Foundations of Probabilities
and Information Theory
for Machine Learning

Random Variables

Some proofs

$E[X+Y] = E[X] + E[Y]$

where X and Y are random variables of the same type (i.e. either discrete or cont.)

The discrete case:

$$\begin{aligned} E[X+Y] &= \sum_{\omega\in\Omega} (X(\omega)+Y(\omega))\cdot P(\omega) \\ &= \sum_{\omega} X(\omega)\cdot P(\omega) + \sum_{\omega} Y(\omega)\cdot P(\omega) = E[X]+E[Y] \end{aligned}$$

The continuous case:

$$\begin{aligned} E[X+Y] &= \int_x \int_y (x+y) p_{XY}(x,y) dy dx \\ &= \int_x \int_y x p_{XY}(x,y) dy dx + \int_x \int_y y p_{XY}(x,y) dy dx \\ &= \int_x x \int_y p_{XY}(x,y) dy dx + \int_y y \int_x p_{XY}(x,y) dx dy \\ &= \int_x x p_X(x) dx + \int_y y p_Y(y) dy = E[X]+E[Y] \end{aligned}$$

X and Y are independent $\Rightarrow E[XY] = E[X] \cdot E[Y]$,

X and Y being random variables of the same type (i.e. either discrete or continuous)

The discrete case:

$$E[XY] = \sum_{x \in Val(X)} \sum_{y \in Val(Y)} xyP(X = x, Y = y) = \sum_{x \in Val(X)} \sum_{y \in Val(Y)} xyP(X = x) \cdot P(Y = y)$$

$$= \sum_{x \in Val(X)} \left(xP(X = x) \sum_{y \in Val(Y)} yP(Y = y) \right) = \sum_{x \in Val(X)} xP(X = x)E[Y] = E[X] \cdot E[Y]$$

The continuous case:

$$E[XY] = \int_x \int_y xy \; p(X = x, Y = y) dy dx = \int_x \int_y xy \; p(X = x) \cdot p(Y = y) dy dx$$

$$= \int_x x \; p(X = x) \left(\int_y y \; p(Y = y) dy \right) \; dx = \int_x x \; p(X = x) E[Y] dx$$

$$= E[Y] \cdot \int_x x \; p(X = x) dx = E[X] \cdot E[Y]$$

Discrete random variables:

independence, conditional independence, conditional probabilities

CMU, 2015 spring, T. Mitchell, N. Balcan, HW2, pr. 1.d

Let X, Y, and Z be random variables taking values in $\{0, 1\}$. The following table lists the probability of each possible assignment of 0 and 1 to the variables X, Y, and Z:

	$Z = 0$		$Z = 1$	
	$X = 0$	$X = 1$	$X = 0$	$X = 1$
$Y = 0$	$1/15$	$1/15$	$4/15$	$2/15$
$Y = 1$	$1/10$	$1/10$	$8/45$	$4/45$

For example,
$P(X = 0, Y = 1, Z = 0) = 1/10$ **and** $P(X = 1, Y = 1, Z = 1) = 4/45$.

a. Is X independent of Y? Why or why not?

b. Is X conditionally independent of Y given Z? Why or why not?

c. Calculate $P(X = 0 \mid X + Y > 0)$.

Answer

a. No.

$$\begin{aligned} P(X=0) &= 1/15+1/10+4/15+8/45 = 11/18, \\ P(Y=0) &= 1/15+1/15+4/15+2/15 = 8/15, \end{aligned}$$

and

$$P(X=0|Y=0) = \frac{P(X=0,Y=0)}{P(Y=0)} = \frac{1/15+4/15}{8/15} = \frac{5}{8}.$$

Since $P(X=0)$ does not equal $P(X=0|Y=0)$, X is not independent of Y.

b. For all pairs $y, z \in \{0, 1\}$, we need to check that $P(X = 0|Y = y, Z = z) = P(X = 0|Z = z)$. That the other probabilities are equal follows from the law of total probability.

$$\begin{aligned} P(X = 0|Y = 0, Z = 0) &= \frac{1/15}{1/15 + 1/15} = 1/2 \\ P(X = 0|Y = 1, Z = 0) &= \frac{1/10}{1/10 + 1/10} = 1/2 \\ P(X = 0|Y = 0, Z = 1) &= \frac{4/15}{4/15 + 2/15} = 2/3 \\ P(X = 0|Y = 1, Z = 1) &= \frac{8/45}{8/45 + 4/15} = 2/3. \end{aligned}$$

and

$$\begin{aligned} P(X = 0|Z = 0) &= \frac{1/15 + 1/10}{1/15 + 1/15 + 1/10 + 1/10} = 1/2 \\ P(X = 0|Z = 1) &= \frac{4/15 + 8/45}{4/15 + 2/15 + 8/45 + 4/45} = 2/3. \end{aligned}$$

This shows that X is independent of Y given Z.

c.

$$P(X = 0|X + Y > 0) = \frac{1/10 + 8/45}{1/15 + 1/10 + 1/10 + 2/15 + 4/45 + 8/45} = 5/12.$$

The *correlation coefficient* of two random variables: two properties

Sheldon Ross

A First Course in Probability, 5th ed., Prentice Hall, 1997, pag. 332

Pentru două variabile aleatoare oarecare X şi Y având $Var\,(X) \neq 0$ şi $Var\,(Y) \neq 0$, *coeficientul de corelaţie* se defineşte astfel:

$$\rho(X,Y) \stackrel{def.}{=} \frac{Cov\,(X,Y)}{\sqrt{Var\,(X) \cdot Var\,(Y)}}.$$

a. Să se demonstreze că $-1 \leq \rho(X,Y) \leq 1$.

***Consecinţă*:** $Cov\,(X,Y) \in [-\sqrt{Var(X)\ Var(Y)}, +\sqrt{Var(X)\ Var(Y)}]$, **dacă** $Var\,(X) \neq 0$ **şi** $Var\,(Y) \neq 0$.

b. Să se arate că dacă $\rho(X,Y) = 1$, atunci $Y = aX+b$, cu $a = Var\,(Y)/Var\,(X) > 0$. Similar, dacă $\rho(X,Y) = -1$, atunci $Y = aX + b$, cu $a = -Var\,(Y)/Var\,(X) < 0$.

***Observaţie*: Aşadar, coeficientul de corelaţie reprezintă o „măsură" a gradului de „dependenţă liniară" dintre X şi Y.**

Indicaţii

1. La punctul a, notând $Var(X) = \sigma_X$ şi $Var(Y) = \sigma_Y$, pentru a demonstra inegalitatea $\rho(X,Y) \geq -1$ vă sugerăm să dezvoltaţi expresia $Var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right)$ folosind următoarele două proprietăţi, valabile pentru orice variabile aleatoare X şi Y:

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

şi

$$Cov(aX, bY) = abCov(X,Y), \text{ pentru orice } a, b \in \mathbb{R}.$$

Apoi veţi proceda similar, dezvoltând expresia $Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right)$, ca să demonstraţi inegalitatea $\rho(X,Y) \leq 1$.

2. La punctul b, veţi ţine cont de faptul că pentru o variabilă aleatoare oarecare X, avem $Var(X) = 0$ dacă şi numai dacă variabila X este constantă. (Mai precis, există $c \in \mathbb{R}$ astfel încât $P(X = c) = 1$, unde P este distribuţia de probabilitate considerată la definirea variabilelor din enunţul problemei.)

Răspuns

a. Pentru a demonstra inegalitatea $\rho(X,Y) \geq -1$, procedăm conform *Indicaţiei* 1:

$$\begin{aligned} Var\left(\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}\right) &= Var\left(\frac{X}{\sigma_X}\right)+Var\left(\frac{Y}{\sigma_Y}\right)+2\,Cov\left(\frac{X}{\sigma_X},\frac{Y}{\sigma_Y}\right) \\ &= \frac{1}{\sigma_X^2}\,Var(X)+\frac{1}{\sigma_Y^2}\,Var(Y)+2\frac{1}{\sigma_X}\frac{1}{\sigma_Y}\,Cov(X,Y) \\ &= 1+1+2\rho(X,Y)=2[1+\rho(X,Y)]. \end{aligned}$$

Întrucât $Var(X) \geq 0$ pentru orice variabilă aleatoare X, rezultă că $Var\left(\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}\right) \geq 0$, deci $1+\rho(X,Y) \geq 0$, adică $\rho(X,Y) \geq -1$.

În mod similar, putem să arătăm că

$$Var\left(\frac{X}{\sigma_X}-\frac{Y}{\sigma_Y}\right)=2[1-\rho(X,Y)] \geq 0,$$

deci $\rho(X,Y) \leq 1$.

b. Dacă $\rho(X,Y) = -1$**, atunci din primul calcul de la punctul** a **va rezulta că** $Var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) = 0$**. Ştim că aceasta se întâmplă dacă** $\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}$ **este o variabilă aleatoare constantă, adică, mai precis, există** $a' \in \mathbb{R}$ **astfel încât** $\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} = a'$ **cu probabilitate 1. Prin urmare, putem scrie** $Y = a'\sigma_Y - \frac{\sigma_Y}{\sigma_X}X$**. Rezultă că** $Y = aX + b$**, unde** $a = -\frac{\sigma_Y}{\sigma_X} < 0$ **şi** $b = a'\sigma_Y$**.**

În mod similar, dacă $\rho(X,Y) = 1$**, atunci din al doilea calcul de la punctul** a **va rezulta că** $Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 0$**, deci există** $a'' \in \mathbb{R}$ **astfel încât** $\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} = a''$ **cu probabilitate 1. Aşadar,** $Y = -a''\sigma_Y + \frac{\sigma_Y}{\sigma_X}X$**. Renotând, obţinem** $Y = aX + b$**, cu** $a = \frac{\sigma_Y}{\sigma_X} > 0$ **şi** $b = -a''\sigma_Y$**.**

Probabilistic Distributions

Some properties

Binomial distribution: $b(r;n,p) \overset{def.}{=} C_n^r\, p^r (1-p)^{n-r}$

Significance: $b(r;n,p)$ **is the probability of drawing** r ***heads*** **in** n **independent flips of a coin having the head probability** p**.**

$b(r;n,p)$ **indeed represents a probability distribution:**

- $b(r;n,p) = C_n^r\, p^r (1-p)^{n-r} \geq 0$ **for all** $p \in [0,1]$**,** $n \in \mathbb{N}$ **and** $r \in \{0,1,\ldots,n\}$**,**
- $\sum_{r=0}^{n} b(r;n,p) = 1$**:**

$$(1-p)^n + C_n^1 p(1-p)^{n-1} + \cdots + C_n^{n-1} p^{n-1}(1-p) + p^n = [p + (1-p)]^n = 1$$

Binomial distribution: calculating the mean

$$E[b(r;n,p)] \stackrel{def.}{=} \sum_{r=0}^{n} r \cdot b(r;n,p) =$$

$$\begin{aligned}
&= 1 \cdot C_n^1 p(1-p)^{n-1} + 2 \cdot C_n^2 p^2 (1-p)^{n-2} + \cdots + (n-1) \cdot C_n^{n-1} p^{n-1}(1-p) + n \cdot p^n \\
&= p\left[C_n^1 (1-p)^{n-1} + 2 \cdot C_n^2 p(1-p)^{n-2} + \cdots + (n-1) \cdot C_n^{n-1} p^{n-2}(1-p) + n \cdot p^{n-1}\right] \\
&= np\left[(1-p)^{n-1} + C_{n-1}^1 p(1-p)^{n-2} + \cdots + C_{n-1}^{n-2} p^{n-2}(1-p) + C_{n-1}^{n-1} p^{n-1}\right] \qquad (1) \\
&= np[p + (1-p)]^{n-1} = np \qquad (2)
\end{aligned}$$

For the (1) equality we used the following property:

$$\begin{aligned}
k\, C_n^k &= k\frac{n!}{k!\,(n-k)!} = \frac{n!}{(k-1)!\,(n-k)!} = \frac{n\,(n-1)!}{(k-1)!\,(n-1-(k-1))!} \\
&= n\, C_{n-1}^{k-1}, \forall k = 1, \ldots, n.
\end{aligned}$$

Binomial distribution: calculating the variance

following www.proofwiki.org/wiki/Variance_of_Binomial_Distribution**, which cites "Probability: An Introduction", by Geoffrey Grimmett and Dominic Welsh, Oxford Science Publications, 1986**

We will make use of the formula $Var[X] = E[X^2] - E^2[X]$**.**

By denoting $q = 1 - p$**, it follows:**

$$
\begin{aligned}
E[b^2(r;n,p)] &\overset{def.}{=} \sum_{r=0}^{n} r^2 C_n^r p^r q^{n-r} = \sum_{r=0}^{n} r^2 \frac{n(n-1)\dots(n-r+1)}{r!} p^r q^{n-r} \\
&= \sum_{r=1}^{n} rn \frac{(n-1)\dots(n-r+1)}{(r-1)!} p^r q^{n-r} = \sum_{r=1}^{n} rn\, C_{n-1}^{r-1}\, p^r q^{n-r} \\
&= np \sum_{r=1}^{n} r\, C_{n-1}^{r-1}\, p^{r-1} q^{(n-1)-(r-1)}
\end{aligned}
$$

Binomial distribution: calculating the variance (cont'd)

By denoting $j = r - 1$ and $m = n - 1$, we'll get:

$$E[b^2(r;n,p)] = np \sum_{j=0}^{m} (j+1)\, C_m^j\, p^j q^{m-j}$$

$$= \quad np\Big[\underbrace{\sum_{j=0}^{m} j\, C_m^j\, p^j q^{m-j}}_{E[b(r;n-1,p)],\ cf.(2)} + \underbrace{\sum_{j=0}^{m} C_m^j\, p^j q^{m-j}}_{1} \Big].$$

Therefore,

$$E[b^2(r;n,p)] = np[(n-1)p + 1] = n^2p^2 - np^2 + np.$$

Finally,

$$Var[X] = E[b^2(r;n,p)] - (E[b(r;n,p)])^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1-p)$$

Binomial distribution: calculating the variance

Another solution

- **se demonstrează relativ uşor că orice variabilă aleatoare urmând distribuţia binomială $b(r; n, p)$ poate fi văzută ca o sumă de n variabile independente care urmează distribuţia Bernoulli de parametru p;**[a]
- **ştim (sau, se poate dovedi imediat) că varianţa distribuţiei Bernoulli de parametru p este $p(1-p)$;**
- **ţinând cont de proprietatea de liniaritate a varianţelor — $Var[X_1 + X_2 \ldots + X_n] = Var[X_1] + Var[X_2] \ldots + Var[X_n]$, dacă $X_1, X_2, \ldots, X_n$ sunt variabile independente —, rezultă că $Var[X] = np(1-p)$.**

[a] Vezi www.proofwiki.org/wiki/Bernoulli_Process_as_Binomial_Distribution, care citează de asemenea ca sursă "Probability: An Introduction" de Geoffrey Grimmett şi Dominic Welsh, Oxford Science Publications, 1986.

The *categorical* distribution:

Computing *probabilities* and *expectations*

CMU, 2009 fall, Geoff Gordon, HW1, pr. 4

Suppose we have n bins and m balls. We throw balls into bins independently at random, so that each ball is equally likely to fall into any of the bins.

a. What is the probability of the first ball falling into the first bin?

b. What is the expected number of balls in the first bin?

***Hint* (1): Define an indicator random variable representing whether the i-th ball fell into the first bin:**

$$X_i = \begin{cases} 1 & \text{if } i\text{-th ball fell into the first bin;} \\ 0 & \text{otherwise.} \end{cases}$$

***Hint* (2): Use linearity of expectation.**

c. What is the probability that the first bin is empty?

d. What is the expected number of empty bins? *Hint* (3): Define an indicator for the event "bin j is empty" and use linearity of expectations.

Answer

a. It is equally likely for a ball to fall in any of the bins, so the probability that first ball falling into the first bin is $1/n$.

b. Define X_i as described in Hint 1. Let Y be the total number of balls that fall into first bin: $Y = X_1 + \ldots + X_m$. The expected number of balls is:

$$E[Y] = \sum_{i=1}^{m} E[X_i] = \sum_{i=1}^{m} 1 \cdot P(X_i = 1) = m \cdot 1/n = m/n$$

c. Let Y and X_i be the same as defined at point b. For the first bin to be empty none of the balls should fall into the first bin: $Y = 0$.

$$P(Y = 0) = P(X_1 = 0, \ldots, X_m = 0) = \Pi_{i=1}^{m} P(X_i = 0) = (1 - 1/n)^m = \left(\frac{n-1}{n}\right)^m$$

d. For each one of the n bins, we define an indicator random variable Y_j for the event "bin j is empty". Let Z be the random variables denoting the number of empty bins. $Z = Y_1 + \ldots + Y_n$. Then the expected number of empty bins is:

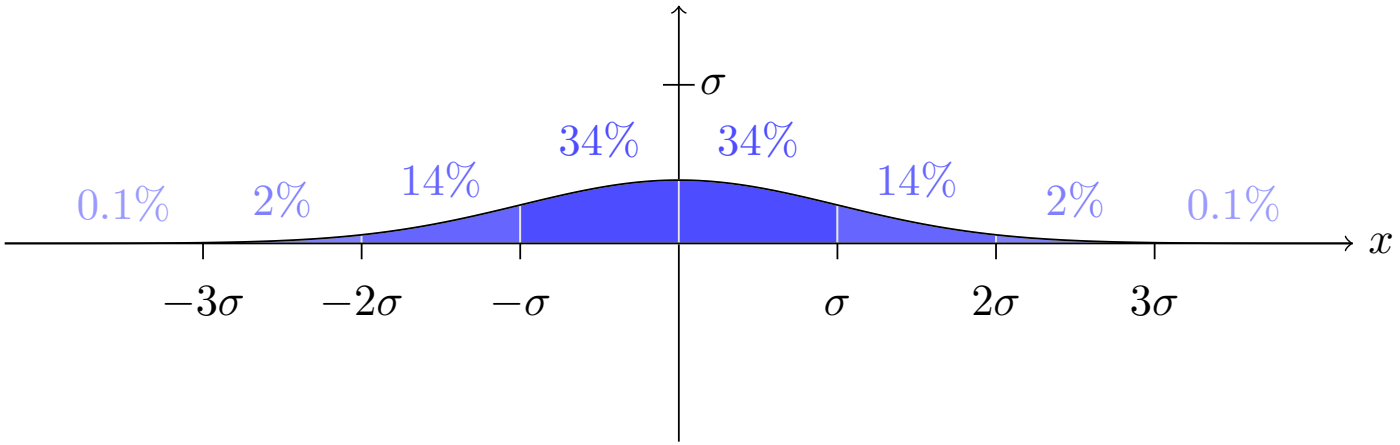
$$E[Z] = E[\sum_{j=1}^{n} Y_j] = \sum_{j=1}^{n} E[Y_j] = \sum_{j=1}^{n} 1 \cdot P(Y_j) = \sum_{j=1}^{n} \left(\frac{n-1}{n}\right)^m = n\left(\frac{n-1}{n}\right)^m$$

The univariate Gaussian distribution:

$$\mathcal{N}_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The plot, when $\mu = 0$:

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



Source: http://www.texample.net/tikz/examples/standard-deviation/

Proving that $\mathcal{N}_{\mu,\sigma}$ is indeed a p.d.f.

1. $\mathcal{N}_{\mu,\sigma}(x) \geq 0 \; \forall x \in \mathbb{R}$ **(true)**

2. $\int_{-\infty}^{\infty} \mathcal{N}_{\mu,\sigma}(x)dx = 1$

Note: **Concerning the second property, it is enough to prove it for the *standard* case ($\mu = 0,\ \sigma = 1$), because the non-standard case can be reduced to this one:**

Using the variable transformation $v = \dfrac{x-\mu}{\sigma}$ **will imply** $x = \sigma v + \mu$ **and** $dx = \sigma dv$**, so:**

$$\int_{-\infty}^{\infty} \mathcal{N}_{\mu,\sigma}(x)dx = \int_{x=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{x=-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{v=-\infty}^{\infty} e^{-\frac{v^2}{2}} \sigma dv = \frac{1}{\sqrt{2\pi}\sigma} \sigma \int_{v=-\infty}^{\infty} e^{-\frac{v^2}{2}} dv = \frac{1}{\sqrt{2\pi}} \int_{x=-\infty}^{\infty} \mathcal{N}_{0,1}(x)dx$$

The *standard* case: proving that $\mathcal{N}_{0,1}$ is indeed a p.d.f.:

$$\left(\int_{v=-\infty}^{\infty} e^{-\frac{v^2}{2}}\,dv\right)^2 = \left(\int_{x=-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx\right)\cdot\left(\int_{y=-\infty}^{\infty} e^{-\frac{y^2}{2}}\,dy\right) = \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dydx$$

$$= \iint_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}}\,dydx$$

By switching from x, y to polar coordinates r, θ (see the *Note* below), it follows:

$$\left(\int_{v=-\infty}^{\infty} e^{-\frac{v^2}{2}}\,dv\right)^2 = \int_{r=0}^{\infty}\int_{\theta=0}^{2\pi} e^{-\frac{r^2}{2}}\,(rdrd\theta) = \int_{r=0}^{\infty} re^{-\frac{r^2}{2}}\left(\int_{\theta=0}^{2\pi} d\theta\right)dr = \int_{r=0}^{\infty} re^{-\frac{r^2}{2}}\,\theta|_0^{2\pi}\,dr$$

$$= 2\pi\int_{r=0}^{\infty} re^{-\frac{r^2}{2}}\,dr = 2\pi\,(-e^{-\frac{r^2}{2}})\Big|_0^{\infty} = 2\pi(1-0) = 2\pi \Rightarrow \int_{v=-\infty}^{\infty} e^{-\frac{v^2}{2}}\,dv = \sqrt{2\pi} \Rightarrow \int_{v=-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{v^2}{2}}\,dv = 1.$$

***Note*: $x = r\cos\theta$ and $y = r\sin\theta$, with $r \geq 0$ and $\theta \in [0, 2\pi)$. Therefore, $x^2 + y^2 = r^2$, and the Jacobian matrix is**

$$\frac{\partial(x,y)}{\partial(r,\theta)} = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r\cos^2\theta + r\sin^2\theta = r \geq 0.$$ **So,** $dxdy = rdrd\theta$.

Calculating the mean

$$E[\mathcal{N}_{\mu,\sigma}(x)] \stackrel{def.}{=} \int_{-\infty}^{\infty} x\,\mathcal{N}_{\mu,\sigma}(x)dx = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} x\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx$$

Using again the variable transformation $v = \dfrac{x-\mu}{\sigma}$ will imply:

$$\begin{aligned}
E[X] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(\sigma v+\mu)e^{-\frac{v^2}{2}}(\sigma dv) = \frac{\sigma}{\sqrt{2\pi}\sigma}\left(\sigma\int_{-\infty}^{\infty} ve^{-\frac{v^2}{2}}\,dv + \mu\int_{-\infty}^{\infty} e^{-\frac{v^2}{2}}\,dv\right)\\
&= \frac{1}{\sqrt{2\pi}}\left(-\sigma\int_{-\infty}^{\infty}(-v)e^{-\frac{v^2}{2}}\,dv + \mu\int_{-\infty}^{\infty} e^{-\frac{v^2}{2}}\,dv\right) = \frac{1}{\sqrt{2\pi}}\left(-\sigma\underbrace{e^{-\frac{v^2}{2}}\Big|_{-\infty}^{\infty}}_{=0} + \mu\int_{-\infty}^{\infty} e^{-\frac{v^2}{2}}\,dv\right)\\
&= \frac{\mu}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{v^2}{2}}\,dv = \frac{\mu}{\sqrt{2\pi}}\sqrt{2\pi} = \mu
\end{aligned}$$

Calculating the variance

We will make use of the formula $Var[X] = E[X^2] - E^2[X]$**.**

$$E[X^2] = \int_{-\infty}^{\infty} x^2 \mathcal{N}_{\mu,\sigma}(x)dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x^2 \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Again, using the transformation $v = \dfrac{x-\mu}{\sigma}$ **will imply** $x = \sigma v + \mu$ **and** $dx = \sigma dv$**. Therefore,**

$$\begin{aligned}
E[X^2] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (\sigma v + \mu)^2 \, e^{-\frac{v^2}{2}} \, (\sigma dv) \\
&= \frac{\sigma}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (\sigma^2 v^2 + 2\sigma\mu v + \mu^2) \, e^{-\frac{v^2}{2}} \, dv \\
&= \frac{1}{\sqrt{2\pi}} \left(\sigma^2 \int_{-\infty}^{\infty} v^2 \, e^{-\frac{v^2}{2}} \, dv + 2\sigma\mu \int_{-\infty}^{\infty} v \, e^{-\frac{v^2}{2}} \, dv + \mu^2 \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} \, dv \right)
\end{aligned}$$

Note that we have already computed $\int_{-\infty}^{\infty} ve^{-\frac{v^2}{2}} \, dv = 0$ **and** $\int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} \, dv = \sqrt{2\pi}$**.**

Calculating the variance (Cont'd)

Therefore, we only need to compute

$$\int_{-\infty}^{\infty} v^2 e^{-\frac{v^2}{2}}\, dv = \int_{-\infty}^{\infty} (-v)\left(-ve^{-\frac{v^2}{2}}\right) dv = \int_{-\infty}^{\infty} (-v)\left(e^{-\frac{v^2}{2}}\right)' dv$$

$$= (-v)\, e^{-\frac{v^2}{2}}\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} (-1)e^{-\frac{v^2}{2}}\, dv = 0 + \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}}\, dv = \sqrt{2\pi}.$$

Here above we used the fact that

$$\lim_{v\to\infty} v\, e^{-\frac{v^2}{2}} = \lim_{v\to\infty} \frac{v}{e^{\frac{v^2}{2}}} \overset{l'Hôpital}{=} \lim_{v\to\infty} \frac{1}{v\, e^{\frac{v^2}{2}}} = 0 = \lim_{v\to-\infty} v\, e^{-\frac{v^2}{2}}$$

So, $E[X^2] = \frac{1}{\sqrt{2\pi}}\left(\sigma^2\sqrt{2\pi} + 2\sigma\mu\cdot 0 + \mu^2\sqrt{2\pi}\right) = \sigma^2 + \mu^2.$

And, finally, $Var[X] = E[X^2] - (E[X])^2 = (\sigma^2 + \mu^2) - \mu^2 = \sigma^2.$

Vectors of random variables.

A property:

The covariance matrix Σ corresponding to such a vector is symmetric and positive semi-definite

Chuong Do, Stanford University, 2008

[adapted by Liviu Ciortuz]

Fie variabilele aleatoare $X_1, \ldots, X_n$, cu $X_i : \Omega \to \mathbb{R}$ pentru $i = 1, \ldots, n$. *Matricea de covarianţă* a *vectorului de variabile aleatoare* $X = (X_1, \ldots, X_n)$ este o matrice pătratică de dimensiune $n \times n$, ale cărei elemente se definesc astfel: $[Cov(X)]_{ij} \stackrel{def.}{=} Cov(X_i, X_j)$, pentru orice $i, j \in \{1, \ldots, n\}$.

Arătaţi că $\Sigma \stackrel{not.}{=} Cov(X)$ este matrice simetrică şi pozitiv semi-definită, cea de-a doua proprietate însemnând că pentru orice vector $z \in \mathbb{R}^n$ are loc inegalitatea $z^\top \Sigma z \geq 0$. (Vectorii $z \in \mathbb{R}^n$ sunt consideraţi vectori-coloană, iar simbolul $\top$ reprezintă operaţia de transpunere de matrice.)

$\boldsymbol{Cov}(X)_{i,j} \stackrel{def.}{=} \boldsymbol{Cov}(X_i, X_j)$**, for all** $i, j \in \{1, \ldots, n\}$**, and**
$\boldsymbol{Cov}(X_i, X_j) \stackrel{def.}{=} E[(X_i - E[X_i])(X_j - E[X_j])] = E[(X_j - E[X_j])(X_i - E[X_i])] = \boldsymbol{Cov}(X_j, X_i)$**,**
therefore $\boldsymbol{Cov}(X)$ **is a symmetric matrix.**

We will show that $z^T \Sigma z \geq 0$ **for any** $z \in \mathbb{R}^n$ **(seen as a column-vector):**

$$
\begin{aligned}
z^T \Sigma z &= \sum_{i=1}^{n} z_i (\sum_{j=1}^{n} \Sigma_{ij} z_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} (z_i \Sigma_{ij} z_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} (z_i \, Cov[X_i, X_j] \, z_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} (z_i \, E[(X_i - E[X_i])(X_j - E[X_j])] \, z_j) = E\left[\sum_{i=1}^{n} \sum_{j=1}^{n} z_i \, (X_i - E[X_i])(X_j - E[X_j]) \, z_j \right] \\
&= E\left[\left(\sum_{i=1}^{n} z_i \, (X_i - E[X_i]) \right) \left(\sum_{j=1}^{n} (X_j - E[X_j]) \, z_j \right) \right] \\
&= E\left[\left(\sum_{i=1}^{n} (X_i - E[X_i]) \, z_i \right) \left(\sum_{j=1}^{n} (X_j - E[X_j]) \, z_j \right) \right] = E[((X - E[X])^T \cdot z)^2] \geq 0
\end{aligned}
$$

Multi-variate Gaussian distributions:

A property:

When the covariance matrix of a multi-variate (d-dimensional) Gaussian distribution is diagonal, then the p.d.f. (probability density function) of the respective multi-variate Gaussian is equal to the product of d independent uni-variate Gaussian densities.

