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:

$$\begin{aligned} E[XY] &= \sum_{x \in Val(X)} \sum_{y \in Val(Y)} xyP(X=x, Y=y) = \sum_{x \in Val(X)} \sum_{y \in Val(Y)} xyP(X=x) \cdot P(Y=y) \\ &= \sum_{x \in Val(X)} \left(xP(X=x) \sum_{y \in Val(Y)} yP(Y=y) \right) = \sum_{x \in Val(X)} xP(X=x)E[Y] = E[X] \cdot E[Y] \end{aligned}$$

The continuous case:

$$\begin{aligned} E[XY] &= \int_x \int_y xy\, p(X=x, Y=y) dydx = \int_x \int_y xy\, p(X=x) \cdot p(Y=y) dydx \\ &= \int_x x\, p(X=x) \left(\int_y y\, p(Y=y) dy \right) dx = \int_x x\, p(X=x) E[Y] dx \\ &= E[Y] \cdot \int_x x\, p(X=x) dx = E[X] \cdot E[Y] \end{aligned}$$

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

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

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

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:**

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

Binomial distribution: calculating the variance (cont'd)

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

$$
\begin{aligned}
E[b^2(r;n,p)] &= np \sum_{j=0}^{m} (j+1)\, C_m^j \, p^j q^{m-j} \\
&= np \left[\sum_{j=0}^{m} j \, C_m^j \, p^j q^{m-j} + \sum_{j=0}^{m} C_m^j \, p^j q^{m-j} \right] \\
&= np \left[\sum_{j=0}^{m} j \frac{m \cdot \ldots \cdot (m-j+1)}{j!} p^j q^{m-j} + (\underbrace{p+q}_{1})^m \right] \\
&= np \left[\sum_{j=1}^{m} m \, C_{m-1}^{j-1} \, p^j q^{m-j} + 1 \right] = np \left[mp \sum_{j=1}^{m} C_{m-1}^{j-1} \, p^{j-1} q^{(m-1)-(j-1)} + 1 \right] \\
&= np[(n-1)p(\underbrace{p+q}_{1})^{m-1} + 1] = np[(n-1)p + 1] = n^2p^2 - np^2 + np
\end{aligned}
$$

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 Gaussian distribution: $p(X=x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Calculating the mean: $E[\mathcal{N}_{\mu,\sigma}(x)] \overset{def.}{=} \int_{-\infty}^{\infty} xp(x)dx = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} x\cdot e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}dx$

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

$$
\begin{aligned}
E[X] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(\sigma v+\mu)e^{-\dfrac{v^2}{2}}(\sigma dv) = \frac{\sigma}{\sqrt{2\pi}\sigma}\left(\sigma\int_{-\infty}^{\infty} ve^{-\dfrac{v^2}{2}}dv+\mu\int_{-\infty}^{\infty} e^{-\dfrac{v^2}{2}}dv\right) \\
&= \frac{1}{\sqrt{2\pi}}\left(-\sigma\int_{-\infty}^{\infty}(-v)e^{-\dfrac{v^2}{2}}dv+\mu\int_{-\infty}^{\infty} e^{-\dfrac{v^2}{2}}dv\right) = \frac{1}{\sqrt{2\pi}}\left(\underbrace{-\sigma\, e^{-\dfrac{v^2}{2}}\Bigg|_{-\infty}^{\infty}}_{=0} +\mu\int_{-\infty}^{\infty} e^{-\dfrac{v^2}{2}}dv\right) \\
&= \frac{\mu}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\dfrac{v^2}{2}}dv \text{ (see the next slide for the computation on this last integral)} \\
&= \frac{\mu}{\sqrt{2\pi}}\sqrt{2\pi}=\mu
\end{aligned}
$$

The Gaussian distribution: calculating the mean (Cont'd)

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

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

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

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

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

***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,** $dxdy = rdrd\theta$.

The Gaussian distribution: calculating the variance

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

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

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

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

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

The Gaussian distribution: calculating the variance (Cont'd)

Therefore, we only need to compute

$$\begin{aligned}\int_{-\infty}^{\infty} v^2 e^{-\frac{v^2}{2}}\, dv &= \int_{-\infty}^{\infty} (-v)\left(-ve^{-\frac{v^2}{2}}\right) dv = \int_{-\infty}^{\infty} (-v)\left(e^{-\frac{v^2}{2}}\right)' dv \\ &= (-v)\, e^{-\frac{v^2}{2}}\Bigg|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} (-1) e^{-\frac{v^2}{2}}\, dv = 0 + \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}}\, dv = \sqrt{2\pi}.\end{aligned}$$

Here above we used the fact that

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

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 $X_i : \Omega \rightarrow \mathbb{R}$ 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: $[Cov(X)]_{ij} \overset{def.}{=} Cov(X_i, X_j)$, pentru orice $i, j \in \{1, \ldots, n\}$.

