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:

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

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

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

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

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

$$
\begin{aligned}
E[b(r;n,p)] &\overset{def.}{=} \sum_{r=0}^{n} r \cdot b(r;n,p) = \\
&= 1 \cdot C_n^1 p(1-p)^{n-1} + 2 \cdot C_n^2 p^2 (1-p)^{n-2} + \cdots + (n-1) \cdot C_n^{n-1} p^{n-1}(1-p) + n \cdot p^n \\
&= p\left[C_n^1 (1-p)^{n-1} + 2 \cdot C_n^2 p(1-p)^{n-2} + \cdots + (n-1) \cdot C_n^{n-1} p^{n-2}(1-p) + n \cdot p^{n-1}\right] \\
&= np\left[(1-p)^{n-1} + C_{n-1}^1 p(1-p)^{n-2} + \cdots + C_{n-1}^{n-2} p^{n-2}(1-p) + C_{n-1}^{n-1} p^{n-1}\right] \qquad (1) \\
&= np[p + (1-p)]^{n-1} = np
\end{aligned}
$$

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:

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

$$= \quad np\left[\sum_{j=0}^{m} j\,C_m^j\,p^j q^{m-j} + \sum_{j=0}^{m} C_m^j\,p^j q^{m-j}\right]$$

$$= \quad np\left[\sum_{j=0}^{m} j\frac{m\cdot\ldots\cdot(m-j+1)}{j!}p^j q^{m-j} + \underbrace{(p+q)}_{1}{}^m\right]$$

$$= \quad np\left[\sum_{j=1}^{m} m\,C_{m-1}^{j-1}\,p^j q^{m-j} + 1\right] = np\left[mp\sum_{j=1}^{m} C_{m-1}^{j-1}\,p^{j-1}q^{(m-1)-(j-1)} + 1\right]$$

$$= \quad np[(n-1)p\underbrace{(p+q)}_{1}{}^{m-1} + 1] = np[(n-1)p+1] = n^2p^2 - np^2 + np$$

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

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

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

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

$$= \frac{1}{\sqrt{2\pi}} \left(-\sigma \int_{-\infty}^{\infty} (-v) e^{-\frac{v^2}{2}} dv + \mu \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} dv \right) = \frac{1}{\sqrt{2\pi}} \left(-\sigma \underbrace{e^{-\frac{v^2}{2}} \Big|_{-\infty}^{\infty}}_{=0} + \mu \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} dv \right)$$

$$= \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} dv \quad \textbf{(see the next slide for the computation on this last integral)}$$

$$= \frac{\mu}{\sqrt{2\pi}} \sqrt{2\pi} = \mu$$

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:

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

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

***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 $Var[X] = E[X^2] - E^2[X]$**.**

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

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** $\int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} \, dv = \sqrt{2\pi}$**.**

The Gaussian distribution: 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

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

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

A mixture of categorical distributions: How to compute the expectation and the 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:

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

The covariance matrix Σ corresponding to a vector X made of n random variables is symmetric and positive semi-definite

a. $\boldsymbol{Cov}(X)_{i,j} \stackrel{def.}{=} \boldsymbol{Cov}(X_i, X_j)$**, 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.**

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

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

If the covariance matrix of a multi-variate Gaussian distribution is diagonal, then the density of this is equal to the product of independent univariate Gaussian densities

Let's consider $X = [X_1 \dots X_n]^T$, $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{S}_+^n$, where $\mathbb{S}_+^n$ 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 $x \in \mathbb{S}^n$, $x \neq \mu$).

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

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

Notation: $X \sim \mathcal{N}(\mu,\Sigma)$.

We will make the proof for $n = 2$ (generalization to $n > 2$ 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 $\Sigma \in \mathbb{S}_+^n$ 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.

A property of multi-variate Gaussians whose covariance matrices are diagonal (Cont'd)

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

Random Variables

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;

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

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

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

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: $Z_n \to N(0;1)$

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

***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)**

***Finally*:**

$\sigma^2 \overset{not.}{=} Var_{real} \approx Var_{sample}$ **due to the above theorem, and**

$Var_{sample} = e_{sample}(1 - e_{sample})$ **because** X_i **are Bernoulli variables.**

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

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

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

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 $X_1, \ldots, X_n$.

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. In the third equality we used the trick $X_i = \mathbb{I}\{X_i = 1\}$.