Chuong Do, Stanford University, 2008

[adapted by Liviu Ciortuz]

Let's consider $X = [X_1 \dots X_d]^T$, $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{S}_+^d$, where $\mathbb{S}_+^d$ is the set of symmetric positive definite matrices (which implies $|\Sigma| \neq 0$ and $(x-\mu)^T\Sigma^{-1}(x-\mu) > 0$, therefore $-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) < 0$, for any $x \in \mathbb{S}^d$, $x \neq \mu$).

The probability density function of a multi-variate Gaussian distribution of parameters μ and Σ is:

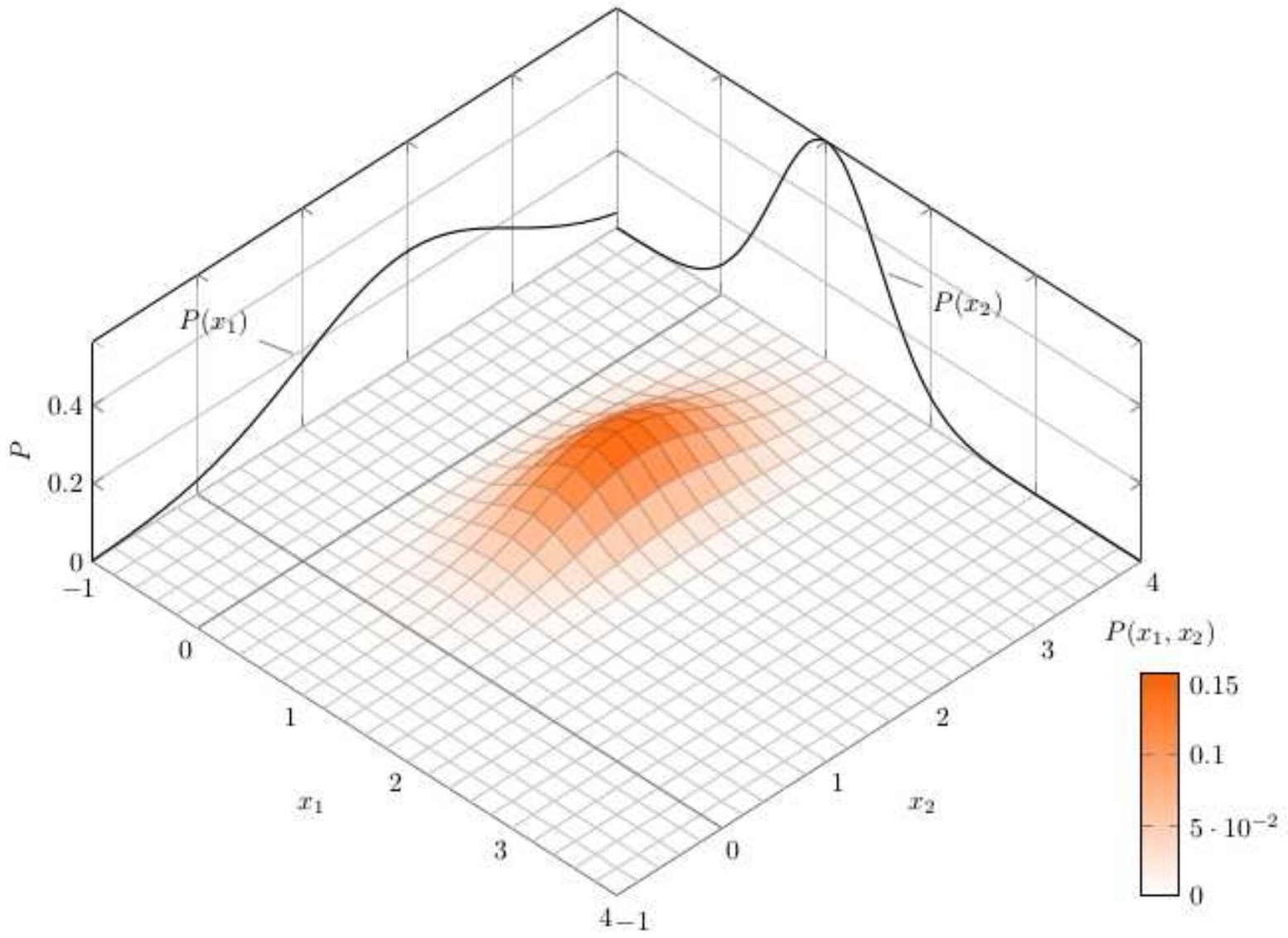
$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),$$

Notation: $X \sim \mathcal{N}(\mu,\Sigma)$.
Show that when the covariance matrice Σ is diagonal, then the p.d.f. (probability density function) of the respective multi-variate Gaussian is equal to the product of d independent uni-variate Gaussian densities.

We will make the proof for $d = 2$ (generalization to $d > 2$ will be easy):

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

Note: It is easy to show that if $\Sigma \in \mathbb{S}_+^d$ is diagonal, the elements on the principal diagonal Σ are indeed strictly positive. (It is enough to consider $z = (1,0)$ and respectively $z = (0,1)$ in formula for *pozitive-definiteness* of Σ.) This is why we wrote these elements of σ as σ_1^2 and σ_2^2.


$P(x_1)$
$P(x_2)$
P
0.4
0.2
0
−1
0
1
2
3
4
x_1
x_2
$P(x_1, x_2)$
0.15
0.1
$5 \cdot 10^{-2}$
0

$$
\begin{aligned}
p(x;\mu,\Sigma) &= \frac{1}{2\pi \left| \begin{array}{cc} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{array} \right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \left[\begin{array}{c} x_1-\mu_1 \\ x_2-\mu_2 \end{array} \right]^T \left[\begin{array}{cc} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{array} \right]^{-1} \left[\begin{array}{c} x_1-\mu_1 \\ x_2-\mu_2 \end{array} \right] \right) \\
&= \frac{1}{2\pi\,\sigma_1\sigma_2} \exp\left(-\frac{1}{2} \left[\begin{array}{c} x_1-\mu_1 \\ x_2-\mu_2 \end{array} \right]^T \left[\begin{array}{cc} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{array} \right] \left[\begin{array}{c} x_1-\mu_1 \\ x_2-\mu_2 \end{array} \right] \right) \\
&= \frac{1}{2\pi\,\sigma_1\sigma_2} \exp\left(-\frac{1}{2} \left[\begin{array}{c} x_1-\mu_1 \\ x_2-\mu_2 \end{array} \right]^T \left[\begin{array}{c} \frac{1}{\sigma_1^2}(x_1-\mu_1) \\ \frac{1}{\sigma_2^2}(x_2-\mu_2) \end{array} \right] \right) \\
&= \frac{1}{2\pi\,\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2-\mu_2)^2 \right) \\
&= p(x_1;\mu_1,\sigma_1^2)\; p(x_2;\mu_2,\sigma_2^2).
\end{aligned}
$$

Factorizing *positive definite matrices*, using eigenvectors

Sebastian Ciobanu,

following Chuong Do (from Stanford University),

The Multivariate Gaussian Distribution, 2008

Fie o variabilă aleatoare $X : \Omega \to \mathbb{R}^d$. În cele ce urmează elementele din $\mathbb{R}^d$ vor fi considerate vectori-coloană. Vom nota cu $\mathbb{S}_+^d$ mulţimea matricelor simetrice pozitiv definite de dimensiune $d \times d$.

***Definiţie*: Fie $A \in \mathbb{R}^{d \times d}$. Se numeşte *valoare proprie* a matricei A un număr complex $\lambda \in \mathbb{C}$ pentru care există un vector nenul $x \in \mathbb{C}^d$ pentru care are loc egalitatea $Ax = \lambda x$. Acest vector x se numeşte *vector propriu* asociat valorii proprii λ.**

a. Demonstraţi că pentru orice matrice $\Sigma \in \mathbb{S}_+^d$ există o matrice $B \in \mathbb{R}^{d \times d}$ astfel încât Σ să poată fi factorizată sub forma următoare: $\Sigma = BB^\top$.

***Observaţie*: Factorizarea aceasta nu este unică, adică există mai multe posibilităţi de a scrie matricea Σ drept $\Sigma = BB^\top$.**

***Indicaţie*: Vă puteţi folosi de următoarele proprietăţi:**

***i.* Orice matrice $A \in \mathbb{R}^{d \times d}$ care este simetrică poate fi scrisă astfel: $A = U \Lambda U^\top$, unde $U \in \mathbb{R}^{d \times d}$ este o matrice ortonormală conţinând vectorii proprii (pentru care impunem să aibă norma 1) ai lui A drept coloane, iar Λ este matricea diagonală conţinând valorile proprii ale lui A în ordinea corespunzătoare coloanelor (adică, a vectorilor proprii) din matricea U.(Faptul că U este matrice ortonormală se poate scrie în mod formal astfel: $U^\top U = U\, U^\top = I$.)**

***ii.* Fie $A \in \mathbb{R}^{d \times d}$ o matrice simetrică. A este pozitiv definită dacă şi numai dacă toate valorile sale proprii sunt (reale şi) pozitive.**

***iii.* $(AB)^\top = B^\top A^\top$, pentru orice matrice $A, B \in \mathbb{R}^{d \times d}$.**

Soluţie

37.

Ştim, prin ipoteză, că matricea $\Sigma \in \mathbb{S}_+^d$, deci este simetrică. Conform proprietăţii $(a.i)$, putem scrie următoarea factorizare pentru Σ:

$$\Sigma = U \Lambda U^\top,$$

unde U este matricea ortonormală conţinând vectorii proprii (cu norma 1) ai lui Σ drept coloane, iar $\Lambda \in \mathbb{R}^{d \times d}$ este matricea diagonală conţinând valorile proprii ale lui Σ, în ordinea corespunzătoare coloanelor matricei U.
Întrucât $\Sigma \in \mathbb{S}_+^d$ este matrice pozitiv definită, conform proprietăţii $(a.ii)$ rezultă că toate valorile proprii ale lui Σ sunt pozitive. Prin urmare, există matricea

$$\Lambda^{1/2} = \begin{bmatrix} \sqrt{\lambda_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda_d} \end{bmatrix} \in \mathbb{R}^{d \times d}.$$

***Observaţii*:**
(1) Este imediat faptul că putem factoriza / descompune matricea Λ în felul următor:

$$\Lambda^{1/2} \cdot \Lambda^{1/2} = \Lambda. \tag{3}$$

(2) $\Lambda^{1/2}$ este matrice diagonală, deci simetrică. Aşadar,

$$(\Lambda^{1/2})^\top = \Lambda^{1/2}. \tag{4}$$

Acum putem relua factorizarea matricei $\Sigma = U\Lambda U^\top$. Elaborând — adică, factorizând matricea Λ — în partea dreaptă a acestei egalităţi, vom putea obţine o nouă factorizare pentru matricea Σ în felul următor:

$$\begin{aligned}
\Sigma &= U\Lambda U^\top \\
&\overset{(3)}{=} U\Lambda^{1/2}\Lambda^{1/2}U^\top \overset{(4)}{=} U\Lambda^{1/2}(\Lambda^{1/2})^\top U^\top \\
&\overset{asoc.}{=} U\Lambda^{1/2}((\Lambda^{1/2})^\top U^\top) \overset{(a.iii)}{=} U\Lambda^{1/2}(U\Lambda^{1/2})^\top \\
&= BB^\top, \text{ unde } B \overset{not.}{=} U\Lambda^{1/2}. \qquad (5)
\end{aligned}$$

***Observaţii*:**

(3) O altă modalitate de a factoriza / descompune matricea Σ sub forma $\Sigma = BB^\top$ este dată de *factorizarea Cholesky*: dacă A este o matrice simetrică şi pozitiv definită, atunci există o unică matrice $L \in \mathbb{R}^{d\times d}$, inferior triunghiulară, astfel încât $A = LL^\top$. (See https://en.wikipedia.org/wiki/Positive-definite_matrix**.)**

(4) Factorizarea (5) este de fapt valabilă şi pentru matrice pozitiv semidefinite. Însă proprietatea de inversabilitate (vedeţi punctul c) *nu* este valabilă decât pentru matrice pozitiv definite.

b. La acest punct vom face o *exemplificare* pentru chestiunile prezentate la punctul a. Considerând matricea

$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix},$$

(deci, $d = 2$) determinaţi o matrice $B \in \mathbb{R}^{2\times 2}$ astfel încât $\Sigma = BB^\top$.

***Indicaţie*: Folosind notaţiile din *Definiţia* dată pentru valorile proprii şi vectorii proprii, vom avea:**

$$Ax = \lambda x \Leftrightarrow (\lambda I_d - A)x = \mathbf{0}, x \neq \mathbf{0} \Leftrightarrow \det(\lambda I_d - A) = 0,$$

unde $\mathbf{0}$ este vectorul coloană nul d-dimensional, iar I_d este matricea identitate d-dimensională.

Soluţie

Mai întâi vom calcula valorile proprii ale matricei Σ din enunţ. Fie deci $\lambda \in \mathbb{R}$ şi $x \in \mathbb{R}^d$, $x \neq 0$. Trebuie să rezolvăm ecuaţia $\Sigma x = \lambda x$ şi, conform *Indicaţiei* din enunţ, aceasta revine la a rezolva ecuaţia $\det(\lambda I_d - \Sigma) = 0$.

$$\begin{aligned} \det(\lambda I_d - \Sigma) = 0 \quad &\Leftrightarrow \quad \det\left(\begin{bmatrix} \lambda - 1 & -0.5 \\ -0.5 & \lambda - 1 \end{bmatrix}\right) = 0 \Leftrightarrow (\lambda - 1)^2 - 0.5^2 = 0 \\ &\Leftrightarrow \quad (\lambda - 1 - 0.5)(\lambda - 1 + 0.5) = 0 \Leftrightarrow (\lambda - 1.5)(\lambda - 0.5) = 0 \\ &\Leftrightarrow \quad \lambda_1 = 0.5,\ \lambda_2 = 1.5. \end{aligned}$$

Deci valorile proprii ale matricei Σ sunt $\lambda_1 = 0.5$ şi $\lambda_2 = 1.5$.
Se observă că $\lambda_1, \lambda_2 > 0$. Aşadar, conform proprietăţii (a, ii), matricea Σ dată în enunţ este pozitiv definită. Conform punctului a, ştim acum că există o matrice B astfel încât matricea Σ să poată fi factorizată sub forma $\Sigma = BB^\top$. Pentru a determina matricea B, trebuie sa calculăm vectorii proprii ai matricei Σ. Îi vom obţine rezolvând următorul sistem, care este compatibil nedeterminat (pentru că $\det(\lambda I_d - \Sigma) = 0$):

$$\begin{cases} (\lambda - 1)v - 0.5u = 0 \Rightarrow v = \dfrac{0.5u}{\lambda - 1} \\ -0.5v + (\lambda - 1)u = 0 \end{cases} .$$

Deci, vectorii proprii sunt de forma:

$$x = (v, u) = \left(\frac{0.5}{\lambda - 1} u, u \right), u \in \mathbb{R}.$$

În mod concret,

$$\lambda_1 = 0.5 \Rightarrow x_1 = (-u, u), u \in \mathbb{R}$$

şi

$$\lambda_2 = 1.5 \Rightarrow x_2 = (u, u), u \in \mathbb{R}.$$

Deoarece vrem ca vectorii x_1 şi x_2 să aibă norma 1, îi vom ,,normaliza“. Mai întâi,

$$\frac{x_1}{\|x_1\|_2} = \frac{(-u, u)}{\sqrt{2u^2}} = \frac{(-u, u)}{\sqrt{2}|u|} \tag{6}$$

şi, considerând $u > 0$, deci $|u| = u$, rezultă că

$$\frac{x_1}{\|x_1\|_2} = \frac{(-u, u)}{\sqrt{2}u} = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right).$$

Apoi, în mod similar obţinem

$$\frac{x_2}{\|x_2\|_2} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right).$$

De la punctul a ştim că putem alege $B = U\Lambda^{1/2}$, deci

$$U = \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}, \quad \Lambda^{1/2} = \begin{bmatrix} \sqrt{0.5} & 0 \\ 0 & \sqrt{1.5} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}}{\sqrt{2}} \end{bmatrix},$$

$$B = \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}.$$

Verificăm ă într-adevăr $\Sigma = BB^\top$:

$$BB^\top = \begin{bmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} \\ \frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} + \frac{3}{4} & -\frac{1}{4} + \frac{3}{4} \\ -\frac{1}{4} + \frac{3}{4} & \frac{1}{4} + \frac{3}{4} \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} = \Sigma.$$

Observaţie

În ton cu *Observaţia* din enunţ, dacă în relaţia (6) **vom considera** $u < 0$**, vom obţine încă o soluţie pentru factorizarea matricei** Σ**:**

$$\frac{x_1'}{\|x_1'\|_2} = \frac{(-u, u)}{\sqrt{2}|u|} = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right) = -\frac{x_1}{\|x_1\|_2}$$
$$\frac{x_2'}{\|x_2'\|_2} = \frac{(u, u)}{\sqrt{2}|u|} = \left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right) = -\frac{x_2}{\|x_2\|_2},$$

deci

$$U' = -U,\ B' = U'\Lambda^{1/2} = -U\Lambda^{1/2} = -B$$

şi, în consecinţă,

$$B'(B')^\top = (-B)(-B)^\top = BB^\top = \Sigma.$$

c. Demonstraţi că matricea B care satisface proprietatea de la punctul a este inversabilă.

***Indicaţie*: Vă puteţi folosi de următoarele proprietăţi:**

i. **Matricea $A \in \mathbb{R}^{d \times d}$ este inversabilă dacă şi numai dacă $\det(A) \neq 0$.**
ii. **$\det(AB) = \det(A)\det(B)$, pentru orice matrice $A, B \in \mathbb{R}^{d \times d}$.**
iii. **$\det(A) = \det(A^\top)$, unde $A \in \mathbb{R}^{d \times d}$.**
iv. **Orice matrice pozitiv definită este inversabilă (iar inversa ei este de asemenea matrice pozitiv definită).**

Soluţie

De la punctul a rezultă că pentru orice matrice $\Sigma \in \mathbb{S}_+^d$ există o descompunere / factorizare de forma $\Sigma = BB^\top$, deci

$$\det(\Sigma) = \det(BB^\top) \overset{(c.ii)}{=} \det(B)\det(B^\top) \overset{(c.iii)}{=} \det(B)^2.$$

Prin urmare,

$$\Rightarrow \det(B) = \pm\sqrt{\det(\Sigma)}. \tag{7}$$

Pe de altă parte, din faptul că Σ este matrice pozitiv definită, rezultă conform proprietăţii $(c.iv)$ că Σ este matrice inversabilă. Mai departe, conform proprietăţii $(c.i)$, vom avea $\det(\Sigma) \neq 0$. Coroborând aceasta cu relaţia (7), rezultă că $\det(B) \neq 0$ şi în consecinţă (din nou, conform proprietăţii $(c.i)$) că matricea B este inversabilă (ceea ce era de demonstrat).

The Gaussian Multivariate Distribution:

Its density function is indeed a p.d.f.

Sebastian Ciobanu,

following Chuong Do (from Stanford University),

***The Multivariate Gaussian Distribution*, 2008**

***Definiţie*: Notăm cu $\mathbb{S}_+^d$ mulţimea matricelor simetrice pozitiv definite de dimensiune $d \times d$. Spunem că variabila aleatoare vectorială $X : \Omega \to \mathbb{R}^d$, reprezentată sub forma $X = (X_1 \dots X_d)^\top$, urmează o distribuţie gaussiană multivariată, având media $\mu \in \mathbb{R}^d$ şi matricea de covarianţă $\Sigma \in \mathbb{S}_+^d$, dacă funcţia ei de densitate [de probabilitate] are forma analitică următoare:**

$$p_X(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right),$$

unde notaţia $\det(\Sigma)$ desemnează determinantul matricei Σ, iar $\exp(\ )$ desemnează funcţia exponenţială având baza e. Pe scurt, vom nota această proprietate de definiţie a lui X sub forma $X \sim \mathcal{N}(\mu, \Sigma)$.

***Observaţie* (1): La problema ... am demonstrat că pentru orice matrice $\Sigma \in \mathbb{R}^d$ simetrică şi pozitiv definită există (însă nu neapărat în mod unic) o matrice $B \in \mathbb{R}^{d \times d}$ cu proprietatea că Σ se poate „factoriza" sub forma $\Sigma = BB^\top$. Mai mult, am demonstrat că orice matrice B care satisface această proprietate este în mod necesar inversabilă.**

a. Demonstraţi că dacă definim $Z = B^{-1}(X - \mu)$, atunci $Z \sim \mathcal{N}(\mathbf{0}, I_d)$, unde $\mathbf{0}$ este vectorul coloană nul d-dimensional, iar I_d este matricea identitate d-dimensională.

***Observaţie* (2): Proprietatea aceasta este o generalizare a metodei de „standardizare" pe care am întâlnit-o deja la problema CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 3, aplicată în cazul distribuţiilor gaussiene *univariate*.**

***Indicaţie* (1): Vă puteţi folosi de următoarele proprietăţi:**

***i.* Fie $X = (X_1 \dots X_d)^\top \in \mathbb{R}^d$ o variabilă aleatoare de tip vector, cu distribuţia comună dată de funcţia de densitate $p_X : \mathbb{R}^d \to \mathbb{R}$. Dacă $Z = H(X) \in \mathbb{R}^d$, unde H este funcţie bijectivă şi diferenţiabilă, atunci Z are distribuţia comună dată de funcţia de densitate $p_Z : \mathbb{R}^d \to \mathbb{R}$, unde**

$$p_Z(z) = p_X(x) \cdot \left| \det \left(\begin{bmatrix} \frac{\partial x_1}{\partial z_1} & \dots & \frac{\partial x_1}{\partial z_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_d}{\partial z_1} & \dots & \frac{\partial x_d}{\partial z_d} \end{bmatrix} \right) \right| .$$

Matricea $\begin{bmatrix} \frac{\partial x_1}{\partial z_1} & \dots & \frac{\partial x_1}{\partial z_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_d}{\partial z_1} & \dots & \frac{\partial x_d}{\partial z_d} \end{bmatrix}$ **se numeşte *matricea jacobiană* a lui x în raport cu z şi se notează, în acest caz, cu** $\frac{\partial x}{\partial z}$**.**

ii. $\frac{\partial Ax + b}{\partial x} = A$**, unde** $x, b \in \mathbb{R}^d$**,** $A \in \mathbb{R}^{d \times d}$**, iar** A **şi** b **nu depind de** x**.**

iii. $(AB)^{-1} = B^{-1}A^{-1}$**, unde** $A, B \in \mathbb{R}^{d \times d}$**.**

iv. $(AB)^\top = B^\top A^\top$**, unde** $A, B \in \mathbb{R}^{d \times d}$**.**

v. $\det(AB) = \det(A)\det(B)$**, unde** $A, B \in \mathbb{R}^{d \times d}$**.**

vi. $\det(A) = \det(A^\top)$**, unde** $A \in \mathbb{R}^{d \times d}$**.**

Soluţie

Sunt imediate următoarele relaţii:

$$z = B^{-1}(x-\mu) \overset{B\cdot}{\Rightarrow} Bz = x-\mu \Rightarrow x = Bz+\mu \tag{8}$$

$$\frac{\partial x}{\partial z} = \frac{\partial(Bz+\mu)}{\partial z} \overset{(a.ii)}{=} B \tag{9}$$

$$(\det(\Sigma))^{1/2} = \sqrt{\det(BB^\top)} \overset{(a.v)}{=} \sqrt{\det(B)\det(B^\top)} \overset{(a.vi)}{=} \sqrt{(\det(B))^2} = |\det(B)|. \tag{10}$$

Vom rescrie acum expresia care constituie argumentul funcţiei $\exp(\,)$ **din definiţia p.d.f.-ului distribuţiei gaussiene multivariate (vedeţi enunţul), în funcţie de vectorul** z**.**

$$\begin{aligned}
&(x-\mu)^\top \Sigma^{-1}(x-\mu) \\
&\overset{(8)}{=} (Bz+\mu-\mu)^\top (BB^\top)^{-1}(Bz+\mu-\mu) = (Bz)^\top (BB^\top)^{-1}(Bz) \\
&\overset{(a.iii)}{=} (Bz)^\top ((B^\top)^{-1}B^{-1})(Bz) \\
&\overset{(a.iv)}{=} (z^\top B^\top)((B^\top)^{-1}B^{-1})(Bz) \\
&\overset{asoc.}{=} z^\top (B^\top (B^\top)^{-1})(B^{-1}B)z = z^\top I_d\, I_d\, z = z^\top z.
\end{aligned} \tag{11}$$

Aplicând acum proprietatea $(a.i)$, vom putea calcula p.d.f.-ul distribuţiei gaussiene asociate variabilei Z:

$$
\begin{aligned}
p_Z(z) &= p_X(x) \cdot \left|\det\left(\frac{\partial x}{\partial z}\right)\right| \\
&= \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)\left|\det\left(\frac{\partial x}{\partial z}\right)\right| \\
&\overset{(9)\,(10)\,(11)}{=} \frac{1}{(2\pi)^{d/2}\cancel{|\det(B)|}} \exp\left(-\frac{1}{2}z^\top z\right)\cancel{|\det(B)|} \\
&= \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}z^\top z\right) \overset{z^\top = z^\top I_d}{=} \frac{1}{(2\pi)^{d/2}(\det(I_d))^{1/2}} \exp\left(-\frac{1}{2}z^\top I_d z\right) \\
&= \mathcal{N}(z; \mathbf{0}, I_d).
\end{aligned}
$$

b. Arătaţi că funcţia $p_X(x; \mu, \Sigma)$ care a fost dată în enunţ este într-adevăr funcţie densitate de probabilitate (p.d.f.).

***Indicaţie* (2): Vă puteţi folosi de următoarele proprietăţi:**

***i.* Pentru cazul $d = 1$, funcţia $p_X(x; \mu, \sigma^2)$ este funcţie densitate de probabilitate.**
***ii.* În cazul în care matricea Σ este diagonală,**

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_d^2 \end{bmatrix}, \text{ iar } \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix},$$

expresia funcţiei de densitate gaussiană multivariată este identică cu produsul a d funcţii de densitate de tip gaussian, univariate şi independente, prima funcţie având media μ_1 şi varianţa σ_1^2, a doua funcţie având media μ_2 şi varianţa σ_2^2, ..., a d-a funcţie având media μ_d şi varianţa σ_d^2.

Soluţie

$p_X(x;\mu,\Sigma)$ este funcţie densitate de probabilitate (p.d.f.) dacă:

- $p_X(x;\mu,\Sigma) \geq 0, \forall x \in \mathbb{R}^d$
- $I \overset{not.}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} p_X(x;\mu,\Sigma)dx_d\ldots dx_2dx_1 = 1$.

Prima condiţie este satisfăcută, pentru că numitorul fracţiei [care este *factorul de normalizare*] din definiţia funcţiei de densitate de probabilitate $p_X(x;\mu,\Sigma)$ este pozitiv, iar $\exp(y) > 0$ pentru orice $y \in \mathbb{R}$.
În continuare vom verifica a doua condiţie.
În integrala I facem substituţia (schimbarea de variabilă): $z = B^{-1}(x-\mu)$, unde $\Sigma = BB^\top, B \in \mathbb{R}^{d\times d}$. Matricea B există şi este inversabilă după cum s-a precizat în enunţ (vedeţi *Observaţia* (1)).
De la punctul a rezultă imediat că:

$$\begin{aligned} I &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} p_Z(z;\mathbf{0},I_d)dz_d\ldots dz_2dz_1 \\ &\overset{(b.ii)}{=} \left(\int_{-\infty}^{\infty} p_{Z_1}(z_1;0,1)dz_1\right)\left(\int_{-\infty}^{\infty} p_{Z_2}(z_2;0,1)dz_2\right)\ldots\left(\int_{-\infty}^{\infty} p_{Z_d}(z_d;0,1)dz_d\right) \overset{(b.i)}{=} 1\cdot 1\cdot\ldots\cdot 1 = 1. \end{aligned}$$

Deci, şi a doua condiţie este satisfăcută, ceea ce înseamnă că funcţia $p_X(x;\mu,\Sigma)$ este într-adevăr funcţie densitate de probabilitate (p.d.f.).