Arătaţi că $\Sigma \overset{not.}{=} Cov(X)$ este matrice simetrică şi pozitiv semidefinită, 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.)

$\boldsymbol{Cov}(X)_{i,j} \overset{def.}{=} \boldsymbol{Cov}(X_i, X_j)$, **for all** $i, j \in \{1, \ldots, n\}$, **and**
$\boldsymbol{Cov}(X_i, X_j) \overset{def.}{=} E[(X_i - E[X_i])(X_j - E[X_j])] = E[(X_j - E[X_j])(X_i - E[X_i])] = \boldsymbol{Cov}(X_j, X_i)$,
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 $X = [X_1 \ldots X_d]^T$, $\mu \in \mathbb{R}^d$ 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.

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

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:

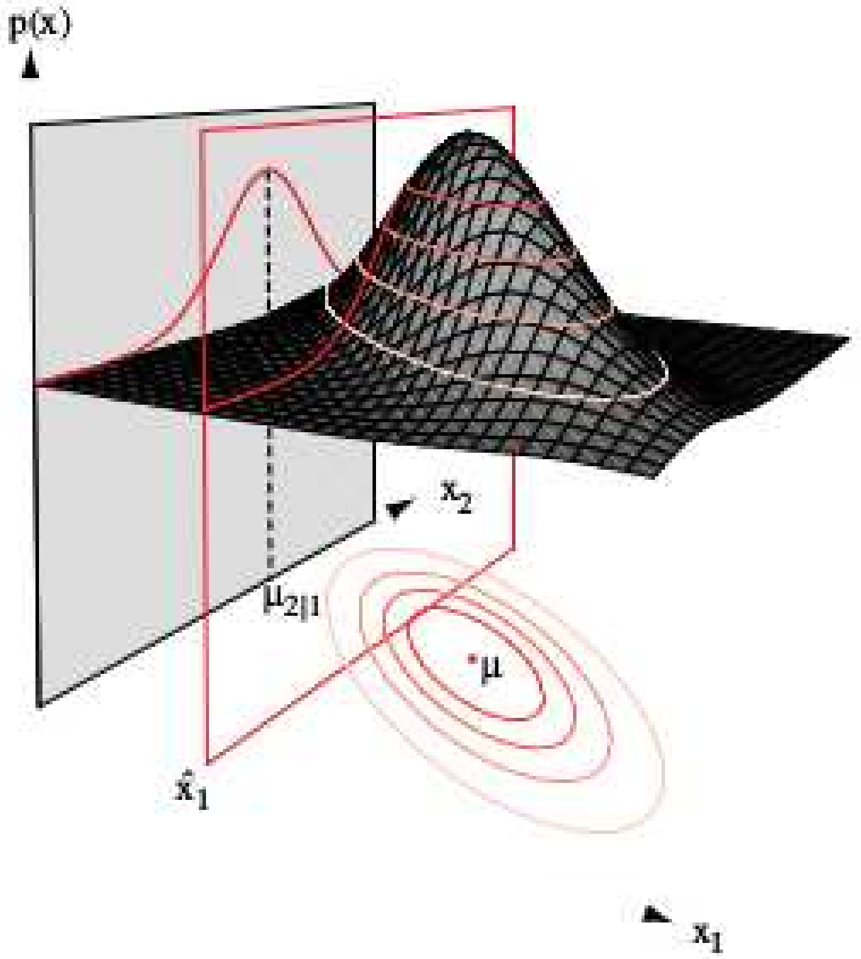
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{2}$$

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 $\mu_{2|1} = \mu_2 + \rho \dfrac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$ **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** $\mu_{1|2} = \mu_1 + \rho \dfrac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$ **şi** $\sigma^2_{1|2} = \sigma^2_1(1 - \rho^2)$**.**



Source:

***Pattern Classification*, Appendix A.5.2,**

Duda, Hart and Stork, 2001

Answer

$$p_{X_2|X_1}(x_2|x_1) \overset{def.}{=} \frac{p_{X_1,X_2}(x_1,x_2)}{p_{X_1}(x_1)}, \tag{3}$$

where

$$\begin{aligned} p_{X_1,X_2}(x_1,x_2) &= \frac{1}{(\sqrt{2\pi})^2\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) \text{ şi} \\ p_{X_1}(x_1) &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2\right). \end{aligned} \tag{4}$$

From (2) 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 $\sqrt{|\Sigma|} = \sigma_1\sigma_2\sqrt{1-\rho^2}$.