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

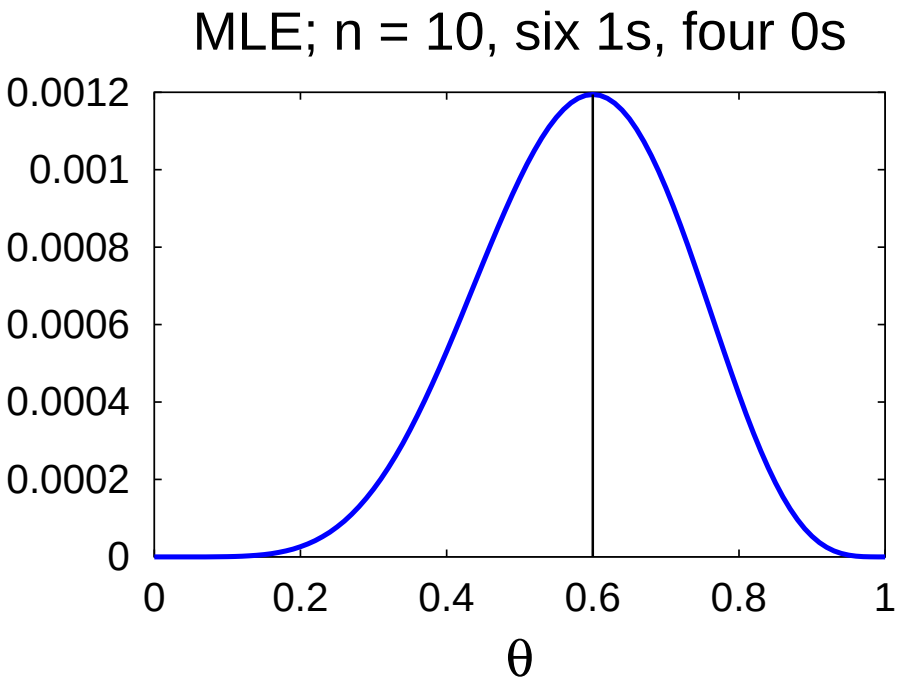
b. Suppose that $n = 10$ and the data set contains six 1s and four 0s.

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 θ that maximizes the likelihood. Using the properties of the $\ln$ function, we can rewrite $l(\hat{\theta})$ as follows:

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

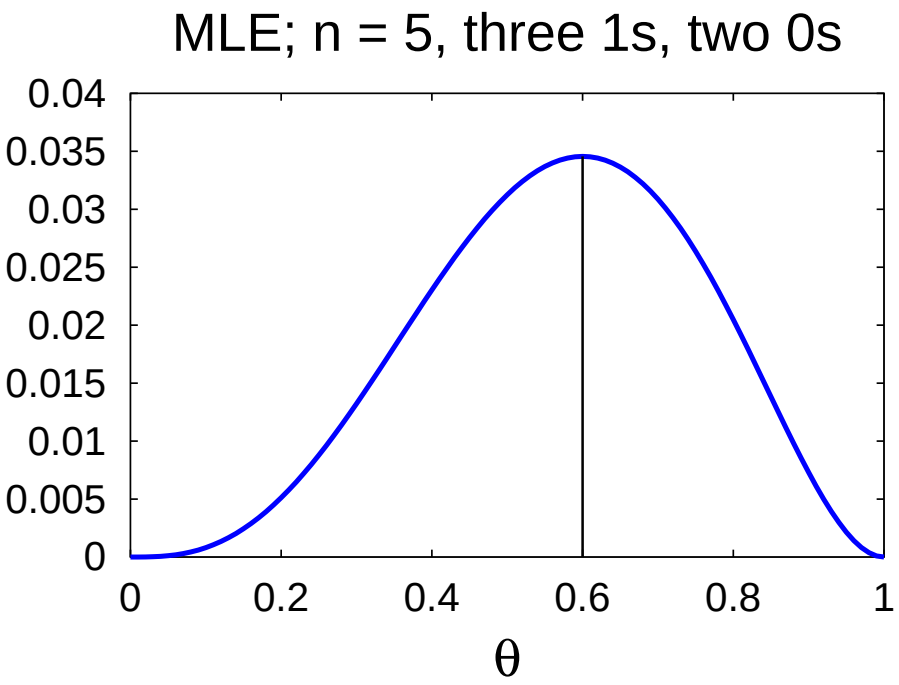
Assuming that $\hat{\theta} \neq 0$ and $\hat{\theta} \neq 1$, the first and second derivatives of l are given by