Bi-variate Gaussian distributions. A property:

The conditional distributions $X_1|X_2$ and $X_2|X_1$ are also Gaussians.

The calculation of their parameters

Duda, Hart and Stork, *Pattern Classification*, 2001, Appendix A.5.2

[adapted by Liviu Ciortuz]

Fie X o variabilă aleatoare care urmează o distribuţie gaussiană bi-variată de parametri μ (vectorul de medii) şi Σ (matricea de covarianţă). Aşadar, $\mu = (\mu_1, \mu_2) \in \mathbb{R}^2$, iar $\Sigma \in \mathcal{M}_{2\times 2}(\mathbb{R})$.

Prin definiţie, $\Sigma = Cov(X, X)$, unde $X \overset{not.}{=} (X_1, X_2)$, aşadar $\Sigma_{ij} = Cov(X_i, X_j)$ pentru $i, j \in \{1, 2\}$. De asemenea, $Cov(X_i, X_i) = Var[X_i] \overset{not.}{=} \sigma_i^2 \geq 0$ pentru $i \in \{1, 2\}$, în vreme ce pentru $i \neq j$ avem $Cov(X_i, X_j) = Cov(X_j, X_i) \overset{not.}{=} \sigma_{ij}$. În sfârşit, dacă introducem „coeficientul de corelare“ $\rho \overset{def.}{=} \dfrac{\sigma_{12}}{\sigma_1 \sigma_2}$, rezultă că putem scrie astfel matricea de covarianţă:

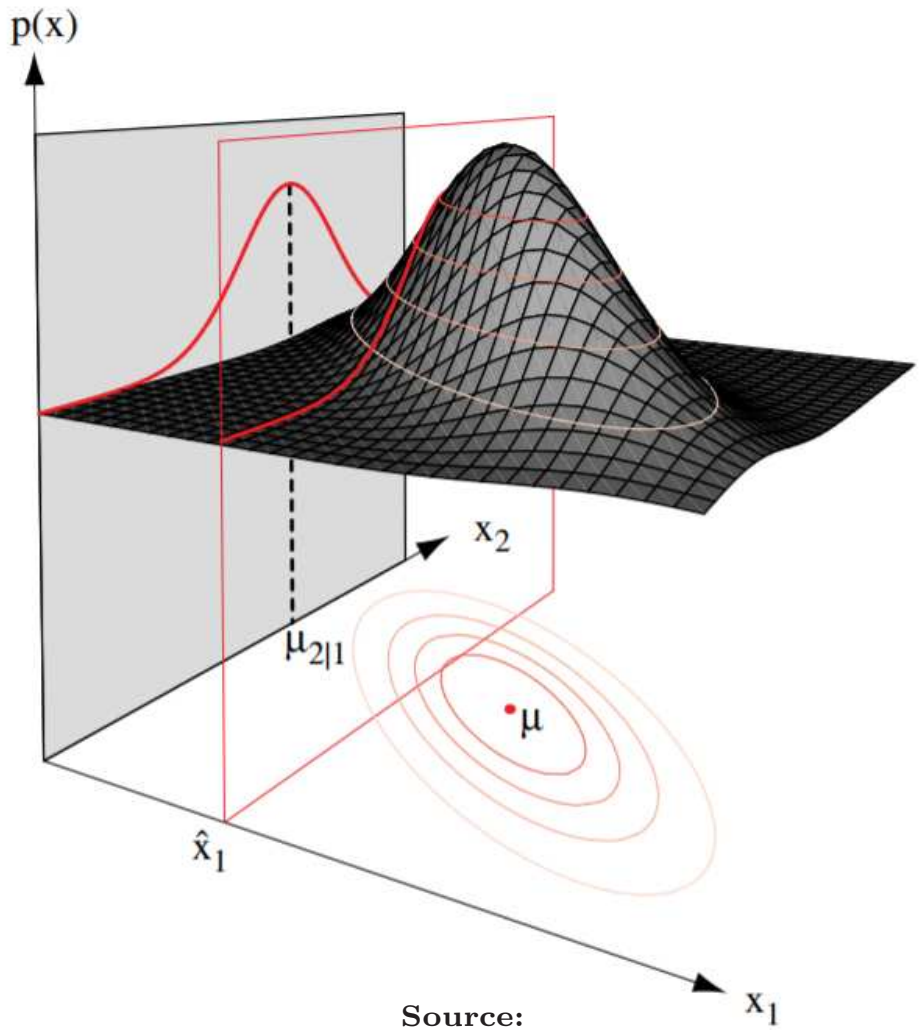
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{12}$$

Demonstraţi că ipoteza $X \sim \mathcal{N}(\mu, \Sigma)$, implică faptul că distribuţia condiţională $X_2|X_1$ este de tip gaussian, şi anume

$$X_2|X_1 = x_1 \sim \mathcal{N}(\mu_{2|1}, \sigma^2_{2|1}),$$

cu $\mu_{2|1} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$ **şi** $\sigma^2_{2|1} = \sigma^2_2(1-\rho^2)$.

***Observaţie*: Pentru $X_1|X_2$, rezultatul este similar:** $X_1|X_2 = x_2 \sim \mathcal{N}(\mu_{1|2}, \sigma^2_{1|2})$, **cu** $\mu_{1|2} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$ **şi** $\sigma^2_{1|2} = \sigma^2_1(1-\rho^2)$.



Source:

***Pattern Classification*, 2nd ed., Appendix A.5.2,**
R. Duda, P. Hart and D. Stork, 2001

Answer

$$p_{X_2|X_1}(x_2|x_1) \overset{def.}{=} \frac{p_{X_1,X_2}(x_1,x_2)}{p_{X_1}(x_1)}, \tag{13}$$

where

$$\begin{aligned} p_{X_1,X_2}(x_1,x_2) &= \frac{1}{(\sqrt{2\pi})^2\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) \text{ şi} \\ p_{X_1}(x_1) &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2\right). \end{aligned} \tag{14}$$

From (12) it follows that $|\Sigma| = \sigma_1^2\sigma_2^2(1-\rho^2)$. In order that $\sqrt{|\Sigma|}$ and Σ^{-1} be defined, it follows that $\rho \in (-1,1)$. Moreover, since $\sigma_1,\ \sigma_2 > 0$, we will have $\sqrt{|\Sigma|} = \sigma_1\sigma_2\sqrt{1-\rho^2}$.

$$\begin{aligned} \Sigma^{-1} &= \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\Sigma^* = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} \\ &= \frac{1}{(1-\rho^2)} \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix} \end{aligned}$$

So,

$$
\begin{aligned}
&p_{X_1,X_2}(x_1,x_2) = \\
&\quad = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho)^2}(x_1-\mu_1, x_2-\mu_2) \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \right) \\
&\quad = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \\
&\qquad \exp\left(-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1} \right) \left(\frac{x_2-\mu_2}{\sigma_2} \right) + \left(\frac{x_2-\mu_2}{\sigma_2} \right)^2 \right] \right) \qquad (15)
\end{aligned}
$$

By substitution (14) and (15) in the definition (13), we will get:

$$
\begin{aligned}
p(x_2|x_1) &= \frac{p_{X_1,X_2}(x_1,x_2)}{p_{X_1}(x_1)} \\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right) \\
&\quad \cdot\sqrt{2\pi}\sigma_1 \exp\left(\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right) \\
&= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x_2-\mu_2}{\sigma_2} - \rho\frac{x_1-\mu_1}{\sigma_1}\right)^2\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2}\left(\frac{x_2-[\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)]}{\sigma_2\sqrt{1-\rho^2}}\right)^2\right]
\end{aligned}
$$

Therefore,

$$X_2|X_1 = x_1 \sim \mathcal{N}(\mu_{2|1}, \sigma^2_{2|1}) \text{ with } \mu_{2|1} \overset{not.}{=} \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1) \text{ and } \sigma^2_{2|1} \overset{not.}{=} \sigma_2^2(1-\rho^2).$$

Using the Central Limit Theorem (the i.i.d. version)

to compute the *real error* of a classifier

CMU, 2008 fall, Eric Xing, HW3, pr. 3.3

Chris recently adopts a new (binary) classifier to filter email spams. He wants to quantitively evaluate how good the classifier is.

He has a small dataset of 100 emails on hand which, you can assume, are randomly drawn from all emails.

He tests the classifier on the 100 emails and gets 83 classified correctly, so the error rate on the small dataset is 17%.

However, the number on 100 samples could be either higher or lower than the real error rate just by chance.

With a confidence level of 95%, what is likely to be the range of the real error rate? Please write down all important steps.

(Hint: You need some approximation in this problem.)

Notations:

Let X_i, $i = 1, \ldots, n = 100$ be defined as:
$X_i = 1$ if the email i was incorrectly classified, and 0 otherwise;

$E[X_i] \overset{not.}{=} \mu \overset{not.}{=} e_{real}$; $\quad Var(X_i) \overset{not.}{=} \sigma^2$

$$e_{sample} \overset{not.}{=} \frac{X_1 + \ldots + X_n}{n} = 0.17$$

$$Z_n = \frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n}\,\sigma} \qquad \textbf{(the standardized form of } X_1 + \ldots + X_n\textbf{)}$$

Key insight:

Calculating the real error of the classifier (more exactly, a symmetric interval around the real error $p \overset{not.}{=} \mu$) with a "confidence" of 95% amounts to finding $a > 0$ such that $P(|Z_n| \leq a) \geq 0.95$.

***Calculus*:**

$$
\begin{aligned}
\mid Z_n \mid \leq a \quad &\Leftrightarrow \quad \left| \frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n}\,\sigma} \right| \leq a \;\Leftrightarrow\; \left| \frac{X_1 + \ldots + X_n - n\mu}{n\sigma} \right| \leq \frac{a}{\sqrt{n}} \\
&\Leftrightarrow \quad \left| \frac{X_1 + \ldots + X_n - n\mu}{n} \right| \leq \frac{a\sigma}{\sqrt{n}} \;\Leftrightarrow\; \left| \frac{X_1 + \ldots + X_n}{n} - \mu \right| \leq \frac{a\sigma}{\sqrt{n}} \\
&\Leftrightarrow \quad |e_{\boldsymbol{sample}} - e_{\boldsymbol{real}}| \leq \frac{a\sigma}{\sqrt{n}} \Leftrightarrow |e_{\boldsymbol{real}} - e_{\boldsymbol{sample}}| \leq \frac{a\sigma}{\sqrt{n}} \\
&\Leftrightarrow \quad -\frac{a\sigma}{\sqrt{n}} \leq e_{\boldsymbol{real}} - e_{\boldsymbol{sample}} \leq \frac{a\sigma}{\sqrt{n}} \\
&\Leftrightarrow \quad e_{\boldsymbol{sample}} - \frac{a\sigma}{\sqrt{n}} \leq e_{\boldsymbol{real}} \leq e_{\boldsymbol{sample}} + \frac{a\sigma}{\sqrt{n}} \\
&\Leftrightarrow \quad e_{\boldsymbol{real}} \in \left[e_{\boldsymbol{sample}} - \frac{a\sigma}{\sqrt{n}}\,,\; e_{\boldsymbol{sample}} + \frac{a\sigma}{\sqrt{n}} \right]
\end{aligned}
$$

***Important facts*:**

The Central Limit Theorem: $Z_n \to \mathcal{N}(0;1)$

Therefore, $P(|Z_n| \leq a) \approx P(|X| \leq a) = \Phi(a) - \Phi(-a)$**, where** $X \sim \mathcal{N}(0;1)$ **and** Φ **is the cumulative function distribution of** $\mathcal{N}(0;1)$**.**

***Calculus*:**

$\Phi(-a) + \Phi(a) = 1 \Rightarrow P(|\ Z_n\ | \leq a) = \Phi(a) - \Phi(-a) = 2\Phi(a) - 1$

$P(|\ Z_n\ | \leq a) = 0.95 \Leftrightarrow 2\Phi(a) - 1 = 0.95 \Leftrightarrow \Phi(a) = 0.975 \Leftrightarrow a \cong 1.97$ **(see** Φ **table)**

$\sigma^2 \overset{not.}{=} Var_{real} = e_{real}(1 - e_{real})$ **because** X_i **are Bernoulli variables. Futhermore, we can approximate** e_{real} **with** e_{semple}**, because**

$$E[e_{sample}] = e_{real} \text{ and } Var_{sample} = \frac{1}{n} Var_{real} \to 0 \text{ for } n \to +\infty,$$

cf. CMU, 2011 fall, T. Mitchell, A. Singh, HW2, pr. 1.ab.

***Finally*:**

$$\Rightarrow \frac{a\sigma}{\sqrt{n}} \approx 1.97 \cdot \frac{\sqrt{0.17(1-0.17)}}{\sqrt{100}} \cong 0.07$$

$$|e_{real} - e_{sample}| \leq 0.07 \ \Leftrightarrow\ |e_{real} - 0.17| \leq 0.07 \ \Leftrightarrow\ -0.07 \leq e_{real} - 0.17 \leq 0.07$$

$$\Leftrightarrow\ e_{real} \in [0.10,\ 0.24]$$

Exemplifying

a mixture of categorical distributions;

how to compute its expectation and variance

CMU, 2010 fall, Aarti Singh, HW1, pr. 2.2.1-2

Suppose that I have two six-sided dice, one is fair and the other one is loaded – having:

$$P(x) = \begin{cases} \frac{1}{2} & x = 6 \\ \frac{1}{10} & x \in \{1, 2, 3, 4, 5\} \end{cases}$$

I will toss a coin to decide which die to roll. If the coin flip is heads I will roll the fair die, otherwise the loaded one. The probability that the coin flip is heads is $p \in (0, 1)$.

a. What is the expectation of the *die roll* (in terms of p).

b. What is the variation of the *die roll* (in terms of p).

Solution:

a.

$$\begin{aligned} E[X] &= \sum_{i=1}^{6} i \cdot [P(i|fair) \cdot p + P(i|loaded) \cdot (1-p)] \\ &= \left[\sum_{i=1}^{6} i \cdot P(i|fair)\right] p + \left[\sum_{i=1}^{6} i \cdot P(i|loaded)\right] (1-p) \\ &= \frac{7}{2}p + \frac{9}{2}(1-p) = \frac{9}{2} - p \end{aligned}$$

b. Recall that we may write $Var(X) = E[X^2] - (E[X])^2$, therefore:

$$\begin{aligned}
E[X^2] &= \sum_{i=1}^{6} i^2 \cdot [P(i|fair) \cdot p + P(i|loaded) \cdot (1-p)] \\
&= \left[\sum_{i=1}^{6} i^2 \cdot P(i|fair)\right] p + \left[\sum_{i=1}^{6} i^2 \cdot P(i|loaded)\right] (1-p) \\
&= \frac{91}{6}p + \left(\frac{36}{2} + \frac{55}{10}\right)(1-p) \\
&= \frac{47}{2} - \frac{25}{3}p
\end{aligned}$$

Combining this with the result of the previous question yields:

$$\begin{aligned}
Var(X) &= E[X^2] - (E[X])^2 = \frac{141}{6} - \frac{50}{6}p - (\frac{9}{2} - p)^2 \\
&= \frac{141}{6} - \frac{50}{6}p - (\frac{81}{4} - 9p + p^2) \\
&= (\frac{141}{6} - \frac{81}{4}) - (\frac{50}{6} - 9)p - p^2 \\
&= \frac{13}{4} + \frac{2}{3}p - p^2
\end{aligned}$$

Estimating the parameters of some probability distributions:

Exemplifications

Estimating the parameter of the Bernoulli distribution: the MLE and MAP approaches

CMU, 2015 spring, Tom Mitchell, Nina Balcan, HW2, pr. 2

Suppose we observe the values of n i.i.d. (independent, identically distributed) random variables $X_1, \ldots, X_n$ drawn from a single Bernoulli distribution with parameter θ. In other words, for each X_i, we know that

$$P(X_i = 1) = \theta \quad \textbf{and} \quad P(X_i = 0) = 1 - \theta.$$

Our *goal* is to estimate the value of θ from the observed values of $X_1, \ldots, X_n$.

Reminder: Maximum Likelihood Estimation

For any hypothetical value $\hat{\theta}$, we can compute the probability of observing the outcome $X_1, \ldots, X_n$ if the true parameter value θ were equal to $\hat{\theta}$.

This probability of the observed data is often called the *data likelihood*, and the function $L(\hat{\theta})$ that maps each $\hat{\theta}$ to the corresponding likelihood is called the *likelihood function*.

A natural way to estimate the unknown parameter θ is to choose the $\hat{\theta}$ that maximizes the likelihood function. Formally,

$$\hat{\theta}_{MLE} = \underset{\hat{\theta}}{\operatorname{argmax}}\, L(\hat{\theta}).$$

a. Write a formula for the likelihood function, $L(\hat{\theta})$.

Your function should depend on the random variables $X_1, \ldots, X_n$ and the hypothetical parameter $\hat{\theta}$.

Does the likelihood function depend on the order of the random variables?

Solution:

Since the X_i are independent, we have

$$\begin{aligned} L(\hat{\theta}) &= P_{\hat{\theta}}(X_1, \ldots, X_n) = \prod_{i=1}^{n} P_{\hat{\theta}}(X_i) = \prod_{i=1}^{n} (\hat{\theta}^{X_i} \cdot (1-\hat{\theta})^{1-X_i}) \\ &= \hat{\theta}^{\#\{X_i=1\}} \cdot (1-\hat{\theta})^{\#\{X_i=0\}}, \end{aligned}$$

where $\#\{\cdot\}$ counts the number of X_i for which the condition in braces holds true.
Note that in the third equality we used the trick $X_i = \mathrm{I}_{\{X_i=1\}}$.

The likelihood function does not depend on the order of the data.

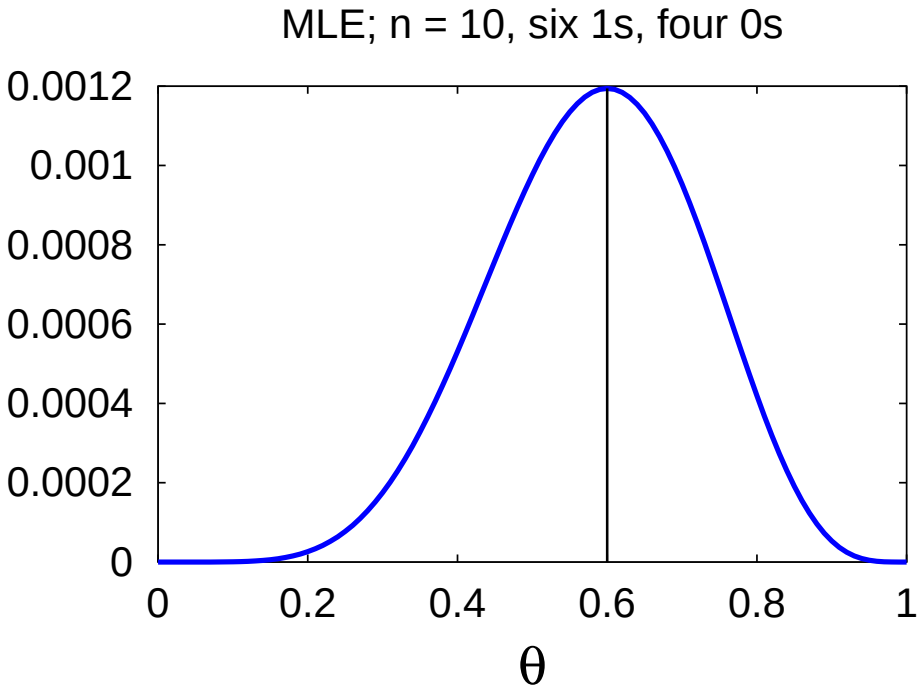
b. Suppose that $n = 10$ and the data set contains six 1s and four 0s.

Write a short computer program that plots the likelihood function of this data.

For the plot, the x-axis should be $\hat{\theta}$, and the y-axis $L(\hat{\theta})$. Scale your y-axis so that you can see some variation in its value.

Estimate $\hat{\theta}_{MLE}$ by marking on the x-axis the value of $\hat{\theta}$ that maximizes the likelihood.

Solution:



c. Find a closed-form formula for $\hat{\theta}_{MLE}$, the MLE estimate of $\hat{\theta}$. Does the closed form agree with the plot?

Solution:

Let's consider $l(\theta) = \ln(L(\theta))$. Since the $\ln$ function is increasing, the $\hat{\theta}$ that maximizes the log-likelihood is the same as the $\hat{\theta}$ that maximizes the likelihood. Using the properties of the $\ln$ function, we can rewrite $l(\hat{\theta})$ as follows:

$$l(\hat{\theta}) = \ln(\hat{\theta}^{n_1} \cdot (1-\hat{\theta})^{n_0}) = n_1 \ln(\hat{\theta}) + n_0 \ln(1-\hat{\theta}).$$

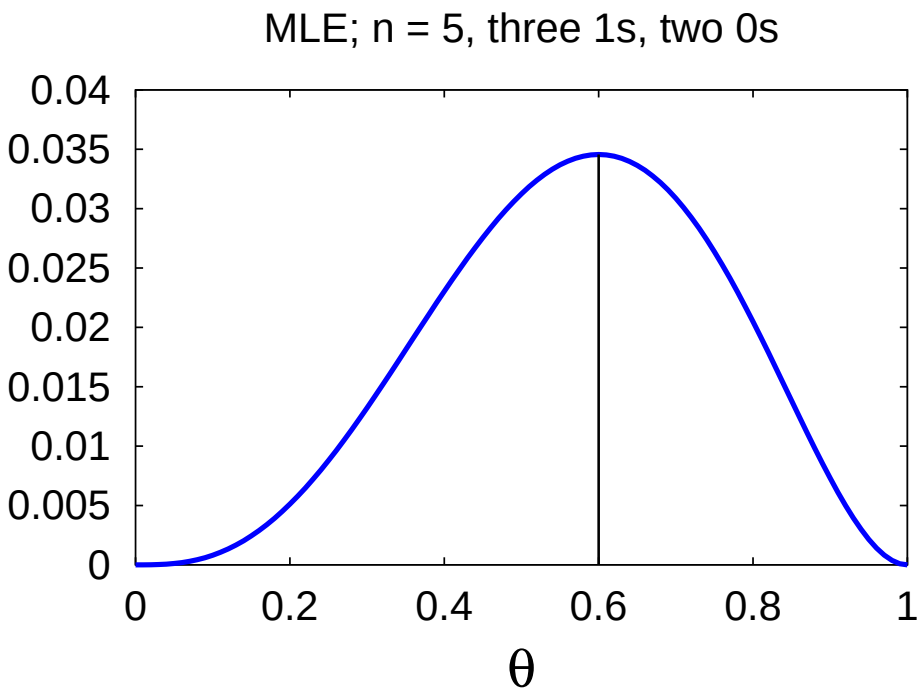
where $n_1 \overset{not.}{=} \#\{X_i = 1\}$, iar $n_0 \overset{not.}{=} \#\{X_i = 0\}$.
Assuming that $\hat{\theta} \neq 0$ and $\hat{\theta} \neq 1$, the first and second derivatives of l are given by

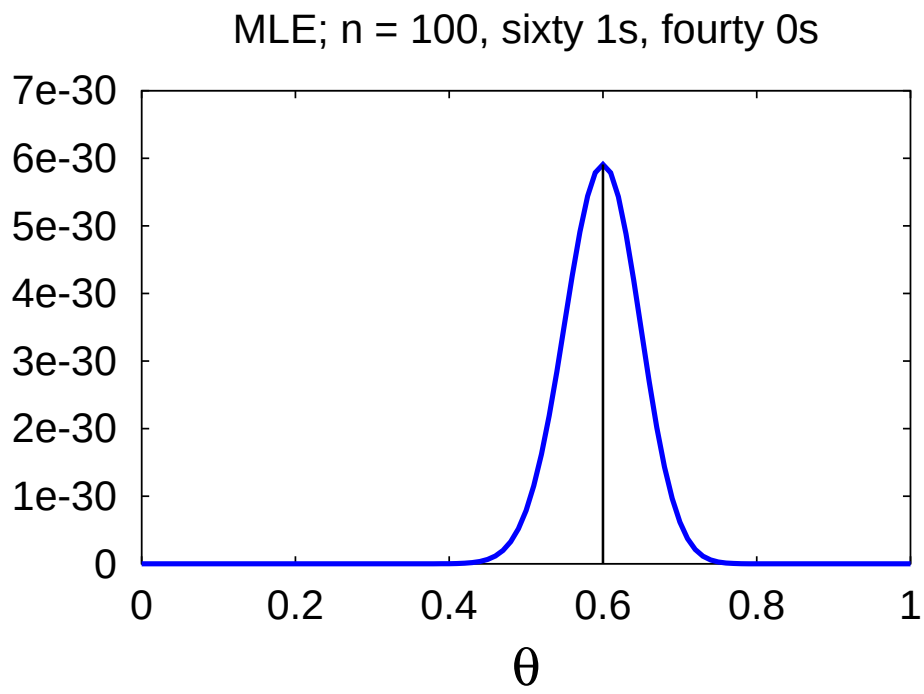
$$l'(\hat{\theta}) = \frac{n_1}{\hat{\theta}} - \frac{n_0}{1-\hat{\theta}} \quad \textbf{and} \quad l''(\hat{\theta}) = -\frac{n_1}{\hat{\theta}^2} - \frac{n_0}{(1-\hat{\theta})^2}$$

Since $l''(\hat{\theta})$ is always negative, the l function is concave, and we can find its maximizer by solving the equation $l'(\theta) = 0$.
The solution to this equation is given by $\hat{\theta}_{MLE} = \dfrac{n_1}{n_1 + n_0}$.

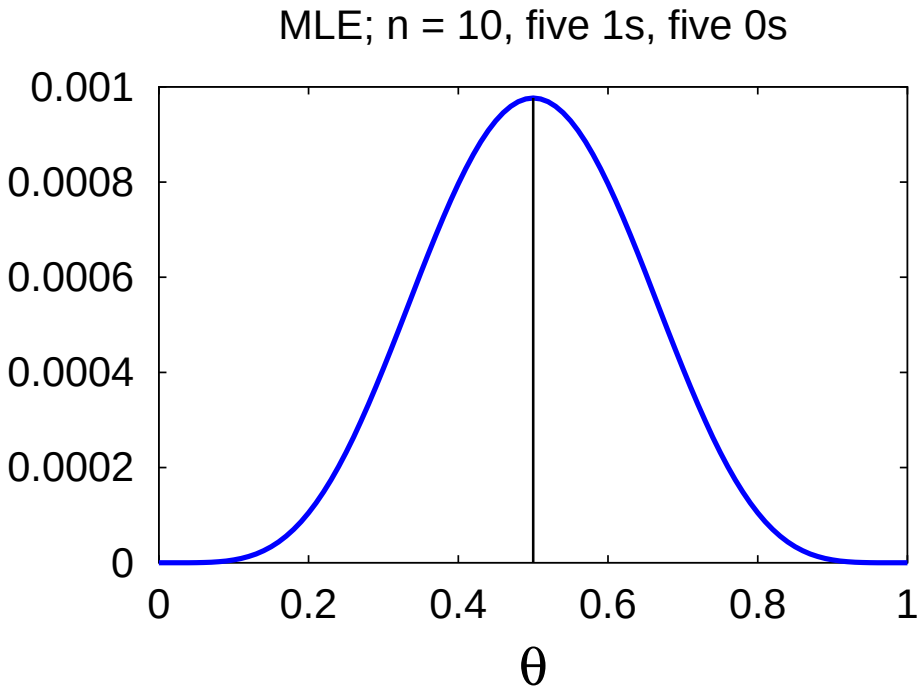
d. Create three more likelihood plots: one where $n = 5$ and the data set contains three 1s and two 0s; one where $n = 100$ and the data set contains sixty 1s and fourty 0s; and one where $n = 10$ and there are five 1s and five 0s.

Solution:





Solution (to part d.):



e. Describe how the likelihood functions and maximum likelihood estimates compare for the different data sets.

Solution (to part e.):

The MLE is equal to the proportion of 1s observed in the data, so for the first three plots the MLE is always at 0.6, while for the last plot it is at 0.5.

As the number of samples n increases, the likelihood function gets more peaked at its maximum value, and the values it takes on decrease.

Reminder: Maximum a Posteriori Probability Estimation

In the maximum likelihood estimate, we treated the true parameter value θ as a fixed (non-random) number. In cases where we have some prior knowledge about θ, it is useful to treat θ itself as a random variable, and express our prior knowledge in the form of a prior probability distribution over θ.

