Slides for the

Machine Learning exercise book

(to be published in Romanian)

by Liviu Ciortuz, Alina Munteanu, Elena Bădărău

Faculty of Computer Science, University of Iaşi, Romania

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 number of *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

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

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

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:**

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

Binomial distribution: calculating the variance (cont'd)

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

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

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

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

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

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

The last integral is computed as shown on the next slide.

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, it follows:

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

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

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

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

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 $v = \dfrac{x-\mu}{\sigma}$ **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 $\int_{-\infty}^{\infty} v e^{-\dfrac{v^2}{2}} dv = 0$ **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

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

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

So,

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

Finally,

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

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

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

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

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

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 $X = [X_1 \ldots X_n]^T$, $\mu \in \mathbb{R}^n$ 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):

$$x = \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] \quad \mu = \left[\begin{array}{c} \mu_1 \\ \mu_2 \end{array}\right] \quad \Sigma = \left[\begin{array}{cc} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{array}\right]$$

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)

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

Random Variables:

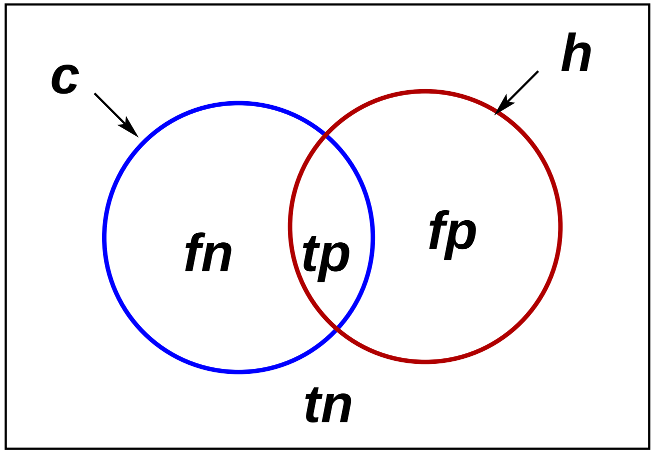
Some exercises

Exemplifying
the computation of *expected values* for random variables
and the [use of] *sensitivity* and *specificity* of a test
in a real-world application

CMU, 2009 fall, Geoff Gordon, HW1, pr. 2

There is a disease which affects 1 in 500 people. A 100.00 dollar blood test can help reveal whether a person has the disease. A positive outcome indicates that the person may have the disease.

The test has perfect sensitivity (*true positive rate*), i.e., a person who has the disease tests positive 100% of the time. However, the test has 99% specificity (*true negative rate*), i.e., a healthy person tests positive 1% of the time.



sensitivity (or: recall): $\dfrac{tp}{tp + fn}$

specificity: $\dfrac{tn}{tn + fp}$

a. A randomly selected individual is tested and the result is positive.

What is the *probability* of the individual having the disease?

b. There is a second more expensive test which costs 10, 000.00 dollars but is exact with 100% *sensitivity* and *specificity*.

If we require all people who test positive with the less expensive test to be tested with the more expensive test, what is the *expected cost* to check whether an individual has the disease?

c. A pharmaceutical company is attempting to decrease the cost of the second (perfect) test.

How much would it have to make the second test cost, so that the first test is no longer needed? That is, at what cost is it cheaper simply to use the perfect test alone, instead of screening with the cheaper test as described in part b?

Answer:

Let's define the following *random variables*:

B: $\begin{cases} \textbf{1/true} & \textbf{for persons affected by that disease,} \\ \textbf{0/false} & \textbf{otherwise;} \end{cases}$

T_1: **the result of the first test: + (in case of disease) or − (otherwise);**

T_2: **the result of the second test: again + or −.**

***Known facts*:**

$$P(B) = \frac{1}{500}$$
$$P(T_1 = + \mid B) = 1,\ P(T_1 = + \mid \bar{B}) = \frac{1}{100},$$
$$P(T_2 = + \mid B) = 1,\ P(T_2 = + \mid \bar{B}) = 0$$

a.

$$P(B \mid T_1 = +) \overset{TBayes}{=} \frac{P(T_1 = + \mid B) \cdot P(B)}{P(T_1 = + \mid B) \cdot P(B) + P(T_1 = + \mid \bar{B}) \cdot P(\bar{B})}$$
$$= \frac{1 \cdot \dfrac{1}{500}}{1 \cdot \dfrac{1}{500} + \dfrac{1}{100} \cdot \dfrac{499}{500}} = \frac{100}{599} \approx 0.1669$$

b.

Let's consider a new random variable:

$$C = \begin{cases} c_1 & \textbf{if the person only takes the first test} \\ c_1 + c_2 & \textbf{if the person takes the two tests} \end{cases}$$

$$\Rightarrow P(C = c_1) = P(T_1 = -) \textbf{ and } P(C = c_1 + c_2) = P(T_1 = +)$$

$$\begin{aligned} \Rightarrow E[C] &= c_1 \cdot (1 - P(T_1 = +)) + (c_1 + c_2) \cdot P(T_1 = +) \\ &= c_1 - c_1 \cdot P(T_1 = +) + c_1 \cdot P(T_1 = +) + c_2 \cdot P(T_1 = +) \\ &= c_1 + c_2 \cdot P(T_1 = +) \\ &= 100 + 10000 \cdot \frac{599}{50000} = 219.8 \approx 220\$ \end{aligned}$$

Note: Here above we used

$$\begin{aligned} P(T_1 = +) &\overset{\textit{total probability form.}}{=} P(T_1 = + \mid B) \cdot P(B) + P(T_1 = + \mid \bar{B}) \cdot P(\bar{B}) \\ &= 1 \cdot \frac{1}{500} + \frac{1}{100} \cdot \frac{499}{500} = \frac{599}{50000} = 0.01198 \end{aligned}$$

c.

