exactitatea Θ . Se poate doar estimă pe baza datelor, i.e. se dorește să aflăm $p_{\Theta|X}$.
- ▶ X este o v.a \Rightarrow asociem un model probabilistic $p_{X|\Theta}$ sau $f_{X|\Theta}$.

Notări:

- ▶ $x = \{x_1, x_2, \dots, x_n\}$ o realizare a lui X .
- ▶ $\hat{\Theta}_x = g(X)$ **un estimator**. Este o v.a
- ▶ $\hat{\theta}_n = g(x)$ **valoarea estimată**. Este un număr.

Inferenta Bayes

Formularea problemei



Cum se gaseste probabilitatea apriorica p_θ ?

- ▶ Simetria: Daca exista o serie de optiuni posibile pentru Θ si exista un motiv sa credem ca toate sunt la fel de probabile, atunci consideratia de simetrie ne ofera o probabilitate a prior uniform.
- ▶ Intervale cunoscute
- ▶ Studii anterioare
- ▶ Arbitrara/subiectiva, reflectand doar convingerile decidentului despre Θ .

Pentru a calcula probabilitatea posterioara $p_{\Theta|X}$ se utilizeaza formula lui Bayes:

$$p_{\Theta|X}(\theta | X = x) = \frac{p_{X|\Theta}(X = x | \theta)p_{\Theta}(\theta)}{p(X = x)}$$

Estimarea de puncte cu inferenta Bayes

Inferenta Bayes produce o probabilitate a posterioara: $p_{\Theta|X}$ sau $f_{\Theta|X}$



Estimarea punctului θ :

1. Regula probabilitatii a posterioare maxime (MAP):

$$p_{\Theta|X}(\theta^* | x) = \max_{\theta} p_{\Theta|X}(\theta | x)$$
$$f_{\Theta|X}(\theta^* | x) = \max_{\theta} f_{\Theta|X}(\theta | x)$$

2. Estimatorul Bayes:

$$\theta^* = \arg \min_{\theta \in \hat{\Theta}_X} E[L(\Theta; \hat{\Theta}_X) | X]$$

- ▶ $L(\Theta; \hat{\Theta}_X)$ s.n functie loss
- ▶ Pentru a gasi θ^* se rezolva ecuatia:

$$E\left[\frac{\partial L}{\partial \theta_x}(\theta, \theta_x)\right] = 0$$

- ▶ Exemplu $L = (\Theta - \hat{\Theta}_X)^2$ (LMS - least mean square)

Proprietati ale estimatorului Bayes

Estimatorul Bayes $\hat{\Theta}_X$ este:

- ▶ biased : $\text{bias}(\hat{\Theta}_X | \Theta) = \mathbb{E}(\hat{\Theta}_X - \theta | \Theta = \theta) = \mathbb{E}(\hat{\Theta}_X | \theta) - \theta$
- ▶ asimptotic unbiased si consistent, i.e pentru o sevenita de estimatori Bayes $\hat{\Theta} = \{\hat{\Theta}_n, n \in \mathbb{N}_+\}$, $\text{bias}(\hat{\Theta}_n | \theta) \rightarrow 0$ cand $n \rightarrow \infty$ pentru fiecare $\theta \in \Theta$

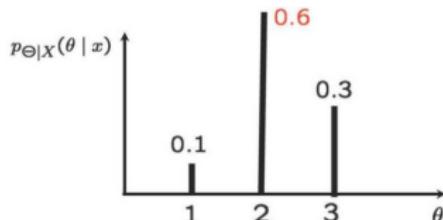
Estimarea MAP

Studiu de caz: Θ si X discrete.

- θ - semnificatia unei ipoteze.

$$p_{\Theta|X}(\theta | x) = \frac{p_{\Theta}(\theta)p_{X|\Theta}(x | \theta)}{p_X(x)}$$

$$p_X(x) = \sum_{\theta_i} p_{\Theta}(\theta_i) p_{X|\Theta}(x | \theta_i)$$



Regula MAP: $\hat{\theta} = 2$

Cat de buna este estimarea?

- Un criteriu bun pentru cazul ipotezelor este probabilitatea conditionata a erorii:

$$P(\hat{\theta} \neq \Theta | X = x) = 0.4$$

mica sub regula MAP

- Probabilitatea erorii in general:

$$P(\hat{\Theta} \neq \Theta) = \sum_x P(\hat{\Theta} \neq \Theta | X = x)p_X(x)$$

$$\stackrel{\text{sau}}{=} \sum_{\theta} P(\hat{\Theta} \neq \Theta | \Theta = \theta)p_{\Theta}(\theta)$$

Obs: Sub regula MAP, probabilitatea erorii este de asemenea mica \Rightarrow regula MAP este o *modalitate optima* de estimare în contextul testării ipotezelor, unde se dorește minimalizarea probabilității de eroare

Estimarea MAP

Studiu de caz: Θ discret si X continuu

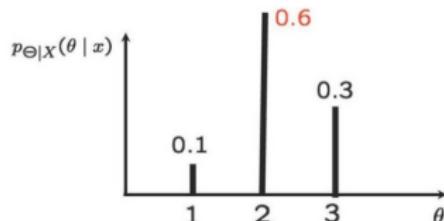
Exemplu standard:

- ▶ Se trimite semnalul $\Theta = \{1, 2, 3\}$
- ▶ Observatia $X = \Theta + W$, unde
- ▶ zgometul $W \sim N(0, \sigma^2)$ si este independent de Θ

$$p_{\Theta|X}(\theta | x) = \frac{p_{\Theta}(\theta) f_{X|\Theta}(x | \theta)}{f_X(x)}$$

$$f_X(x) = \sum_{\theta_i} p_{\Theta}(\theta_i) f_{X|\Theta}(x | \theta_i)$$

$$f_{X|\Theta}(x | \theta) = f_W(x - \theta)$$



Regula MAP: $\hat{\theta} = 2$

Cat de buna este estimarea?

- ▶ Un criteriu bun pentru cazul ipotezelor este probabilitatea conditionata a erorii:

$$P(\hat{\theta} \neq \Theta | X = x) = 0.4$$

mica sub regula MAP

- ▶ Probabilitatea erorii in general:

$$\begin{aligned} P(\hat{\Theta} \neq \Theta) &= \int P(\hat{\Theta} \neq \Theta | X = x) f_X(x) dx \\ &= \sum_{\theta} P(\hat{\Theta} \neq \theta | \Theta = \theta) p_{\Theta}(\theta) \end{aligned}$$

Obs:

- ▶ Sub regula MAP, probabilitatea erorii este de asemenea mica
- ▶ Similar se trateaza si cazul cand Θ si X sunt continue.

Estimarea MAP

$$\theta^* = \arg \max_{\theta} p_{\Theta|X}(\theta|x) = \frac{p_{X|\Theta}(X=x|\theta)p_{\Theta}(\theta)}{p(X=x)} = \arg \max_{\theta} p_{X|\Theta}(X=x|\theta)p_{\Theta}(\theta) = \\ = \arg \max_{\theta} \prod_i p(x_i|\theta)p_{\Theta}(\theta)$$

* - $p(X=x)$ este un factor de normare si se poate renunta la el

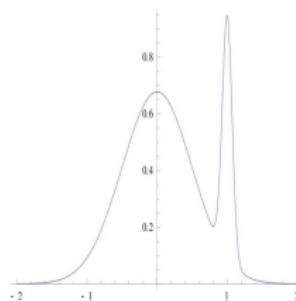
Pentru a maximiza mai usor, vom utiliza expresia logaritmizata:

$$\theta^* = \arg \max_{\theta} \ln p_{\Theta|X}(\theta|x) = \arg \max_{\theta} \ln p_{\Theta}(\theta) + \sum_i \ln p(x_i|\theta)$$

Limitari

In multe tipuri de modele, cum ar fi amestecul de modele, probabilitatea posteriora poate fi multimodala.

Intr-un astfel de caz, recomandarea obisnuita este aceea de a alege modul cel mai inalt: acest lucru nu este intotdeauna fezabil (optimizarea globala este o problema dificila) sau posibil in unele cazuri (cum ar fi atunci cand apar probleme de identificare). Mai mult, cel mai inalt mod poate fi necaracteristic pentru majoritatea posteriora.



Estimatorul MAP

Exemplu 1

Fie X_1, X_2, \dots, X_n v.a independente cu o distributie Bernoulli avand pmf-ul:

$$f(x | \theta) = \theta^x (1 - \theta)^{1-x}, \quad x = 0, 1$$

unde θ este necunoscut. Presupunem ca θ este ales dintr-o distributie uniforma pe $(0, 1)$. Utilizati regula MAP pentru a gasi estimarea $\hat{\theta}$

Solutie: Stim ca $p(\theta) = 1$, $0 \leq \theta \leq 1$. Asadar,

$$\begin{aligned}\hat{\theta}_n &= \arg \max_{\theta} \ln p_{\Theta|X}(\theta|x) = \arg \max_{\theta} \ln p_{\Theta}(\theta) + \sum_i \ln f(x_i|\theta) \\ &= \arg \max_{\theta} \sum_{i=1}^n \ln [\theta^{x_i} (1 - \theta)^{1-x_i}] = \arg \max_{\theta} \sum_{i=1}^n [\ln \theta^{x_i} + \ln(1 - \theta)^{1-x_i}] \\ &= \arg \max_{\theta} \sum_{i=1}^n [x_i \ln \theta + (1 - x_i) \ln(1 - \theta)]\end{aligned}$$

Estimatorul MAP

Exemplu 1 cont.