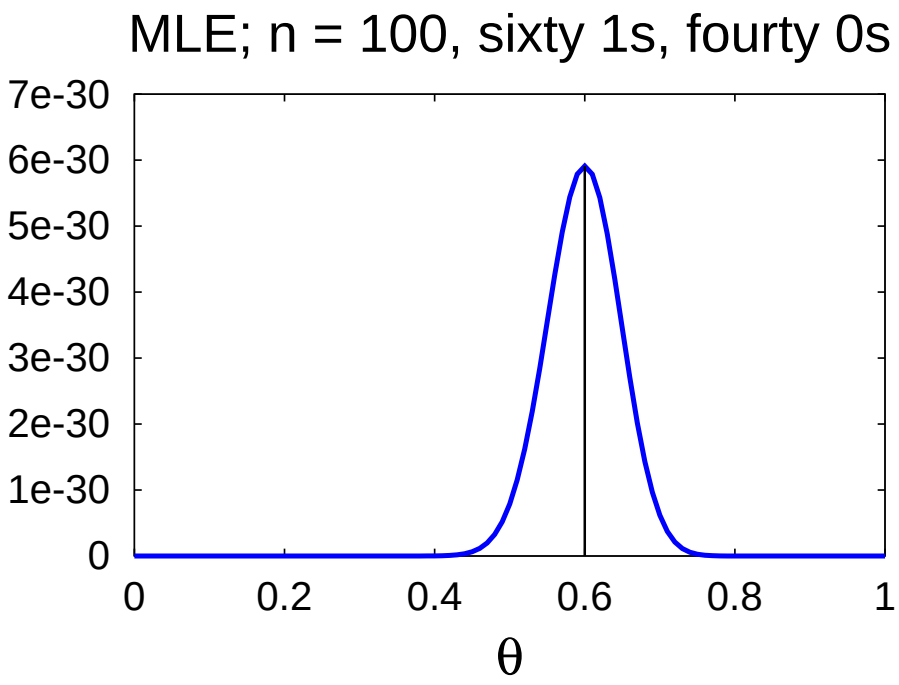
$$l'(\hat{\theta}) = \frac{n_1}{\hat{\theta}} - \frac{n_0}{1-\hat{\theta}} \quad \textbf{and} \quad l''(\hat{\theta}) = -\frac{n_1}{\hat{\theta}^2} - \frac{n_0}{(1-\hat{\theta})^2}$$

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 $\hat{\theta}_{MLE} = \dfrac{n_1}{n_1 + n_0}$.

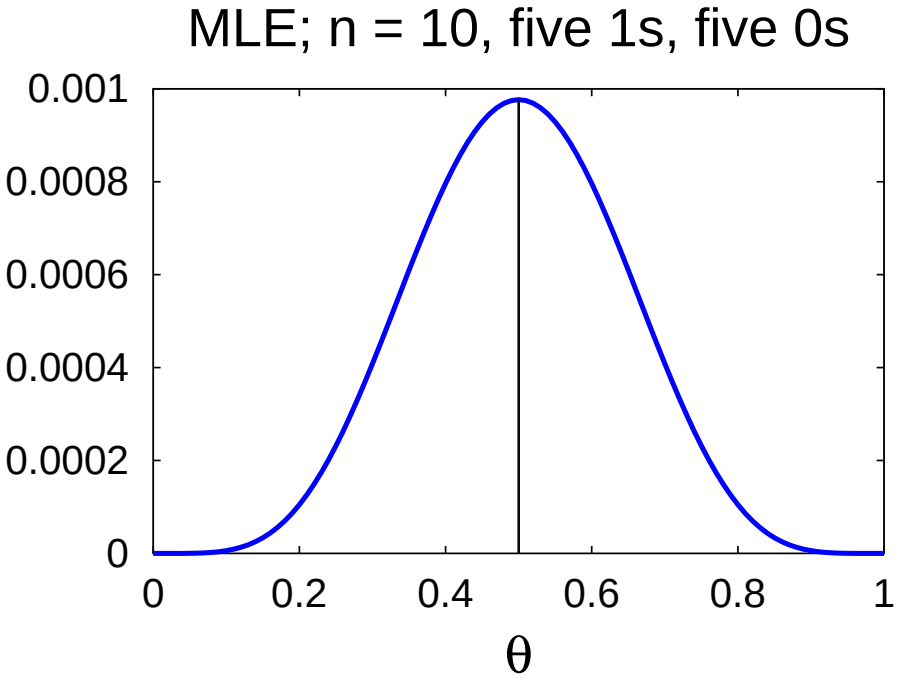
d. Create three more likelihood plots: one where $n = 5$ and the data set contains three 1s and two 0s; one where $n = 100$ and the data set contains sixty 1s and fourty 0s; and one where $n = 10$ and there are five 1s and five 0s.