For *example*, suppose that the $X_1, \ldots, X_n$ are generated in the following way:

- **First, the value of θ is drawn from a given prior probability distribution**
- **Second, $X_1, \ldots, X_n$ are drawn independently from a Bernoulli distribution using this value for θ.**

Since both θ and the sequence $X_1, \ldots, X_n$ are random, they have a joint probability distribution. In this setting, a natural way to estimate the value of θ is to simply choose its most probable value given its prior distribution plus the observed data $X_1, \ldots, X_n$.

Definition:

$$\hat{\theta}_{MAP} = \underset{\hat{\theta}}{\operatorname{argmax}}\, P(\theta = \hat{\theta} | X_1, \ldots, X_n).$$

This is called the maximum a posteriori probability (MAP) estimate of θ.

Reminder (cont'd)

Using Bayes rule, we can rewrite the posterior probability as follows:

$$P(\theta = \hat{\theta}|X_1, \ldots, X_n) = \frac{P(X_1, \ldots, X_n|\theta = \hat{\theta})P(\theta = \hat{\theta})}{P(X_1, \ldots, X_n)}.$$

Since the probability in the denominator does not depend on $\hat{\theta}$, the MAP estimate is given by

$$\begin{aligned}\hat{\theta}_{MAP} &= \underset{\hat{\theta}}{\operatorname{argmax}}\, P(X_1, \ldots, X_n|\theta = \hat{\theta})P(\theta = \hat{\theta}) \\ &= \underset{\hat{\theta}}{\operatorname{argmax}}\, L(\hat{\theta})P(\theta = \hat{\theta}).\end{aligned}$$

In words, the MAP estimate for θ is the value $\hat{\theta}$ that maximizes the likelihood function multiplied by the prior distribution on θ.

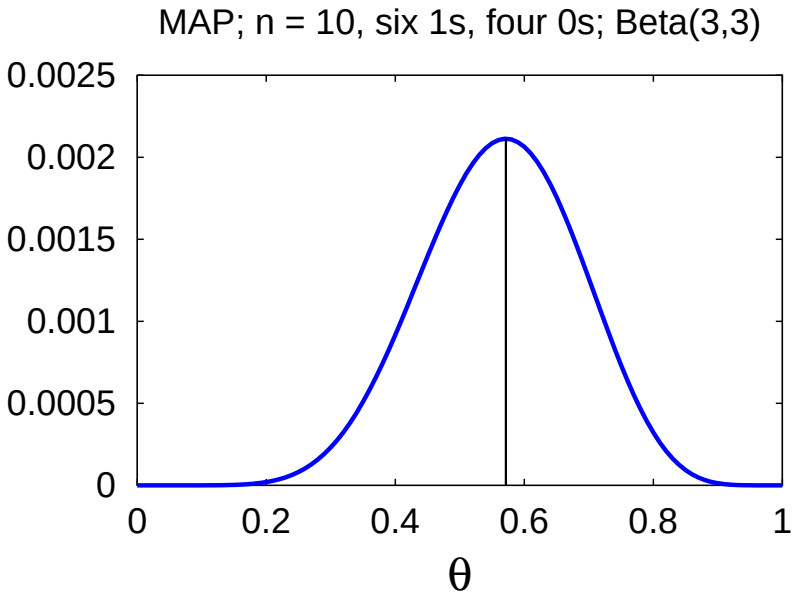
We will consider a $Beta(3,3)$ prior distribution for θ, which has the density function given by $p(\hat{\theta}) = \dfrac{\hat{\theta}^2(1-\hat{\theta})^2}{B(3,3)}$, where $B(\alpha,\beta)$ is the beta function and $B(3,3) \approx 0.0333$.

f. Suppose, as in part b, that $n = 10$ and we observed six 1s and four 0s.

Write a short computer program that plots the function $\hat{\theta} \mapsto L(\hat{\theta})\, p(\hat{\theta})$ for the same values of $\hat{\theta}$ as in part b.

Estimate $\hat{\theta}_{MAP}$ by marking on the x-axis the value of $\hat{\theta}$ that maximizes the function.

Solution:



g. Find a closed form formula for $\hat{\theta}_{MAP}$, the MAP estimate of $\hat{\theta}$. Does the closed form agree with the plot?

Solution:

As in the case of the MLE, we will apply the ln function before finding the maximizer. We want to maximize the function

$$l(\hat{\theta}) = \ln(L(\hat{\theta}) \cdot p(\hat{\theta})) = \ln(\hat{\theta}^{n_1+2} \cdot (1-\hat{\theta})^{n_0+2}) - \ln(B(3,3)).$$

The normalizing constant for the prior appears as an additive constant and therefore the first and second derivatives are identical to those in the case of the MLE (except with $n_1 + 2$ and $n_0 + 2$ instead of n_1 and n_0, respectively).

It follows that the closed form formula for the MAP estimate is given by

$$\hat{\theta}_{MAP} = \frac{n_1 + 2}{n_1 + n_0 + 4}.$$

This formula agrees with the plot obtained in part f.

h. Compare the MAP estimate to the MLE computed from the same data in part b. Briefly explain any significant difference.

Solution:

The MAP estimate is equal to the MLE with four additional virtual random variables, two that are equal to 1, and two that are equal to 0. This pulls the value of the MAP estimate closer to the value 0.5, which is why $\hat{\theta}_{MAP}$ is smaller than $\hat{\theta}_{MLE}$.

i. Comment on the relationship between the MAP and MLE estimates as n goes to infinity, while the ratio $\#\{X_i = 1\}/\#\{X_i = 0\}$ remains constant.

Solution:

It is obvious that as n goes to infinity, the influence of the 4 virtual random variables diminishes, and the two estimators become equal.

The MLE estimator for the parameter of the Bernoulli distribution: the bias and [an example of] inadmissibility

CMU, 2004 fall, Tom Mitchell, Ziv Bar-Joseph, HW2, pr. 1

Suppose X is a binary random variable that takes value 0 with probability p and value 1 with probability $1-p$. Let $X_1, \ldots, X_n$ be i.i.d. samples of X.

a. Compute an MLE estimate of p (denote it by $\hat{p}$).

Answer:

By way of definition,

$$\hat{p} = \operatorname*{argmax}_p P(X_1, \ldots, X_n | p) \overset{i.i.d.}{=} \operatorname*{argmax}_p P(X_i | p) = \operatorname*{argmax}_p p^k (1-p)^{n-k}$$

where k is the number of 0's in $x_1, \ldots, x_n$.
Furthermore, since $\ln$ is a monotonic (strictly increasing) function,

$$\hat{p} = \operatorname*{argmax}_p \ln p^k (1-p)^{n-k} = \operatorname*{argmax}_p (k \ln p + (n-k) \ln(1-p))$$

Computing the first derivative of $k \ln p + (n-k)\ln(1-p)$ w.r.t. p leads to:

$$\frac{\partial}{\partial p}(k \ln p + (n-k)\ln(1-p)) = \frac{k}{p} - \frac{n-k}{1-p}.$$

Hence,

$$\frac{\partial}{\partial p}(k \ln p + (n-k)\ln(1-p)) = 0 \Leftrightarrow \frac{k}{p} = \frac{n-k}{1-p} \Leftrightarrow \hat{p} = \frac{k}{n}.$$

b. Is $\hat{p}$ an unbiased estimate of p? Prove the answer.

Answer:

Since k can be seen as a sum of n [independent] Bernoulli variables of parameter p, we can write:

$$\begin{aligned} E[\hat{p}] &= E\left[\frac{k}{n}\right] = \frac{1}{n}E[k] = \frac{1}{n}E[n - \sum_{i=1}^{n} X_i] = \frac{1}{n}(n - \sum_{i=1}^{n} E[X_i]) \\ &= \frac{1}{n}(n - \sum_{i=1}^{n}(1-p)) = \frac{1}{n}(n - n(1-p)) = \frac{1}{n}np = p. \end{aligned}$$

Therefore, $\hat{p}$ is an *unbiased estimator* for the parameter p.

c. Compute the expected square error of $\hat{p}$ in terms of p.

Answer:

$$\begin{aligned} E[(\hat{p}-p)^2] &= E[\hat{p}^2] - 2E[\hat{p}]\,p + p^2 = \frac{E[k^2]}{n^2} - 2p^2 + p^2 \\ &= \frac{Var(k) + E^2[k]}{n^2} - p^2 = \frac{np(1-p) + (np)^2}{n^2} - p^2 = \frac{p}{n}(1-p) \end{aligned}$$

We used the fact that $Var(k) = E[k^2] - E^2[k]$, and also $Var(k) = np(1-p)$ because, as already said, k can be seen as a sum of n independent Bernoulli variables of parameter p.

Note that $E[\hat{p}] = p$ (cf. part b), therefore

$$E[(\hat{p}-p)^2] = E[(\hat{p}-E[\hat{p}])^2] = Var[\hat{p}] = \frac{p}{n}(1-p).$$

This implies that $Var[\hat{p}] \to 0$ for $n \to \infty$.

d. Prove that if you know that p lies in the interval $[1/4; 3/4]$ and you are given only $n = 3$ samples of X, then $\hat{p}$ is an *inadmissible estimator* of p when minimizing the expected square error of estimation.

***Note*: An estimator δ of a parameter θ is said to be *inadmissible* when there exists a different estimator δ' such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all θ and $R(\theta, \delta') < R(\theta, \delta)$ for some θ, where $R(\theta, \delta)$ is a *risk function* and in this problem it is the expected square error of the estimator.**

Answer:

Consider another estimator, $\tilde{p} = 1/2$.
$E[(\tilde{p} - p)^2] = (1/2 - p)^2$.

For $p = 1/2$ we have $E[(\tilde{p} - p)^2] = 0 < E[(\hat{p} - p)^2] = 1/12$.
We now need to show that $E[(\tilde{p} - p)^2] \leq E[(\hat{p} - p)^2]$ over $p \in [1/4; 3/4]$.

$$E[(\tilde{p} - p)^2] - E[(\hat{p} - p)^2] = \left(\frac{1}{2} - p\right)^2 - \frac{1}{3} \cdot p(1-p) = \frac{1}{4} - \frac{4}{3}p + \frac{4}{3}p^2.$$

This is a parabola going up, so we need to show that it lies below or equal to zero for $p \in [1/4; 3/4]$.

It is equivalent to showing that it is below or equal to 0 at boundary points.
In fact it is: $\frac{1}{4} - \frac{4}{3}p + \frac{4}{3}p^2 = 0$ for both $p = 1/4$ and $p = 3/4$.

Estimating the parameters of the categorical distribution: the MLE approach

CMU, 2009 spring, Ziv Bar-Joseph, HW1, pr. 2.3

In this problem we will derive the MLE for the parameters of a *categorical distribution* where the variable of interest, X, can take on k values, namely $a_1, a_2, \ldots, a_k$, the probability of seeing an event of type j being θ_j for $j = 1, \ldots, k$.

a. Given data describing n independent identically distributed *observations* of X, namely $d_1, \ldots, d_n$, each of which can be one of k values, express the *likelihood* of the data given $k-1$ parameters for the distribution over X. Let n_i represent the number of times X takes on value i in the data.

Answer:

The verosimility of the data is:

$$
\begin{aligned}
L(\theta) &= P(d_1, \ldots, d_n | \theta) \overset{i.i.d.}{=} \prod_{j=1}^{n} \sum_{i=1}^{k} (\theta_i I_{d_j = a_i}) && \text{(I is the indicator function)} \\
&= \prod_{i=1}^{k} \theta_i^{n_i} && (n_i \overset{not.}{=} \textstyle\sum_{j=1}^{n} I_{d_j = a_i}) \\
&= (\underbrace{1 - \sum_{i=1}^{k-1} \theta_i}_{\theta_k})^{n_k} \prod_{i=1}^{k-1} \theta_i^{n_i} && \text{(since the thetas sum to one)}
\end{aligned}
$$

b. Find $\hat{\theta}_j$, the MLE for θ_j, one of the $k-1$ parameters, by setting the partial derivative of the likelihood in part a with respect to θ_j equal to zero and solving for it.

***Hint*: You may want to start by first taking the log of the likelihood from part a before taking its derivative.**

Answer:

$$\ln L(\theta) = n_k \ln(1 - \sum_{i=1}^{k-1} \theta_i) + \sum_{i=1}^{k-1} n_i \ln \theta_i \Rightarrow \frac{\partial \ln L(\theta)}{\partial \theta_j} = -\frac{n_k}{1 - \sum_{i=1}^{k-1} \theta_i} + \frac{n_j}{\theta_j}$$

$$\frac{\partial \ln L(\theta)}{\partial \theta_j} = 0 \Leftrightarrow -\frac{n_k}{1 - \hat{\theta}_j - \sum_{i \neq j,k}^{k-1} \theta_i} + \frac{n_j}{\hat{\theta}_j} = 0 \Leftrightarrow \frac{n_j}{\hat{\theta}_j} = \frac{n_k}{1 - \hat{\theta}_j - \sum_{i \neq j,k}^{k-1} \theta_i}$$

$$\Leftrightarrow n_j(1 - \sum_{i \neq j,k}^{k-1} \theta_i) = (n_k + n_j)\hat{\theta}_j$$

$$\Leftrightarrow \hat{\theta}_j = \frac{n_j}{n_j + n_k}(1 - \sum_{i \neq j,k}^{k-1} \theta_i) \textbf{ for all } j \in \{1, \ldots, k-1\}.$$

c. At this point you should have $k-1$ equations describing MLEs of different parameters. Show how those equations imply that the MLE for a parameter θ_j representing the probability that X takes on value j is equal to $\frac{n_j}{n}$.

***Hint*: In order to remove the k-th parameter from the likelihood in part a you had to represent it with an equation, $\theta_k = f(\ )$. At this point you may find it helpful to replace all occurrences of $f(\ )$ with θ_k. After replacing $f(\ )$ with θ_k you can substitute all occurrences of each other parameter in $f(\ )$ with its MLE from part b. This should allow you to solve for the MLE of θ_k, which can then be used to simplify all of the other equations.**

Answer

As the likelihood function is uniquely optimal for the vector θ, the last equation in part b can be written as:

$$\hat{\theta}_j = \frac{n_j}{n_j + n_k}\left(1 - \sum_{i \neq j,k}^{k-1} \hat{\theta}_i\right) \Leftrightarrow \hat{\theta}_j = \frac{n_j}{n_j + n_k}(\hat{\theta}_j + \hat{\theta}_k) \text{ (because } \hat{\theta}_k = 1 - \sum_{i=1}^{k-1} \hat{\theta}_i)$$

$$\Leftrightarrow \quad \hat{\theta}_j\left(1 - \frac{n_j}{n_j + n_k}\right) = \frac{n_j}{n_j + n_k}\hat{\theta}_k \Leftrightarrow \hat{\theta}_j \frac{n_k}{n_j + n_k} = \frac{n_j}{n_j + n_k}\hat{\theta}_k$$

$$\Leftrightarrow \quad \hat{\theta}_j n_k = n_j \hat{\theta}_k$$

$$\Leftrightarrow \quad \hat{\theta}_j = \frac{n_j}{n_k}\hat{\theta}_k \text{ for all } j \in \{1, \ldots, k-1\}.$$

Finally,

$$\hat{\theta}_k = 1 - \hat{\theta}_1 - \ldots - \hat{\theta}_{k-1} = 1 - \frac{n_1}{n_k}\hat{\theta}_k - \ldots - \frac{n_{k-1}}{n_k}\hat{\theta}_k$$

$$\Rightarrow n_k\hat{\theta}_k = n_k - (n_1 + \ldots + n_{k-1})\hat{\theta}_k$$

$$\Rightarrow \hat{\theta}_k(\underbrace{n_1 + \ldots + n_{k-1} + n_k}_{n}) = n_k$$

$$\Rightarrow \hat{\theta}_k = \frac{n_k}{n}$$

$$\Rightarrow \hat{\theta}_j = \frac{n_j}{n_k} \cdot \frac{n_k}{n} = \frac{n_j}{n} \textbf{ for all } j \in \{1, \ldots, k-1\}.$$

***Note*: Even though [here] we can go from the non-hatted (θ_i) to the hatted form ($hat\theta_i$) of the equation in the first step of c, this will generally not be possible. To solve for a maximum likelihood criterion under additional constraints like the thetas summing to one, a generic and useful method is the method of Lagrange multipliers.**

The Gaussian [uni-variate] distribution: estimating μ when σ^2 is known

CMU, 2011 fall, Tom Mitchell, Aarti Singh, HW2, pr. 1

CMU, 2010 fall, Ziv Bar-Joseph, HW1, pr. 1.2-3

Assume we have n samples, $x_1, \ldots, x_n$, independently drawn from a normal distribution with *known* variance σ^2 and *unknown* mean μ.

a. Derive the MLE estimator for the mean μ.

Solution:

$$P(x_1, \ldots, x_n|\mu) = \prod_{i=1}^{n} P(x_i|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\Rightarrow \ln P(x_1, \ldots, x_n|\mu) = \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i-\mu)^2}{2\sigma^2} \right)$$

$$\Rightarrow \frac{\partial}{\partial \mu} P(x_1, \ldots, x_n|\mu) = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2}$$

$$\frac{\partial}{\partial \mu} P(x_1, \ldots, x_n|\mu) = 0 \Leftrightarrow \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \Leftrightarrow \sum_{i=1}^{n} (x_i - \mu) = 0 \Leftrightarrow \sum_{i=1}^{n} x_i = n\mu$$

$$\Rightarrow \mu_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$$

***Remark*: It can be easily shown that $\ln P(x_1, \ldots, x_n|\mu)$ indeed reaches its maximum for $\mu = \mu_{MLE}$.**

b. Show that $E[\mu_{MLE}] = \mu$.

Solution:

The sample $x_1, \ldots, x_n$ can be seen as the realization of n independent random variables $X_1, \ldots, X_n$ of Gaussian distribution of mean μ and variance σ^2. Then, due to the property of linearity for the expectation of random variables, we get:

$$E[\mu_{MLE}] = E\left[\frac{X_1 + \ldots + X_n}{n}\right] = \frac{E[X_1] + \ldots + E[X_n]}{n} = \frac{n\mu}{n} = \mu$$

Therefore, the μ_{MLE} estimator is unbiased.

c. What is $Var[\mu_{MLE}]$?

Solution:

$$Var[\mu_{MLE}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \overset{(\star)}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] \overset{i.i.d.}{=} n\frac{1}{n^2} Var[X_1] = \frac{\sigma^2}{n}$$

Therefore, $Var[\mu_{MLE}] \to 0$ as $n \to \infty$.

($\star$) **Remember that** $Var[aX] = a^2\, Var[X]$.

d. Now derive the MAP estimator for the mean μ. Assume that the prior distribution for the mean is itself a normal distribution with mean ν and variance β^2.

Solution 1:

$$P(\mu|x_1,\ldots,x_n) \overset{T.\ Bayes}{=} \frac{P(x_1,\ldots,x_n|\mu)\,P(\mu)}{P(x_1,\ldots,x_n)} \tag{16}$$

$$= \frac{\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \cdot \frac{1}{\sqrt{2\pi}\beta} e^{-\frac{(\mu-\nu)^2}{2\beta^2}}}{C} \tag{17}$$

where $C \overset{not.}{=} P(x_1,\ldots,x_n)$**.**

$$\Rightarrow \quad \ln P(\mu|x_1,\ldots,x_n) = -\sum_{i=1}^{n}\left(\ln\sqrt{2\pi}\sigma + \frac{(x_i-\mu)^2}{2\sigma^2}\right) - \ln\sqrt{2\pi}\beta - \frac{(\mu-\nu)^2}{2\beta^2} - \ln C$$

$$\Rightarrow \quad \frac{\partial}{\partial\mu}\ln P(\mu|x_1,\ldots,x_n) = \sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2} - \frac{\mu-\nu}{\beta^2}$$

$$\frac{\partial}{\partial\mu}\ln P(\mu|x_1,\ldots,x_n) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2} = \frac{\mu-\nu}{\beta^2} \Leftrightarrow \mu\left(\frac{1}{\beta^2}+\frac{n}{\sigma^2}\right) = \frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{\nu}{\beta^2}$$

$$\Rightarrow \quad \mu_{MAP} = \frac{\sigma^2\nu + \beta^2\sum_{i=1}^{n} x_i}{\sigma^2 + n\beta^2}$$

Solution 2:

Instead of computing the derivative of the posterior distribution $P(\mu|x_1,\ldots,x_n)$,
we will first show that the right hand side of (17) is itself a Gaussian, and then we will use the fact that the mean of a Gaussian is where it achieves its maximum value.

$$
\begin{aligned}
P(\mu|x_1,\ldots,x_n) &= \frac{1}{C}\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)\cdot\frac{1}{\sqrt{2\pi}\beta}e^{-\frac{(\mu-\nu)^2}{2\beta^2}} \\
&= const\ \cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}-\frac{(\mu-\nu)^2}{2\beta^2}} \\
&= const\ \cdot e^{-\frac{\beta^2\sum_{i=1}^{n}(x_i-\mu)^2+\sigma^2(\mu-\nu)^2}{2\sigma^2\beta^2}} \\
&= const\ \cdot e^{-\frac{n\beta^2+\sigma^2}{2\sigma^2\beta^2}\mu^2+\frac{\beta^2\sum_{i=1}^{n}x_i+\nu\sigma^2}{\sigma^2\beta^2}\mu-\frac{\beta^2\sum_{i=1}^{n}x_i^2+\nu^2\sigma^2}{2\sigma^2\beta^2}}
\end{aligned}
$$

$$
\begin{aligned}
P(\mu|x_1,\ldots,x_n) &= \\
&= const \cdot \exp\left(-\frac{\mu^2 - 2\mu\frac{\beta^2\sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2+\sigma^2} + \frac{\beta^2\sum_{i=1}^n x_i^2 + \nu^2\sigma^2}{n\beta^2+\sigma^2}}{\frac{2\sigma^2\beta^2}{n\beta^2+\sigma^2}} \right) \\
&= const \cdot \exp\left(-\frac{(\mu - \frac{\beta^2\sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2+\sigma^2})^2 - \left(\frac{\beta^2\sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2+\sigma^2}\right)^2 + \frac{\beta^2\sum_{i=1}^n x_i^2 + \nu^2\sigma^2}{n\beta^2+\sigma^2}}{2\frac{\sigma^2\beta^2}{n\beta^2+\sigma^2}} \right) \\
&= const \cdot \exp\left(-\frac{\left(\mu - \frac{\beta^2\sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2+\sigma^2}\right)^2}{2\frac{\sigma^2\beta^2}{n\beta^2+\sigma^2}} \right) \cdot \exp\left(\frac{\left(\frac{\beta^2\sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2+\sigma^2}\right)^2 - \frac{\beta^2\sum_{i=1}^n x_i^2 + \nu^2\sigma^2}{n\beta^2+\sigma^2}}{2\frac{\sigma^2\beta^2}{n\beta^2+\sigma^2}} \right) \\
&= const' \cdot \exp\left(-\frac{\left(\mu - \frac{\beta^2\sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2+\sigma^2}\right)^2}{2\frac{\sigma^2\beta^2}{n\beta^2+\sigma^2}} \right)
\end{aligned}
$$

The exp **term in the last equality being a Gaussian of mean** $\frac{\beta^2 \sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2 + \sigma^2}$ **and variance** $\frac{\sigma^2\beta^2}{n\beta^2 + \sigma^2}$**, it follows that its maximum is obtained for** $\mu = \frac{\beta^2 \sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2 + \sigma^2} = \mu_{MAP}$**.**

e. Please comment on what happens to the MLE and MAP estimators for the mean μ as the number of samples n goes to infinity.

Solution:

$$\begin{aligned}
\mu_{MLE} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
\mu_{MAP} &= \frac{\sigma^2\nu + \beta^2 \sum_{i=1}^{n} x_i}{\sigma^2 + n\beta^2} = \frac{\sigma^2\nu}{\sigma^2 + n\beta^2} + \frac{\beta^2 \sum_{i=1}^{n} x_i}{\sigma^2 + n\beta^2} \\
&= \frac{\sigma^2\nu}{\sigma^2 + n\beta^2} + \frac{\frac{1}{n}\sum_{i=1}^{n} x_i}{1 + \frac{\sigma^2}{n\beta^2}} = \frac{\sigma^2\nu}{\sigma^2 + n\beta^2} + \frac{\mu_{MLE}}{1 + \frac{\sigma^2}{n\beta^2}}
\end{aligned}$$

$$n \to \infty \Rightarrow \frac{\sigma^2\nu}{\sigma^2 + n\beta^2} \to 0 \textbf{ and } \frac{\sigma^2}{n\beta^2} \to 0 \Rightarrow \mu_{MAP} \to \mu_{MLE}$$

The Gaussian [uni-variate] distribution: estimating σ^2 when $\mu = 0$

CMU, 2009 spring, Ziv Bar-Joseph, HW1, pr. 2.1

Let X be a random variable distributed according to a Normal distribution with 0 mean, and σ^2 variance, i.e. $X \sim N(0, \sigma^2)$.

a. Find the maximum likelihood estimate for σ^2, i.e. σ^2_{MLE}.

Solution:

Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. $\sim N(0, \sigma^2)$. Let f be the density function corresponding to X. Then we can write the likelihood function as:

$$\begin{aligned} L(X_1, X_2, \ldots, X_n|\sigma^2) &= \prod_{i=1}^{n} f(X_i; \mu = 0, \sigma^2) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{(X_i - 0)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^{n} X_i^2}{2\sigma^2}\right) \\ \Rightarrow \ln L &= \text{constant} - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n} X_i^2 \\ \Rightarrow \frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} X_i^2. \text{ Therefore, } \frac{\partial \ln L}{\partial \sigma^2} = 0 \Leftrightarrow \sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{aligned}$$

Note: **It can be easily shown that $L(X_1, X_2, \ldots, X_n|\sigma^2)$ indeed reaches its maximum for $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$.**

b. Is the estimator you obtained biased?

Solution:

It is unbiased, since:

$$\begin{aligned} E[\tfrac{1}{n}\sum_{i=1}^{n} X_i^2] &= \frac{n}{n}E[X^2] && \text{since i.i.d.} \\ &= Var\,[X] + (E[X])^2 \\ &= Var\,[X] = \sigma^2 && \text{since } E[X] = 0 \end{aligned}$$

The Gaussian [uni-variate] distribution: estimating σ^2 (without restrictions on μ)

CMU, 2010 fall, Ziv Bar-Joseph, HW1, pr. 2.1.1-2

Let $\mathbf{x} \overset{not.}{=} (x_1, \ldots, x_n)$ be observed i.i.d. samples from a Gaussian distribution $N(x|\mu, \sigma^2)$.

a. Derive σ^2_{MLE}, the MLE for σ^2.

Solution:

The p.d.f. for $N(x|\mu, \sigma^2)$ has the form $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}$.

The log likelihood function of the data x is:

$$\begin{aligned} \ln \mathcal{L}(\mathbf{x} \mid \mu, \sigma^2) &= \ln \prod_{i=1}^{n} f(x_i) = \sum_{i=1}^{n} \left(-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\ &= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

The partial derivative of $\ln \mathcal{L}$ w.r.t. σ^2: $\dfrac{\partial \ln \mathcal{L}(\mathbf{x} \mid \mu, \sigma^2)}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2$.

Solving the equation $\dfrac{\partial \ln \mathcal{L}(\mathbf{x} \mid \mu, \sigma^2)}{\partial \sigma^2} = 0$, **we get:** $\sigma^2_{MLE} = \dfrac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{MLE})^2$.

Note that we had to take into account the optimal value of μ (see problem CMU, 2011 fall, T. Mitchell, A. Singh, HW2, pr. 1)

b. Show that $E[\sigma^2_{MLE}] = \frac{n-1}{n}\sigma^2$.