$c_n \overset{not.}{=}$ **the new price for the second test** (T_2')

$$\begin{aligned} c_n \leq E[C'] &= c_1 \cdot P(C = c_1) + (c_1 + c_n) \cdot P(C = c_1 + c_n) \\ &= c_1 + c_n \cdot P(T_1 = +) = 100 + c_n \cdot \frac{599}{50000} \end{aligned}$$

$c_n = 100 + c_n \cdot 0.01198 \Rightarrow c_n \approx 101.2125$**.**

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;

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

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

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

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*:**

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

***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*:**

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

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

***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, in sense MLE and MAP

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.

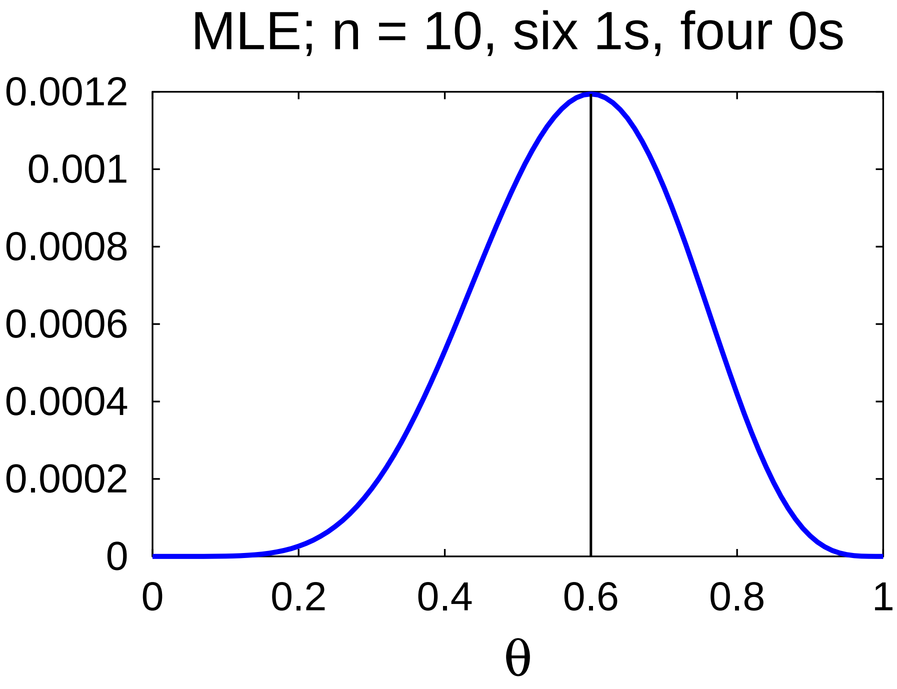
b. Suppose that $n = 10$ and the data set contains six 1s and four 0s.

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:

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

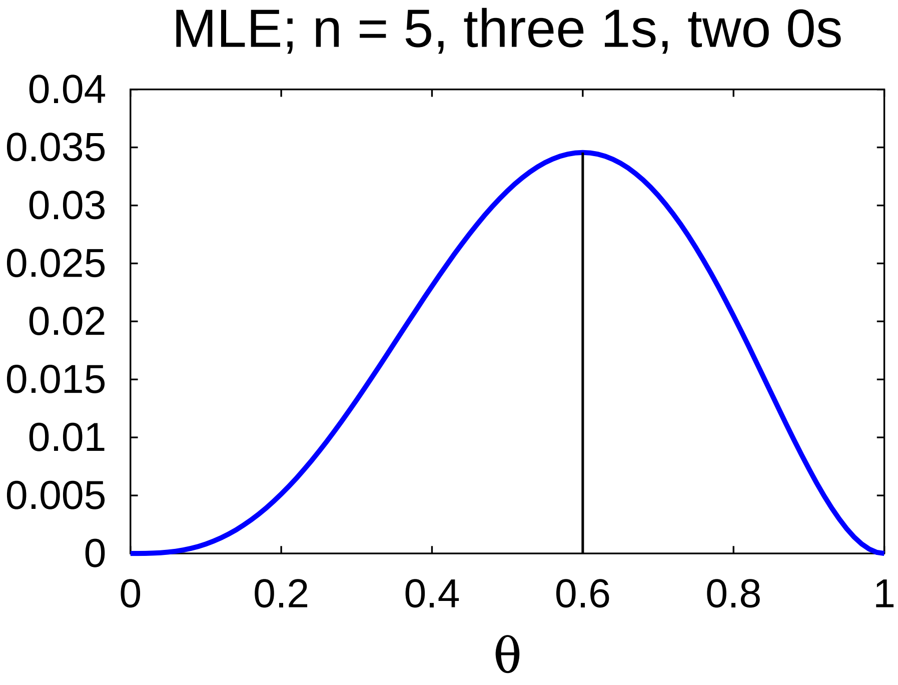
Assuming that $\hat{\theta} \neq 0$ and $\hat{\theta} \neq 1$, the first and second derivatives of l are given by