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.

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 $X_1, \ldots, X_n$.

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

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

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. The MAP estimate for θ is given by

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

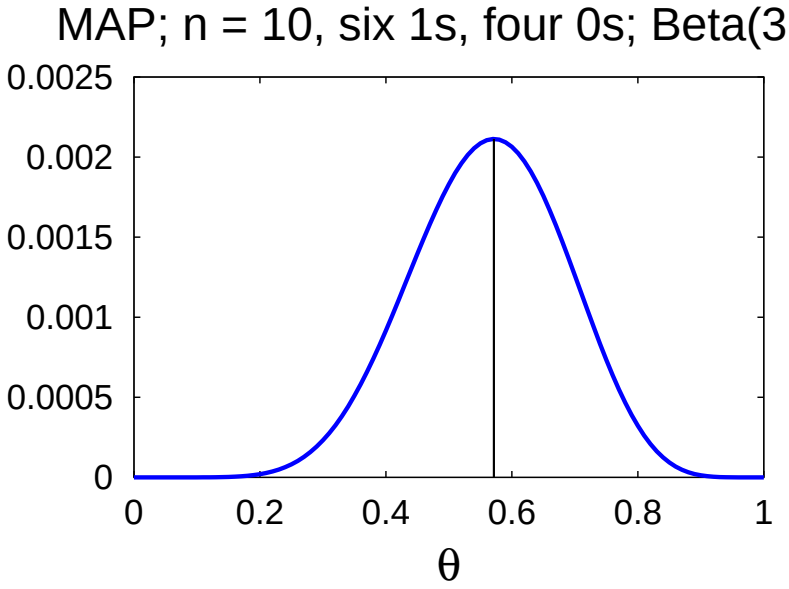
We will consider a $Beta(3,3)$ prior distribution for θ, which has the density function given by $p(\hat{\theta}) = \dfrac{\hat{\theta}^2(1-\hat{\theta})^2}{B(3,3)}$, where $B(\alpha,\beta)$ is the beta function and $B(3,3) \approx 0.0333$.

f. Suppose, as in part c, 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 c.

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

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

h. Compare the MAP estimate to the MLE computed from the same data in part c. 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:

As n goes to infinity, the influence of the 4 virtual random variables diminishes, and the two estimators become equal.

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 $a_1, a_2, \ldots, a_k$.

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:

Since the probability of seeing an event of type j is θ_j, and we are given an ordered list of events and not an unordered bag of events, the verosimility of the data is:

$$
\begin{aligned}
L(\theta) &= P(d_1, \ldots, d_n|\theta) \overset{i.i.d.}{=} \textstyle\prod_{i=1}^{n} \sum_{i=1}^{k} (\theta_j I_{d_i=a_j}) && \text{(}I\text{ is the indicator function)} \\
&= \textstyle\prod_{i=1}^{k} \theta_i^{n_i} \\
&= \underbrace{(1 - \sum_{i=1}^{k-1} \theta_i)}_{\theta_k}{}^{n_k} \textstyle\prod_{i=1}^{k-1} \theta_i^{n_i} && \text{(since the thetas sum to one)}
\end{aligned}
$$

b. Find the MLE for one of the $k-1$ parameters, θ_j, 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:

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

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

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

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

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 $\frac{n_j}{n}$.

Answer:

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

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

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

Therefore, $\frac{\hat{\theta}_j}{n_j} - \frac{\hat{\theta}_k}{n_k} = 0$ for all j. So, $\hat{\theta}_j = \frac{n_j}{n_k}\hat{\theta}_k$.