Solution:

$$
\begin{aligned}
E[\sigma^2_{MLE}] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{MLE})^2\right] = E[(x_1 - \mu_{MLE})^2] = E\left[(x_1 - \frac{1}{n}\sum_{i=1}^{n} x_i)^2\right] \\
&= E\left[x_1^2 - \frac{2}{n}x_1\sum_{i=1}^{n} x_i + \frac{1}{n^2}(\sum_{i=1}^{n} x_i)^2\right] \\
&= E\left[x_1^2 - \frac{2}{n}x_1\sum_{i=1}^{n} x_i + \frac{1}{n^2}\sum_{i=1}^{n} x_i^2 + \frac{2}{n^2}\sum_{i<j} x_i x_j\right] \\
&= E[x_1^2] + \frac{1}{n^2}\sum_{i=1}^{n} E[x_i^2] - \frac{2}{n}\sum_{i=1}^{n} E[x_1 x_i] + \frac{2}{n^2}\sum_{i<j} E[x_i x_j] \\
&= E[x_1^2] + \frac{1}{n^2} n E[x_1^2] - \frac{2}{n}E[x_1^2] - \frac{2}{n}(n-1)E[x_1 x_2] + \frac{2}{n^2}\frac{n(n-1)}{2}E[x_1 x_2] \\
&= \frac{n-1}{n}E[x_1^2] - \frac{n-1}{n}E[x_1 x_2]
\end{aligned}
$$

$$\sigma^2 = Var(x_1) = E[x_1^2] - (E[x_1])^2 = E[x_1^2] - \mu^2 \Rightarrow E[x_1^2] = \sigma^2 + \mu^2$$

Because x_1 and x_2 are independent, it follows that $Cov(x_1, x_2) = 0$. Therefore,

$$\begin{aligned} 0 &= Cov(x_1, x_2) = E[(x_1 - E[x_1])(x_2 - E[x_2])] = E[(x_1 - \mu)(x_2 - \mu)] \\ &= E[x_1 x_2] - \mu E[x_1 + x_2] + \mu^2 = E[x_1 x_2] - \mu(E[x_1] + E[x_2]) + \mu^2 \\ &= E[x_1 x_2] - \mu(2\mu) + \mu^2 = E[x_1 x_2] - \mu^2 \end{aligned}$$

So, $E[x_1 x_2] = \mu^2$.

By substituting $E[x_1^2] = \sigma^2 + \mu^2$ and $E[x_1 x_2] = \mu^2$ into the previously obtained equality ($E[\sigma_{MLE}] = \frac{n-1}{n} E[x_1^2] - \frac{n-1}{n} E[x_1 x_2]$), we get:

$$E[\sigma_{MLE}^2] = \frac{n-1}{n}(\sigma^2 + \mu^2) - \frac{n-1}{n}\mu^2 = \frac{n-1}{n}\sigma^2$$

c. Find an unbiased estimator for σ^2.

Solution:

It can be immediately proven that $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_{MLE})^2$ is an unbiased estimator of σ^2.

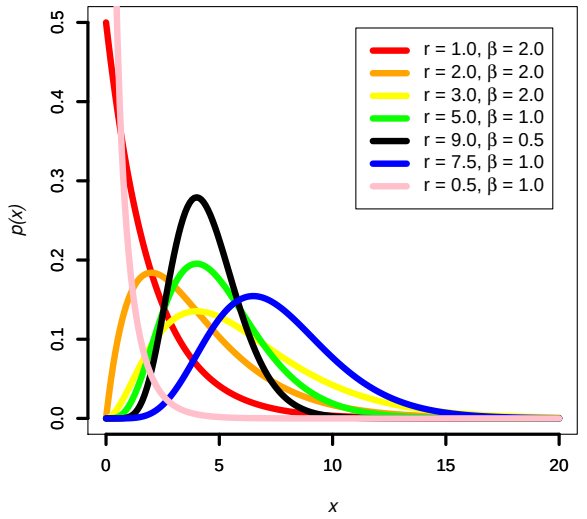
The Gamma distribution:
Maximum Likelihood Estimation of parameters

Liviu Ciortuz, 2017

The *Gamma distribution* of parameters $r > 0$ and $\beta > 0$ has the following *density function*:

$$Gamma(x|r,\beta) = \frac{1}{\beta^r \Gamma(r)} x^{r-1} e^{-\frac{x}{\beta}}, \text{ for all } x > 0,$$

where the Γ symbol designates Euler's *Gamma function*.



***Notes*:**

1. In the above definition, $\frac{1}{\beta^r \Gamma(r)}$ is the so-called *normalization factor*, since it does not depend on x, and $\int_{x=-\infty}^{+\infty} x^{r-1} e^{-\frac{x}{\beta}} dx = \beta^r \Gamma(r)$.

2. Euler's Gamma function is defined as follows: $\Gamma(r) = \int_0^{+\infty} t^{r-1} e^{-t} dt$, for all $r \in \mathbb{R}$, except for the negative integers. Starting from the definition of Γ, it can be easily shown that $\Gamma(r+1) = r\Gamma(r)$ for any $r > 0$, and also $\Gamma(1) = 1$. Therefore, $\Gamma(r+1) = r \cdot \Gamma(n) = r \cdot (r-1) \cdot \Gamma(r-1) = \ldots = r \cdot (r-1) \cdot \ldots \cdot 2 \cdot 1 = r!$, which means that the Γ function generalizes the *factorial function*.

3. The exponential distribution is a member of the Gamma family of distributions. (Just set r to 1 in Gamma's density function.)

Consider $x_1, \ldots, x_n \in \mathbb{R}^+$, all of them [having been] generated by one component of the above family of distributions.

Find the maximum likelihood estimation of the parameters r and β.

Solution

- **The verosimility function:**

$$\begin{aligned} L(r,\beta) &\overset{def.}{=} P(x_1,\ldots,x_n|r,\beta) \overset{i.i.d.}{=} \prod_{i=1}^{n} P(x_i|r,\beta) \\ &= \beta^{-rn}(\Gamma(r))^{-n}\left(\prod_{i=1}^{n} x_i\right)^{r-1} e^{-\frac{1}{\beta}\sum_{i=1}^{n} x_i} \end{aligned}$$

- **The log-verosimility function:**

$$\ell(r,\beta) \overset{def.}{=} \ln L(r,\beta) = -rn\ln\beta - n\ln\Gamma(r) + (r-1)\sum_{i=1}^{n}\ln x_i - \frac{1}{\beta}\sum_{i=1}^{n} x_i.$$

Now we will calculate the partial derivative of $\ell(r,\beta)$ w.r.t. β, and then equate it to 0:

$$\frac{\partial}{\partial\beta}\ell(r,\beta) = -\frac{rn}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} x_i = \frac{1}{\beta^2}\left[\sum_{i=1}^{n} x_i - rn\beta\right]$$

$$\frac{\partial}{\partial\beta}\ell(r,\beta) = 0 \Leftrightarrow \hat{\beta} = \frac{1}{rn}\sum_{i=1}^{n} x_i > 0.$$

By substituting $\hat{\beta}$ into $\ell(r,\beta)$, we will get:

$$\begin{aligned}
\ell(r,\hat{\beta}) &= -rn\ln\hat{\beta} - n\ln\Gamma(r) + (r-1)\sum_{i=1}^{n}\ln x_i - \frac{1}{\hat{\beta}}\sum_{i=1}^{n} x_i \\
&= rn\ln(rn) - rn\ln\sum_{i=1}^{n} x_i - n\ln\Gamma(r) + (r-1)\sum_{i=1}^{n}\ln x_i - \frac{rn}{\sum_{i=1}^{n} x_i}\cdot\sum_{i=1}^{n} x_i \\
&= rn\ln(rn) - rn\left(\ln\sum_{i=1}^{n} x_i + 1\right) - n\ln\Gamma(r) + (r-1)\sum_{i=1}^{n}\ln x_i
\end{aligned}$$

Therefore, by computing the partial derivative of $\ell(r, \hat{\beta})$ with respect to r, and then equating this derivative to 0, we will get:

$$\frac{\partial}{\partial r}\ell(r,\hat{\beta}) = 0 \Leftrightarrow n\ln(nr) + n - n\left(\ln\sum_{i=1}^{n} x_i + 1\right) - n\cdot\frac{\Gamma'(r)}{\Gamma(r)} + \sum_{i=1}^{n}\ln x_i = 0 \Leftrightarrow$$

$$n(\ln r - \psi(r)) = -n\ln n - \sum_{i=1}^{n}\ln x_i + n\ln\sum_{i=1}^{n} x_i \Leftrightarrow$$

$$\ln r - \psi(r) = -\ln n - \frac{1}{n}\sum_{i=1}^{n}\ln x_i + \ln\sum_{i=1}^{n} x_i.$$

The solution of the last equation is $\hat{r}$, the maximum likelihood estimation of the parameter r.

The Gaussian multi-variate distribution:

ML estimation of

the mean and the *precision matrix*, Λ

(Λ is the inverse of the *covariance matrix*, Σ)

CMU, 2010 fall, Aarti Singh, HW1, pr. 3.2.1

The density function of a d-dimensional Gaussian distribution is as follows:

$$\mathcal{N}(x \mid \mu, \Lambda^{-1}) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^\top \Lambda (x-\mu)\right)}{(2\pi)^{d/2}\sqrt{|\Lambda^{-1}|}},$$

where Λ is the inverse of the covariance matrix, or the so-called precision matrix. Let $\{x_1, x_2, \ldots, x_n\}$ be an i.i.d. sample from a d-dimensional Gaussian distribution.

Suppose that $n \gg d$. Derive the MLE estimates $\hat{\mu}$ and $\hat{\Lambda}$.

Hint

You may find useful the following formulas (taken from *Matrix Identities*, by Sam Roweis, 1999):

(2b) $|A^{-1}| = \dfrac{1}{|A|}$

(2e) $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$;[a]
more generally, $\mathrm{Tr}(ABC\ldots) = \mathrm{Tr}(BC\ldots A) = \mathrm{Tr}(C\ldots AB) = \ldots$

(3b) $\dfrac{\partial}{\partial X}\mathrm{Tr}(XA) = \dfrac{\partial}{\partial X}\mathrm{Tr}(AX) = A^\top$

(4b) $\dfrac{\partial}{\partial X}\ln|X| = (X^{-1})^\top = (X^\top)^{-1}$

(5c) $\dfrac{\partial}{\partial X}a^\top Xb = ab^\top$

(5g) $\dfrac{\partial}{\partial X}(Xa+b)^\top C(Xa+b) = (C+C^\top)(Xa+b)a^\top$

$\mathrm{Tr}(A)$**, the *trace* of an n-by-n square matrix** A**, is defined as the sum of the elements on the main diagonal (the diagonal from the upper left to the lower right) of** A**, i.e.,** $a_{11} + \ldots + a_{nn}$**.**

[a] See Theorem 1.3.d from *Matrix Analysis for Statistics*, 2017, James R. Schott.

Given the $x_1, \ldots, x_n$ data, the log-likelihood function is: 118.

$$\begin{aligned} l(\mu, \Lambda) &\overset{i.i.d.}{=} \ln \prod_{i=1}^{n} \mathcal{N}(x_i|\mu, \Lambda^{-1}) = \sum_{i=1}^{n} \ln \mathcal{N}(x_i|\mu, \Lambda^{-1}) \\ &= -\frac{nd}{2} \ln(2\pi) - \frac{n}{2} \ln |\Lambda^{-1}| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^\top \Lambda (x_i - \mu) \\ &\overset{(2b)}{=} -\frac{nd}{2} \ln(2\pi) + \frac{n}{2} \ln |\Lambda| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^\top \Lambda (x_i - \mu). \end{aligned}$$

For any fixed positive definite precision matrix Λ, the log-likelihood is a quadratic function of μ with a negative leading coefficient, hence a strictly concave function of μ. We then solve

$$\nabla_\mu l(\mu, \Lambda) = 0 \overset{(5g)}{\iff} -\frac{1}{2}(\Lambda + \Lambda^\top) \sum_{i=1}^{n} (x_i - \mu)(-1) = 0 \iff \Lambda \sum_{i=1}^{n} (x_i - \mu) = 0 \iff \Lambda \sum_{i=1}^{n} x_i = n\Lambda\mu. \tag{18}$$

From (18) we get, by the assumption that Λ is invertible, the following estimate of μ:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n},$$

which coincides with the *sample mean* $\bar{x}$ and is constant w.r.t. Λ.

Now that we have

$$l(\mu, \Lambda) \leq l(\hat{\mu}, \Lambda) \quad \forall \mu \in \mathbb{R}^d, \ \Lambda \textbf{ being positive definite,}$$

we continue to consider Λ by first plugging $\hat{\mu}$ back in the log-likelihood function (18):

$$l(\hat{\mu}, \Lambda) = -\frac{nd}{2}\ln(2\pi) + \frac{n}{2}\ln|\Lambda| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^\top \Lambda (x_i - \bar{x}) \tag{19}$$

$$\overset{(2e)}{=} -\frac{nd}{2}\ln(2\pi) + \frac{n}{2}(\ln|\Lambda| - \mathrm{Tr}(S\Lambda)), \tag{20}$$

where S is the *sample covariance matrix*: $S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$.

***Explanation*:**

$(x_i - \bar{x})^\top \Lambda (x_i - \bar{x})$ is a 1-by-1 matrix, therefore $(x_i - \bar{x})^\top \Lambda (x_i - \bar{x}) = \mathrm{Tr}((x_i - \bar{x})^\top \Lambda (x_i - \bar{x}))$, and using the (2e) rule, it can be further written as $\mathrm{Tr}((x_i - \bar{x})(x_i - \bar{x})^\top \Lambda)$.
Using another simple rule, $\mathrm{Tr}(A + B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$ (which can be easily proven), it follows that

$$\sum_{i=1}^{n}(x_i - \bar{x})^\top \Lambda (x_i - \bar{x}) = \sum_{i=1}^{n}\mathrm{Tr}((x_i - \bar{x})^\top \Lambda (x_i - \bar{x})) = \sum_{i=1}^{n}\mathrm{Tr}((x_i - \bar{x})(x_i - \bar{x})^\top \Lambda)$$
$$= \mathrm{Tr}\Big(\sum_{i=1}^{n}((x_i - \bar{x})(x_i - \bar{x})^\top \Lambda)\Big) = \mathrm{Tr}\Big(\Big(\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top\Big)\Lambda\Big) = \mathrm{Tr}((nS)\Lambda) = n\mathrm{Tr}(S\Lambda).$$

By the fact that $\ln|\Lambda|$ is strictly concave on the domain of positive definite Λ,[a] and that $\mathrm{Tr}(S\Lambda)$ is linear in Λ, we are able to find the maximum of the expression (20) by solving

$$\nabla_\Lambda l(\hat{\mu}, \Lambda) = 0,$$

which can be proven[b] to be equivalent to

$$\Lambda^{-1} - S = 0. \qquad (\textbf{Therefore, } \hat{\Sigma} = \hat{\Lambda}^{-1} = S.)$$

Since $n \gg d$, we can safely assume that S is invertible and get the following estimate:

$$\hat{\Lambda} = S^{-1}.$$

[a] See, for example, Section 3.1.5, *Convex Optimization*: `http://www.stanford.edu/~boyd/cvxbook/`.

[b] Applying (4b) and (3b) on (20) you'll get $(\Lambda^\top)^{-1} - S^\top$. Then $(\Lambda^\top)^{-1} - S^\top = 0 \Leftrightarrow (\Lambda^{-1})^\top = S^\top \Leftrightarrow \Lambda^{-1} = S$.

Notes

1. In the above derivation, we have ensured that the estimates μ and $\hat{\Lambda}$ are in the parameter space and satisfy

$$l(\mu, \Lambda) \leq l(\hat{\mu}, \Lambda) \leq l(\hat{\mu}, \hat{\Lambda}) \quad \forall \mu \in \mathbb{R}^d, \ \Lambda \text{ being positive definite,}$$

so they are the MLE estimates.

2. Instead of using the relation (20), i.e., working with the Tr functional, one could directly compute the partial derivative of $l(\hat{\mu}, \Lambda)$ in (19):

$$\begin{aligned} \nabla_\Lambda l(\hat{\mu}, \Lambda) &\overset{(4b),(5c)}{=} \frac{n}{2}(\Lambda^\top)^{-1} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^\top = \frac{n}{2}\Lambda^{-1} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^\top \\ &= \frac{n}{2}\Lambda^{-1} - \frac{n}{2}S. \end{aligned}$$

So,

$$\nabla_\Lambda l(\hat{\mu}, \Lambda) = 0 \Leftrightarrow \hat{\Lambda}^{-1} \overset{not.}{=} \hat{\Sigma} = S \overset{not.}{=} \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^\top.$$

Elements of Information Theory:

Some examples and then some useful proofs

Computing entropies and specific conditional entropies for discrete random variables

CMU, 2012 spring, R. Rosenfeld, HW2, pr. 2

On the roll of two six-sided fair dice,

a. Calculate the distribution of the sum (S) of the total.

b. The amount of *information* (or *surprise*) when seeing the outcome x for a random variable X is defined as $\log_2 \frac{1}{P(X=x)} = -\log_2 P(X=x)$. How surprised are you (in bits) to observe $S=2$, $S=11$, $S=5$, $S=7$?

c. Calculate the *entropy* of S [as the *expected value* of the random variable $-\log_2 P(X=x)$].

d. Let's say you throw the die one by one, and the first die shows 4. What is the entropy of S after this observation? Was any information gained / lost in the process? If so, calculate how much information (in bits) was lost or gained.

a.

S	2	3	4	5	6	7	8	9	10	11	12
$P(S)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

b.

$$
\begin{aligned}
\mathit{Information}(S=2) &= -\log_2(1/36) = \log_2 36 = 2\log_2 6 = 2(1+\log_2 3) \\
&= 5.169925001 \text{ bits} \\
\mathit{Information}(S=11) &= -\log_2 2/36 = \log_2 18 = 1 + 2\log_2 3 = 4.169925001 \text{ bits} \\
\mathit{Information}(S=5) &= -\log_2 4/36 = \log_2 9 = 2\log_2 3 = 3.169925001 \text{ bits} \\
\mathit{Information}(S=7) &= -\log_2 6/36 = \log_2 6 = 1 + \log_2 3 = 2.584962501 \text{ bits}
\end{aligned}
$$

c.

$$
\begin{aligned}
H(S) &= -\sum_{i=1}^{n} p_i \log_2 p_i \\
&= -\left(2 \cdot \frac{1}{36} \log_2 \frac{1}{36} + 2 \cdot \frac{2}{36} \log_2 \frac{2}{36} + 2 \cdot \frac{3}{36} \log_2 \frac{3}{36} + 2 \cdot \frac{4}{36} \log_2 \frac{4}{36} + \right. \\
&\quad \left. 2 \cdot \frac{5}{36} \log_2 \frac{5}{36} + \frac{6}{36} \log_2 \frac{6}{36}\right) \\
&= \frac{1}{36}(2 \log_2 36 + 4 \log_2 18 + 6 \log_2 12 + 8 \log_2 9 + 10 \log_2 \frac{36}{5} + 6 \log_2 6) \\
&= \frac{1}{36}(2 \log_2 6^2 + 4 \log_2 6 \cdot 3 + 6 \log_2 6 \cdot 2 + 8 \log_2 3^2 + 10 \log_2 \frac{6^2}{5} + 6 \log_2 6) \\
&= \frac{1}{36}(40 \log_2 6 + 20 \log_2 3 + 6 - 10 \log_2 5) \\
&= \frac{1}{36}(60 \log_2 3 + 46 - 10 \log_2 5) = 3.274401919 \text{ bits.}
\end{aligned}
$$

d.

S	2	3	4	5	6	7	8	9	10	11	12
$P(S\|...)$	0	0	0	1/6	1/6	1/6	1/6	1/6	1/6	0	0

$$H(S|\textit{First-die-shows-4}) = -6 \cdot \frac{1}{6} \log_2 \frac{1}{6} = \log_2 6 = 2.58 \text{ bits},$$

$$IG(S; \textit{First-die-shows-4}) = H(S) - H(S|\textit{First-die-shows-4}) = 3.27 - 2.58 = 0.69 \text{ bits}.$$

Computing entropies and average conditional entropies for discrete random variables

CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 3

A doctor needs to diagnose a person having cold (C). The primary factor he considers in his diagnosis is the outside temperature (T). The random variable C takes two values, *yes* / *no*, and the random variable T takes 3 values, *sunny*, *rainy*, *snowy*. The joint distribution of the two variables is given in following table.

	$T = sunny$	$T = rainy$	$T = snowy$
$C = no$	0.30	0.20	0.10
$C = yes$	0.05	0.15	0.20

a. Calculate the *marginal probabilities* $P(C)$, $P(T)$.

***Hint*: Use $P(X = x) = \sum_Y P(X = x; Y = y)$. For example,**

$$P(C = no) = P(C = no, T = sunny) + P(C = no, T = rainy) + P(C = no, T = snowy).$$

b. Calculate the *entropies* $H(C)$, $H(T)$.

c. Calculate the *average conditional entropies* $H(C|T)$, $H(T|C)$.

a. $P_C = (0.6,\, 0.4)$ **şi** $P_T = (0.35,\, 0.35,\, 0.30)$**.**

b.

$$\begin{aligned} H(C) &= 0.6 \log_2 \frac{5}{3} + 0.4 \log_2 \frac{5}{2} = \log_2 5 - 0.6 \log_2 3 - 0.4 = 0.971 \quad \text{bits} \\ H(T) &= 2 \cdot 0.35 \log_2 \frac{20}{7} + 0.3 \log_2 \frac{10}{3} \\ &= 0.7(2 + \log_2 5 - \log_2 7) + 0.3(1 + \log_2 5 - \log_2 3) \\ &= 1.7 + \log_2 5 - 0.7 \log_2 7 - 0.3 \log_2 3 = 1.581 \quad \text{bits.} \end{aligned}$$

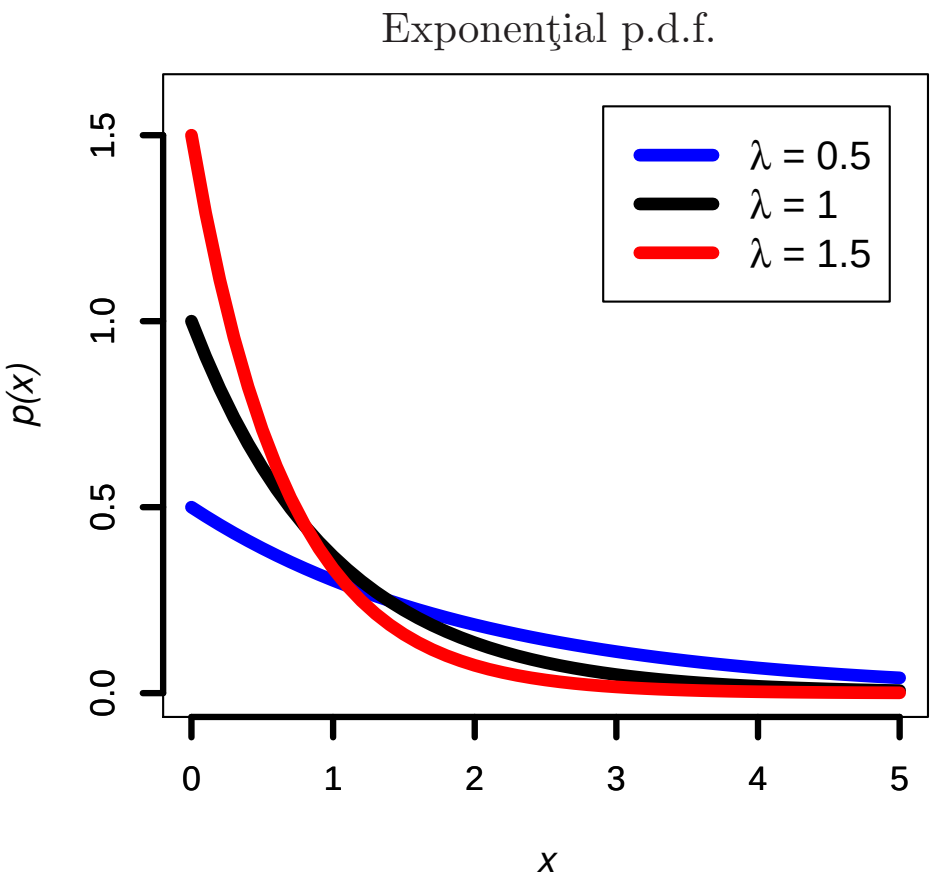
c.

$$
\begin{aligned}
H(C|T) &\overset{def.}{=} \sum_{t \in Val(T)} P(T=t) \cdot H(C|T=t) \\
&= P(T=sunny) \cdot H(C|T=sunny) + P(T=rainy) \cdot H(C|T=rainy) + \\
&\quad P(T=snowy) \cdot H(C|T=snowy) \\
&= 0.35 \cdot H\left(\frac{0.30}{0.30+0.05}, \frac{0.05}{0.30+0.05}\right) + 0.35 \cdot H\left(\frac{0.20}{0.20+0.15}, \frac{0.15}{0.20+0.15}\right) + \\
&\quad 0.30 \cdot H\left(\frac{0.10}{0.10+0.20}, \frac{0.20}{0.20+0.10}\right) \\
&= \frac{7}{20} \cdot H\left(\frac{6}{7}, \frac{1}{7}\right) + \frac{7}{20} \cdot H\left(\frac{4}{7}, \frac{3}{7}\right) + \frac{3}{10} \cdot H\left(\frac{1}{3}, \frac{2}{3}\right) \\
&= \frac{7}{20} \cdot \left(\frac{6}{7}\log_2\frac{7}{6} + \frac{1}{7}\log_2 7\right) + \frac{7}{20} \cdot \left(\frac{4}{7}\log_2\frac{7}{4} + \frac{3}{7}\log_2\frac{7}{3}\right) + \frac{3}{10} \cdot \left(\frac{1}{3}\log_2 3 + \frac{2}{3}\log_2\frac{3}{2}\right) \\
&= \frac{7}{20} \cdot \left(\log_2 7 - \frac{6}{7} - \frac{6}{7}\log_2 3\right) + \frac{7}{20} \cdot \left(\log_2 7 - \frac{8}{7} - \frac{3}{7}\log_2 3\right) + \frac{3}{10} \cdot \left(\log_2 3 - \frac{2}{3}\right) \\
&= \frac{7}{10}\log_2 7 - \left(\frac{3}{10} + \frac{4}{10} + \frac{2}{10}\right) - \left(\frac{6}{20} + \frac{3}{20} - \frac{3}{10}\right) \cdot \log_2 3 = \frac{7}{10}\log_2 7 - \frac{3}{20}\log_2 3 - \frac{9}{10} \\
&= 0.82715 \quad \text{bits.}
\end{aligned}
$$

$$
\begin{aligned}
H(T|C) &\stackrel{def.}{=} \sum_{c \in Val(C)} P(C=c) \cdot H(T|C=c) \\
&= P(C=no) \cdot H(T|C=no) + P(C=yes) \cdot H(T|C=yes) \\
&= 0.60 \cdot H\left(\frac{0.30}{0.30+0.20+0.10}, \frac{0.20}{0.30+0.20+0.10}, \frac{0.10}{0.30+0.20+0.10}\right) + \\
&\quad 0.40 \cdot H\left(\frac{0.05}{0.05+0.15+0.20}, \frac{0.15}{0.05+0.15+0.20}, \frac{0.20}{0.05+0.15+0.20}\right) \\
&= \frac{3}{5} \cdot H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) + \frac{2}{5} \cdot H\left(\frac{1}{8}, \frac{3}{8}, \frac{1}{2}\right) \\
&= \frac{3}{5}\left(\frac{1}{2} + \frac{1}{3}\log_2 3 + \frac{1}{6}(1+\log_2 3)\right) + \frac{2}{5}\left(\frac{1}{8}\cdot 3 + \frac{3}{8}(3-\log_2 3) + \frac{1}{2}\right) \\
&= \frac{3}{5}\left(\frac{2}{3} + \frac{1}{2}\log_2 3\right) + \frac{2}{5}\left(2 - \frac{3}{8}\log_2 3\right) \\
&= \frac{6}{5} + \frac{3}{20}\log_2 3 = 1.43774 \ \text{ bits.}
\end{aligned}
$$

Computing the entropy of the exponential distribution

CMU, 2011 spring, R. Rosenfeld, HW2, pr. 2.c



Pentru o distribuţie de probabilitate continuă P, entropia se defineşte astfel:

$$H(P) = \int_{-\infty}^{+\infty} P(x) \log_2 \frac{1}{P(x)} dx$$

Calculaţi entropia *distribuţiei* continue *exponenţiale* de parametru $\lambda > 0$. Definiţia acestei distribuţii este următoarea:

$$P(x) = \begin{cases} \lambda e^{-\lambda x}, & \textbf{dacă } x \geq 0; \\ 0, & \textbf{dacă } x < 0. \end{cases}$$