Calculam derivata paritala in raport cu θ si egalam cu zero:

$$\sum_{i=1}^n \frac{\partial}{\partial \theta} [x_i \ln \theta + (1 - x_i) \ln(1 - \theta)] = 0$$

$$\sum_{i=1}^n \left[\frac{x_i}{\theta} - \frac{1 - x_i}{1 - \theta} \right] = 0$$

Notam cu $\bar{x} = \sum_{i=1}^n x_i$ si revenim in ultima ecuatie:

$$\frac{\bar{x}}{\theta} - \frac{n - \bar{x}}{1 - \theta} = 0 \Leftrightarrow \bar{x} - \bar{x}\theta = n\theta - \bar{x}\theta \Rightarrow \theta = \frac{\bar{x}}{n}$$

De exemplu, daca am avea $n = 10$ incercari, din care 6 au succes, atunci utilizand rezultatul de mai sus, gasim estimarea $\theta = \frac{6}{10}$

Estimatorul Bayes

Exemplu 2: Clasificarea Binara

Un exemplu clasic de clasificare cu estimatorul Bayes o reprezinta filtrarea email-urilor, adica dandu-se un e-mail dorim sa decidem daca un email este "spam" sau "ham". Pentru a modela acest lucru, se considera:

- ▶ spatiul parametrilor $\Theta = \{0, 1\}$, unde
 - ▶ 0: e-mail "ham"
 - ▶ 1: e-mail "spam".
- ▶ e-mailul X este extras fie din
 - ▶ distributia de "ham" f_0 (notatie pentru $f_{X|0}$)
 - ▶ distributia de spam f_1 (notatie pentru $f_{X|1}$).
- ▶ spatiul de decizie (estimarea) $\hat{\Theta}_X = \{0, 1\}$

Pentru aceasta clasificare binara, o alegere naturala a functiei loss este:

$$L(\theta, \hat{\theta}) = \begin{cases} 0, & \hat{\theta} = \theta \\ 1, & \hat{\theta} \neq \theta \end{cases}$$

Definim o probabilitate apriori pentru e-mailuri:

$$P(\hat{\Theta}_X = 0) = 1 - p \text{ si } P(\hat{\Theta}_X = 1) = p$$

Alegera lui p se bazeaza pe cunostintele anterioare si p este egal cu proportia de e-mailuri primite anterior care erau spam.

Estimatorul Bayes

Exemplu 2 cont.

Estimatorul Bayes minimizeaza valoara medie a riscului:

$$\min E[L(\Theta; \hat{\Theta}_X) | X] = P(\hat{\Theta}_X \neq \Theta | X) = 1 - P(\Theta = \hat{\Theta}_X | X)$$

A se observa ca minimizarea riscului in medie este echivalent cu a maximiza probabilitatea de clasificare corecta $P(\Theta = \hat{\Theta}_X | X)$.

Ca exercitiu de gandire, considerati incercarea de a prezice eticheta unui e-mail primit inainte de a fi vizualizat. Cum nu exista date pentru conditionare, singura optiune valabila o reprezinta parametrul estimatorului θ . Atunci riscul mediu al acestuia este:

$$r(\pi, \delta_1) = \pi(1)R(1, \delta_1) + \pi(0)R(0, \delta_1) = 1 - p$$

si

$$r(\pi, \delta_0) = \pi(1)R(1, \delta_0) + \pi(0)R(0, \delta_0) = p$$

Deci $\theta = 1$ are un risk mediu mic cand $p > 1/2$, si $\theta = 0$ are un risk mediu mic cand $p < 1/2$.

Estimatorul Bayes

Exemplu 2 cont.

Dupa observarea datelor $X = x$, estimam $\hat{\theta} = 1$ daca are o probabilitate posteroioara mare $\mathbb{P}(\Theta = \hat{\theta}_x | X = x)$, aceasta fiind proportionala cu produsul dintre probabilitatea a priori si likelihood:

$$\begin{aligned}\mathbb{P}(\Theta = 1 | X = x) &\propto f_1(x) \cdot P(\hat{\theta}_x = 1) = f_1(x)p \\ \mathbb{P}(\Theta = 0 | X = x) &\propto f_0(x) \cdot P(\hat{\theta}_x = 0) = f_0(x)(1 - p)\end{aligned}$$

De observat ca ambele probabilitati posteriorare au un factor de normalizare comun $1 / [f_1(x)p + f_0(x)(1 - p)]$, deci este suficient sa se considere doar numaratorul. Deci estimatorul Bayes va prezice 1 daca

$$\frac{f_1(x)p}{f_0(x)(1 - p)} > 1$$

adica regula de clasificare "email bun" este:

$$\frac{f_1(x)}{f_0(x)} > \frac{1 - p}{p}$$

Estimatorul celor mai mici patrate (LMS)

Fara date observabile

Fie o v.a Θ si probabilitatea ei $p_\Theta(\theta)$. Cautam estimarea punctului $\hat{\theta}$.

► Introducem un *criteriu de performanta*, i.e media erorii la patrat (MSE):

$$E[(\Theta - \hat{\theta})^2].$$

► Dorim ca eroarea sa fie cat mai mica.

→ Abordarea pe baza probabilitatilor - folosind proprietatile variantei avem:

$$E[(\Theta - \hat{\theta})^2] = Var(\Theta - \hat{\theta}) + (E[\Theta - \hat{\theta}])^2$$

$\hat{\theta}$ este un numar, o constanta $\Rightarrow Var(\Theta - \hat{\theta}) = Var(\Theta)$.

Eroarea este minima cand $E[\Theta - \hat{\theta}] = 0 \Rightarrow \hat{\theta} = E[\Theta]$

Deci MSE optim este: $E[(\Theta - \hat{\theta})^2] = Var(\Theta)$

Definitie

Estimatorul celor mai mici patrate

$$\hat{\theta} = \arg \min_{\theta \in \Theta} E[(\Theta - \theta)^2]$$

→ Abordarea pe baza optimizarii - pentru a gasi solutia (explicita) avem:

$$E[(\Theta - \theta)^2] = E[\Theta^2] - 2E[\Theta]\theta + \theta^2$$

$$\frac{\partial E[(\Theta - \theta)^2]}{\partial \theta} = 0 \Leftrightarrow -2E[\Theta] + 2\theta = 0 \Rightarrow \hat{\theta} = E[\Theta]$$

Estimatorul celor mai mici patrate (LMS)

Cu date observabile

- ▶ Fie o v.a Θ si probabilitatea ei $p_\Theta(\theta)$.
- ▶ Fie X date observabile si modelul $p_{X|\Theta}(x|\theta)$.
- ▶ Pentru $X = x$ cautam estimarea punctului $\hat{\theta}$.

In acest caz, estimatorul LMS este conditionat:

$$\hat{\theta} = \arg \min_{\theta \in \Theta} E[(\Theta - \theta)^2 | X = x]$$

- ▶ Valoarea estimata cu LMS: $\hat{\theta} = E[\Theta | X = x]$
- ▶ Estimatorul LMS: $\hat{\Theta} = E[\Theta | X]$

$$E[(\Theta - \hat{\theta})^2 | X = x] = E[(\Theta - E[\Theta | X = x])^2 | X = x] \leq E[(\Theta - g(x))^2 | X = x] \quad \forall g(x)$$

Rescriem relatia de mai sus in termeni de v.a:

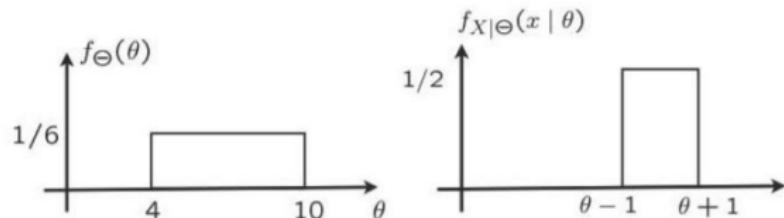
$$\begin{aligned} E[(\Theta - E[\Theta | X])^2 | X] &\leq E[(\Theta - g(X))^2 | X] \stackrel{\text{luand } E}{\Rightarrow} \\ E[(\Theta - E[\Theta | X])^2] &\leq E[(\Theta - g(X))^2] \quad \forall g(X) \end{aligned}$$

Intr-un curs anterior (vezi cursul 3 despre media conditionata) am aratat ca $\hat{\Theta}_{LMS} = E[\Theta | X]$ minimizeaza $E[(\Theta - g(X))^2]$ pentru $\forall \hat{\Theta} = g(X)$

Estimatorul celor mai mici patrate (LMS)

Exemplu 1

- ▶ Fie Θ v.a si pdf $f_\Theta(\theta)$
- ▶ Fie X obs. si modelul $f_{X|\Theta}(x|\theta)$
- ▶ Gasiti $\hat{\Theta}_{LMS}$



Observatii:

- ▶ Θ este uniform distribuita pe intervalul $[4, 10]$.
- ▶ Modelul observatiei poate fi privit ca valoarea lui θ perturbata de un zgomot z ce are o distributie uniforma pe intervalul $z \in [-1, 1]$: $x = \theta + z$
- ▶ z este independent de θ

Dat fiind ca $\hat{\Theta}_{LMS} = E[\Theta|X] \Rightarrow$ trebuie sa determinam $f_{\Theta|X}(\theta|x)$.

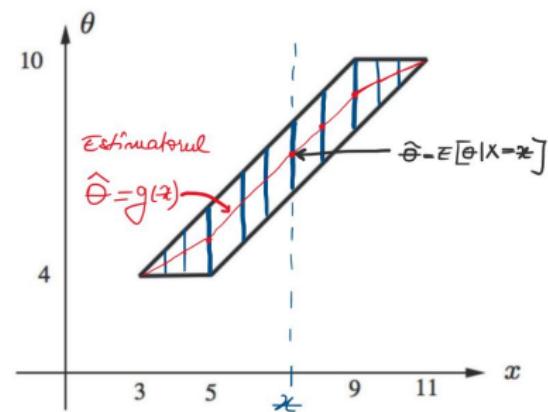
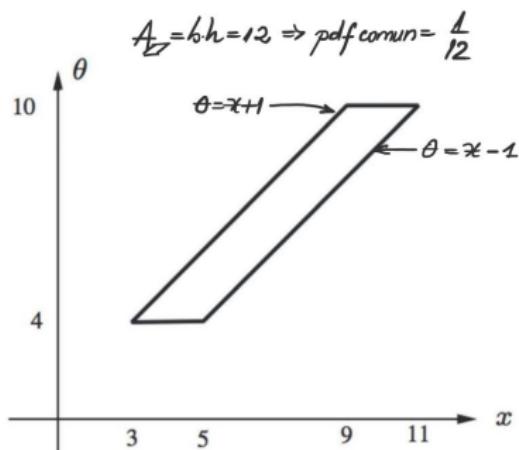
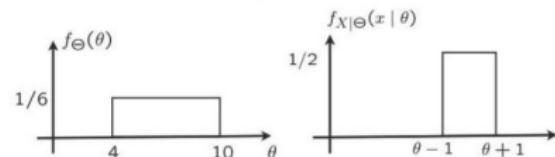
Conform regulii lui Bayes cautam distributia comună, i.e $f_\Theta(\theta)f_{X|\Theta}(x|\theta)$.