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

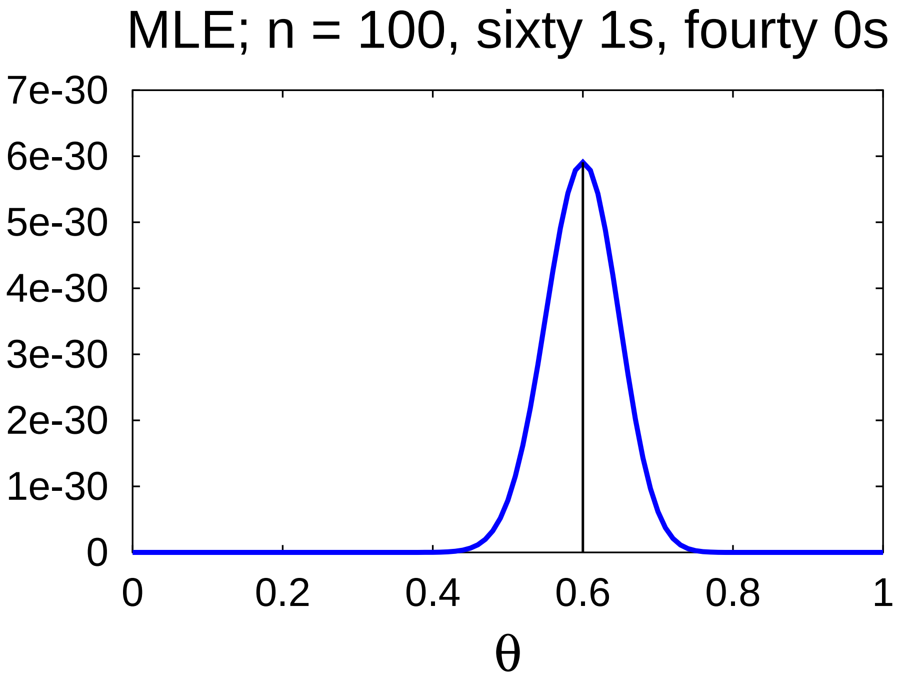
Since $l''(\hat{\theta})$ is always negative, the l function is concave, and we can find its maximizer by solving the equation $l'(\theta) = 0$. The solution to this equation is given by $\hat{\theta}_{MLE} = \dfrac{n_1}{n_1 + n_0}$.

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:

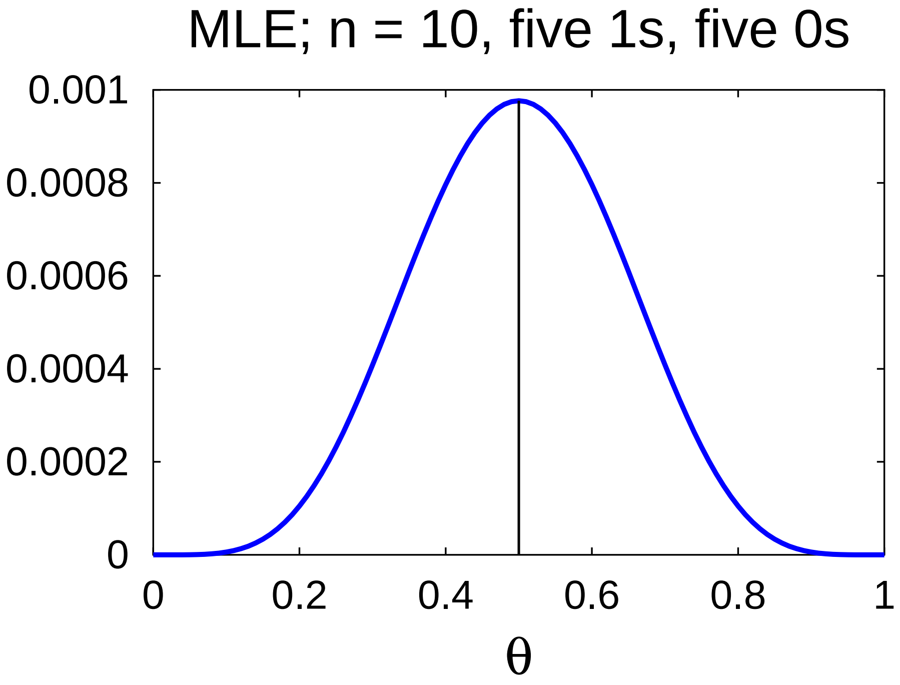


MLE; n = 5, three 1s, two 0s



MLE; n = 100, sixty 1s, fourty 0s

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

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

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

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

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 $B(3,3) \approx 0.0333$.

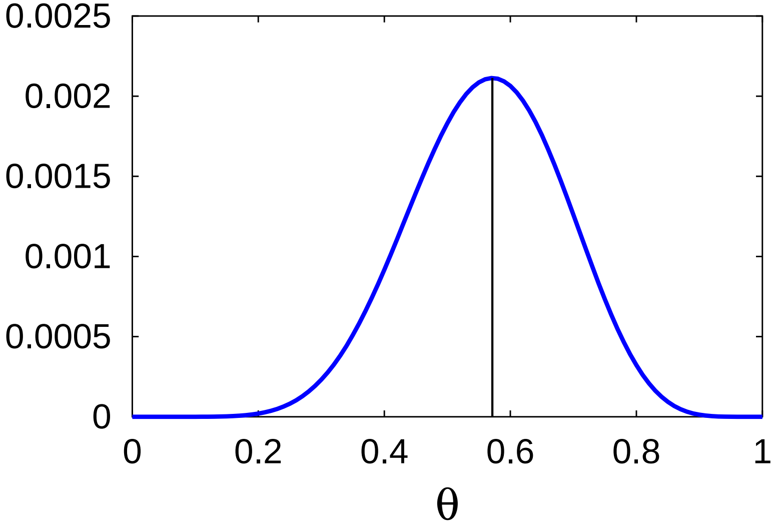
f. Suppose, as in part c, that $n = 10$ and we observed six 1s and four 0s.

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:

MAP; n = 10, six 1s, four 0s; Beta(3,3)



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.

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:

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

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

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

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

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

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

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:

$$P(\mu|x_1,\ldots,x_n) \overset{T.\ Bayes}{=} \frac{P(x_1,\ldots,x_n|\mu)\,P(\mu)}{P(x_1,\ldots,x_n)} \tag{1}$$

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

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

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

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

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

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:

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

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

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:

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

Note: **It can be easily shown that $L(\sigma^2 | X_1, X_2, \ldots, X_n)$ 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:

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

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 $f(x) = \dfrac{1}{\sqrt{2\pi\sigma}} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}$.