***Indicaţie*: Dacă $P(x) = 0$, veţi presupune că $-P(x) \log_2 P(x) = 0$.**

Answer

$$
\begin{aligned}
H(P) &= \int_{-\infty}^{0} P(x)\log_2\frac{1}{P(x)}\,dx + \int_{0}^{\infty} P(x)\log_2\frac{1}{P(x)}\,dx \\
&\overset{\text{def. } P}{=} \underbrace{\int_{-\infty}^{0} 0\log_2 0\,dx}_{0} + \int_{0}^{\infty} \lambda e^{-\lambda x}\log_2\frac{1}{\lambda e^{-\lambda x}}\,dx = \int_{0}^{\infty} \lambda e^{-\lambda x}\log_2\frac{1}{\lambda e^{-\lambda x}}\,dx \\
\Rightarrow H(P) &= \frac{1}{\ln 2}\int_{0}^{\infty} \lambda e^{-\lambda x}\ln\frac{1}{\lambda e^{-\lambda x}}\,dx = \frac{1}{\ln 2}\int_{0}^{\infty} \lambda e^{-\lambda x}\left(\ln\frac{1}{\lambda} + \ln\frac{1}{e^{-\lambda x}}\right)\,dx \\
&= \frac{1}{\ln 2}\int_{0}^{\infty} \lambda e^{-\lambda x}\left(-\ln\lambda + \ln e^{\lambda x}\right)\,dx \\
&= \frac{1}{\ln 2}\int_{0}^{\infty} \lambda e^{-\lambda x}\left(-\ln\lambda + \lambda x\right)\,dx \\
&= \frac{1}{\ln 2}\int_{0}^{\infty} \lambda e^{-\lambda x}(-\ln\lambda)\,dx + \frac{1}{\ln 2}\int_{0}^{\infty} \lambda e^{-\lambda x}\lambda x\,dx \\
&= \frac{-\ln\lambda}{\ln 2}\int_{0}^{\infty} \lambda e^{-\lambda x}\,dx + \frac{\lambda}{\ln 2}\int_{0}^{\infty} \lambda e^{-\lambda x} x\,dx \\
&= \frac{\ln\lambda}{\ln 2}\int_{0}^{\infty} \left(e^{-\lambda x}\right)'\,dx - \frac{\lambda}{\ln 2}\int_{0}^{\infty} \left(e^{-\lambda x}\right)' x\,dx
\end{aligned}
$$

Prima integrală se rezolvă foarte uşor:

$$\int_0^\infty \left(e^{-\lambda x}\right)' dx = e^{-\lambda x}\Big|_0^\infty = e^{-\infty} - e^0 = 0 - 1 = -1$$

Pentru a rezolva cea de-a doua integrală se poate folosi *formula de integrare prin părţi*:

$$\int_0^\infty \left(e^{-\lambda x}\right)' x\, dx = e^{-\lambda x}x\Big|_0^\infty - \int_0^\infty e^{-\lambda x}x'\, dx = e^{-\lambda x}x\Big|_0^\infty - \int_0^\infty e^{-\lambda x}\, dx$$

Integrala definită $e^{-\lambda x}x\Big|_0^\infty$ nu se poate calcula direct (din cauza conflictului $0 \cdot \infty$ care se produce atunci când lui x i se atribuie valoarea-limită ∞), ci se calculează folosind *regula lui l'Hôpital*:

$$\lim_{x\to\infty} xe^{-\lambda x} = \lim_{x\to\infty} \frac{x}{e^{\lambda x}} = \lim_{x\to\infty} \frac{x'}{\left(e^{\lambda x}\right)'} = \lim_{x\to\infty} \frac{1}{\lambda e^{\lambda x}} = \frac{1}{\lambda} \lim_{x\to\infty} e^{-\lambda x} = e^{-\infty} = 0,$$

deci

$$e^{-\lambda x}x\Big|_0^\infty = 0 - 0 = 0.$$

Integrala $\int_0^\infty e^{-\lambda x}\,dx$ se calculează uşor:

$$\int_0^\infty e^{-\lambda x}\,dx = -\frac{1}{\lambda}\int_0^\infty \left(e^{-\lambda x}\right)'\,dx = -\frac{1}{\lambda}\,e^{-\lambda x}\Big|_0^\infty = -\frac{1}{\lambda}(0-1) = \frac{1}{\lambda}$$

Prin urmare,

$$\int_0^\infty \left(e^{-\lambda x}\right)' x\,dx = 0 - \frac{1}{\lambda} = -\frac{1}{\lambda},$$

ceea ce conduce la rezultatul final:

$$H(P) = \frac{\ln\lambda}{\ln 2}(-1) - \frac{\lambda}{\ln 2}\left(-\frac{1}{\lambda}\right) = -\frac{\ln\lambda}{\ln 2} + \frac{1}{\ln 2} = \frac{1-\ln\lambda}{\ln 2}.$$

Entropie, entropie comună, entropie condiţională, câştig de informaţie: definiţii şi proprietăţi imediate

CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 2

Definiţii

- **Entropia variabilei X:**
  $H(X) \stackrel{def.}{=} -\sum_i P(X = x_i) \log P(X = x_i) \stackrel{not.}{=} E_X[-\log P(X)]$.

- **Entropia condiţională specifică a variabilei Y în raport cu valoarea x_k a variabilei X:**
  $H(Y \mid X = x_k) \stackrel{def.}{=} -\sum_j P(Y = y_j \mid X = x_k) \log P(Y = y_j \mid X = x_k)$
  $\stackrel{not.}{=} E_{Y|X=x_k}[-\log P(Y \mid X = x_k)]$.

- **Entropia condiţională medie a variabilei Y în raport cu variabila X:**
  $H(Y \mid X) \stackrel{def.}{=} \sum_k P(X = x_k) H(Y \mid X = x_k) \stackrel{not.}{=} E_X[H(Y \mid X)]$.

- **Entropia comună a variabilelor X şi Y:**
  $H(X, Y) \stackrel{def.}{=} -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(X = x_i, Y = y_j)$
  $\stackrel{not.}{=} E_{X,Y}[-\log P(X, Y)]$.

- **Informaţia mutuală a variabilelor X şi Y, numită de asemenea *câştigul de informaţie* al variabilei X în raport cu variabila Y (sau invers):**
  $MI(X, Y) \stackrel{not.}{=} IG(X, Y) \stackrel{def.}{=} H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$
  (Observaţie: ultima egalitate de mai sus are loc datorită rezultatului de la punctul c de mai jos.)

a.
$H(X) \geq 0$.

$$H(X) = -\sum_i P(X = x_i) \log P(X = x_i) = \sum_i \underbrace{P(X = x_i)}_{\geq 0} \underbrace{\log \frac{1}{P(X = x_i)}}_{\geq 0} \geq 0$$

Mai mult, $H(X) = 0$ dacă şi numai dacă variabila X este constantă:

„⇒“ Presupunem că $H(X) = 0$, adică $\sum_i P(X = x_i) \log \frac{1}{P(X = x_i)} = 0$. Datorită faptului că fiecare termen din această sumă este mai mare sau egal cu 0, rezultă că $H(X) = 0$ doar dacă pentru $\forall i$, $P(X = x_i) = 0$ sau $\log \frac{1}{P(X = x_i)} = 0$, adică dacă pentru $\forall i$, $P(X = x_i) = 0$ sau $P(X = x_i) = 1$. Cum însă $\sum_i P(X = x_i) = 1$ rezultă că există o singură valoare x_1 pentru X astfel încât $P(X = x_1) = 1$, iar $P(X = x) = 0$ pentru orice $x \neq x_1$. Altfel spus, variabila aleatoare discretă X este constantă.

„⇐“ Presupunem că variabila X este constantă, ceea ce înseamnă că X ia o singură valoare x_1, cu probabilitatea $P(X = x_1) = 1$. Prin urmare, $H(X) = -1 \cdot \log 1 = 0$.

b.

$$H(Y \mid X) = -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(Y = y_j \mid X = x_i)$$

$$\begin{aligned}
H(Y \mid X) &= \sum_i P(X = x_i) H(Y \mid X = x_i) \\
&= \sum_i P(X = x_i) \left[-\sum_j P(Y = y_j \mid X = x_i) \log P(Y = y_j \mid X = x_i) \right] \\
&= -\sum_i \sum_j \underbrace{P(X = x_i) P(Y = y_j \mid X = x_i)}_{= P(X = x_i, Y = y_j)} \log P(Y = y_j \mid X = x_i) \\
&= -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(Y = y_j \mid X = x_i)
\end{aligned}$$

c.

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

$$\begin{aligned}
H(X,Y) &= -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j) \\
&= -\sum_i \sum_j p(x_i) \cdot p(y_j \mid x_i) \log[p(x_i) \cdot p(y_j \mid x_i)] \\
&= -\sum_i \sum_j p(x_i) \cdot p(y_j \mid x_i)[\log p(x_i) + \log p(y_j \mid x_i)] \\
&= -\sum_i \sum_j p(x_i) \cdot p(y_j \mid x_i) \log p(x_i) - \sum_i \sum_j p(x_i) \cdot p(y_j \mid x_i) \log p(y_j \mid x_i) \\
&= -\sum_i p(x_i) \log p(x_i) \cdot \underbrace{\sum_j p(y_j \mid x_i)}_{=1} - \sum_i p(x_i) \sum_j p(y_j \mid x_i) \log p(y_j \mid x_i) \\
&= H(X) + \sum_i p(x_i) H(Y \mid X = x_i) = H(X) + H(Y \mid X)
\end{aligned}$$

Mai general (regula de înlănţuire):

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1, \ldots, X_{n-1})$$

$$\begin{aligned}
H(X_1, \ldots, X_n) = & \; E\left[\log \frac{1}{p(x_1, \ldots, x_n)}\right] \\
= & \; -E_{p(x_1,\ldots,x_n)}[\log \underbrace{p(x_1, \ldots, x_n)}_{p(x_1)\cdot p(x_2|x_1)\cdot \ldots \cdot p(x_n|x_1,\ldots,x_{n-1})}] \\
= & \; -E_{p(x_1,\ldots,x_n)}[\log p(x_1) + \log p(x_2 \mid x_1) + \ldots + \log p(x_n \mid x_1, \ldots, x_{n-1})] \\
= & \; -E_{p(x_1)}[\log p(x_1)] - E_{p(x_1,x_2)}[\log p(x_2 \mid x_1)] - \ldots \\
& \; -E_{p(x_1,\ldots,x_n)}[\log p(x_n \mid x_1, \ldots, x_{n-1})] \\
= & \; H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1, \ldots, X_{n-1})
\end{aligned}$$

An upper bound for the entropy of a discrete distribution

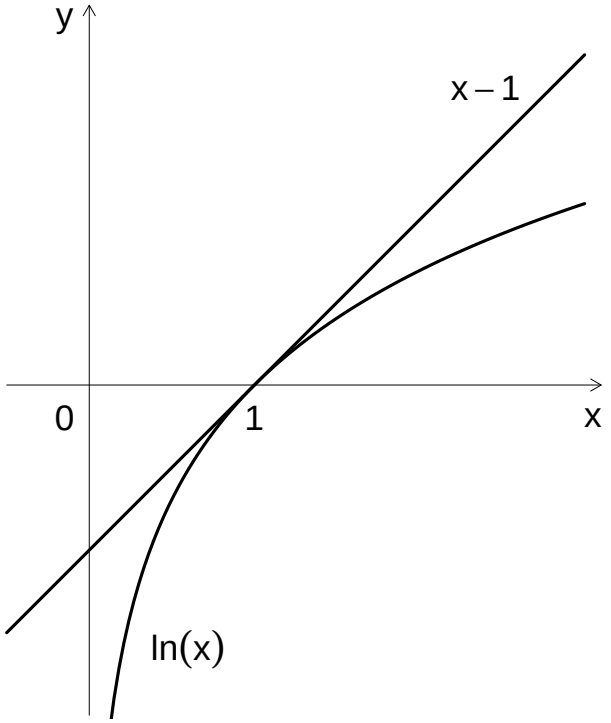
CMU, 2003 fall, T. Mitchell, A. Moore, HW1, pr. 1.1

Fie X o variabilă aleatoare discretă care ia n valori şi urmează distribuţia probabilistă P. Conform definiţiei, entropia lui X este

$$H(X) = -\sum_{i=1}^{n} P(X = x_i) \log_2 P(X = x_i).$$

Arătaţi că $H(X) \leq \log_2 n$.

***Sugestie*: Puteţi folosi inegalitatea $\ln x \leq x-1$, care are loc pentru orice $x > 0$.**



Answer

$$H(X) = \frac{1}{\ln 2}\left(-\sum_{i=1}^{n} P(X = x_i) \ln P(X = x_i)\right)$$

Aşadar,

$$\begin{aligned}
H(X) \leq \log_2 n \quad &\Leftrightarrow \quad \frac{1}{\ln 2}\left(-\sum_{i=1}^{n} P(X = x_i) \ln P(X = x_i)\right) \leq \log_2 n \\
&\Leftrightarrow \quad -\sum_{i=1}^{n} P(x_i) \ln P(x_i) \leq \ln n \\
&\Leftrightarrow \quad \sum_{i=1}^{n} P(x_i) \ln \frac{1}{P(x_i)} - \underbrace{\left(\sum_{i=1}^{n} P(x_i)\right)}_{1} \ln n \leq 0 \\
&\Leftrightarrow \quad \sum_{i=1}^{n} P(x_i) \ln \frac{1}{P(x_i)} - \sum_{i=1}^{n} P(x_i) \ln n \leq 0 \\
&\Leftrightarrow \quad \sum_{i=1}^{n} P(x_i) \left(\ln \frac{1}{P(x_i)} - \ln n\right) \leq 0 \\
&\Leftrightarrow \quad \sum_{i=1}^{n} P(x_i) \ln \frac{1}{n\, P(x_i)} \leq 0
\end{aligned}$$

Aplicând inegalitatea $\ln x \leq x - 1$ **pentru** $x = \dfrac{1}{n\,P(x_i)}$**, vom avea:**

$$\sum_{i=1}^{n} P(x_i) \ln \frac{1}{n\,P(x_i)} \leq \sum_{i=1}^{n} P(x_i) \left(\frac{1}{n\,P(x_i)} - 1 \right) = \sum_{i=1}^{n} \frac{1}{n} - \underbrace{\sum_{i=1}^{n} P(x_i)}_{1} = 1 - 1 = 0$$

***Observaţie*: Această margine superioară chiar este „atinsă". De exemplu, în cazul în care o variabilă aleatoare discretă** X **având** n **valori urmează distribuţia uniformă, se poate verifica imediat că** $H(X) = \log_2 n$**.**

A doua soluţie: folosind inegalitatea lui Jensen

Prezentăm acum o *altă demonstraţie* pentru inegalitatea $H(X) \leq \log_2 n$, pornind de la inegalitatea lui Jensen în formă probabilistă (vedeţi pr. CMU, 2003 fall, T. Mitchell, A. Moore, HW1, pr. 1.1.c):

Dacă X este o variabilă aleatoare şi f este o funcţie convexă, atunci $f(E[X]) \leq E[f(X)]$, iar dacă f este funcţie concavă, atunci $f(E[X]) \geq E[f(X)]$.

Considerând $x_1, \ldots, x_n \in \mathbb{R}$ şi înlocuind în a doua formă a inegalităţii lui Jensen dată mai sus funcţia f cu funcţia concavă $\log_2 x$ şi variabila X cu variabila aleatoare $\dfrac{1}{P(X = x)}$, obţinem inegalitatea:

$$\log_2 \Bigg(\sum_{i=1}^{n} \underbrace{\cancel{P(X = x_i)} \cdot \frac{1}{\cancel{P(X = x_i)}}}_{1} \Bigg) \geq \sum_{i=1}^{n} P(X = x_i) \log_2 \frac{1}{P(X = x_i)},$$

care se rescrie imediat sub forma $\log_2 n \geq H(X)$.

Folosind *Observaţia 1* din enunţul pr. CMU, 2003 fall, T. Mitchell, A. Moore, HW1, pr. 1.1 (observaţie valabilă şi pentru cazul funcţiilor concave) rezultă că $\log_2 n \geq H(X)$ dacă şi numai dacă $\dfrac{1}{P(X = x_1)} = \ldots = \dfrac{1}{P(X = x_n)}$, adică $P(X = x_i) = \dfrac{1}{n}$ pentru $i = 1, \ldots, n$.

A treia soluţie: folosind metoda multiplicatorilor lui Lagrange

Conform enunţului, fie X o variabilă aleatoare discretă n-ară ale cărei valori sunt luate cu probabilităţile $p_1, \ldots, p_n$ respectiv.

Ştim că $H(X) \overset{not.}{=} H(p_1, \ldots, p_n) \overset{def.}{=} \sum_{i=1}^{n} - \sum_i p_i \log_2 p_i$**.**

Vom arăta că $H(X) \leq \log_2 n$ **rezolvând următoarea *problemă de optimizare cu restricţii*, pe care o vom nota cu (P):**

$$\begin{array}{l} \max_{p_i \geq 0} H(p_1, \ldots, p_n) \\ \sum_i p_i = 1 \end{array} \quad \Leftrightarrow \quad \begin{array}{l} \min_{p_i \geq 0} -H(p_1, \ldots, p_n) \\ \sum_i p_i = 1 \end{array}$$

Facem *observaţia* că în continuare vom lăsa de o parte restricţiile $p_i \geq 0$ **pentru** $i = 1, \ldots, n$**, întrucât după rezolvarea problemei [simplificată astfel] va rezulta că aceste restricţii (şi, de asemenea, restricţiile** $p_i \leq 1$ **pentru** $i = 1, \ldots, n$**) sunt satisfăcute de către soluţia obţinută.**

Condiţiile Karush-Kuhn-Tucker **pentru problema (P) se scriu astfel:**

$$\textbf{fezabilitate primală: } \sum_i p_i = 1;$$

$$\textbf{fezabilitate duală: } \lambda \in \mathbb{R};$$

$$\textbf{complementaritate: } \lambda\big(\sum_i p_i - 1\big) = 0;$$

$$\textbf{staţionaritate / optimalitate: } \frac{\partial}{\partial p_i} L_P(\lambda, p_1, \ldots, p_n) = 0 \textbf{ pentru } i = 1, \ldots, n,$$

unde L_P **este** ***funcţia lagrangeană*****, definită astfel:**

$$L_P(\lambda, p_1, \ldots, p_n) = \sum_i p_i \log_2 p_i + \lambda\big(\sum_i p_i - 1\big).$$

Conform ***teoremei Karush-Kuhn-Tucker*****, dacă** $\lambda^*, p_1^*, \ldots, p_n^*$ **este soluţie a sistemului format de condiţiile Karush-Kuhn-Tucker enunţate mai sus [iar** $-H$ **este funcţie convexă, ceea ce vom demonstra mai jos], atunci** $p_1^*, \ldots, p_n^*$ **este soluţie a problemei (P).**

Rezolvând condiţiile de staţionaritate / optimalitate enunţate mai sus, vom avea:

$$\frac{\partial}{\partial p_i} L_P(\lambda, p_1, \ldots, p_n) = \log_2 p_i + \cancel{p_i} \cdot \frac{1}{\ln 2} \cdot \frac{1}{\cancel{p_i}} + \lambda = \log_2 p_i + \frac{1}{\ln 2} + \lambda$$

$$\frac{\partial L_P}{\partial p_i} = 0 \Leftrightarrow \lambda = -\log_2 p_i - \frac{1}{\ln 2}, \text{ constantă care nu depinde de } i.$$

Prin urmare, am obţinut $p_1^* = \ldots = p_n^* = 2^{-\lambda - \frac{1}{\ln 2}}$ **(constant). Cum** $\sum_i p_i^* = 1$**, rezultă că** $p_i^* = \frac{1}{n}$ **pentru** $i = 1, \ldots, n$ **(şi toate *condiţiile* de mai sus sunt satisfăcute).**

Este imediat că $H(p_1^*, \ldots, p_n^*) = n \cdot \frac{1}{n} \log_2 n = \log_2 n$**.**

Rămâne să mai arătăm că soluţia $p_1^* = \ldots = p_n^* = \dfrac{1}{n}$ corespunde maximului funcţiei $H(p_1, \ldots, p_n)$. Pentru aceasta, vom arăta că $-H$ este funcţie strict convexă în raport cu argumentele $p_1, \ldots, p_n$:

$$\frac{\partial(-H)}{\partial p_i} = (p_i \log_2 p_i)' = \log_2 p_i + \frac{1}{\ln 2} = \frac{\ln p_i}{\ln 2} + \frac{1}{\ln 2}.$$

Prin urmare, $\Rightarrow \dfrac{\partial^2(-H)}{\partial p_i^2} = \dfrac{1}{\ln 2} \cdot \dfrac{1}{p_i} > 0$**, iar** $\dfrac{\partial^2 H}{\partial p_j \partial p_i} = 0$ **pentru orice** $i \neq j$**.**

Aşadar, matricea hessiană a funcţiei $-H$ este [simetrică şi] pozitiv definită, ceea ce justifică faptul că $-H$ este funcţie strict convexă,[a] deci soluţia $p_1^* = \ldots = p_n^* = \dfrac{1}{n}$ corespunde unicului punct de minim al ei. Altfel spus, $p_1^* = \ldots = p_n^* = \dfrac{1}{n}$ corespunde unicului punct de maxim al funcţiei H, entropia variabilei aleatoare X.

[a] Vedeţi *Observaţia* de la finalul rezolvării punctului a de la problema CMU, 2015 fall, A. Smola, B. Poczos, HW1, pr. 3.1.

The particular form of the chain rule for entropies when X and Y are independent random variables:

$$H(X,Y) = H(X) + H(Y)$$

CMU, 2012 spring, R. Rosenfeld, HW2, pr. 7.b

Prove that if X and Y are independent random variables, the following property holds: $H(X,Y) = H(X) + H(Y)$.

Answer:

$$
\begin{aligned}
H(X,Y) &= -\sum_{x,y} P(X=x, Y=y) \log P(X=x, Y=y) \\
&\overset{indep.}{=} -\sum_{x,y} P(X=x)P(Y=y) \log(P(X=x)P(Y=y)) \\
&= -\sum_{x,y} P(X=x)P(Y=y)(\log P(X=x) + \log P(Y=y)) \\
&= -\Big(\sum_{x,y} P(X=x)P(Y=y) \log P(X=x)\Big) - \Big(\sum_{x,y} P(X=x)P(Y=y) \log P(Y=y)\Big) \\
&= -\Big(\sum_{y} P(Y=y) \sum_{x} P(X=x) \log P(X=x)\Big) - \Big(\sum_{x} P(X=x) \sum_{y} P(Y=y) \log P(Y=y)\Big) \\
&= -\Big(\sum_{y} P(Y=y)\Big) \cdot \Big(\sum_{x} P(X=x) \log P(X=x)\Big) - \Big(\sum_{x} P(X=x)\Big) \cdot \Big(\sum_{y} P(Y=y) \log P(Y=y)\Big) \\
&= -\sum_{x} P(X=x) \log P(X=x) - \sum_{y} P(Y=y) \log P(Y=y) \\
&= H(X) + H(Y)
\end{aligned}
$$

If X, Y are continue and independent random variables, then

$$p_{X,Y}(X = x, Y = y) = p_X(X = x)p_Y(Y = y)$$

for any x and y, where p denotes the p.d.f. corresponding to the [c.d.f. of] P. Therefore,

$$\begin{aligned}
H(X,Y) &\overset{def.}{=} -\int_X \int_Y p_{X,Y}(X = x, Y = y) \log p_{X,Y}(X = x, Y = y)\, dx\, dy \\
&\overset{indep.}{=} -\int_X \int_Y p_X(X = x)\, p_Y(Y = y)(\log p_X(X = x) + \log p_Y(Y = y))\, dx\, dy \\
&\overset{not.}{=} -\int_X \int_Y p_X(x)\, p_Y(y)(\log p_X(x) + \log p_Y(y))\, dx\, dy \\
&= -\int_X \int_Y p_Y(y)\, p_X(x) \log p_X(x) dx\, dy - \int_X \int_Y p_X(x)\, p_Y(y) \log p_Y(y) dx\, dy \\
&= -\int_X p_X(x) \log p_X(x) \underbrace{\left(\int_Y p_Y(y)\, dy \right)}_{1} dx - \int_X p_X(x) \underbrace{\left(\int_Y p_Y(y) \log p_Y(y)\, dy \right)}_{H(Y)} dx \\
&= -\int_X p_X(x) \log p_X(x)\, dx + H(Y) \underbrace{\left(\int_X p_X(x)\, dx \right)}_{1} \\
&= H(X) + H(Y)
\end{aligned}$$

The conditional form of

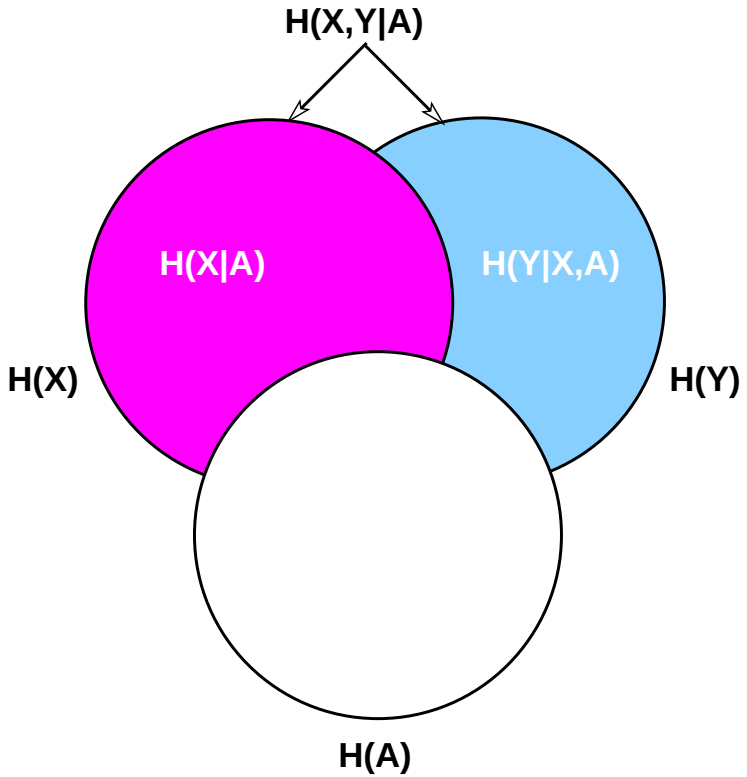
the simplest case of the chain rule for entropies ($n = 2$):

$$H(X, Y|A) = H(X|A) + H(Y|X, A)$$

CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 4.b

Prove that for any 3 discrete random variables X, Y and A, the following property holds:

$$\begin{aligned} H(X,Y|A) &= H(X|A) + H(Y|X,A) \\ &= H(Y|A) + H(X|Y,A). \end{aligned}$$



Answer:

$$\begin{aligned} H(X,Y|A) &= H(X,Y,A) - H(A) = H(X,Y,A) - H(Y,A) + H(Y,A) - H(A) \\ &= (H(X,Y,A) - H(Y,A)) + (H(Y,A) - H(A)) = H(X|Y,A) + H(Y|A) \\ &= H(Y|A) + H(X|Y,A). \end{aligned}$$

We've used the fact that $H(X,Y) = H(X|Y) + H(Y)$ (see CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 2).

Proving [in a direct manner] that the Information Gain is always positive or 0 and that $IG(X, Y) = 0 \Leftrightarrow X$ is independent of Y

(an indirect proof is made at CMU, 2007 fall, Carlos Guestrin, HW1, pr. 1.2)

Liviu Ciortuz, 2017

Definiţia *câştigului de informaţie* (sau: a *informaţiei mutuale*) al unei variabile aleatoare X în raport cu o altă variabilă aleatoare Y este

$$IG(X,Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X).$$

La CMU, 2007 fall, Carlos Guestrin, HW1, pr. 1.2 s-a demonstrat — pentru cazul în care X şi Y sunt discrete — că $IG(X,Y) = KL(P_{X,Y} || P_X\, P_Y)$, unde KL desemnează *entropia relativă* (sau: *divergenţa Kullback-Leibler*), P_X şi P_Y sunt distribuţiile variabilelor X şi, respectiv, Y, iar $P_{X,Y}$ este distribuţia comună a acestor variabile. Tot la CMU, 2007 fall, Carlos Guestrin, HW1, pr. 1.2 s-a arătat că divergenţa KL este întotdeauna ne-negativă. În consecinţă, $IG(X,Y) \geq 0$ pentru orice X şi Y.

a. Aici vă cerem să demonstraţi inegalitatea $IG(X,Y) \geq 0$ în manieră directă, plecând de la prima definiţie dată mai sus, fără a [mai] apela la divergenţa Kullback-Leibler.

b. Arătaţi tot într-o manieră directă că atunci când $IG(X,Y) = 0$ rezultă că X şi Y sunt independente. (Într-o manieră indirectă, acest rezultat a fost demonstrat la CMU, 2007 fall, Carlos Guestrin, HW1, pr. 1.2.c.)