Conform graficului $f_{X|\Theta}(x|\theta) \Rightarrow \theta - 1 \leq x \leq \theta + 1 \Rightarrow \begin{cases} \theta \leq x + 1 \\ \theta \geq x - 1 \end{cases}$

Estimatorul celor mai mici patrate (LMS)

Exemplu 1 cont.

- ▶ $\begin{cases} \theta \leq x + 1 \\ \theta \geq x - 1 \end{cases}$
- ▶ in afara paralelogramului pdf-ul comun este 0
- ▶ Observam ca pdf-ul comun este produs de pdf-uri uniforme si constante
- ▶ Pdf-ul comun va fi uniform si constant mai exact $\frac{1}{12}$



Dand X estimatorul LMS este uniform $\hat{\Theta}_{LMS} = E[\Theta|X]$ pe intervalul $[X - 1, X + 1]$ si deci $\hat{\theta}_{LMS} = E[\Theta|x] = x/2$ (va amintesc ca pentru o variabila distribuita uniform pe intervalul $[a, b]$ media este $(a + b)/2$)

Estimatorul celor mai mici patrate (LMS)

Exemplu 2

Fie X_1, X_2, \dots, X_n v.a independente cu o distributie Bernoulli avand pmf-ul:

$$f(x | \theta) = \theta^x (1 - \theta)^{1-x}, \quad x = 0, 1$$

unde θ este necunoscut. Presupunem ca θ este ales dintr-o distributie uniforma pe $(0, 1)$. Gasiti estimatorul Bayes a lui θ (coincide cu estimatorul LMS cand functia loss se alege cea patratica: $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$)

Solutie: Trebuie sa calculam $E[L(\theta, \hat{\theta}) | X_1, \dots, X_n] = E[\theta | X_1, \dots, X_n]$. Stim ca $p(\theta) = 1$, $0 \leq \theta \leq 1$. Deci trebuie sa calculam probabilitatea posterioara:

$$\begin{aligned} f(\theta | x_1, \dots, x_n) &= \frac{f(x_1, \dots, x_n, \theta)}{f(x_1, \dots, x_n)} \\ &= \frac{f(x_1, \dots, x_n | \theta) p(\theta)}{\int_0^1 f(x_1, \dots, x_n | \theta) p(\theta) d\theta} \\ &= \frac{\theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}}{\int_0^1 \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i} d\theta} \end{aligned} \tag{1}$$

Se poate arata ca:

$$\int_0^1 \theta^m (1 - \theta)^r d\theta = \frac{m! r!}{(m + r - 1)!} \tag{2}$$

Estimatorul celor mai mici patrate

Exemplu 2 cont.

Notam $\bar{x} = \sum_{i=1}^n x_i$ si utilizand (2) in (1) obtinem:

$$f(\theta | x_1, \dots, x_n) = \frac{(n+1)! \theta^{\bar{x}} (1-\theta)^{n-\bar{x}}}{\bar{x}!(n-\bar{x})!}$$

Astfel, avem

$$\begin{aligned} E[\theta | x_1, \dots, x_n] &= \frac{(n+1)!}{\bar{x}!(n-\bar{x})!} \int_0^1 \theta^{1+\bar{x}} (1-\theta)^{n-\bar{x}} d\theta \\ &\stackrel{(2)}{=} \frac{(n+1)!}{\bar{x}!(n-\bar{x})!} \frac{(1+\bar{x})!(n-\bar{x})!}{(n+2)!} \\ &= \frac{\bar{x}+1}{n+2} \end{aligned}$$

Deci estimatorul Bayes pentru functia loss $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ este:

$$E[\theta | X_1, \dots, X_n] = \frac{\sum_{i=1}^n X_i + 1}{n+2}$$

De exemplu, daca am avea $n = 10$ incercari, din care 6 au succes, atunci utilizand rezultatul de mai sus, gasim estimarea $\theta = \frac{7}{12}$. Observa ca estimatorul Bayes coincide cu estimatorul LMS cand functia loss se alege cea patratica!!!

Estimatroul celor mai mici patrate

Exemplu 3

Fie X_1, X_2, \dots, X_n v.a independente cu o distributie Normala, fiecare avand media θ necunoscuta si varianta cunoscuta σ_0^2 . Daca presupunem ca θ este ales si el dintr-o distributie normala cu μ si σ^2 cunoscute. Calculati estimatorul Bayes pentru θ cand functia loss este $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$?

Solutie: Cautam estimatorul Bayes pentru care alegem functia loss $L = (\theta - \hat{\theta})^2$, i.e vrem sa calculam $E[L(\theta, \hat{\theta})|X_1, \dots, X_n] = E[\theta|X_1, \dots, X_n]$. Deci trebuie mai intai sa calculam probabilitatea lui θ conditionata de X_1, \dots, X_n :

$$f(\theta | x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n | \theta) p(\theta)}{f(x_1, \dots, x_n)},$$

unde

$$f(x_1, \dots, x_n | \theta) = \frac{1}{(2\pi)^{n/2} \sigma_0^n} \exp \left\{ - \sum_{i=1}^n (x_i - \theta)^2 / 2\sigma_0^2 \right\}$$

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -(\theta - \mu)^2 / 2\sigma^2 \right\}$$

$$f(x_1, \dots, x_n) = \int_{-\infty}^{\infty} f(x_1, \dots, x_n | \theta) p(\theta) d\theta$$

Estimatorul celor mai mici patrate

Exemplu 3 cont.

Calculand acum media conditionata, obtinem estimatorul:

$$E[\theta | X_1, \dots, X_n] = \frac{n\sigma^2}{n\sigma^2 + n\sigma_0^2} \bar{X} + \frac{\sigma_0^2}{n\sigma^2 + \sigma_0^2} \mu = \frac{\frac{n}{\sigma_0^2}}{\frac{n}{\sigma_0^2} + \frac{1}{\sigma^2}} \bar{X} + \frac{\frac{1}{\sigma^2}}{\frac{n}{\sigma_0^2} + \frac{1}{\sigma^2}} \mu,$$

unde $\bar{X} = \sum_i^n X_i$. Deci initial consideram ca o estimare a parametrului θ este μ , dar apoi din observatii (date) ajungem la concluzia ca o mai buna estimare a lui θ ar fi o combinatie convexa intre $\bar{x} = \sum_i^n x_i$ si μ . Cu cat observam mai multe date (n mare) cu atat ne incredem ca o buna estimare a lui θ este \bar{x} .

Varianta este:

$$\text{Var}(\theta | X_1, \dots, X_n) = \frac{\sigma_0^2 \sigma^2}{n\sigma^2 + \sigma_0^2}$$

Remarca: In multe aplicatii statistice, varianta distributiei normale este mai stabila decat media, astfel incat presupunerea ca varianta este cunoscuta nu este in intregime artificiala.

Estimatorul celor mai mici patrate

Exemplu 4: Filtru Kalman

- ▶ Filtru Kalman - numit si estimatorul liniar patratice - este o metoda care foloseste masuratori (date) pentru a estima niste variabile necunoscute.
- ▶ Filtru Kalman - minimizeaza eroarea patratica medie
- ▶ Filtru Kalman se foloseste in multe aplicatii: automatica (guidance, navigation, and control of vehicles; robotic motion planning); economie, procesare de semnal (analyza seriilor de timp), etc
- ▶ Filtru Kalman a fost folosit pentru estimarea traectoriei rachetei in programul Apollo
- ▶ Filtru Kalman foloseste un model dinamic al sistemului, un controller (inputs) dat; si masuratori ale iesirii obtinute de la sensori pentru a estima intreg vectorul de stari a sistemului.
- ▶ Este o metoda iterativa care foloseste doar masuratorile curente (un singur pas din trecut) pentru a calcula noua estimare a starii.
- ▶ Filtru Kalman considera atat "noisy measurements" cat si "process noise"
- ▶ Produce o estimare a starii curente ca o combinatie liniara dintre stare prezisa de sistem si noile valori masurate.

Estimatorul celor mai mici patrate

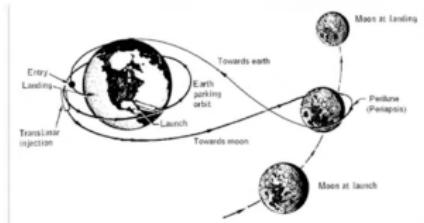
Exemplu 4: Filtru Kalman



The Kalman Filter



Apollo Guidance Computer



The (extended) Kalman Filter became widely known after its use in the Apollo Guidance Computer for circumlunar navigation.