The log likelihood function of the data x is:

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

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

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

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

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

Solution:

$$
\begin{aligned}
E[\sigma_{MLE}] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_{MLE})^2\right] = E[(x_1-\mu_{MLE})^2] = E\left[(x_1-\frac{1}{n}\sum_{i=1}^{n}x_i)^2\right] \\
&= E\left[x_1^2-\frac{2}{n}x_1\sum_{i=1}^{n}x_i+\frac{1}{n^2}(\sum_{i=1}^{n}x_i)^2\right] \\
&= E\left[x_1^2-\frac{2}{n}x_1\sum_{i=1}^{n}x_i+\frac{1}{n^2}\sum_{i=1}^{n}x_i^2+\frac{2}{n^2}\sum_{i<j}x_ix_j\right] \\
&= E[x_1^2]+\frac{1}{n^2}\sum_{i=1}^{n}E[x_i^2]-\frac{2}{n}\sum_{i=1}^{n}E[x_1x_i]+\frac{2}{n^2}\sum_{i<j}E[x_ix_j] \\
&= E[x_1^2]+\frac{1}{n^2}nE[x_1^2]-\frac{2}{n}E[x_1^2]-\frac{2}{n}(n-1)E[x_1x_2]+\frac{2}{n^2}\frac{n(n-1)}{2}E[x_1x_2] \\
&= \frac{n-1}{n}E[x_1^2]-\frac{n-1}{n}E[x_1x_2]
\end{aligned}
$$

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

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 ($E[\sigma_{MLE}] = \dfrac{n-1}{n} E[x_1^2] - \dfrac{n-1}{n} E[x_1 x_2]$), 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 $\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_{MLE})^2$ **is an unbiased estimator of** σ^2**.**

Elements of Information Theory

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

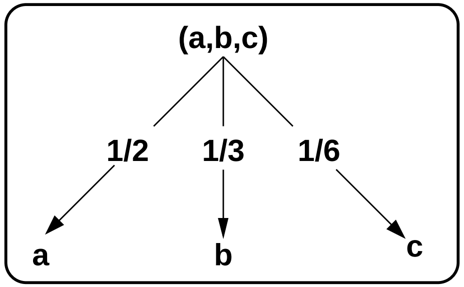
A1. H_n **should be continuous in** p_i **and symmetric in its arguments;**

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.)

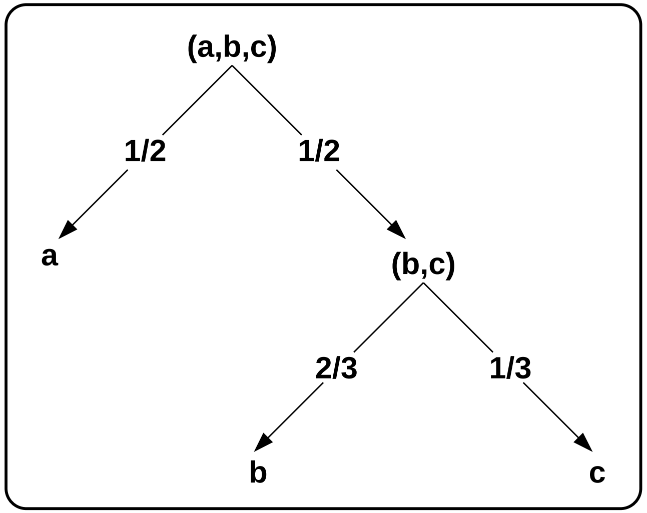
A3. if a choice among N **events is broken down into successive choices, then entropy should be the weighted sum of the entropy at each stage;**

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**

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

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**

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

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

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

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

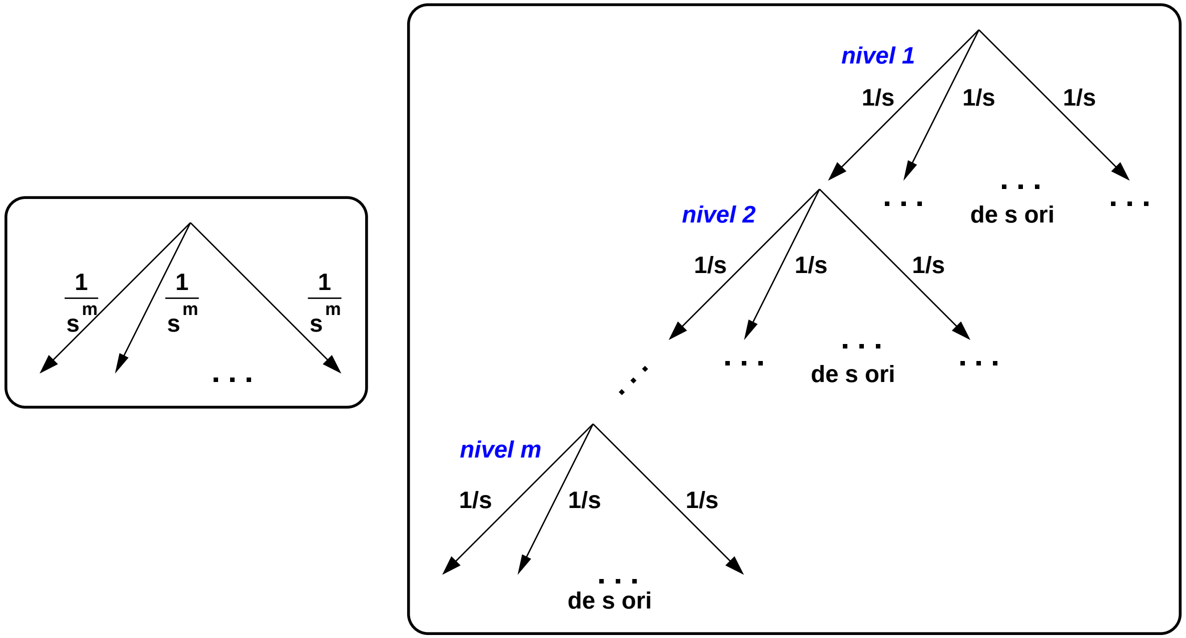
$$\mid \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \mid \leq \frac{2}{n} \text{ pentru } s \neq 1 \qquad (4)$$