***Sugestie*: Puteţi folosi următoarea formă a inegalităţii lui Jensen:**

$$\sum_{i=1}^{n} a_i \log x_i \leq \log \left(\sum_{i=1}^{n} a_i x_i \right)$$

unde baza logaritmului se consideră supraunitară, $a_i \geq 0$ **pentru** $i = 1, \ldots, n$ **şi** $\sum_{i=1}^{n} a_i = 1$**.**

***Observaţie*: Avantajul la această problemă, comparativ cu CMU, 2007 fall, Carlos Guestrin, HW1, pr. 1.2.a, este că aici se lucrează cu o singură distribuţie (p), nu cu două distribuţii (p şi q). Totuşi, demonstraţia de aici va fi mai laborioasă.**

Answer (in Romanian)

a. Presupunem că valorile variabilei X sunt $x_1, x_2, \ldots, x_n$, iar valorile variabilei Y sunt $y_1, y_2, \ldots, y_m$. Avem:

$$\begin{aligned} IG(X,Y) &\overset{def.}{=} H(X) - H(X|Y) \\ &\overset{def.}{=} \sum_{i=1}^{n} -P(x_i) \log_2 P(x_i) - \sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} (-P(x_i|y_j) \log_2 P(x_i|y_j)) \end{aligned}$$

$$
\begin{aligned}
-IG(X,Y) &= \sum_{i=1}^{n} P(x_i)\log_2 P(x_i) - \sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} P(x_i|y_j)\log_2 P(x_i|y_j) \\
&\overset{def.}{=} \\
&\overset{prob.\ marg.}{=} \sum_{i=1}^{n} \left(\sum_{j=1}^{m} P(x_i,y_j) \right) \log_2 P(x_i) - \sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} P(x_i|y_j)\log_2 P(x_i|y_j) \\
&\overset{distrib.\cdot,+}{=} \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j)\log_2 P(x_i) - \sum_{j=1}^{m}\sum_{i=1}^{n} P(y_j)P(x_i|y_j)\log_2 P(x_i|y_j) \\
&\overset{def.}{=} \\
&\overset{prob.\ cond.}{=} \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j)\log_2 P(x_i) - \sum_{j=1}^{m}\sum_{i=1}^{n} P(x_i,y_j)\log_2 P(x_i|y_j) \\
&\overset{distrib.\cdot,+}{=} \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j)(\log_2 P(x_i) - \log_2 P(x_i|y_j)) \\
&\overset{prop.}{=} \\
&\overset{log.}{=} \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j)\log_2 \frac{P(x_i)}{P(x_i|y_j)} \overset{reg.\ de\ multipl.}{=} \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i|y_j)P(y_j)\log_2 \frac{P(x_i)}{P(x_i|y_j)} \\
&\overset{distrib.\cdot,+}{=} \sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} \underbrace{P(x_i|y_j)}_{a_i} \log_2 \frac{P(x_i)}{P(x_i|y_j)}
\end{aligned}
$$

Întrucât pe de o parte $P(x_i|y_j) \geq 0$ şi pe de altă parte $\sum_{i=1}^{n} P(x_i|y_j) = 1$ pentru fiecare valoare y_j a lui Y în parte, putem aplica inegalitatea lui Jensen pentru cea de-a doua sumă din ultima expresie de mai sus — mai exact, pentru fiecare valoare a indicelui j în parte — şi obţinem:

$$\sum_{i=1}^{n} P(x_i|y_j) \log_2 \frac{P(x_i)}{P(x_i|y_j)} \leq \log_2 \underbrace{\left(\sum_{i=1}^{n} P(x_i) \right)}_{1} = 0.$$

Aşadar,

$$-IG(X,Y) = \sum_{j=1}^{m} P(y_j) \log_2 \left(\sum_{i=1}^{n} P(x_i|y_j) \frac{P(x_i)}{P(x_i|y_j)} \right) \leq \sum_{j=1}^{m} P(y_j) \cdot 0 = 0$$

Prin urmare, $IG(X,Y) \geq 0$.

b. Din relaţia $IG(X,Y) = -\sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} P(x_i|y_j) \log_2 \frac{P(x_i)}{P(x_i|y_j)}$ care a fost obţinută pe slide-ul precedent, decurge imediat următoarea consecinţă: atunci când X este indepenedent de Y, întrucât $P(x_i) = P(x_i|y_j)$ pentru orice i şi j, urmează că $\log_2 \frac{P(x_i)}{P(x_i|y_j)} = 0$, deci $IG(X,Y) = 0$.

Invers, vom demonstra acum că
$IG(X,Y)=0$ **implică faptul că**
X **este indepenedent de** Y**.**

Pentru aceasta vom folosi teorema lui Jensen.

Într-adevăr, din egalitatea

$$-IG(X,Y)=\sum_{j=1}^{m}P(y_j)\underbrace{\sum_{i=1}^{n}P(x_i|y_j)\log_2\frac{P(x_i)}{P(x_i|y_j)}}_{\leq 0}$$

avem

$$IG(X,Y)=0 \overset{P(y_j)>0}{\Rightarrow} \sum_{i=1}^{n}P(x_i|y_j)\log_2\frac{P(x_i)}{P(x_i|y_j)}=0,\forall j=1,\ldots,m.$$

Este însă adevărat şi că

$$\log_2\sum_{i=1}^{n}\cancel{P(x_i|y_j)}\frac{P(x_i)}{\cancel{P(x_i|y_j)}}=\log_2\underbrace{\sum_{i=1}^{n}P(x_i)}_{1}=0,\forall j=1,\ldots,m.$$

Prin urmare,

$$\sum_{i=1}^{n}\underbrace{P(x_i|y_j)}_{a_i}\log_2\frac{P(x_i)}{P(x_i|y_j)}=\log_2\sum_{i=1}^{n}P(x_i|y_j)\frac{P(x_i)}{P(x_i|y_j)}=0,\forall j=1,\ldots,m.$$

Aplicând pentru fiecare valoare a indicelui j în parte [ultima parte din] teorema lui Jensen — ceea ce este posibil, întrucât $a_i\geq 0$ şi $\sum_{i=1}^{n}a_i=1$ pentru fiecare valoare fixată a lui j —, rezultă că

$$\frac{P(x_i)}{P(x_i|y_j)}=\alpha\ (\textbf{const.}),\forall i=1,\ldots,n.$$

Aşadar,

$$P(x_i)=\alpha P(x_i|y_j)\ \forall i=1,\ldots,n.$$

Însumând aceste egalităţi pentru $i=1,\ldots,n$, rezultă $\sum_{i=1}^{n}P(x_i)=\alpha\sum_{i=1}^{n}P(x_i|y_j)$, deci $1=\alpha$.
Deci $P(x_i)=P(x_i|y_j)$, $\forall i=1,\ldots,n$ [şi pentru fiecare $j=1,\ldots,m$], deci X este independent de Y.

Corollary

Since

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y),$$

and also due to the definition of the Information Gain,

$$IG(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X),$$

it follows that

$$\begin{aligned} X \textbf{ independent of } Y &\Leftrightarrow H(X) = H(X|Y) \\ &\Leftrightarrow H(Y) = H(Y|X) \\ &\Leftrightarrow H(X,Y) = H(X) + X(Y) \end{aligned}$$

Cross-entropy (CH): definition, some basic properties, and exemplifications

CMU, 2011 spring, Roni Rosenfeld, HW2, pr. 3.c

***Cross-entropy*, *CH(p,q)*, measures the average number of bits needed to encode an event from a set of possibilities, if a coding scheme is used based on a given probability distribution q, rather than the "true" distribution p.**

For discrete random variables p and q this means

$$CH(p,q) = -\sum_x p(x)\, \log q(x)$$

The situation for continuous random variable distributions is analogous:

$$CH(p,q) = -\int_X p(x)\, \log q(x)\, dx.$$

a. Can cross-entropy be negative? Either prove or give a counter-example.

Answer

a. No. Here follows the *proof*:
For every probability value $p(x)$ and $q(x)$, we know from defnition that $0 \leq p(x) \leq 1$ and $0 \leq q(x) \leq 1$. From $q(x) \leq 1$, we conclude that $\log q(x) \leq 0$. Given that $0 \leq p(x)$, and $-\log q(x) \geq 0$, we conclude that $0 \leq -p(x) \log q(x)$. Thus, the sum of these terms will also be greater or equal to 0, so cross-entropy is never negative.

***Note*, however that unlike entropy, cross-entropy is not bounded, so it can grow to infinity if for an x, $q(x)$ is zero and $p(x)$ is not zero.**

b. In many experiments, the quality of different hypothesis models are compared on a data set. Imagine you derived two diferent models to predict the probabilities of 7 diferent possible outcomes, and the probability distributions predicted by the models are q_1, and q_2 as follows:

$$q_1 = (0.1, 0.1, 0.2, 0.3, 0.2, 0.05, 0.05) \text{ and } q_2 = (0.05, 0.1, 0.15, 0.35, 0.2, 0.1, 0.05).$$

The experiments are done on a data set with the following *empirical* distribution:

$$p_{empirical} = (0.05, 0.1, 0.2, 0.3, 0.2, 0.1, 0.05).$$

Compute the cross-entropies, $CH(p_{empirical}, q_1)$ and $CH(p_{empirical}, q_2)$.
Which model has a lower cross-entropy? Is this model guaranteed to be a better one? Explain your answer. Can cross-entropy be negative? Either prove or give a counter-example.

Answer

b. Using the cross-entropy formula we see that:

$$\begin{aligned} CH(p_{empirical}, q_1) &= -(0.05\log 0.1 + 0.1\log 0.1 + 0.2\log 0.2 + 0.3\log 0.3 + \\ &\quad 0.2\log 0.2 + 0.1\log 0.05 + 0.05\log 0.05) = 2.596 \text{ bits} \\ CH(p_{empirical}, q_2) &= -(0.05\log 0.05 + 0.1\log 0.1 + 0.2\log 0.15 + 0.3\log 0.35 + \\ &\quad 0.2\log 0.2 + 0.1\log 0.1 + 0.05\log 0.05) = 2.56 \text{ bits.} \end{aligned}$$

Also,

$$\begin{aligned} H(p_{empiric}) &= 2\cdot\frac{1}{20}\log_2 20 + 2\cdot\frac{1}{10}\log_2 10 + 2\cdot\frac{1}{5}\log_2 5 + \frac{3}{10}\log_2\frac{10}{3} \\ &= 0.7 + 0.9\log_2 5 - 0.3\log_2 3 = 2.314 \textbf{ bits.} \end{aligned}$$

It can be seen that the $p_{empirical}$ distribution has a lower cross-entropy with the model q_2, so it is reasonable to say that q_2 is a better choice.

***However*, it is not guaranteed that this model is always the better model, because we are working with an "empirical" distribution here, and the "true" distribution might not be reflected in this empirical distribution completely. Usually, sampling bias and insufficient training data samples will widen the gap between the true distribution and the empirical one, so in practice when designing the experiment, you should always have that in mind, and if possible use techniques that minimize these risks.**

Relative entropy a.k.a. the Kulback-Leibler divergence, and the [relationship to] information gain; some basic properties

CMU, 2007 fall, C. Guestrin, HW1, pr. 1.2

[adapted by Liviu Ciortuz]

The *relative entropy* — also known as the *Kullback-Leibler (KL) divergence* — from a distribution p to a distribution q is defined as

$$KL(p||q) \stackrel{def.}{=} -\sum_{x \in X} p(x) \log \frac{q(x)}{p(x)}$$

From an information theory perspective, the KL-divergence specifies the number of additional bits required on average to transmit values of X if the values are distributed with respect to p but we encode them assuming the distribution q.

Notes

1. KL **is not a *distance measure*, since it is not symmetric (i.e., in general** $KL(p||q) \neq KL(q||p)$**).**

Another measure, which is defined as $JSD(p||q) = \frac{1}{2}(KL(p||(p+q)/2) + KL(q||(p+p)/2))$**, and is called the Jensen-Shannon divergence is symmetric.**

2. The quantity

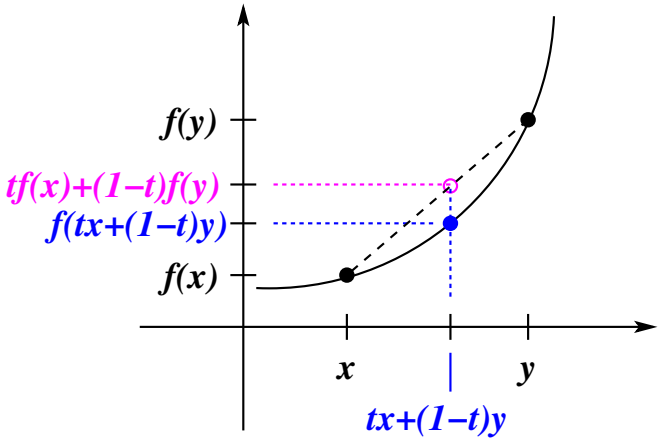
$$\begin{aligned} d(X,Y) &\stackrel{def.}{=} H(X,Y) - IG(X,Y) = H(X) + H(Y) - 2IG(X,Y) \\ &= H(X \mid Y) + H(Y \mid X) \end{aligned}$$

known as variation of information, is a distance metric, i.e., it is non-negative, symmetric, implies indiscernability, and satisfies the triangle inequality.

a. Show that $KL(p||q) \geq 0$, and $KL(p||q) = 0$ iff $p(x) = q(x)$ for all x.
(More generally, the smaller the KL-divergence, the more similar the two distributions.)

Indicaţie:

Pentru a demonstra punctul acesta puteţi folosi inegalitatea lui Jensen:



Dacă $f : \mathbb{R} \to \mathbb{R}$ este o funcţie convexă, atunci pentru orice $t \in [0, 1]$ şi orice $x_1, x_2 \in \mathbb{R}$ urmează

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2).$$

Dacă f este funcţie strict convexă, atunci egalitatea are loc doar dacă $x_1 = x_2$.

Mai general, pentru orice $a_i \geq 0,\ i = 1, \ldots, n$ cu $\sum_i a_i \neq 0$ şi orice $x_i \in \mathbb{R},\ i = 1, \ldots, n$, avem $f\left(\frac{\sum_i a_i x_i}{\sum_j a_j}\right) \leq \frac{\sum_i a_i f(x_i)}{\sum_j a_j}$.

Dacă f este strict convexă, atunci egalitatea are loc doar dacă $x_1 = \ldots = x_n$.

Evident, rezultate similare pot fi formulate şi pentru funcţii concave.

Answer

Vom dovedi inegalitatea $KL(p||q) \geq 0$ folosind inegalitatea lui Jensen, în expresia căreia vom înlocui f cu funcţia convexă $-\log_2$, pe a_i cu $p(x_i)$ şi pe x_i cu $\frac{q(x_i)}{p(x_i)}$.

(Pentru convenienţă, în cele ce urmează vor renunţa la indicele variabilei x.)

Vom avea:

$$\begin{aligned} KL\left(p \,||\, q\right) &\overset{def.}{=} -\sum_x p(x) \log \frac{q(x)}{p(x)} \\ &\overset{Jensen}{\geq} -\log(\sum_x p(x)\frac{q(x)}{p(x)}) = -\log(\underbrace{\sum_x q(x)}_{1}) = -\log 1 = 0 \end{aligned}$$

Aşadar, $KL\left(p \,||\, q\right) \geq 0$, oricare ar fi distribuţiile (discrete) p şi q.

Vom demonstra acum că $KL(p||q) = 0 \Leftrightarrow p = q$**.**

„⇐“

Egalitatea $p(x) = q(x)$ **implică** $\frac{q(x)}{p(x)} = 1$**, deci** $\log \frac{q(x)}{p(x)} = 0$ **pentru orice** x**, de unde rezultă imediat** $KL(p||q) = 0$**.**

„⇒“

Ştim că în inegalitatea lui Jensen are loc egalitatea doar în cazul în care $x_i = x_j$ **pentru orice** i **şi** j**.**

În cazul de faţă, această condiţie se traduce prin faptul că raportul $\frac{q(x)}{p(x)}$ **este acelaşi** (α) **pentru orice valoare a lui** x**.**

Ţinând cont că $\sum_x p(x) = 1$ **şi** $\sum_x q(x) = \sum_x p(x)\frac{q(x)}{p(x)} = 1$**, rezultă că** $\alpha = \frac{q(x)}{p(x)} = 1$**, deci** $p(x) = q(x)$ **pentru orice** x**, ceea ce înseamnă că distribuţiile** p **şi** q **sunt identice.**

b. We can define the *information gain* as the KL-divergence from the observed joint distribution of X and Y to the product of their observed marginals:

$$IG(X,Y) \overset{def.}{=} KL(p_{X,Y} \,||\, (p_X\, p_Y)) = -\sum_x \sum_y p_{X,Y}(x,y) \log \left(\frac{p_X(x)p(y)}{p_Y(x,y)} \right)$$

$$\overset{not.}{=} -\sum_x \sum_y p(x,y) \log \left(\frac{p(x)p(y)}{p(x,y)} \right)$$

Prove that this definition of information gain is equivalent to the one given in problem CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 2. That is, show that $IG(X,Y) = H[X] - H[X|Y] = H[Y] - H[Y|X]$, starting from the definition in terms of KL-divergence.

Remark:

It follows that

$$IG(X,Y) = \sum_y p(y) \sum_x p(x \mid y) \log \frac{p(x \mid y)}{p(x)} = \sum_y p(y) KL(p_{X|Y} \,||\, p_X)$$

$$= E_Y[KL(p_{X|Y} \,||\, p_X)]$$

Answer

By making use of the multiplication rule, namely $p(x,y) = p(x \mid y)p(y)$, we will have:

$$
\begin{aligned}
& KL(p_{X,Y} \,||\, (p_X\, p_Y)) \\
\overset{def.\ KL}{=} \quad & -\sum_x \sum_y p(x,y) \log \left(\frac{p(x)p(y)}{p(x,y)} \right) \\
= \quad & -\sum_x \sum_y p(x,y) \log \left(\frac{p(x)}{p(x \mid y)} \right) = -\sum_x \sum_y p(x,y)[\log p(x) - \log p(x \mid y)] \\
= \quad & -\sum_x \sum_y p(x,y) \log p(x) - \left(-\sum_x \sum_y p(x,y) \log p(x \mid y) \right) \\
= \quad & -\sum_x \log p(x) \underbrace{\sum_y p(x,y)}_{=p(x)} - H[X \mid Y] = \sum_x p(x) \log p(x) - H[X \mid Y] \\
= \quad & H[X] - H[X \mid Y] = IG(X,Y)
\end{aligned}
$$

c.
A direct consequence of parts a. and b. is that $IG(X,Y) \geq 0$ (and therefore $H(X) \geq H(X|Y)$ and $H(Y) \geq H(Y|X)$) for any discrete random variables X and Y.

Prove that $IG(X,Y) = 0$ iff X and Y are independent.

Answer:

This is also an immediate consequence of parts a. and b. already proven:

$$IG(X,Y) = 0 \overset{(b)}{\Leftrightarrow} KL(p_{X,Y}||p_X\, p_Y) = 0 \overset{(a)}{\Leftrightarrow} X \text{ and } Y \text{ are independent.}$$

Remark (in Romanian)

Putem (re)demonstra inegalitatea $IG(X,Y) \geq 0$, folosind (doar!) rezultatul de la punctul b. (nu şi cel de la punctul a.), şi anume că $IG(Y,X) = -\sum_x \sum_y p(x,y) \log \left(\frac{p(x)p(y)}{p(x,y)} \right)$. Ideea este să aplicăm inegalitatea lui Jensen într-o formă uşor generalizată, şi anume:

- **în locul unui singur indice, se vor considera doi indici (aşadar în loc de a_i şi x_i vom avea a_{ij} şi respectiv x_{ij});**
- **vom lua $f = -\log_2$ iar $a_{ij} \leftarrow p(x_i, y_j)$ şi $x_{ij} \leftarrow \frac{p(x_i)p(x_j)}{p(x_i,x_j)}$;**
- **în fine, vom ţine cont că $\sum_i \sum_j p(x_i, y_j) = 1$.**

Prin urmare,

$$
\begin{aligned}
IG(X,Y) &= -\sum_i \sum_j p(x_i,y_j) \log \frac{p(x_i)\cdot p(y_j)}{p(x_i,y_j)} = \sum_i \sum_j p(x_i,y_j) \Big[-\log \frac{p(x_i)\cdot p(y_j)}{p(x_i,y_j)} \Big] \\
&\geq -\log \Big(\sum_i \sum_j p(x_i,y_j) \frac{p(x_i)\cdot p(y_j)}{p(x_i,y_j)} \Big) = -\log \Big(\sum_i \sum_j p(x_i)\cdot p(y_j) \Big) \\
&= -\log(\underbrace{\sum_i p(x_i)}_{1} \cdot \underbrace{\sum_j p(y_j)}_{1}) = -\log 1 = 0
\end{aligned}
$$

În concluzie, $IG(X,Y) \geq 0$.

Remark (cont'd)

În ce priveşte echivalenţa $IG(X,Y)=0 \Leftrightarrow X$ şi Y sunt independente:

„⇐"

Dacă X şi Y sunt variabilele independente,
atunci $p(x_i,y_j)=p(x_i)p(y_j)$ pentru orice i şi j.
În consecinţă, toţi logaritmii din partea dreaptă a primei egalităţi din calculul de mai sus sunt 0 şi rezultă $IG(X,Y)=0$.

„⇒"

Invers, presupunând că $IG(X,Y)=0$, vom ţine cont de faptul că putem exprima câştigul de informaţie cu ajutorul divergenţei KL şi vom aplica un raţionament similar cu cel de la punctul a.

Rezultă că $\dfrac{p(x_i)p(y_j)}{p(x_i,y_j)}=1$ şi deci $p(x_i)p(y_j)=p(x_i,y_j)$ pentru orice i şi j.
Aceasta echivalează cu a spune că variabilele X şi Y sunt independente.

Using Information Gain / Mutual Information for doing Feature Selection

CMU, 2009 spring, Ziv Bar-Joseph, HW5, pr. 6

Given the following observations for input binary features X_1, X_2, X_3, X_4, X_5, and output binary label Y, we would like to use a filter approach to reduce the feature space of $\{X_1, X_2, X_3, X_4, X_5\}$.

X_1	X_2	X_3	X_4	X_5	Y
0	1	1	0	1	0
1	0	0	0	1	0
0	1	0	1	0	1
1	1	1	1	0	1
0	1	1	0	0	1
0	0	0	1	1	1
1	0	0	1	0	1
1	1	1	0	1	1

a. Calculate the mutual information $MI(X_i, Y)$ for each i.

b. Accordingly choose the smallest subset of features such that the best classifier trained on the reduced feature set will perform at least as well as the best classifier trained on the whole feature set. Explain the reasons behind your choice.

Answer

183.

a. As shown at CMU, 2007 fall, Carlos Guestrin, HW1, pr. 1.2, the mutual information can be calculated as:

$$MI(X,Y) \stackrel{def.}{=} -\sum_x \sum_y P(x,y) \log \left(\frac{P(x)\,P(y)}{P(x,y)} \right).$$

The marginal probabilities, estimated (in the MLE sense) from the given data are shown in the nearby tables.

	$x_i = 0$	$x_i = 1$
$P(X_1 = x_1)$	1/2	1/2
$P(X_2 = x_2)$	3/8	5/8
$P(X_3 = x_3)$	1/2	1/2
$P(X_4 = x_4)$	1/2	1/2
$P(X_5 = x_5)$	1/2	1/2

y	$P(Y = y)$
0	1/4
1	3/4

The joint probabilities $P(X_i = x_i, Y = y)$, also estimated from the data, are:

	$x_i = 0, y = 0$	$x_i = 0, y = 1$	$x_i = 1, y = 0$	$x_i = 1, y = 1$
$P(x_1, y)$	1/8	3/8	1/8	3/8
$P(x_2, y)$	1/8	1/4	1/8	1/2
$P(x_3, y)$	1/8	3/8	1/8	3/8
$P(x_4, y)$	1/4	1/4	0	1/2
$P(x_5, y)$	0	1/2	1/4	1/4

***Note*: Columns 3 and 5 can be easily computed using columns 2 and respectively 4 (and also the first table from above), as $P(x_i = 0, y = 1) = P(x_i = 0) - P(x_i = 0, y = 0)$ and $P(x_i = 1, y = 1) = P(x_i = 1) - P(x_i = 1, y = 0)$.**

Using these probabilities and the given formula, the mutual information for each feature is: $MI(X_1, Y) = 0$, $MI(X_2, Y) = 0.01571$, $MI(X_3, Y) = 0$, $MI(X_4, Y) = 0.3113$ **and** $MI(X_5, Y) = 0.3113$.

***Note*: One could easily check, using the previous tables, that X_1 is independent of Y, and X_3 is also independent of Y (so, we can determine in this way too that $MI(X_1, Y)$ and $MI(X_3, Y)$ are 0).**

b. In order to select a set of features, we can prioritize the ones with more mutual information because they are less independent to Y.

By looking at the results of the previous question, we can see that X_5 and X_4 are the features with more mutual information (0.3113), followed by X_2 (0.0157) and, finally, X_1 and X_3 that do not have mutual information with Y.

By inspection of the data, we can see that if we select X_5, X_4 and X_2 there are two samples of different classes with the same features ($X_2 = 1$, $X_4 = 0$, $X_5 = 1$). To avoid this problem, we can add X_1 as an extra feature.

***Note*: Although the mutual information of X_1 with Y is zero, it does not mean that the combination of X_1 with other features will also have zero mutual information [LC: w.r.t. Y].**

The chain rule for the KL divergence / relative entropy

Stanford, 2015 fall, Andrew Ng, HW3, pr. 5.b

Remember that the Kullback-Leibler (KL) divergence between two discrete-valued distributions $P(X)$, $Q(X)$ is defined as follows:[a]

$$KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

For notational convenience, we assume $P(x) > 0$ and $Q(x) > 0$, $\forall x$. Sometimes, we also write the KL divergence as $KL(P||Q) = KL(P(X)||Q(X))$.

The KL divergence between two conditional distributions $P(X|Y)$, $Q(X|Y)$ is defined as follows:

$$KL(P(X|Y)||Q(X|Y)) = \sum_y P(y) \Big(\sum_x P(x|y) \log \frac{P(x|y)}{Q(x|y)} \Big).$$

This can be thought of as the expected KL divergence between the corresponding conditional distributions on x (that is, between $P(X|Y = y)$ and $Q(X|Y = y)$), where the expectation is taken over the random y.

Prove the following *chain rule* for KL divergence:

$$KL(P(X,Y)||Q(X,Y)) = KL(P(X)||Q(X)) + KL(P(Y|X)||Q(Y|X)).$$

---

[a] If P and Q are densities for continuous-valued random variables, then the sum is replaced by an integral, and everything stated in this problem works fine as well. But for the sake of simplicity, in this problem we'll just work with this form of KL divergence for probability mass functions / discrete-valued distributions.

Answer

$$
\begin{aligned}
& KL(P(X,Y)||Q(X,Y)) \\
&= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{Q(x,y)} \\
&= \sum_{x,y} P(x,y) \log \frac{P(x)\,P(y|x)}{Q(x)\,Q(y|x)} \\
&= \sum_{x,y} \left(P(x,y) \log \frac{P(x)}{Q(x)} + P(x,y) \log \frac{P(y|x)}{Q(y|x)} \right) \\
&= \sum_{x} P(x) \log \frac{P(x)}{Q(x)} + \sum_{x,y} P(x)P(y|x) \log \frac{P(y|x)}{Q(y|x)} \Big) \\
&= \sum_{x} P(x) \log \frac{P(x)}{Q(x)} + \sum_{x} P(x) \sum_{y} P(y|x) \log \frac{P(y|x)}{Q(y|x)} \Big) \\
&= KL(P(X)||Q(X)) + KL(P(Y|X)||Q(Y|X)),
\end{aligned}
$$

where we applied (in order): the definition of KL, the definition of conditional probability, log of product is sum of logs, splitting the summation, $\sum_y P(x,y) = P(x)$, and the definition of KL.

The maximization of the (log-)verosimility function

is equivalent to the minimization of

the KL divergence / relative entropy

Stanford, 2015 fall, Andrew Ng, HW3, pr. 5.c

Consider a density estimation problem, and suppose we are given a training dataset $\{x^{(i)}; i = 1, \ldots, m\}$. Let the empirical distribution be

$$\hat{P}(x) = \frac{1}{m} \sum_{i=1}^{m} 1_{\{x^{(i)}=x\}}.$$

($\hat{P}$ is just the uniform distribution over the training set; i.e., sampling from the empirical distribution is the same as picking a random example from the training set.)

Suppose we have some family of distributions P_θ parametrized by θ. (If you like, think of $P_\theta(x)$ as an alternative notation for $P(x; \theta)$.)