Estimatorul celor mai mici patrate

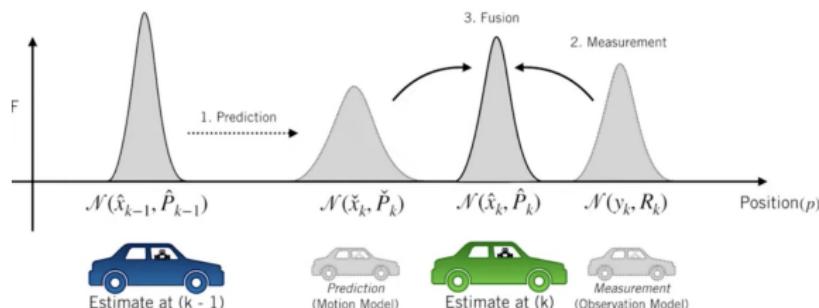
Exemplu 4: Filtru Kalman

- ▶ Filtru Kalman - estimeaza starea interna a unui proces dinamic (time-varying) dat de ecuatiiile:

$$x_k = F_k x_{k-1} + B_k u_k + w_k, \quad y_k = H_k x_k + v_k$$

- ▶ unde x_k este starea (state), u_k este intrarea (input) si y_k este iesirea (output), w_k process noise, v_k measurement noise
- ▶ starea initiala si zgomotele $x_0, w_1 \dots w_k, v_1 \dots v_k$ sunt presupuse independente; w_k, v_k white noise: $w_k \approx \mathcal{N}(0, Q_k)$ si $v_k \approx \mathcal{N}(0, R_k)$, i.e. distribuite Gaussian cu medie 0 si matrici de covarianta Q_k si R_k cunoscute!
- ▶ notam cu $\hat{x}_{k|k}$ valoarea medie a starii aposteriori estimata
- ▶ $P_{k|k}$ matricea de covarianta aposteriori estimata

The Kalman Filter | Prediction and Correction



Estimatorul celor mai mici patrate

Exemplu 4: Filtru Kalman

Filtru Kalman - are două faze:

Predict:

- ▶ predict a priori state estimate using the model:

$$\hat{x}_{k|k-1} = F_k \hat{x}_{k-1|k-1} + B_k u_k$$

- ▶ predict a priori covariance estimate (obs.: w_k independent de $x_k, \hat{x}_{k|k-1}$):

$$P_{k|k-1} = \text{cov}(x_k - \hat{x}_{k|k-1}) = F_k P_{k-1|k-1} F_k^T + Q_k$$

Update:

- ▶ measurement prefit residual: $e_k = y_k - H_k \hat{x}_{k|k-1}$
- ▶ covariance prefit residual: $E_k = \text{cov}(e_k) = H_k P_{k|k-1} H_k^T + R_k$
- ▶ Optimal Kalman gain: $K_k = P_{k|k-1} H_k^T E_k^{-1}$
- ▶ Updated (a posteriori) state estimate:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k e_k \rightarrow \text{liniar interpolation} := (I - K_k H_k) \hat{x}_{k|k-1} + K_k y_k$$

i.e., estimam starea curentă ca o combinare liniară dintre stare prezisă de sistem ($\hat{x}_{k|k-1}$) și noile valori măsurate (y_k)!

- ▶ Updated (a posteriori) covariance estimate

$$P_{k|k} = P_{k|k-1} - K_k (H_k P_{k|k-1})$$

Estimatorul celor mai mici patrate

Exemplu 4: Filtru Kalman

- daca avem un model bun al sistemului dinamic si daca $\hat{x}_{0|0}, P_{0|0}$ reflecta distributia starii initiale (media si covariantă), atunci avem invariantii:

$$\mathbb{E}[x_k - \hat{x}_{k|k}] = 0, \quad \mathbb{E}[e_k] = 0 \quad \forall k$$

- De asemenea, avem invariantii:

$$P_{k|k} = \text{cov}(x_k - \hat{x}_{k|k}), \quad E_k = \text{cov}(e_k)$$

- matricile $P_{k|k-1} = F_k P_{k-1|k-1} F_k^T + Q_k$, $E_k = H_k P_{k|k-1} H_k^T + R_k$, $K_k = P_{k|k-1} H_k^T E_k^{-1}$ si $P_{k|k} = P_{k|k-1} - K_k (H_k P_{k|k-1})$ nu depind de masuratori (i.e. y_k) - pot fi calculate offline.
- daca matrix gain K_k converge la un K_∞ si daca matricile sistemului sunt independente de timp ($F_k = F$, $B_k = B$, $H_k = H$) atunci Filtru Kalman devine un filtru liniar invariant in timp:

$$\hat{x}_k = F\hat{x}_{k-1} + Bu_k + K_\infty(y_k - H(F\hat{x}_{k-1} + Bu_k))$$

unde $K_\infty = P_\infty H^T (R + HP_\infty H^T)^{-1}$ si P_∞ este solutia ecuatiei Riccati:

$$P_\infty = F(P_\infty + P_\infty H^T (R + HP_\infty H^T)^{-1} H P_\infty) F^T + Q$$

Estimatorul celor mai mici patrate

Exemplu 4: Filtru Kalman

Derivarea filtrului Kalman:

- matricea de covarianta aposteriori

$$\begin{aligned} P_{k|k} &= \text{cov}(x_k - \hat{x}_{k|k}) \stackrel{\text{definitia lui } \hat{x}}{=} \text{cov}(x_k - (\hat{x}_{k|k-1} + K_k e_k)) \\ &\stackrel{\text{definitia lui } e}{=} \text{cov}(x_k - [\hat{x}_{k|k-1} + K_k(y_k - H_k \hat{x}_{k|k-1})]) \\ &\stackrel{\text{definitia lui } y}{=} \text{cov}(x_k - [\hat{x}_{k|k-1} + K_k(H_k x_k + v_k - H_k \hat{x}_{k|k-1})]) \\ &= \text{cov}((I - K_k H_k)(x_k - \hat{x}_{k|k-1}) - K_k v_k) \\ &\stackrel{v \text{ independent de } x}{=} \text{cov}((I - K_k H_k)(x_k - \hat{x}_{k|k-1})) + \text{cov}(K_k v_k) \\ &\stackrel{\text{proprietati cov}}{=} (I - K_k H_k) \text{cov}(x_k - \hat{x}_{k|k-1})(I - K_k H_k)^T + K_k \text{cov}(v_k) K_k^T \\ &= (I - K_k H_k) P_{k|k-1} (I - K_k H_k)^T + K_k R_k K_k^T \\ &\stackrel{\text{desfacem parantezele}}{=} P_{k|k-1} - K_k H_k P_{k|k-1} - P_{k|k-1} H_k^T K_k^T + K_k E_k K_k^T \end{aligned}$$

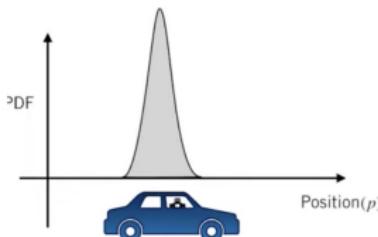
- Filtru Kalman minimizeaza eroarea aposteriori a starii estimate, i.e. $\min \mathbb{E}[\|x_k - \hat{x}_{k|k}\|^2]$ care este echivalenta (obs. $\|x\|^2 = x^T x = \text{trace}(xx^T)$) cu minimizarea trace-ului $P_{k|k}$. Derivand asadar expresia lui $P_{k|k}$ w.r.t. K_k :

$$\partial \text{trace}(P_{k|k}) / \partial K_k = -2(H_k P_{k|k-1})^T + 2K_k E_k = 0 \rightarrow K_k = P_{k|k-1} H_k^T E_k^{-1}$$

Estimatorul celor mai mici patrate

Filtru Kalman - aplicatie

The Kalman Filter | Short Example



Motion/Process Model

$$\mathbf{x}_k = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \mathbf{x}_{k-1} + \begin{bmatrix} 0 \\ \Delta t \end{bmatrix} \mathbf{u}_{k-1} + \mathbf{w}_{k-1}$$

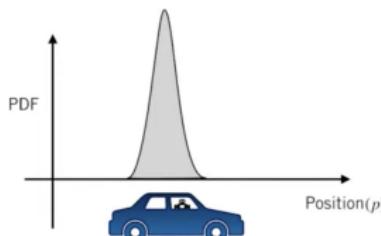
Position Observation

$$y_k = [1 \ 0] \mathbf{x}_k + v_k$$

$$\mathbf{x} = \begin{bmatrix} p \\ \frac{dp}{dt} = \dot{p} \end{bmatrix} \quad \mathbf{u} = a = \frac{d^2 p}{dt^2}$$

Noise Densities

$$v_k \sim \mathcal{N}(0, 0.05) \quad \mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, (0.1)\mathbf{I}_{2 \times 2})$$



Data

$$\hat{\mathbf{x}}_0 \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 5 \end{bmatrix}, \begin{bmatrix} 0.01 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

$$\Delta t = 0.5\text{s}$$

$$u_0 = -2 \text{ [m/s}^2\text{]} \quad y_1 = 2.2 \text{ [m]}$$

Estimatorul celor mai mici patrate

Filtru Kalman - aplicatie

Prediction

$$\begin{aligned}\check{\mathbf{x}}_k &= \mathbf{F}_{k-1} \mathbf{x}_{k-1} + \mathbf{G}_{k-1} \mathbf{u}_{k-1} \\ \begin{bmatrix} \check{p}_1 \\ \check{p}_1 \end{bmatrix} &= \begin{bmatrix} 1 & 0.5 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 5 \end{bmatrix} + \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} (-2) = \begin{bmatrix} 2.5 \\ 4 \end{bmatrix}\end{aligned}$$

$$\check{\mathbf{P}}_k = \mathbf{F}_{k-1} \hat{\mathbf{P}}_{k-1} \mathbf{F}_{k-1}^T + \mathbf{Q}_{k-1}$$

$$\check{\mathbf{P}}_1 = \begin{bmatrix} 1 & 0.5 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.01 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0.5 \\ 0 & 1 \end{bmatrix}^T + \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} = \begin{bmatrix} 0.36 & 0.5 \\ 0.5 & 1.1 \end{bmatrix}$$