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

So,

$$p_{X_1,X_2}(x_1,x_2) =$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}(x_1-\mu_1)\begin{bmatrix} \dfrac{1}{\sigma_1^2} & -\dfrac{\rho}{\sigma_1\sigma_2} \\ -\dfrac{\rho}{\sigma_1\sigma_2} & \dfrac{1}{\sigma_2^2} \end{bmatrix}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\right)$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot$$

$$\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right) \qquad (5)$$

By substitution (4) and (5) in the definition (3), we will get:

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

Therefore,

$X_2|X_1 = x_1 \sim \mathcal{N}(\mu_{2|1}, \sigma^2_{2|1})$ **with** $\mu_{2|1} \overset{not.}{=} \mu_2 + \rho\dfrac{\sigma_2}{\sigma_1}(x_1-\mu_1)$ **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** $a > 0$ **sunch 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: $Z_n \rightarrow \mathcal{N}(0;1)$

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*:**

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

$P(\mid Z_n \mid \leq a) = 0.95 \Leftrightarrow 2\Phi(a) - 1 = 0.95 \Leftrightarrow \Phi(a) = 0.975 \Leftrightarrow a \cong 1.97$ **(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**

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

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:

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

Combining this with the result of the previous question yields:

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

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.

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

c.

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

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

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

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

Computing entropies and mean conditional entropies for discrete random variables

CMU, 2012 spring, R. 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 *mean 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.

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

c.

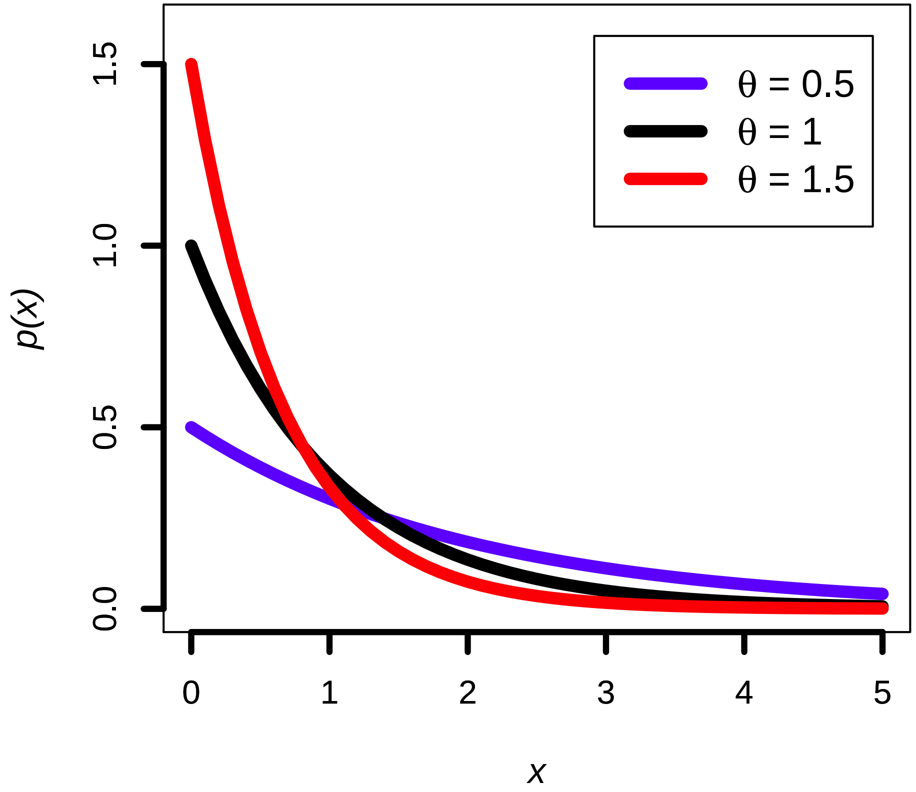
$$
\begin{aligned}
H(C|T) &\overset{def.}{=} \sum_{t \in Val(T)} P(T=t) \cdot H(C|T=t) \\
&= P(T=sunny) \cdot H(C|T=sunny) + P(T=rainy) \cdot H(C|T=rainy) + \\
&\quad P(T=snowy) \cdot H(C|T=snowy) \\
&= 0.35 \cdot H\left(\frac{0.30}{0.30+0.05}, \frac{0.05}{0.30+0.05}\right) + 0.35 \cdot H\left(\frac{0.20}{0.20+0.15},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

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

Computing the entropy of the exponential distribution

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

Exponenţial p.d.f.