Prove that finding the maximum likelihood estimate for the parameter θ is equivalent to finding P_θ with minimal KL divergence from $\hat{P}$. I.e., prove:

$$\arg\min_\theta KL(\hat{P}||P_\theta) = \operatorname*{argmax}_\theta \sum_{i=1}^{m} \ln P_\theta(x^{(i)})$$

Answer

$$\arg\min_\theta KL(\hat{P}||P_\theta) = \arg\min_\theta \sum_x (\hat{P}(x)\log\hat{P}(x) - \hat{P}(x)\log P_\theta(x))$$

$$= \arg\min_\theta \sum_x -\hat{P}(x)\log P_\theta(x) = \operatorname*{argmax}_\theta \sum_x \hat{P}(x)\log P_\theta(x)$$

$$= \operatorname*{argmax}_\theta \sum_x \frac{1}{m}\sum_{i=1}^m 1_{\{x^{(i)}=x\}}\log P_\theta(x) = \operatorname*{argmax}_\theta \frac{1}{m}\sum_x\sum_{i=1}^m 1_{\{x^{(i)}=x\}}\log P_\theta(x)$$

$$= \operatorname*{argmax}_\theta \frac{1}{m}\sum_{i=1}^m \log P_\theta(x^{(i)}) = \operatorname*{argmax}_\theta \sum_{i=1}^m \log P_\theta(x^{(i)}),$$

where we used in order: the definition of KL, leaving out terms independent of θ, flip sign and correspondingly flip min-max, definition of $\hat{P}$, switching order of summation, definition of the indicator, and simplification.

Remark

Consider the relationship between the prezent exercise and multi-variate Bernoulli Naive Bayes parameter estimation. In the Naive Bayes model we assumed P_θ is of the following form: $P_\theta(x, y) = p(y) \prod_{i=1}^{n} p(x_i|y)$. By the chain rule for KL divergence (see Stanford, 2015 fall, Andrew Ng, HW3, pr. 5.b), we therefore have:

$$KL(\hat{P}||P_\theta) = KL(\hat{P}(y)||p(y)) + \sum_{i=1}^{n} KL(\hat{P}(x_i|y)||p(x_i|y)).$$

This shows that finding the maximum likelihood / minimum KL-divergence estimate of the parameters decomposes into $2n + 1$ independent optimization problems: one for the class prior distributions $p(y)$ and one for each of the conditional distributions $p(x_i|y)$ for each feature x_i given each of the two possible labels for y.

Specifically, finding the maximum likelihood estimates for each of these problems individually results in also maximizing the likelihood of the joint distribution. (A similar remark applies to parameter estimation for Bayesian networks.)

Derivation of entropy definition, starting from a set of desirable properties

CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 2.2

Remark: **The definition we gave for entropy $-\sum_{i=1}^{n} p_i \log p_i$ is not very intuitive.**

Theorem:

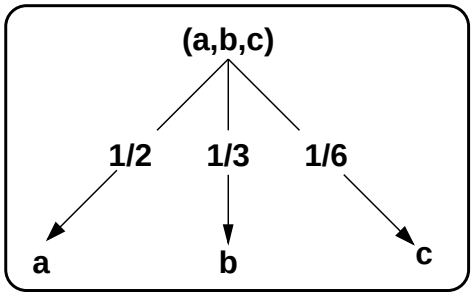
If $\psi_n(p_1, \ldots, p_n)$ satisfies the following axioms

- **A0.** **[LC:] $\psi_n(p_1, \ldots, p_n) \geq 0$ for any $n \in \mathbb{N}^*$ and $p_1, \ldots, p_n$, since we view ψ_n is a measure of *disorder*; also, $\psi_1(1) = 0$ because in this case there is no disorder;**
- **A1.** **ψ_n should be continuous in p_i and symmetric in its arguments;**
- **A2.** **if $p_i = 1/n$ then ψ_n should be a monotonically increasing function of n; (If all events are equally likely, then having more events means being more uncertain.)**
- **A3.** **if a choice among N events is broken down into successive choices, then the entropy should be the weighted sum of the entropy at each stage;**

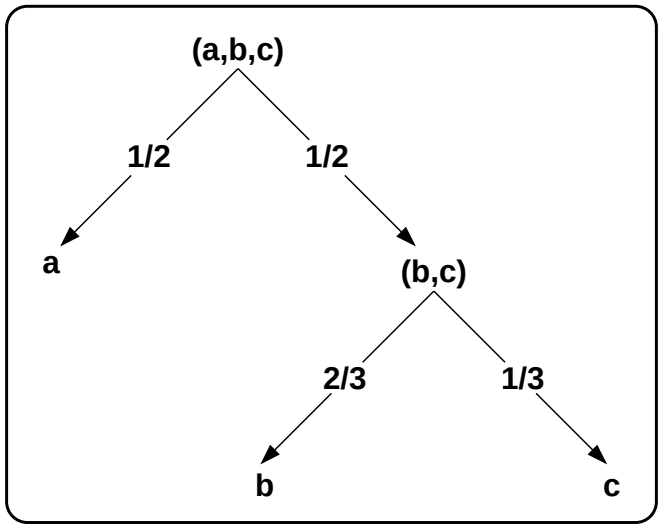
then $\psi_n(p_1, \ldots, p_n) = -K \sum_i p_i \log p_i$ where K is a positive constant.
(As we'll see, K depends however on $\psi_s\left(\frac{1}{s}, \ldots, \frac{1}{s}\right)$ for a certain $s \in \mathbb{N}^*$).

Remark: **We will prove the theorem firstly for uniform distributions ($p_i = 1/n$) and secondly for the case $p_i \in \mathbb{Q}$ (only!).**

Example for the axiom A3:



Encoding 1



Encoding 2

$$H\left(\frac{1}{2},\frac{1}{3},\frac{1}{6}\right) = \frac{1}{2}\log 2+\frac{1}{3}\log 3+\frac{1}{6}\log 6 = \left(\frac{1}{2}+\frac{1}{6}\right)\log 2+\left(\frac{1}{3}+\frac{1}{6}\right)\log 3 = \frac{2}{3}+\frac{1}{2}\log 3$$

$$H\left(\frac{1}{2},\frac{1}{2}\right)+\frac{1}{2}H\left(\frac{2}{3},\frac{1}{3}\right) = 1+\frac{1}{2}\left(\frac{2}{3}\log\frac{3}{2}+\frac{1}{3}\log 3\right) = 1+\frac{1}{2}\left(\log 3-\frac{2}{3}\right) = \frac{2}{3}+\frac{1}{2}\log 3$$

The next 3 slides:

Case 1: $p_i = 1/n$ for $i = 1,\ldots,n$; proof steps:

a. $A(n) \overset{not.}{=} \psi(1/n, 1/n, \ldots, 1/n)$ **implies**

$$A(s^m) = m\, A(s) \text{ for any } s, m \in \mathbb{N}^*. \tag{1}$$

b. If $s, m \in \mathbb{N}^\star$ **(fixed),** $s \neq 1$**, and** $t, n \in \mathbb{N}^\star$ **such that** $s^m \leq t^n \leq s^{m+1}$**, then**

$$\left| \frac{m}{n} - \frac{\log t}{\log s} \right| \leq \frac{1}{n}. \tag{2}$$

c. For $s^m \leq t^n \leq s^{m+1}$ **as above, due to A2 it follows (immediately)**

$$\psi_{s^m}\left(\frac{1}{s^m}, \ldots, \frac{1}{s^m}\right) \leq \psi_{t^n}\left(\frac{1}{t^n}, \ldots, \frac{1}{t^n}\right) \leq \psi_{s^{m+1}}\left(\frac{1}{s^{m+1}}, \ldots, \frac{1}{s^{m+1}}\right)$$

i.e. $A(s^m) \leq A(t^n) \leq A(s^{m+1})$

Show that

$$\left| \frac{m}{n} - \frac{A(t)}{A(s)} \right| \leq \frac{1}{n} \quad \textbf{for} \quad s \neq 1. \tag{3}$$

d. Combining (2) + (3) immediately gives

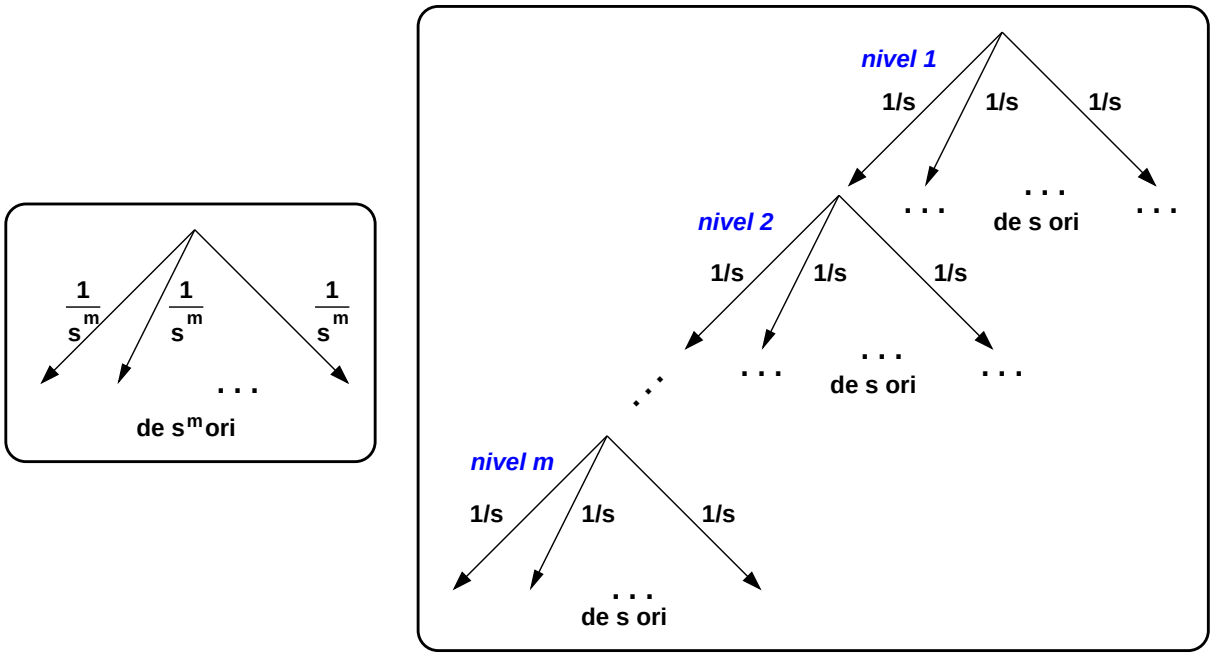
$$\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \leq \frac{2}{n} \quad \textbf{for} \quad s \neq 1 \tag{4}$$

Show that this inequation implies

$$A(t) = K \log t \quad \textbf{with } K > 0 \textbf{ (due to A2).} \tag{5}$$

Proof

a.



Applying the axion A3 on the right encoding from above gives:

$$\begin{aligned} A(s^m) &= A(s) + s \cdot \frac{1}{s} A(s) + s^2 \cdot \frac{1}{s^2} A(s) + \ldots + s^{m-1} \cdot \frac{1}{s^{m-1}} A(s) \\ &= \underbrace{A(s) + A(s) + A(s) + \ldots + A(s)}_{m \text{ times}} = mA(s) \end{aligned}$$

Proof (cont'd)

b.

$$s^m \le t^n \le s^{m+1} \Rightarrow m \log s \le n \log t \le (m+1) \log s \Rightarrow$$

$$\frac{m}{n} \le \frac{\log t}{\log s} \le \frac{m}{n} + \frac{1}{n} \Rightarrow 0 \le \frac{\log t}{\log s} - \frac{m}{n} \le \frac{1}{n} \;\Rightarrow\; \left| \frac{\log t}{\log s} - \frac{m}{n} \right| \le \frac{1}{n}$$

c.

$$A(s^m) \le A(t^n) \le A(s^{m+1}) \overset{(1)}{\Rightarrow} m\, A(s) \le n\, A(t) \le (m+1)\, A(s) \overset{s \ne 1}{\Rightarrow}$$

$$\frac{m}{n} \le \frac{A(t)}{A(s)} \le \frac{m}{n} + \frac{1}{n} \Rightarrow 0 \le \frac{A(t)}{A(s)} - \frac{m}{n} \le \frac{1}{n} \;\Rightarrow\; \left| \frac{A(t)}{A(s)} - \frac{m}{n} \right| \le \frac{1}{n}$$

d. Consider again $s^m \le t^n \le s^{m+1}$ with s, t fixed. If $m \to \infty$ then $n \to \infty$ and from $\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \le \frac{2}{n}$ it follows that $\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \to 0$.

Therefore $\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| = 0$ and so $\frac{A(t)}{A(s)} = \frac{\log t}{\log s}$.

Finally, $A(t) = \frac{A(s)}{\log s} \log t = K \log t$, where $K = \frac{A(s)}{\log s} > 0$ (if $s \ne 1$).

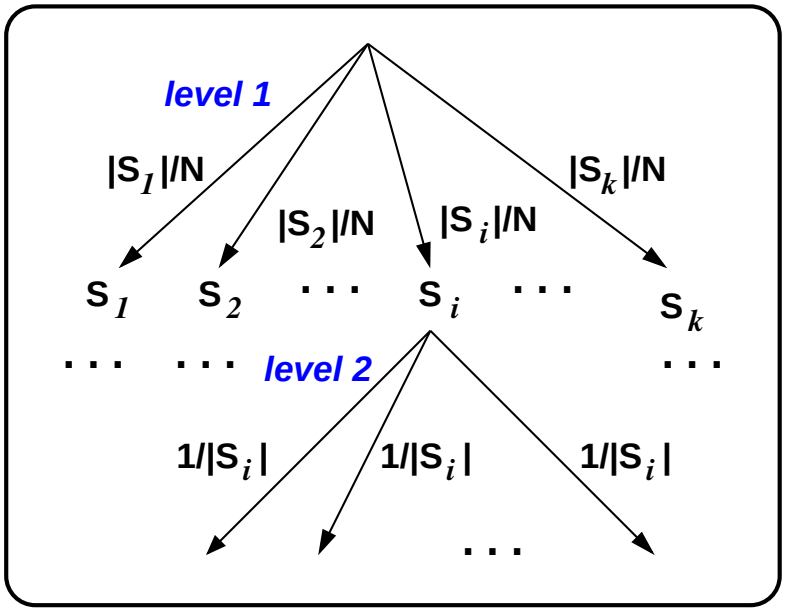
Case 2: $p_i \in \mathbb{Q}$ for $i = 1, \ldots, n$

Let's consider a set of $N \geq 2$ equiprobable random events, and $\mathcal{P} = (S_1, S_2, \ldots, S_k)$ a partition of this set. Let's denote $p_i = \mid S_i \mid /N$.

A "natural" two-step ecoding (as shown in the nearby figure) leads to $A(N) = \psi_k(p_1, \ldots, p_k) + \sum_i p_i A(\mid S_i \mid)$, based on the axiom A3.

Finally, using the result $A(t) = K \log t$, gives:

$$K \log N = \psi_k(p_1, \ldots, p_k) + K \sum_i p_i \log \mid S_i \mid$$



$$\Rightarrow \psi_k(p_1, \ldots, p_k) = K[\, \log N - \sum_i p_i \log \mid S_i \mid]$$

$$= K[\, \log N \sum_i p_i - \sum_i p_i \log \mid S_i \mid] = -K \sum_i p_i \log \frac{\mid S_i \mid}{N} = -K \sum_i p_i \log p_i$$

Jensen's inequality

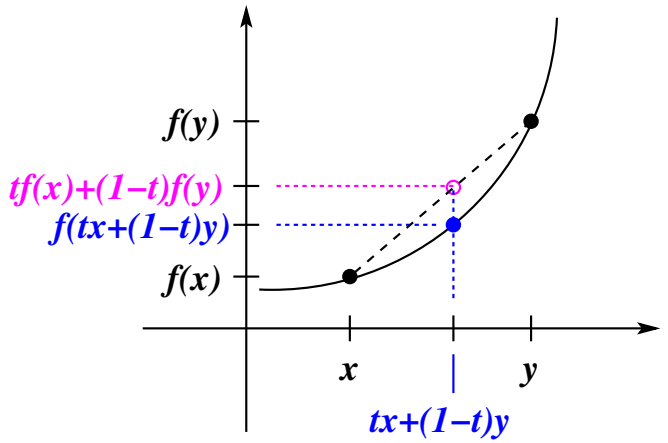
and some consequences

Liviu Ciortuz, 2020

Dacă $f : \mathbb{R} \to \mathbb{R}$ este o *funcţie convexă*, atunci, conform *definiţiei*, pentru orice $t \in [0,1]$ şi orice $x_1, x_2 \in \mathbb{R}$ urmează

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2). \qquad (21)$$

Dacă f este funcţie strict convexă, atunci egalitatea are loc doar dacă $x_1 = x_2$.



a. Folosind definiţia de mai sus, demontraţi *inegalitatea lui Jensen:*[a]
Pentru orice $a_i \geq 0,\ i = 1, \ldots, n$ **cu** $\sum_i a_i = 1$ **şi orice** $x_i \in \mathbb{R},\ i = 1, \ldots, n$**, dacă** f **este funcţie convexă, atunci**

$$f\Big(\sum_i a_i x_i\Big) \leq \sum_i a_i f(x_i). \tag{22}$$

Mai general, pentru orice $a_i' \geq 0$**, cu** $i = 1, \ldots, n$ **şi** $\sum_i a_i' \neq 0$ **avem**

$$f\left(\frac{\sum_i a_i' x_i}{\sum_j a_j'}\right) \leq \frac{\sum_i a_i' f(x_i)}{\sum_j a_j'}. \tag{23}$$

***Observaţii*:**
1. Dacă f **este strict convexă, atunci în relaţiile de mai sus egalitatea are loc doar dacă** $x_1 = \ldots = x_n$**.**
2. Evident, rezultate similare cu cele de mai sus pot fi formulate şi pentru funcţii concave, înlocuind în relaţiile (22) **şi** (23) **semnul** $\leq$ **cu** $\geq$**.**

[a] Johan Jensen, inginer şi matematician danez (18591925).

b. Demonstraţi *inegalitatea mediilor* folosind inegalitatea lui Jensen:

$$\frac{x_1 + x_2 + \ldots + x_n}{n} \geq \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \quad \textbf{pentru orice } x_i \geq 0,\ i = 1, \ldots, n.$$

c. În contextul teoriei probabilităţilor, inegalitatea lui Jensen este exprimată astfel: dacă X este o variabilă aleatoare şi f este o funcţie convexă, atunci $f(E[X]) \leq E[f(X)]$. (Similar, dacă f este funcţie concavă, atunci $f(E[X]) \geq E[f(X)]$.)

Demontraţi această inegalitate în cazul în care X este variabilă aleatoare discretă cu un număr finit de valori (adică, $|Val\,(X)| < \infty$).

Soluţie

Vom privi inegalitatea (22) **ca pe o *ipoteză inductivă*, pe care o vom desemna cu** $P(n)$**, unde** $n \in \mathbb{N}$**,** $n \geq 2$**.**[a] **Vom demonstra că** $P(n)$ **este adevărată, folosind *principiul inducţiei matematice*:**

$P(2)$**: Dacă** $x_1, x_2 \in \mathbb{R}$ **şi** $a_1, a_2 \in [0,1]$ **astfel încât** $a_1 + a_2 = 1$**, iar** f **este funcţie convexă, atunci inegalitatea** $f(a_1x_1 + a_2x_2) \leq a_1 f(x_1) + a_2 f(x_2)$ **se rescrie echivalent ca** $f(a_1x_1 + (1-a_1)x_2) \leq a_1 f(x_1) + (1-a_1)f(x_2)$**.**

Această ultimă inegalitate coincide practic cu relaţia (21) **din definiţia funcţiei convexe, deci este adevărată.**

Presupunem că proprietatea $P(n)$ **este adevărată şi vom demonstra** $P(n+1)$**, adică: dacă** $x_1, \ldots, x_n, x_{n+1} \in \mathbb{R}$ **şi** $a_1, \ldots, a_n, a_{n+1} \in [0,1]$ **astfel încât** $\sum_{i=1}^{n+1} a_i = 1$**, iar** f **o funcţie convexă, atunci rezultă că**

$$f\left(\sum_{i=1}^{n+1} a_i x_i\right) \leq \sum_{i=1}^{n+1} a_i f(x_i). \tag{24}$$

[a] Inegalitatea (22) se verifică şi pentru $n = 1$, fiindcă $f(x_1) = f(x_1)$.

Vom rescrie acum membrul stâng al acestei inegalităţi într-o formă convenabilă:

$$\begin{aligned} f\left(\sum_{i=1}^{n+1} a_i x_i\right) &= f\left(\sum_{i=1}^{n} a_i x_i + a_{n+1}x_{n+1}\right) \\ &= f\left(\left(\sum_{i=1}^{n} a_i\right)\cdot \sum_{i=1}^{n} \frac{a_i}{\sum_{i=1}^{n} a_i} x_i + a_{n+1}x_{n+1}\right) \\ &= f\left((1-a_{n+1})\cdot \sum_{i=1}^{n} \frac{a_i}{1-a_{n+1}} x_i + a_{n+1}x_{n+1}\right) \end{aligned} \tag{25}$$

Este imediat că dacă vom considera $A_i \overset{not.}{=} \dfrac{a_i}{1-a_{n+1}}$**, pentru** $i = 1, \ldots, n$**, atunci rezultă că** $A_i \geq 0$ **pentru** $i = 1, \ldots, n$ **şi** $\sum_{i=1}^{n} A_i = \dfrac{\sum_{i=1}^{n} a_i}{1-a_{n+1}} = \dfrac{1-a_{n+1}}{1-a_{n+1}} = 1$**.**

Prin urmare,

$$\begin{aligned}
f\Big(\sum_{i=1}^{n+1} a_i x_i\Big) &\overset{(25)}{=} f\Big((1-a_{n+1})\cdot \underbrace{\sum_{i=1}^{n} A_i x_i}_{x_1'} + a_{n+1}\underbrace{x_{n+1}}_{x_2'}\Big)\\
&\overset{f\ convexă}{\leq} (1-a_{n+1})\cdot f\Big(\sum_{i=1}^{n} A_i x_i\Big) + a_{n+1}\cdot f(x_{n+1})\\
&\overset{P(n)}{\leq} (1-a_{n+1})\cdot \sum_{i=1}^{n} A_i f(x_i) + a_{n+1}\cdot f(x_{n+1})\\
&= (1-a_{n+1})\cdot \left(\sum_{i=1}^{n} \frac{a_i}{1-a_{n+1}} f(x_i)\right) + a_{n+1}\cdot f(x_{n+1})\\
&= \sum_{i=1}^{n} a_i f(x_i) + a_{n+1} f(x_{n+1}) = \sum_{i=1}^{n+1} a_i f(x_i).
\end{aligned}$$

Aşadar, $f\Big(\sum_{i=1}^{n+1} a_i x_i\Big) \leq \sum_{i=1}^{n+1} a_i f(x_i)$**, deci proprietatea** $P(n+1)$ **este adevărată.**

Sumarizând, din faptul că $P(2)$ **este adevărată, iar implicaţia** $P(n) \Rightarrow P(n+1)$ **este adevărată, conform principiul inducţiei complete rezultă că proprietatea** $P(n)$ **este adevărată pentru orice** $n \in \mathbb{N}^* \setminus \{1\}$**.**

b. Fie numerele $x_1, x_2, \ldots, x_n$, toate mai mari sau egale cu 0. Considerăm $a_1 = a_2 = \ldots = a_n = \dfrac{1}{n}$ (ceea ce implică $\sum_{i=1}^{n} a_i = 1$).

Conform inegalităţii lui Jensen, în care vom alege pe postul funcţiei f funcţia $\ln$ (logaritmul având ca bază numărul e) care este concavă, putem scrie:

$$\ln\Big(\sum_{i=1}^{n} a_i x_i\Big) \geq \sum_{i=1}^{n} a_i \ln(x_i) \Leftrightarrow \ln\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) \geq \frac{1}{n}\sum_{i=1}^{n} \ln(x_i) \Leftrightarrow$$

$$\ln\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) \geq \frac{1}{n}\ln\Big(\prod_{i=1}^{n} x_i\Big) \Leftrightarrow \ln\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) \geq \ln\Big(\prod_{i=1}^{n} x_i\Big)^{\frac{1}{n}} \Leftrightarrow$$

$$\ln\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) \geq \ln(\sqrt[n]{x_1 x_2 \ldots x_n}) \Leftrightarrow \frac{\sum_{i=1}^{n} x_i}{n} \geq \sqrt[n]{x_1 x_2 \ldots x_n}.$$

Ultima echivalenţă are loc întrucât funcţia $\ln$ este strict crescătoare.

c. Fie $f : \mathbb{R} \to \mathbb{R}$ o funcţie convexă şi X o variabilă aleatoare discretă având următorul *tablou de repartiţie*:

$$X : \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}.$$

Conform inegalităţii lui Jensen (pe care o putem aplica luând în locul „ponderilor“ a_i probabilităţile p_i, întrucât $p_i \geq 0$ şi $\sum_{i=1}^{n} p_i = 1$), rezultă:

$$f\Big(\sum_{i=1}^{n} p_i x_i\Big) \leq \sum_{i=1}^{n} p_i f(x_i) \Leftrightarrow f(E_P[X]) \leq E_P[f(X)].$$

În mod similar, dacă f este funcţie concavă, rezultă că $f(E_P[X]) \geq E_P[f(X)]$.

Exemplifying

The gradient descent method:

finding the minimum of a real function of second degree

University of Utah, 2008 fall, Hal Daumé III, HW4, pr. 1

Suppose we are trying to find the minimum of the function $f(x) = 3x^2 - 2x + 1$, for uni-variate (scalar) x.

First, verify that this function is convex (and therefore it has a *global* minimum).

Second, find the minimum of this function using calculus.

Finally, perform (by hand – show your work) three steps of gradient descent with $\eta = 0.1$ and the initial point $x_0 = 1$. How close does it get to the true solution?

Solution (in Romanian)

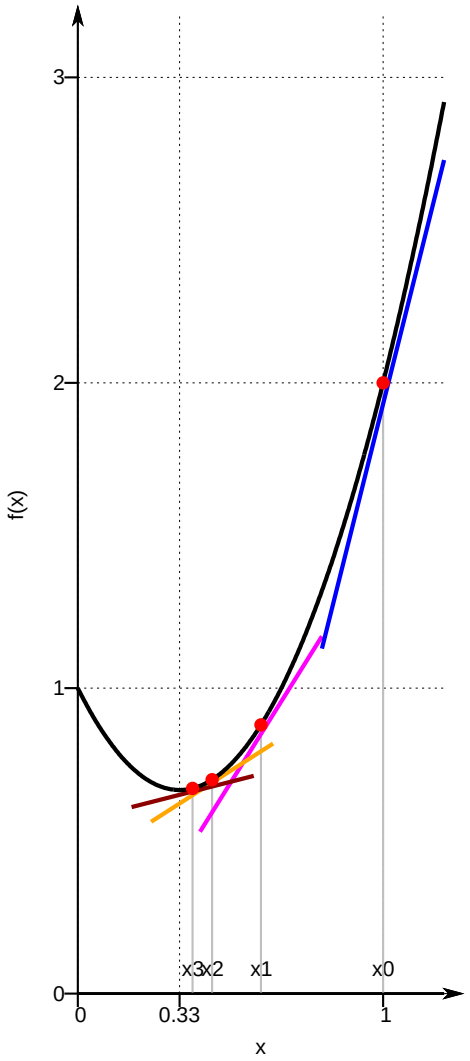
Pentru a studia convexitatea funcţiei $f(x)$ se calculează derivata a doua:

$$f''(x) = 6 > 0, \forall x \in \mathbb{R}.$$

Rezultă că funcţia f este convexă pe întreg domeniul ei de definiţie, deci are un (singur) punct de minim.
Pentru a calcula minimul funcţiei $f(x) = 3x^2 - 2x + 1$ utilizăm derivata de ordinul întâi:

$$f'(x) = 6x - 2.$$

Punctul de minim este dat de soluţia ecuaţiei $f'(x) = 0$, şi anume: $x = \frac{1}{3} \approx 0.33$.



Observaţii

Se constată uşor că

1. apropierea de punctul de optim este [mai] rapidă atâta timp cât valoarea primei derivate (i.e., panta tangentei la graficul funcţiei) este mare în valoare absolută;

2. dacă rata de învăţare η a fost fixată la o valoare prea mare, atunci este posibil ca la un moment dat să depăşim punctul de optim şi apoi să „pendulăm" în jurul lui. Acest *punct slab* al metodei gradientului poate fi contracarat reducând în mod dinamic mărimea lui η.

Altminteri metoda gradientului are *avantajul* de a fi o *tehnică de optimizare* foarte simplă din punct de vedere conceptual şi uşor de implementat.

Alte două *puncte slabe* ale metodei gradientului descendent sunt:
imposibilitatea de a garanta găsirea optimul global şi
numărul mare de iteraţii care trebuie executate pe unele seturi de date reale.