Correction

$$\begin{aligned}\mathbf{K}_1 &= \check{\mathbf{P}}_1 \mathbf{H}_1^T (\mathbf{H}_1 \check{\mathbf{P}}_1 \mathbf{H}_1^T + \mathbf{R}_1)^{-1} \\ &= \begin{bmatrix} 0.36 & 0.5 \\ 0.5 & 1.1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \left([1 \ 0] \begin{bmatrix} 0.36 & 0.5 \\ 0.5 & 1.1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 0.05 \right)^{-1} \\ &= \begin{bmatrix} 0.88 \\ 1.22 \end{bmatrix}\end{aligned}$$

$$\hat{\mathbf{x}}_1 = \check{\mathbf{x}}_1 + \mathbf{K}_1 (\mathbf{y}_1 - \mathbf{H}_1 \check{\mathbf{x}}_1)$$

$$\begin{bmatrix} \hat{p}_1 \\ \hat{p}_1 \end{bmatrix} = \begin{bmatrix} 2.5 \\ 4 \end{bmatrix} + \begin{bmatrix} 0.88 \\ 1.22 \end{bmatrix} (2.2 - [1 \ 0] \begin{bmatrix} 2.5 \\ 4 \end{bmatrix}) = \begin{bmatrix} 2.24 \\ 3.63 \end{bmatrix}$$

$$\begin{aligned}\text{Bonus!} \\ \hat{\mathbf{P}}_1 &= (\mathbf{I} - \mathbf{K}_1 \mathbf{H}_1) \check{\mathbf{P}}_1 \\ &= \begin{bmatrix} 0.04 & 0.06 \\ 0.06 & 0.49 \end{bmatrix}\end{aligned}$$

Obs.: covarianta obtinuta pe baza de model este mult mai mare decat cea obtinuta dupa masuratori !

Estimatorul celor mai mici patrate (LMS)

Proprietati

Definim estimatorul LMS: $\hat{\Theta} = E[\Theta|X]$ si eroarea acestuia $\tilde{\Theta} = \hat{\Theta} - \Theta$

Din proprietatile expectantei obtinem ca

$$E[\hat{\Theta}] = E[E[\Theta|X]] = E[\Theta] \Rightarrow E[\tilde{\Theta}] = 0 \quad (3)$$

1) Media erorii este zero si in prezenta unor masuratori: $E[\tilde{\Theta}|X = x] = 0$

Dem: $E[\tilde{\Theta}|X = x] = E[\hat{\Theta} - \Theta|X = x] = \hat{\Theta} - E[\Theta|X = x] = 0$

Estimatorul este impartial (unbiased).

2) Covarianta: $Cov(\tilde{\Theta}, \hat{\Theta}) = 0$

Dem:

$$Cov(\tilde{\Theta}, \hat{\Theta}) = E[\tilde{\Theta}\hat{\Theta}] - E[\tilde{\Theta}]E[\hat{\Theta}] \stackrel{(3)}{\Leftrightarrow} Cov(\tilde{\Theta}, \hat{\Theta}) = E[\tilde{\Theta}\hat{\Theta}]$$

Termenul obtinut pare dificil de calculat. Ne uitam la versiunea conditionata:

$$E[\tilde{\Theta}\hat{\Theta}|X = x] = \hat{\Theta}E[\tilde{\Theta}|X = x] = 0$$

Luand expectanta in ultima relatie obtinem: $E[\tilde{\Theta}\hat{\Theta}] = 0 \Rightarrow Cov(\tilde{\Theta}, \hat{\Theta}) = 0$

3) $Var(\Theta) = Var(\tilde{\Theta}) + Var(\hat{\Theta})$

Dem:

$$Var(\Theta) = Var(\tilde{\Theta} + \hat{\Theta}) = Var(\tilde{\Theta}) + Var(\hat{\Theta}) + 2Cov(\tilde{\Theta}, \hat{\Theta}) = Var(\tilde{\Theta}) + Var(\hat{\Theta})$$

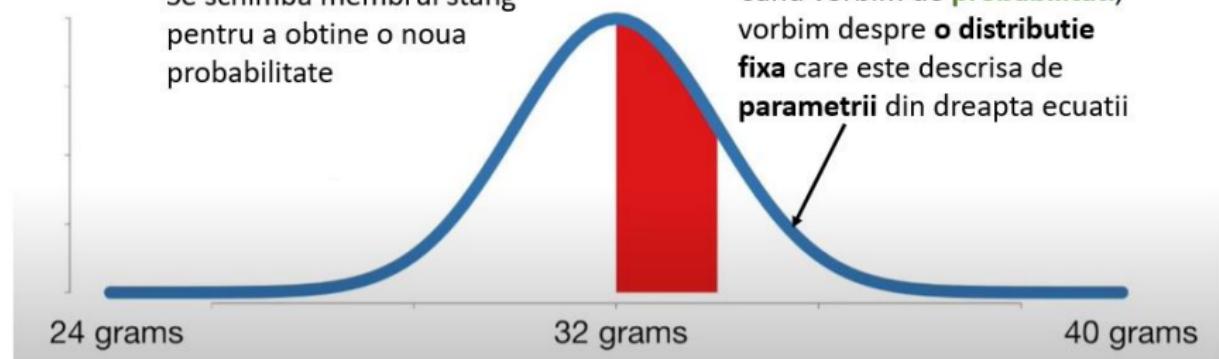
Probabilitate versus verosimilitate (engl. Likelihood) -Exemplu

Amintim: O distributie se poate caracteriza prin medie si deviatie standard.
Exemplu: Se masoara greutatea unor soareci de laborator. Cel mai slab soarece are 24 de grame si in cel mai greu are 40 de grame. Consideram ca aceste masuratori au o distributie uniforma.

$$P(32 \leq \text{greutate} \leq 34 | \mu = 32 \text{ si } \sigma = 2.5) = 0.29$$

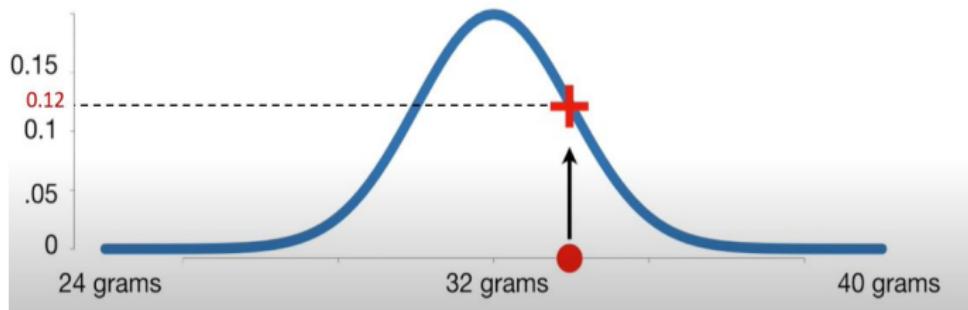
Se schimba membrul stang pentru a obtine o noua probabilitate

Cand vorbim de **probabilitati**, vorbim despre **o distributie fixa** care este descrisa de **parametrii** din dreapta ecuatiei



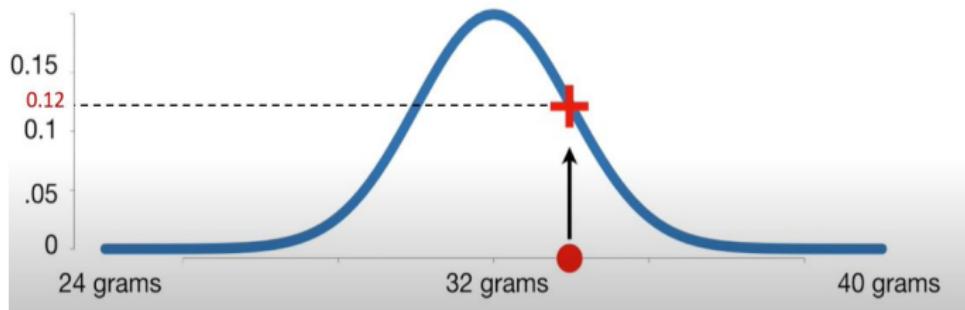
Probabilitate versus verosimilitate (engl. Likelihood) -Exemplu

$$L(\mu = 32 \text{ si } \sigma = 2.5 | \text{greutate} = 34) = 0.12$$



Probabilitate versus verosimilitate (engl. Likelihood) -Exemplu

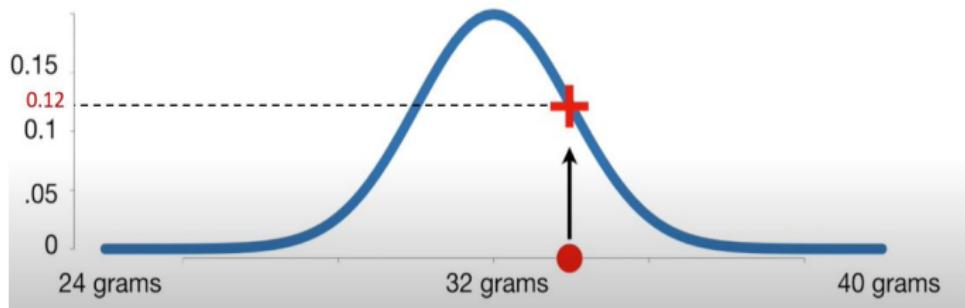
$$L(\mu = 32 \text{ si } \sigma = 2.5 | \text{greutate} = 34) = 0.12$$



... daca deplasam distributia astfel incat $\mu = 34$ atunci.

Probabilitate versus verosimilitate (engl. Likelihood) -Exemplu

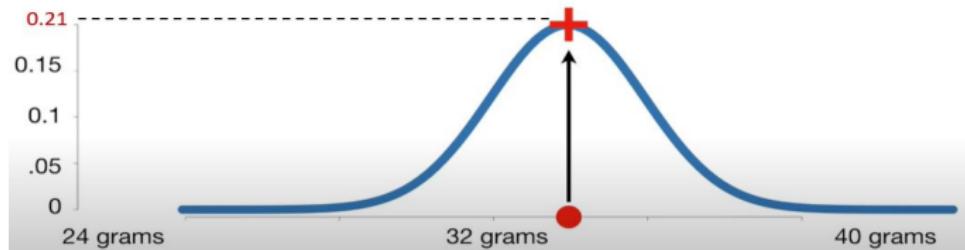
$$L(\mu = 32 \text{ si } \sigma = 2.5 | \text{greutate} = 34) = 0.12$$



... daca deplasam distributia astfel incat $\mu = 34$ atunci.