Finally,

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

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

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

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

$$\Rightarrow \hat{\theta}_j = \frac{n_j}{n_k} \cdot \frac{n_k}{n} = \frac{n_j}{n}$$

***Note*: Even though we can go from the non-hatted to hatted form 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:

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

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

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:

$$P(\mu|x_1,\ldots,x_n) \overset{T.\ Bayes}{=} \frac{P(x_1,\ldots,x_n|\mu)\,P(\mu)}{P(x_1,\ldots,x_n)} \tag{2}$$

$$= \frac{\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\right)e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\;\frac{1}{\sqrt{2\pi}\beta}e^{-\frac{(\mu-\nu)^2}{2\beta^2}}}{C} \tag{3}$$

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 (3) is itself a Gaussian, and then we will use the fact that the mean of a Gaussian is where it achieves its maximum value.

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

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

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:

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

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 $\mathbf{x} = (x_1, \ldots, x_n)$ 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 $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

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: $\frac{\partial \ln \mathcal{L}(\mathbf{x} \mid \mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2$.

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

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 $E[\sigma^2_{MLE}] = \dfrac{n-1}{n}\sigma^2$.

Solution:

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

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

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:

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

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 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.a

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

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

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 $\hat{\mu}_{mle}$ and $\hat{\Lambda}_{mle}$.

Hint

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

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

(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., $a_{11} + \ldots + a_{nn}$.

The log-likelihood is as follows:

$$l(\mu, \Lambda) = -\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^\top \Lambda (x_i - \mu) - \frac{nd}{2}\ln(2\pi) - \frac{n}{2}\ln|\Lambda^{-1}|.$$

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

$$\nabla_\mu l(\mu, \Lambda) = 0 \overset{(5g)}{\iff} -\Lambda \sum_{i=1}^{n}(\mu - x_i) = 0,$$

and 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 with respect to 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:

$$\begin{aligned} l(\hat{\mu}, \Lambda) &\overset{(2b)}{=} -\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^\top \Lambda (x_i - \bar{x}) - \frac{nd}{2}\ln(2\pi) + \frac{n}{2}\ln|\Lambda| \\ &= -\frac{n}{2}(\mathrm{Tr}(S\Lambda) - \ln|\Lambda|) - \frac{nd}{2}\ln(2\pi), \end{aligned} \tag{4}$$

where S is the *sample covariance matrix*:

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top.$$

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 (4) by solving

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

which can be shown[b] as being equivalent to

$$S - \Lambda^{-1} = 0.$$

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

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

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

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

so they are the MLE estimates.

---

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

[b] Use (3b) and (4b) or (without using Tr) (5c) and (4b).

Elements of Information Theory:

Some proofs

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**

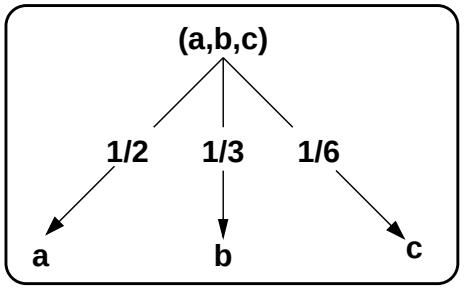
A1. H_n **should be continuous in** p_i **and symmetric in its arguments;**

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.)**

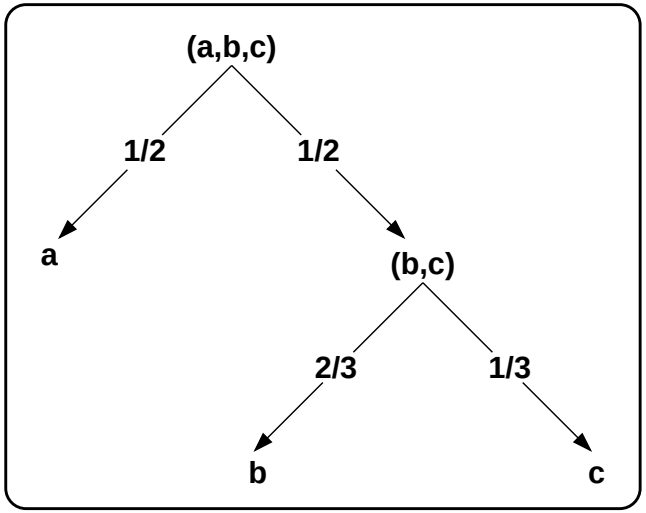
A3. if a choice among N **events is broken down into successive choices, then entropy should be the weighted sum of the entropy at each stage;**

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

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

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

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, 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