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:

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

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** $\frac{A(t)}{A(s)} = \frac{\log t}{\log s}$**.**

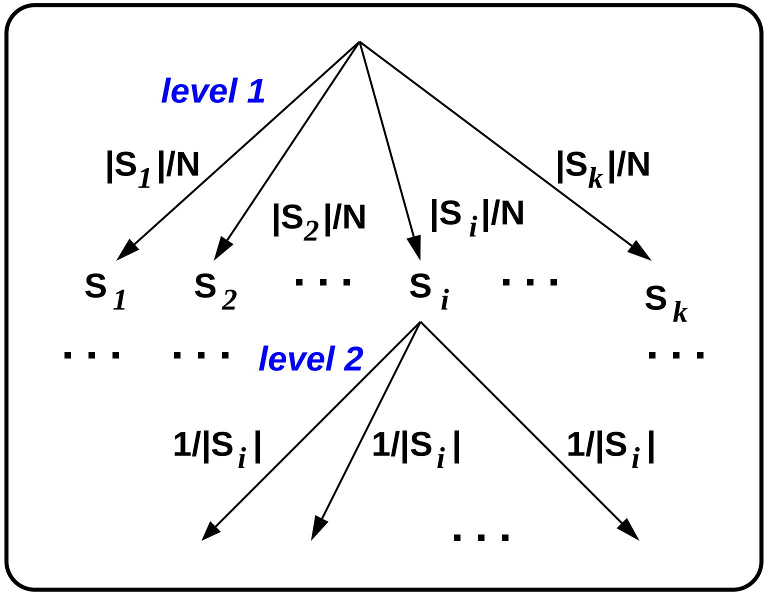
Finally, $A(t) = \frac{A(s)}{\log s} \log t = K \log t$**, where** $K = \frac{A(s)}{\log s} > 0$ **(if** $s \neq 1$**).**

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:



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

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

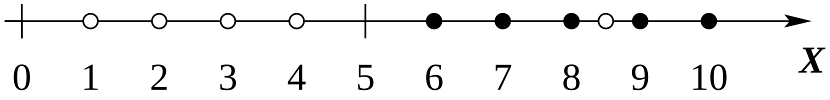
Some exercises

Exemplifying

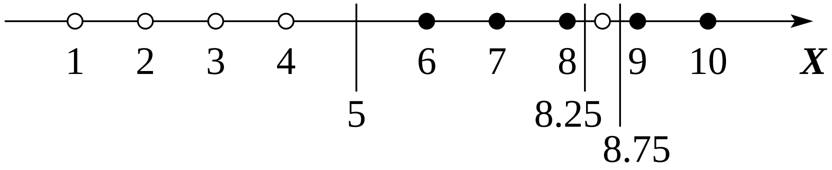
The application of the ID3 algorithm on continuous attributes; Decision surfaces; decision boundaries; The computation of the CVLOO error

CMU, 2002 fall, Andrew Moore, midterm, pr. 3

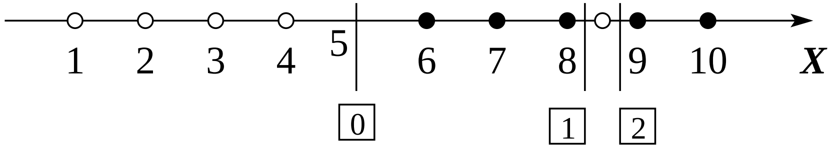
- **training data:**



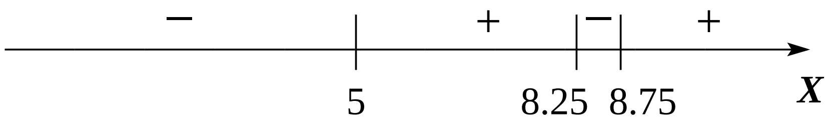
- **discretization / decision thresholds:**



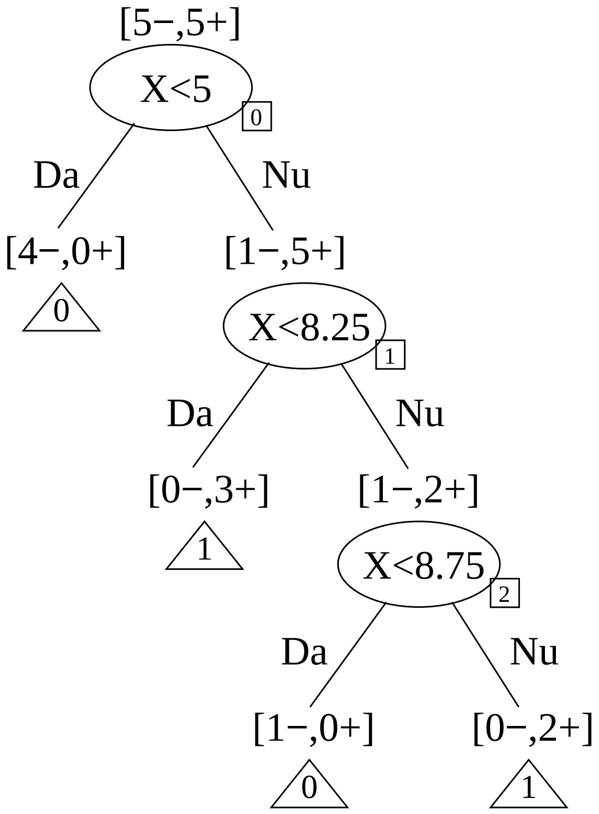
- **compact representation of the ID3 tree:**