In likelihood, masuratorile din partea dreapta sunt fixe

$$L(\mu = 34 \text{ si } \sigma = 2.5 | \text{greutate} = 34) = 0.21$$



Functia de verosimilitate

Fie setul de esantioane pentru niste variabile aleatoare $Y = \{y_1, \dots, y_n\}$ alese in concordanta cu o functie de probabilitate f din familia de probabilitati \mathbb{P}_θ .

Definitie

Functia de verosimilitate likelihood este functia de densitate privita ca o functie in θ

$$L(\theta | y) = f(y | \theta) \quad \theta \in \Theta$$

- ▶ Valoarea exacta a oricarui likelihood este neimportanta
- ▶ Valoarea relativa, numita si raportul likelihood, este mai informativa

$$\frac{L(\theta_0 | y)}{L(\theta_1 | y)}$$

Pe un esantion de dimensiune n , functia likelihood se defineste:

$$L(\theta | y) = f(Y | \theta) = \prod_{i=1}^n f(y_i | \theta)$$

daca y_1, \dots, y_n sunt independente...

Estimatorul Maximum Likelihood (MLE)

Definitie

Estimatorul maximum likelihood se defineste ca

$$\hat{\theta}(y) = \arg \max_{\theta \in \Theta} L(\theta | y)$$

Maximizarea lui $L(\theta|y)$ poate fi dificila \rightarrow se alege :

$$\ell(\theta) = \ln L(\theta|y) = \sum_i^n \ln f(y_i|\theta)$$

In general este mai usor de maximizat $\sum_i \ln(f_i)$ decat produsul $\prod_i f_i$

Ne amintim de la optimizare ca pentru a gasi estimatorul in cazul $\Theta = \mathbb{R}^m$, trebuie sa rezolvam ecuatia (conditia de optimalitate de ordinul intai in cazul neconstrans):

$$\nabla \ell(\theta) = 0 \tag{4}$$

Pentru a garanta ca $\hat{\theta}$ este punct de maxim trebuie sa verificam daca hesiana este negativ definita:

$$\nabla^2 \ell(\hat{\theta}) < 0.$$

In general acest sistem de ecuatii nu are solutie explicita si pentru a gasi $\hat{\theta}$ se aplica algoritmi de optimizare iterativi (ex. gradient sau Newton):

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \alpha_k \nabla \ell(\hat{\theta}_k) \quad (\text{maximizam!})$$

Estimatorul Maximum Likelihood (MLE)

Definitie

Estimatorul maximum likelihood se defineste ca

$$\hat{\theta}(y) = \arg \max_{\theta \in \Theta} L(\theta | y)$$

Maximizarea lui $L(\theta|y)$ poate fi dificila \rightarrow se alege :

$$\ell(\theta) = \ln L(\theta|y) = \sum_i^n \ln f(y_i|\theta)$$

In multe cazuri insa avem constrangeri pe parametrii θ , ex.:

$$\theta \in \Theta = \{\theta : A\theta \leq b\}$$

Atunci conditiile de optimalitate (conditiile KKT) se schimba (vezi cursul de optimizare 5):

$$\nabla \ell(\theta) + A^T \lambda = 0, \quad A\theta \leq b, \quad \lambda \geq 0.$$

Si in acest caz, sistemul de ecuatii nu are solutie explicita si pentru a gasi $\hat{\theta}$ se aplica algoritmi de optimizare iterativi (ex. metode de punct interior).

Estimatorul Maximum Likelihood (MLE)

Definitie

Estimatorul maximum likelihood se defineste ca

$$\hat{\theta}(y) = \arg \max_{\theta \in \Theta} L(\theta | y)$$

Maximizarea lui $L(\theta|y)$ poate fi dificila \rightarrow se alege :

$$\ell(\theta) = \ln L(\theta|y) = \sum_i^n \ln f(y_i|\theta)$$

Pentru a gasi estimatorul trebuie sa rezolvam ecuatia:

$$u(\theta) = \frac{d}{d\theta} \ell(\theta) = 0 \quad (5)$$

$u(\theta)$ - **functia scor** a lui Fisher.

Scorul este un vector random care are urmatoarele proprietati:

- ▶ Pentru θ adevarat: $E[u(\theta)] = 0$
- ▶ Matricea informatie este: $I(\theta) = E [u(\theta)u^T(\theta)] = \text{var}[u(\theta)]$

În condiții de regularitate, matricea informatie se poate defini si ca:

$$I_O(\theta|y) = -\frac{d^2}{d\theta^2} L(\theta|y).$$

Media informatiei este: $I(\theta) = E [I_O(\theta|y)]$

Cu aceasta cantitate putem afla varianta estimatorului: $\text{var}(\hat{\theta}) \approx I(\theta)^{-1}$

Estimatorul Maximum Likelihood (MLE)

Exemplu 1

Rata de succes la inseminarea artificiala.

Consideram pentru acest scenariu functia likelihood (distributie geometrica):

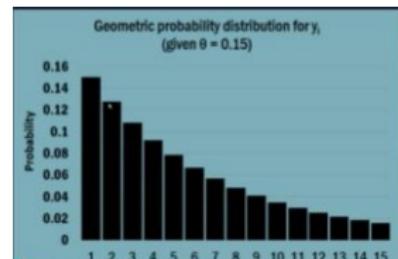
$$L(\theta|y) = \theta^n(1 - \theta)^{y-n}$$

- ▶ n - nr de cupluri
- ▶ y - nr total de incercari pentru toate cuplurile.
- ▶ θ - probabilitatea de succes pt fiecare cuplu in parte.

Gasiti MLE pentru $n = 20$ si $y = 100$?

$$\ell(\theta) = \ln L(\theta|y) = n \ln \theta + (y - n) \ln(1 - \theta)$$

$$\frac{d}{d\theta} \ell(\theta) = \frac{n}{\theta} - \frac{y - n}{1 - \theta} = 0 \quad \rightarrow \quad \hat{\theta} = \frac{n}{y} = \frac{20}{100} = 0.2$$



Estimatorul Maximum Likelihood

Exemplu 2

Fie X_1, \dots, X_n v.a. cu distributie $N(\mu, \sigma)$ si $f_X(x|\mu, \sigma) =$

$$\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}. \text{ Gasiti } \mu \text{ si } \sigma \text{ folosind MLE?}$$

Problema de optimizare MLE este (aici $\theta = (\mu, \sigma)$):

$$\text{minimize } -\ln f_X(x|\mu, \sigma) = \frac{n}{2} \ln \sigma + \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} = \ell(\mu, \sigma)$$

Derivarea dupa μ :

$$\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = 0 \Leftrightarrow \frac{1}{\sigma} \sum_i^n (x_i - \mu) = 0 \Leftrightarrow \sum_i^n x_i = n\mu \Rightarrow \hat{\mu} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Derivarea dupa σ :

$$\frac{\partial \ell(\mu, \sigma)}{\partial \sigma} = 0 \Leftrightarrow \frac{n}{2\sigma} - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} = 0 \Leftrightarrow n - \frac{1}{\sigma} \sum_{i=1}^n (x_i - \mu)^2 = 0$$

$$\Rightarrow \hat{\sigma} = \sqrt{\frac{1}{n} \sum_i^n (x_i - \hat{\mu})^2} = \sqrt{\frac{1}{n} \sum_i^n x_i^2 - \frac{1}{n^2} \sum_i^n \sum_j^n x_i x_j}$$

Estimatorul Maximum Likelihood

Exemplu 3

In multe cazuri variabilele $y_1 \dots y_n$ sunt dependente. Fie $Y_1 \dots Y_m$ cutii in numar de m si consideram un numar n de bile care cad in aceste cutii:

$$y_1 + \dots + y_m = n \quad (y_i = \text{nr. bile din cutia } i)$$

Probabilitatea sa cada intr-o cutie este p_i , deci avem constrangerea:

$$p_1 + \dots + p_m = 1$$

In aceste caz Y_i -urile nu sunt independente si probabilitatea comună a vectorului $(y_1 \dots y_m)$ este multinomială:

$$f(y|p) = \binom{n}{y_1, y_2, \dots, y_m} p_1^{y_1} p_2^{y_2} \cdots p_m^{y_m}$$

Log-likelihood-ul este:

$$\ell(p) = \log L(p|y) = \log f(y|p) = \log n! - \sum_i \log y_i! + \sum_i y_i \log p_i$$

Avem conditiile de optimalitate (KKT):

$$(i) \nabla \ell(p) - \lambda \nabla(p_1 + \dots + p_m - 1) = \nabla_p \ell(p) - \lambda e = 0, \quad (ii) p_1 + \dots + p_m = 1 \Rightarrow$$

din (i) obtinem $y_i/p_i - \lambda = 0, \forall i = 1 : m$, i.e. $p_i = y_i/\lambda$, combinat cu

$p_1 + \dots + p_m = 1$ conduce la solutia explicită $\lambda = \sum_j y_j = n$ si deci $p_i = y_i/n$.



Legatura cu estimatorul Bayes

Estimatorul MLE coincide cu estimatorul Bayes de tip MAP dand probabilitate uniforma a priori. Intr-adevar, estimatorul aposteriori este acel parametru θ care maximizeaza probabilitatea lui θ dand datele:

$$\max_{\theta} P(\theta|y_1 \cdots y_n) = \frac{f(y_1 \cdots y_n|\theta)P(\theta)}{P(y_1 \cdots y_n)}$$

Deoarece numitorul nu depinde de θ , il putem neglaja in maximizare. Cum la numarator, $P(\theta)$ este constanta (probabilitate a priori uniforma), atunci obtinem ca

$$\max_{\theta} P(\theta|y_1 \cdots y_n) = \max_{\theta} f(y_1 \cdots y_n|\theta) = \max_{\theta} \ell(\theta)$$

Estimatorul Maximum Likelihood - Exemplu 4

Regresia logistica - Clasificarea binara

Fie setul de date X_1, X_2, \dots, X_n si etichetele asociate y_1, \dots, y_n , unde $y_i \in \{0, 1\}, \forall i = 1 : n$.