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

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

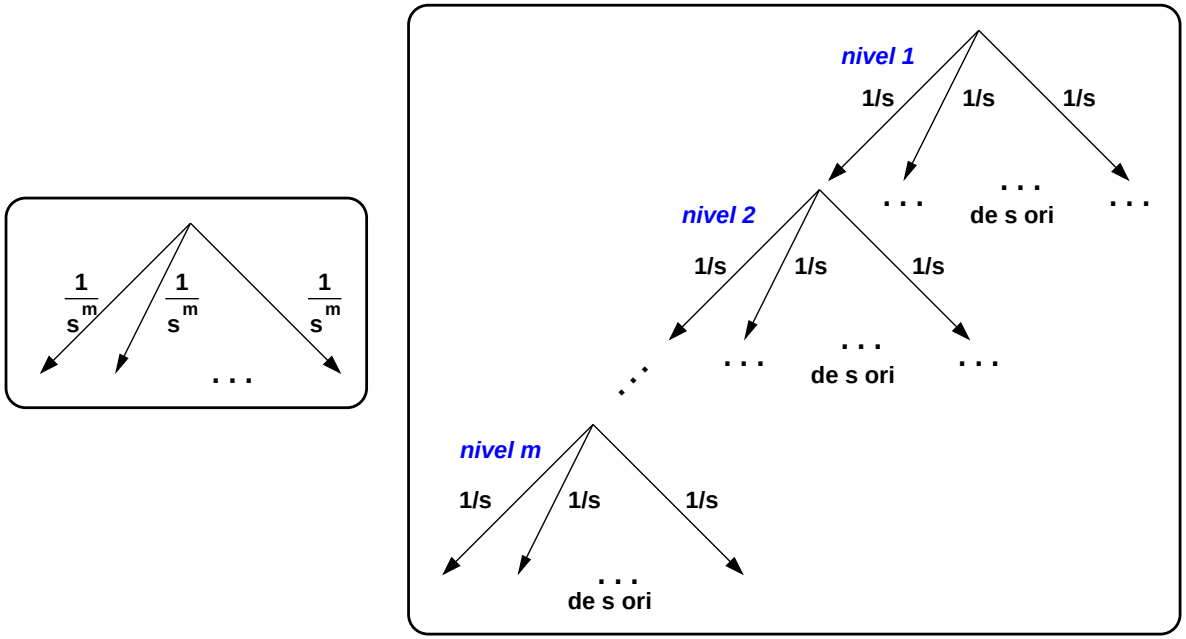
$$\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \leq \frac{2}{n} \quad \text{pentru} \quad s \neq 1 \tag{4}$$

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.

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

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

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** $\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \le \frac{1}{n}$ **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** $\frac{A(t)}{A(s)} = \frac{\log t}{\log s}$.

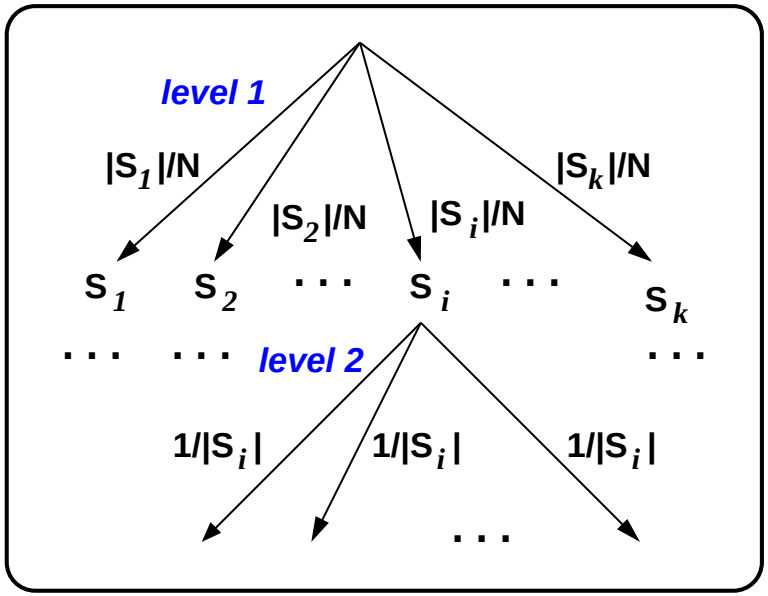
Finally, $A(t) = \frac{A(s)}{\log s} \log t = K \log t$, **where** $K = \frac{A(s)}{\log s} > 0$ **(if** $s \ne 1$**).**