- **decision “surfaces”:**

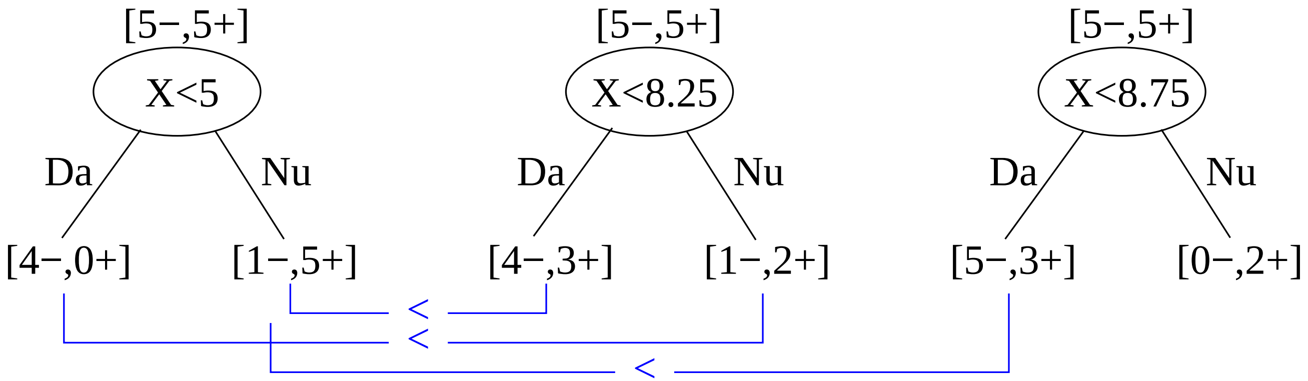


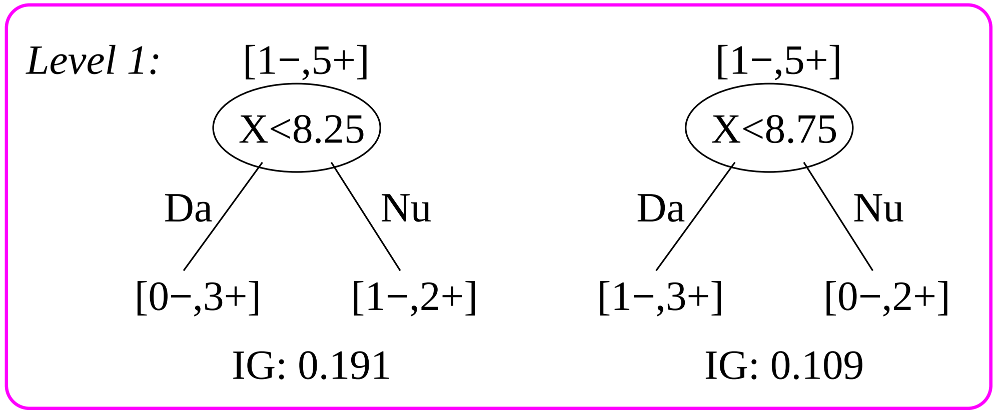
ID3 tree:



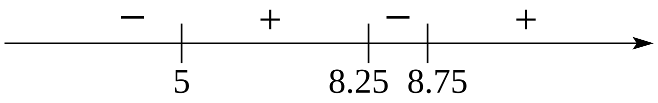
ID3: IG computations

Level 0:



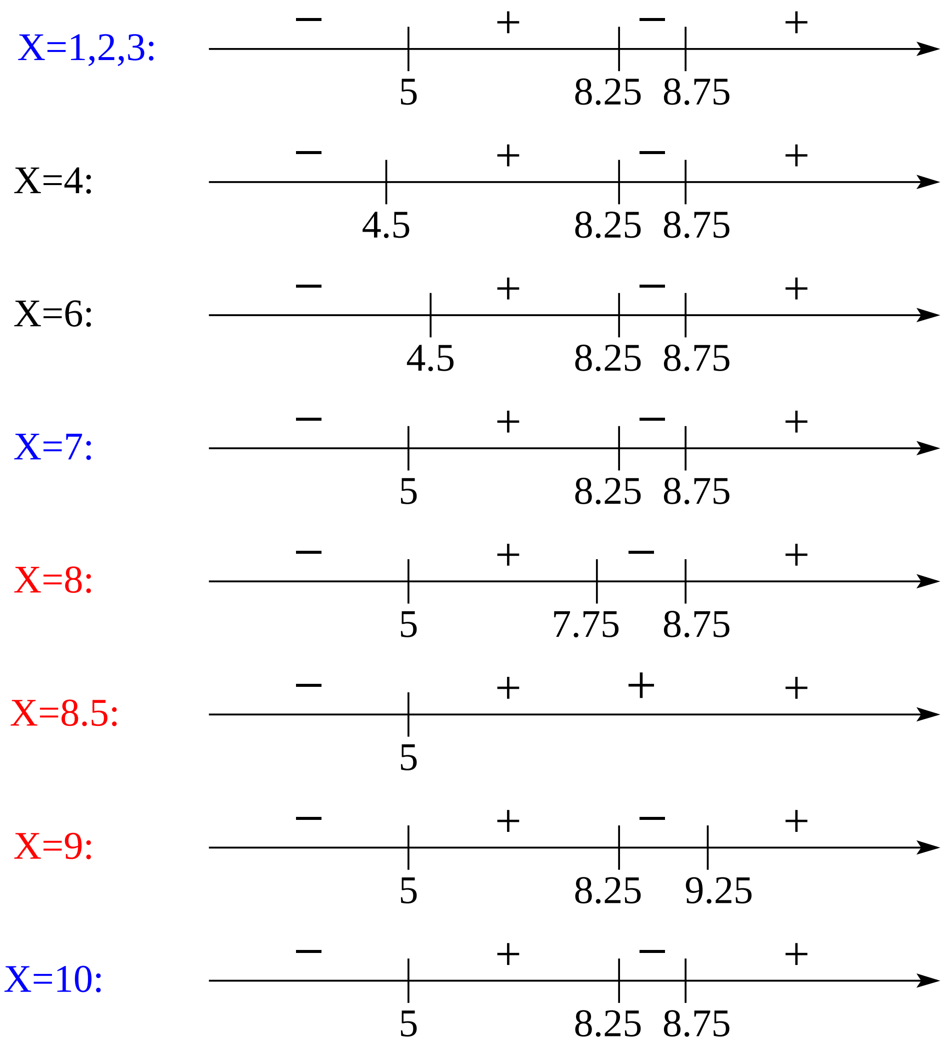


Decision "surfaces":

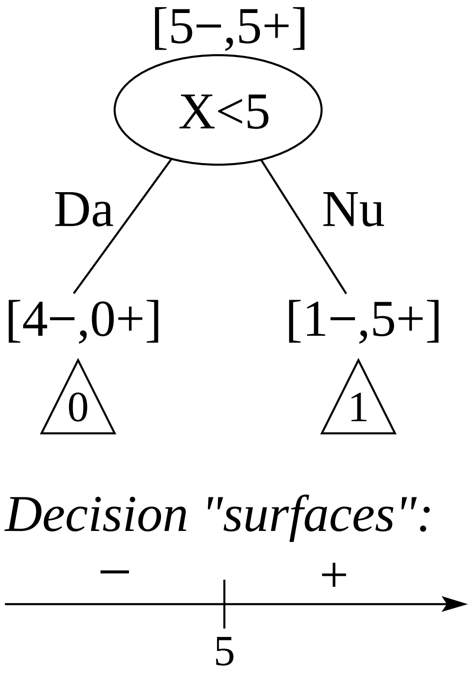


ID3, CVLOO: Decision surfaces

CVLOO error: 3/10

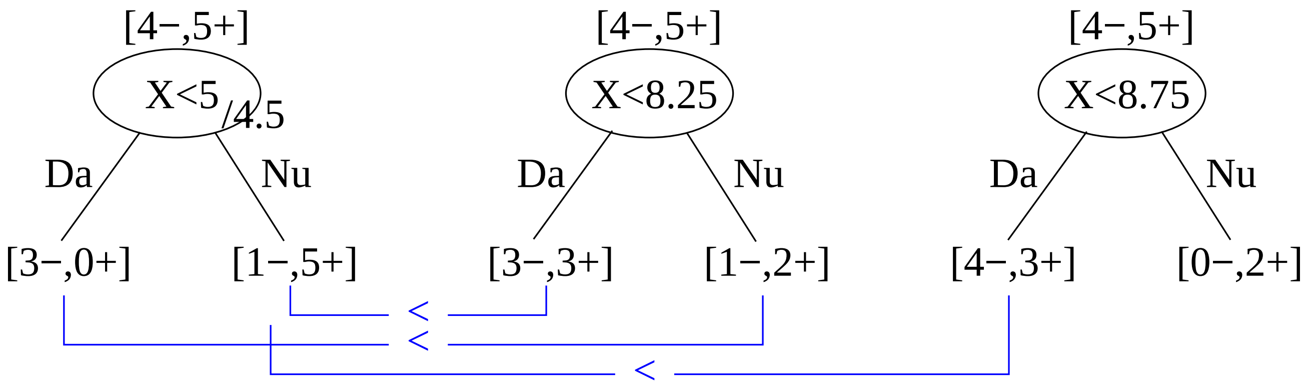

X=1,2,3: − 5 + 8.25 − 8.75 +
X=4: − 4.5 + 8.25 − 8.75 +
X=6: − 4.5 + 8.25 − 8.75 +
X=7: − 5 + 8.25 − 8.75 +
X=8: − 5 + 7.75 − 8.75 +
X=8.5: − 5 + + +
X=9: − 5 + 8.25 − 9.25 +
X=10: − 5 + 8.25 − 8.75 +