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:

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

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 $H_n(X) = -\sum_i p_i \log p_i$ is not very intuitive.

Theorem:

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

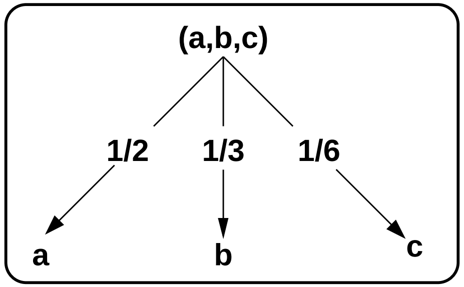
A1. H_n **should be continuous in** p_i **and symmetric in its arguments;**

A2. if $p_i = 1/n$ **then** H_n **should be a monotonically increasing function of** n**;**
(If all events are equally likely, then having more events means being more uncertain.)

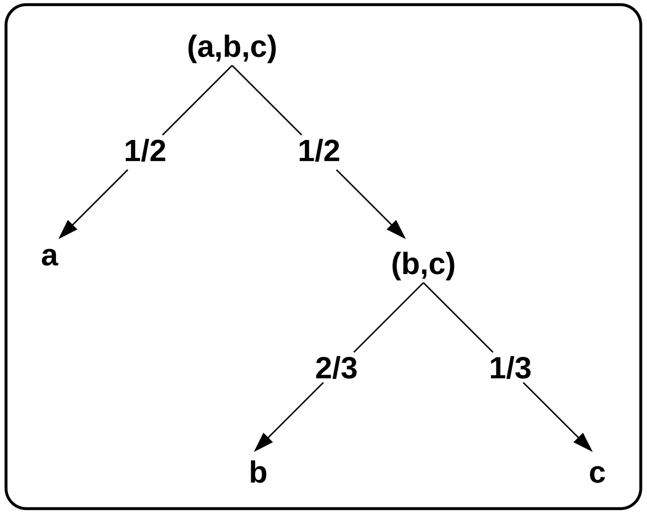
A3. if a choice among N **events is broken down into successive choices, then entropy should be the weighted sum of the entropy at each stage;**

then $\psi_n(p_1, \ldots, p_n) = -K \sum_i p_i \log p_i$ **where** K **is a positive constant.**

Example for the axiom A3:



Encoding 1



Encoding 2

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

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

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**

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

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**

$$\mid \frac{m}{n} - \frac{\log t}{\log s} \mid \leq \frac{1}{n}. \qquad \textbf{(2)}$$

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

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

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

Show that

$$\mid \frac{m}{n} - \frac{A(t)}{A(s)} \mid \leq \frac{1}{n} \text{ for } s \neq 1. \qquad \textbf{(3)}$$

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

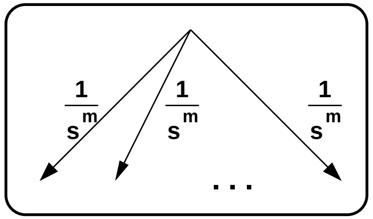
$$\mid \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \mid \leq \frac{2}{n} \text{ pentru } s \neq 1 \qquad \textbf{(4)}$$