Ca si in SVM, dorim sa calculam un hiperplan de separare a doua clase parametrizat in w si b :

$$y_i \leftarrow f(x_i) = w^T x_i + b.$$

Intrebare: Ce fel de valori returneaza functia $f(x_i)$ si ce fel de variabila este eticheta?

Observam ca eticheta este o variabila discrete binara, iar functia returneaza valori reale. Deci este necesar sa "ducem" (engl. *to map*) functia $f(x_i)$ cel putin in intervalul $[0, 1]$.

Introducem functia sigmoid:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Interpretare: Mapeaza o valoare de predictie (o valoare reala) intr-o probabilitate (o valoare intre 0 si 1), a se observa figura alaturata.

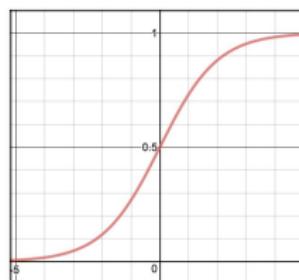


Figure: Functia sigmoid standard $\sigma(z)$ cu $z \in [-6, 6]$ si $\sigma(z) \in [0, 1]$.

Estimatorul Maximum Likelihood - Exemplu 4

Regresia logistica - Clasificarea binara

In final utilizam functia h , care este o compunere intre functia sigmoid si una liniara:

$$h(x_i) = \sigma(f(x_i)) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}_i - b}}$$

In continuare interpretam

$$h(X) = P(Y = 1|X, w, b),$$

Tinând cont de teoria fundamentală a probabilitatii (i.e. suma probabilitatilor este 1, $\sum P(Y|X, w) = 1$), avem ca:

$$P(Y = 0|X, w, b) = 1 - h(X).$$

Presupunem că toate observațiile (etichetele) sunt independente, Bernoulli distribuite. Amintim funcția de masă a probabilității f a distribuției Bernoulli, pentru posibilele rezultate y , este

$$f(y; p) = \begin{cases} p, & \text{dacă } y = 1, \\ 1 - p, & \text{dacă } y = 0. \end{cases}$$

Aceasta poate fi exprimată și ca

$$f(y; p) = p^y(1 - p)^{1-y} \quad \text{pentru } y \in \{0, 1\}$$

Estimatorul Maximum Likelihood - Exemplu 4

Regresia logistica - Clasificarea binara

Acum putem defini urmatoarea functie de verosimilitate:

$$p \approx h(X)$$

$$L(w, b|y, x) = P(Y|X, w, b) = \prod_{i=1}^N P(y_i|x_i, w, b) = \prod_{i=1}^N h(x_i)^{y_i} (1 - h(x_i))^{(1-y_i)}.$$

Maximizam si logaritmam functia de verosimilitate:

$$\max_{w,b} \log L(w, b|y, x) = \sum_{i=1}^N \left[y^{(i)} \log \left(h(x^{(i)}) \right) + (1 - y^{(i)}) \log \left(1 - h(x^{(i)}) \right) \right]$$

Aceasta functie se numeste *functia logistica binara* si problema se numeste problema de **Regresie Logistică**

In practica se maximizeaza media:

$$\max_{w,b} \frac{1}{N} \log L(w, b|y, x)$$

ce poate fi interpretat si ca *estimatorul Bayes cu probabilitatea apriori fixata* $\frac{1}{N}$.

Pentru problema de clasificare, asiguram caracterul categorial functiei $h(x)$ pentru data de test x prin selectarea unui prag de decizie (de exemplu 0.5):

$$h(x) = \begin{cases} p \geq 0.5, & y = 1 \\ p < 0.5, & y = 0. \end{cases}$$

Estimatorul Maximum Likelihood (MLE)

Tipuri de MLE:

- ▶ Explicit: cand se poate calcula explicit din conditiile de optimalitate
- ▶ Implicit: estimatorul este generat de algoritmi de optimizare (vezi cursul de optimizari unde am prezentat astfel de metode)

Avantaje:

- ▶ nu este neaparat impartial
- ▶ este asimptotic impartial
- ▶ este asimptotic consistent: $\hat{\theta}_n \rightarrow \theta$
- ▶ Distributia tinde asimptotic spre normala: $\frac{\hat{\theta}_n - \theta}{\sigma(\hat{\theta}_n)} \rightarrow N(0, 1)$
- ▶ Parametrizare invarianta, adica daca $\hat{\theta}(x)$ este MLE pentru θ , atunci $g(\hat{\theta})$ este MLE pentru $g(\theta)$.
- ▶ Foarte eficient pe seturi de date suficient de mari.

Dezavantaje:

- ▶ Nu performeaza bine pentru situatii particulare.
- ▶ Necesa cunoasterea unei distributii pentru date (i.e. $f(y|\theta)$)

Inferenta clasica

Estimatorul celor mai mici patrate: regresia liniara

Fitting de date: se dau urmatoarele

- ▶ date sau masuratori $(u_i, b_i) \quad \forall i = 1, \dots, m$ (de obicei $m \gg n$)
- ▶ dorim sa gasim un model care potriveste datele:

$$\mathcal{M}(u) = b \iff \mathcal{M}(u_i) = b_i \quad \forall i = 1 : m \quad \text{HARD PROBLEM!}$$

- ▶ SIMPLER PROBLEM: dand functii f_1, \dots, f_n numiti regresori sau functii de baza consideram $\mathcal{M} = \theta_1 f_1 + \dots + \theta_n f_n$
- Problema: gasiti coeficientii reali $\theta_1, \dots, \theta_n$ s.t.

$$b_i = \theta_1 f_1(u_i) + \dots + \theta_n f_n(u_i) \quad \forall i = 1, \dots, m$$

- Obtinem sistemul linear:

$$A\theta = b \quad \text{unde } A \in \mathbb{R}^{m \times n}, \quad A_{ij} = f_j(u_i)$$

- Solutie:

- ▶ A patratica si inversabila, atunci solutie unica
- ▶ A cu $n \gg m$ (subdeterminata) - o infinitate de solutii, atunci se cauta solutia de norma minima (CMMP)
- ▶ A cu $m \gg n$ (supradeterminata) - nici o solutie, atunci se cauta solutia care minimizeaza expresia patratica (CMMP)

$$\sum_{i=1}^m (\theta_1 f_1(u_i) + \dots + \theta_n f_n(u_i) - b_i)^2 = \|A\theta - b\|^2$$

- ▶ solutia problemei de regresie liniara:

$$\theta^* = (A^T A)^{-1} A^T b$$

Regresia liniara

With four parameters I can fit an elephant, with five I can make him wiggle his trunk (Von Neumann)

- Problema: fitting un polinom de grad $< n$

$$p(t) = \theta_0 + \theta_1 t + \cdots + \theta_{n-1} t^{n-1}$$

la datele (t_i, b_i) , $i = 1, \dots, m$

- ▶ functii de baza $p_j(t) = t^{j-1}$ pentru $j = 1, \dots, n$
- ▶ matricea A cu intrarile $A_{ij} = t_i^{j-1}$ (matrice Vandermonde)

$$A = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{n-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^{n-1} \end{pmatrix}$$

Regresie & Predictie

- **Regresie liniara:** fitting un polinom de grad 2 pentru estimarea populatiei planetei.
- Se cunoaste populatia (in miliarde) intre anii 1950 si 1985:

anul	1950	1955	1960	1965	1970	1975	1980	1985
populatie	2.53	2.77	3.05	3.36	3.72	4.1	4.47	4.87

Exemplu: se porneste de la $t_1 = -4$ pana la $t_8 = 3$ si polinomul:

$$p(t) = a_2 t^2 + a_1 t + a_0$$

- regresia presupune rezolvarea sistemului $Ax = b$ in sens cmmp (reamintim matricea A este de tip Vandermonde)!

$$A = \begin{pmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_8 & t_8^2 \end{pmatrix}, \theta = [a_0; a_1; a_2], b = [b_1; \dots; b_8]$$

- ex. $b_1 = 2.53$ etc...

Regresie & Predictie cont.

- ▶ proble de predictie formulata ca problema cmmp: $Ax = b$
- ▶ solutia CMMP este $a_2 = 0.013$, $a_1 = 0.351$, $a_0 = 3.7126$
- ▶ **predictia** permite estimarea populatia planetei din anul 1990

$$t_9 = 4 \rightarrow \text{populatie 1990} = a_0 + a_1 t_9 + a_2 t_9^2$$

$$\text{populatie 1990} = 5.32 \text{ mil.}$$

coincide "aproape" exact cu cifra oficiala de 5.3 miliarde de oameni pe planeta in 1990!

Estimatorul celor mai mici patrate (LMS): regresia liniara "ridge"

Exemplu: Estimare/Regresie

Se doreste estimarea unui vector de parametrii θ pe baza unor masuratori:

$$b = A\theta + v, \quad b - \text{masuratori}, v - \text{zgomot}$$

Solutia problemei regresiei liniare $\min_{\theta} \|A\theta - b\|^2$ este:

$$\theta^* = (A^T A)^{-1} A^T b$$

Daca $A^T A$ nu este inversabila se considera urmatoarea regularizare:

$$\min_{\theta \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^N \frac{1}{2} (a_i^T \theta - b_i)^2 + \frac{\lambda}{2} \|\theta\|^2 \quad \left(:= \|A\theta - b\|^2 + \lambda \|\theta\|^2 \right)$$

least-square + Tikhonov regularization ($\lambda \geq 0$)

Se mai numeste si **regresia Ridge**. Se mai poate interpreta ca gasirea valorii minime a unui sistem de ecuatii lineare (constrangerile se muta in functia cost printr-un termen de penalitate)

$$\min_{\theta} \|\theta\|^2 \quad \text{s.t.} \quad A\theta = b$$

- ▶ Regresia Ridge penalizeaza putin rezidurile mici si mult pe cele mari \rightarrow are solutie explicita $\theta^* = (A^T A + \lambda I_n)^{-1} A^T b$.
- ▶ Astfel ca aceasta formulare produce multe reziduri mici si putine mari, gasind un vector de parametri θ cu valori in general nenule!