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

Let's consider a set of N 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:



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

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

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

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

- **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 corelată 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):

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

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

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 $JSD(p||q) = \frac{1}{2}(KL(p||q) + KL(q||p))$**, 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 $KL(p||q) = 0$ 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:

$$KL(p \,||\, q) \overset{def.}{=} -\sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$\overset{Jensen}{\geq} -\log(\sum_x p(x)\frac{q(x)}{p(x)}) = -\log(\underbrace{\sum_x q(x)}_{1}) = -\log 1 = 0$$

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

$\Leftarrow$

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 $KL(p||q) = 0$.

$\Rightarrow$

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

Ţinând cont că $\sum_x p(x) = 1$ şi $\sum_x p(x)\frac{q(x)}{p(x)} = \sum_x q(x) = 1$, rezultă că $\frac{q(x)}{p(x)} = 1$ 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:

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

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

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

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.
Show that $IG(X, Y) \geq 0$ for any discrete random variables X and Y.
Moreover, $IG(X, Y) = 0$ iff X and Y are independent.

Answer:

The two statements are immediate consequences of parts a. and b. already proven.

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)

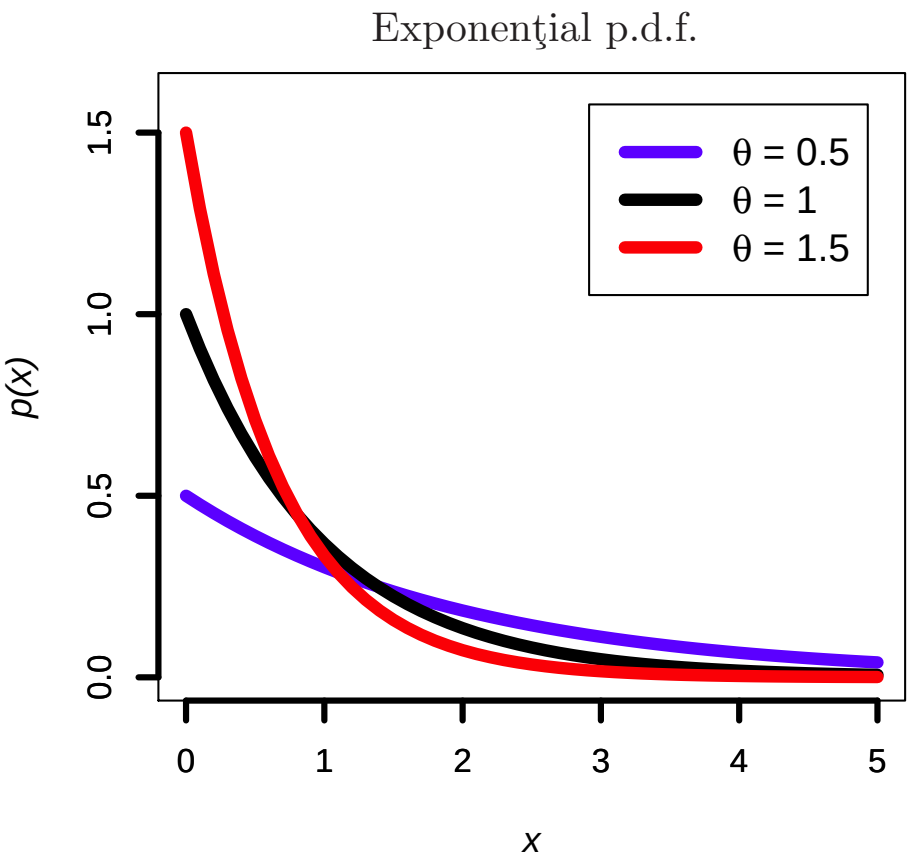
Dacă X şi Y sunt variabilele independente,
atunci $p(x_i, y_j) = p(x_i)p(y_j)$ pentru orice i şi j.
În consecinţă, toţi logaritmii din partea dreaptă a primei egalităţi din calculul de mai sus sunt 0 şi rezultă $IG(X, Y) = 0$.

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.

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 $\lambda > 0$. Vă reamintim că definiţia acestei distribuţii este următoarea:

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

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*:

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

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:

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

An upper bound for the entropy of a discrete distribution

CMU, 2003 fall, 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** $H(X) = \log_2 n$**.**