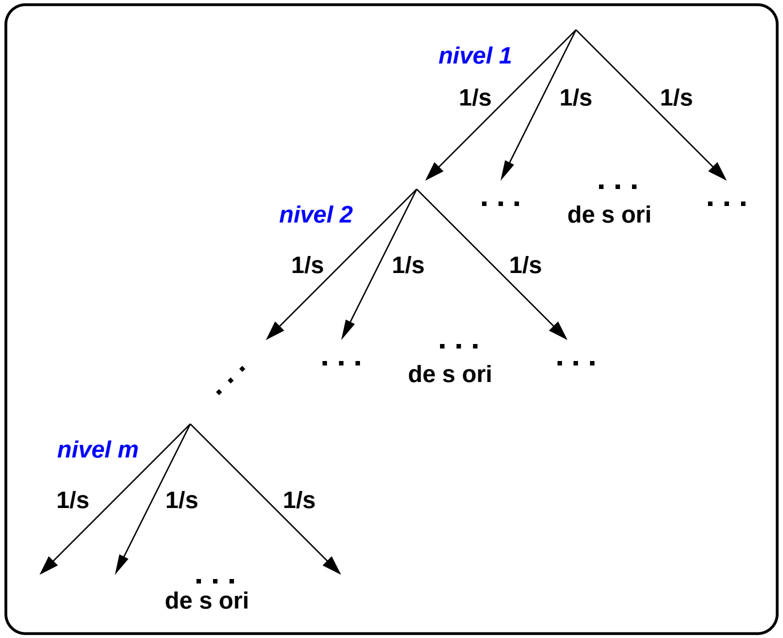
Show that this inequation implies

$A(t) = K \log t$ **with** $K > 0$ **(due to A2).** **(5)**

Proof

a.





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

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

Proof (cont'd)

b.

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

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

c.

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

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

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

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

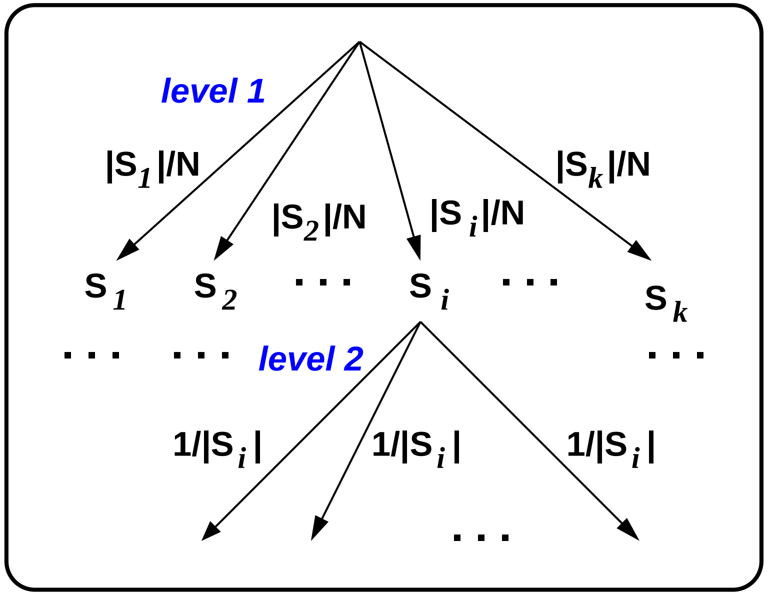
Finally, $A(t) = \frac{A(s)}{\log s} \log t = K \log t$**, where** $K = \frac{A(s)}{\log s} > 0$ **(if** $s \neq 1$**).**

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

Let's consider a set of N 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:



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

$$\begin{aligned}
\Rightarrow \psi_k(p_1, \ldots, p_k) &= K[\, \log N - \sum_i p_i \log \mid S_i \mid] \\
&= K[\, \log N \sum_i p_i - \sum_i p_i \log \mid S_i \mid] = -K \sum_i p_i \log \frac{\mid S_i \mid}{N} = -K \sum_i p_i \log p_i
\end{aligned}$$

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

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

Definiţii

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

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

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

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

- **Informaţia mutuală a variabilelor X şi Y, numită de asemenea *câştigul de informaţie* al variabilei X în raport cu variabila Y (sau invers):**
  $MI(X,Y) \overset{not.}{=} IG(X,Y) \overset{def.}{=} H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$
  (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 $\sum_i P(X=x_i) \log \dfrac{1}{P(X=x_i)} = 0$. 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 $\log \dfrac{1}{P(X=x_i)} = 0$, 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):

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

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

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** $H(X) = \log_2 n$**.**

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

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

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

2. The quantity

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

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 $KL(p||q) = 0$ iff $p(x) = q(x)$ for all x.
(More genrally, the smaller the KL-divergence, the more similar the two distributions.)