Estimatorul celor mai mici patrate (LMS): regresia liniara lasso

Dorim sa estimam vectorul de parametri θ pe baza unor masuratori:

$$b = A\theta + v, \quad b - \text{masuratori}, v - \text{zgomot}$$

O alta formulare este problema **Lasso**:

$$\min_{\theta \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^N \frac{1}{2} (a_i^T \theta - b_i)^2 + \lambda \|\theta\|_1 \quad \left(:= \|A\theta - b\|^2 + \lambda \|\theta\|_1 \right)$$

least-square + 1-norm

Se poate interpreta ca gasirea unei solutii rare pentru un sistem linear (se muta constrangerea in functia cost printr-un termen de penalitate)

$$\min_{\theta} \|\theta\|_1 \quad \text{s.t.} \quad A\theta = b$$

- ▶ Lasso va gasi un vector rar de parametrii $\theta \rightarrow$ nu are solutie explicita.

$$0 \in A^T (A\theta^* - b) + \lambda \partial \|\theta^*\|_1$$

$$\partial|y| = 1(y > 0); -1(y < 0); [-1, 1](y = 0)$$

- ▶ Se foloseste de obicei in contextul de overfitting.

Estimatorul celor mai mici patrate (LMS): regresia liniara huber

Dorim sa estimam vectorul de parametri x pe baza unor masuratori:

$$b = A\theta + v, \quad b - \text{masuratori}, v - \text{zgomot}$$

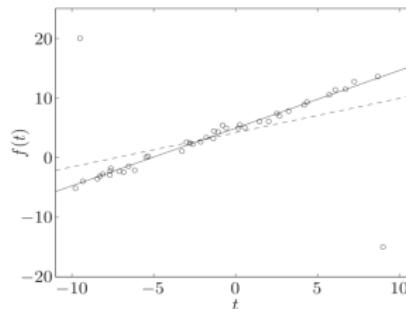
Pentru robustete la valori aberante (outliers) se utilizeaza de obicei formularea

$$\min_{\theta \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^N \phi(a_i^T \theta - b_i) + \lambda \|\theta\|^2$$

unde ϕ este **functia Huber**:

$$\phi(u) = \begin{cases} u^2, & |u| \leq M \\ M(2|u| - M), & |u| > M \end{cases}$$

Datele pot fi aproximate printr-o functie afina, mai putin doua puncte ce au valori aberante. Linia punctata este aproximarea cu metoda celor mai mici patrate. Linia solidă este metoda celor mai mici patrate robusta, obtinuta prin minimizarea functiei Huber. Produce o aproximare mai buna a datelor.



- ▶ Performeaza similar cu metoda celor mai mici patrate pentru rezidurile mici, dar penalizeaza mai putin rezidurile mari. Astfel ca este robust la valori aberante.

Aplicatie: compressed sensing

- ▶ Este o tehnica de procesare de semnal pentru reconstructia unui semnal dintr-o serie de esantioane (masuratori)
- ▶ Se bazeaza pe principiul ca un semnal rar poate fi recuperat din mai putine esantioane decat cere testul Shanon
- ▶ Testul Shanon: semnal continuu $x(t)$ avand frecvente mai mici de B hertz se recupereaza perfect dintr-un semnal discret $z_n = x(n \cdot \Delta T)$ daca perioada de esantionare satisface

$$\Delta T < \frac{1}{2B} \quad \rightarrow \quad \text{nr esantioane} = T/\Delta T$$

unde T este lungimea semnalului continuu.

- ▶ Shanon spune ca daca a priori avem informatie despre frecventa maxima a semnalului ($< B$), atunci mai putine esantioane sunt necesare pentru reconstructie ($T/\Delta T > 2BT$). Dar in aplicatii reale rata de esantionare Shanon ($2BT$) este imposibil de realizat (ex. imagini CT/MRI)
- ▶ in 2004 Candez, Tao si Donoho au demonstrat ca daca a priori stim ca semnalul este rar atunci el poate fi recuperat din mult mai putine esantioane decat spune testul Shanon

Compressed sensing cont.

- ▶ compresia se bazeaza pe gasirea unei baze unde semnalul poate fi reprezentat rar
- ▶ reprezentare rara: avem un semnal (discret) $x \in \mathbb{R}^T$ ce poate fi reprezentat cu numai $s \ll T$ coeficienti in aceasta baza (ex. JPEG, MPEG, MP3)
- ▶ compressed sensing: in loc sa esantioneze semnalul la momente de timp, masoara semnalul prin proiectarea lui intr-o baza cunoscuta
- ▶ compressed sensing are importante aplicatii, ex. in prelucrarea imaginilor medicale (CT or MRI scans)
- ▶ scanarea imaginilor MRI (magnetic resonance image) poate fi accelerata daca se masoara putini coeficienti Fourier
- ▶ **Teorema** (Candes et al): sub anumite conditii rezonabile (RIP - satisfacute in practica) numarul de masuratori (esantioane) necesare este $m > \mathcal{O}(s \log(T))$
- ▶ observati ca aici lungimea semnalului intra sub logaritm in timp ce in Shannon intra liniar $\log(T)$ vrs. $T \iff \log(10^6) \approx 30$ vrs. 10^6 !

Compressed sensing cont.

Formulare matematica (**sistem liniar subdeterminat**):

- ▶ se da un semnal (discret) $x \in \mathbb{R}^T$, o matrice de "sensing" (de achizitie) $A \in \mathbb{R}^{m \times T}$ cu $m \leq T$ si un vector de masuratori $b = Ax$
- ▶ scopul este recuperarea semnalului x din masuratorile b
- ▶ Candes et al. aratat ca daca un semnal admite o reprezentare rara in raport cu un dictionar (matrice de reprezentare) atunci el poate fi recuperat exact: $m > \mathcal{O}(s \log(T))$
- ▶ daca $\mathcal{D} \in \mathbb{R}^{T \times T}$ este acest dictionar a.i. $x = \mathcal{D}c$, cu c avand putini coeficienti nenuli $s \ll T$ atunci semnalul este recuperat:

$$\textbf{lasso : } x = \mathcal{D}c^*, \quad c^* = \arg \min_c \|c\|_1 \text{ s.l. } A\mathcal{D}c = b$$

- ▶ alegerea dictionarelor: ex. imaginile reale tind sa fie rare in "wavelet domain", imaginile MRI in DFT domain, etc
- ▶ ex. imagini MRI au $T = 100$ mil. voxeli. Scaner MRI ia $m \ll T$ esantioane din domeniul Fourier de reprezentare a imaginii (i.e. anumite linii din matricea DFT): aici A este o parte din matricea identitate si \mathcal{D} este matricea DFT de reprezentare a imaginii. Scanarea MRI s-a redus de la 8 min la 70 sec.!

Compressed sensing cont.

Diferenta dintre cmmp (regresia liniara ridge) si lasso:

$$\min_x \|x\|_2 \text{ s.l. } Ax = b \quad \text{vrs.} \quad \min_x \|x\|_1 \text{ s.l. } Ax = b$$

Sistem liniar subdeterminat: $A \in \mathbb{R}^{m \times T}$ cu $m \ll T \rightarrow$ admite o infinitate de solutii. In cmmp o cautam pe solutia cu componentele x_i cele mai mici. In realitate (vezi exemplele anterioare) multe x_i -uri sunt foarte aproape de zero dar sol. cmmp x nu este rara. Componentele aproape de 0 in cmmp devin 0 in problema lasso.

Foarte multe metode de rezolvare a problemei lasso!

Reamintesc ca pentru regresia liniara (ridge) putem utiliza algoritmul Kaczmarz, la fiecare iteratie k alegem o linie i (ciclic, aleator) si actualizam:

$$x^{k+1} = x^k - \frac{a_i^T x^k - b_i}{\|a_i\|^2} a_i \quad (\text{operatii cu vectori})$$

Convergenta - vezi ex. Necoara 2019 (SIAM Journal on Matrix Analysis and Applications) - pentru convergenta acestui alg.

Compressed sensing cont.

Diferenta dintre cmmp (regresia liniara ridge) si lasso:

$$\min_x \|x\|_2 \text{ s.l. } Ax = b \quad \text{vs.} \quad \min_x \|x\|_1 \text{ s.l. } Ax = b$$

Alg. **Sparse Kaczmarz**: la fiecare iteratie se alege o linie i (ciclic, aleator) si se face urmatoarea actualizare pentru x^k :

$$x^{k+1/2} = x^k - \frac{a_i^T x^k - b_i}{\|a_i\|^2} a_i, \quad x^{k+1} = \mathcal{T}_\lambda(x^{k+1/2})$$

unde $\mathcal{T}_\lambda(x) = \text{sign}(x) \cdot \max(|x| - \lambda, 0)$ (soft thresholding)

Teorema (Lorentz 2016, Necoara 2023): iteratia x^k generata de Sparse Kaczmarz converge la solutia problemei "basis pursuit":

$$\min_x \|x\|_2^2 + \lambda \|x\|_1 \text{ s.l. } Ax = b$$

λ mic obtinem solutia cmmp; λ mare obtinem solutia problemei lasso! Multe alte metode de optimizare exista (vezi cursul de Optimizare - anul II, sem II)...

vezi ex. lucrarea: Necoara et al. 2023 - Linear Algebra and Its Applications - pentru convergenta acestui alg.