Indicaţie:

Pentru a demonstra punctul acesta puteţi folosi inegalitatea lui Jensen:

Dacă $\varphi : \mathbb{R} \to \mathbb{R}$ este o funcţie convexă, atunci pentru orice $t \in [0, 1]$ şi orice $x_1, x_2 \in \mathbb{R}$ urmează $\varphi(tx_1 + (1 - t)x_2) \leq t\varphi(x_1) + (1 - t)\varphi(x_2)$.
Dacă φ 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
$$\varphi\left(\frac{\sum_i a_i x_i}{\sum_j a_j}\right) \leq \frac{\sum_i a_i \varphi(x_i)}{\sum_j a_j}.$$
Dacă φ 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 φ 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:

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

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

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

$\Leftarrow$

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

$\Rightarrow$

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 $\dfrac{q(x)}{p(x)}$ este acelaşi pentru orice valoare a lui x.

Ţinând cont că $\sum_x p(x) = 1$ şi $\sum_x p(x)\dfrac{q(x)}{p(x)} = \sum_x q(x) = 1$, rezultă că $\dfrac{q(x)}{p(x)} = 1$ sau, altfel spus, $p(x) = q(x)$ 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:

$$\begin{aligned} IG(X,Y) &\overset{def.}{=} KL(p_{X,Y} \,||\, (p_X\, p_Y)) = -\sum_x \sum_y p_{X,Y}(x,y) \log \left(\frac{p_X(x) p(y)}{p_Y(x,y)} \right) \\ &\overset{not.}{=} -\sum_x \sum_y p(x,y) \log \left(\frac{p(x)p(y)}{p(x,y)} \right) \end{aligned}$$

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

$$\begin{aligned} IG(X,Y) &= \sum_y p(y) \sum_x p(x \mid y) \log \frac{p(x \mid y)}{p(x)} = \sum_y p(y) KL(p_{X|Y} \,||\, p_X) \\ &= E_Y[KL(p_{X|Y} \,||\, p_X)] \end{aligned}$$

Answer

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

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

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:

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

Remark

Putem demonstra inegalitatea $IG(X,Y) \geq 0$ şi în manieră directă, folosind rezultatul de la punctul b. şi aplicând inegalitatea lui Jensen în forma generalizată, cu următoarele ,,amendamente“:

- **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 $\varphi = -\log_2$ iar $a_{ij} \leftarrow p(x_i, y_j)$ şi $x_{ij} \leftarrow \dfrac{p(x_i)p(x_j)}{p(x_i, x_j)}$;**
- **în fine, vom ţine cont că $\sum_i \sum_j p(x_i, y_j) = 1$.**

Prin urmare,

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

În concluzie, $IG(X,Y) \geq 0$.

Remark (cont'd)

**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.**

Proving [in a direct manner] that the Information Gain is always positive or 0

(an indirect proof was 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 corelată 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.

La acest exerciţiu 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.

***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)

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:

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

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

$$\underset{\textit{prob. marg.}}{\overset{\textit{def.}}{=}} \quad \sum_{i=1}^{n}\left(\sum_{j=1}^{m} P(x_i,y_j)\right)\log_2 P(x_i) - \sum_{j=1}^{m} P(y_j)\sum_{i=1}^{n} P(x_i|y_j)\log_2 P(x_i|y_j)$$

$$\overset{\textit{distrib.}\cdot,+}{=} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j)\log_2 P(x_i) - \sum_{j=1}^{m}\sum_{i=1}^{n} P(y_j)P(x_i|y_j)\log_2 P(x_i|y_j)$$

$$\underset{\textit{prob. cond.}}{\overset{\textit{def.}}{=}} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j)\log_2 P(x_i) - \sum_{j=1}^{m}\sum_{i=1}^{n} P(x_i,y_j)\log_2 P(x_i|y_j)$$

$$\overset{\textit{distrib.}\cdot,+}{=} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j)(\log_2 P(x_i) - \log_2 P(x_i|y_j))$$

$$\underset{\textit{log.}}{\overset{\textit{prop.}}{=}} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j)\log_2\frac{P(x_i)}{P(x_i|y_j)} \underset{\textit{multipl.}}{\overset{\textit{reg. de}}{=}} \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i|y_j)P(y_j)\log_2\frac{P(x_i)}{P(x_i|y_j)}$$

$$\overset{\textit{distrib.}\cdot,+}{=} \quad \sum_{j=1}^{m} P(y_j)\sum_{i=1}^{n}\underbrace{P(x_i|y_j)}_{a_i}\log_2\frac{P(x_i)}{P(x_i|y_j)}$$

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:

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

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