DT2

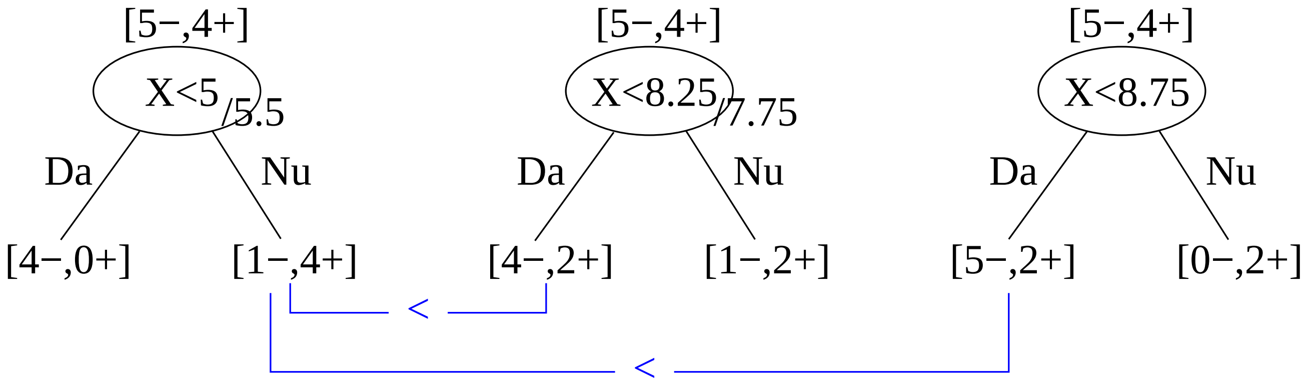


DT2, CVLOO
IG computations

Case 1: X=1, 2, 3, 4

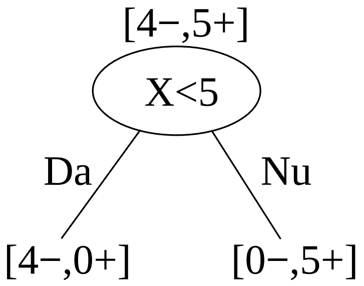


Case 2: X=6, 7, 8

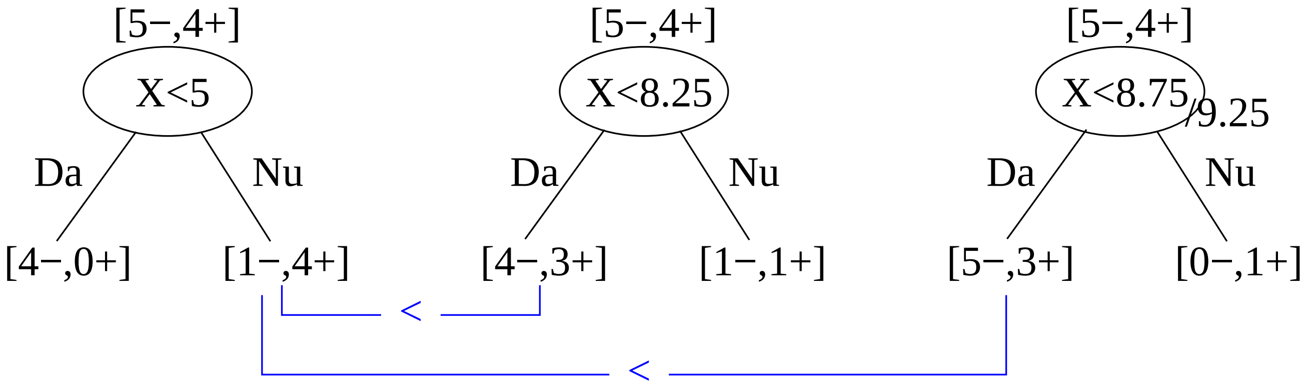


DT2, CVLOO IG computations (cont'd)

Case 3: X=8.5



Case 2: X=9, 10



CVLOO error: 1/10

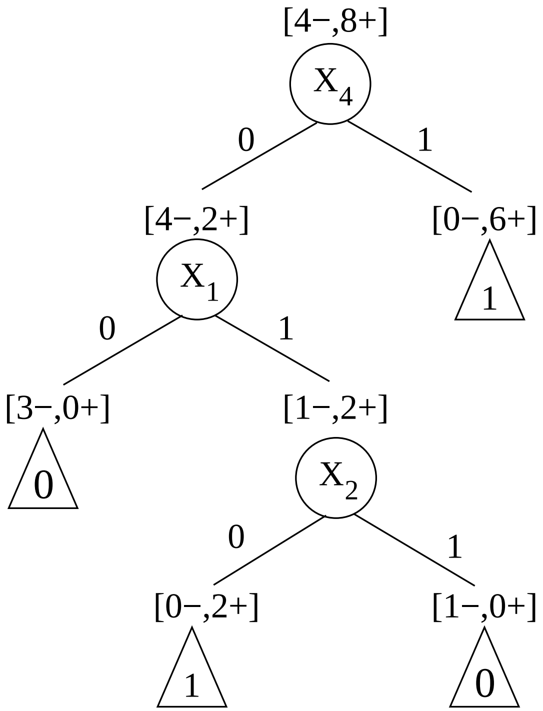
Exemplifying

χ^2-Based Pruning of Decision Trees

CMU, 2010 fall, Ziv Bar-Joseph, HW2, pr. 2.1

Input:

X_1	X_2	X_3	X_4	***Class***
1	1	0	0	0
1	0	1	0	1
0	1	0	0	0
1	0	1	1	1
0	1	1	1	1
0	0	1	0	0
1	0	0	0	1
0	1	0	1	1
1	0	0	1	1
1	1	0	1	1
1	1	1	1	1
0	0	0	0	0



Idea

While traversing the ID3 tree [usually in bottom-up manner],
remove the nodes for which
there is not enough ("significant") statistical evidence that
there is a dependence between
the values of the input attribute tested in that node and the values of the output attribute (the labels),
supported by the set of instances assigned to that node.

Contingency tables

O_{X_4}	$X_4 = 0$	$X_4 = 1$
Class = 0	4	0
Class = 1	2	6

$$\overset{N=12}{\Longrightarrow} \begin{cases} P(\boldsymbol{Class} = 0) = \frac{4}{12} = \frac{1}{3},\ P(\boldsymbol{Class} = 1) = \frac{2}{3} \\ P(X_4 = 0) = \frac{6}{12} = \frac{1}{2},\ P(X_4 = 1) = \frac{1}{2} \end{cases}$$

$O_{X_1 \mid X_4=0}$	$X_1 = 0$	$X_1 = 1$
Class = 0	3	1
Class = 1	0	2

$$\overset{N=6}{\Longrightarrow} \begin{cases} P(\boldsymbol{Class} = 0 \mid X_4 = 0) = \frac{4}{6} = \frac{2}{3} \\ P(\boldsymbol{Class} = 1 \mid X_4 = 0) = \frac{1}{3} \\ P(X_1 = 0 \mid X_4 = 0) = \frac{3}{6} = \frac{1}{2} \\ P(X_1 = 1 \mid X_4 = 0) = \frac{1}{2} \end{cases}$$

$O_{X_2 \mid X_4=0, X_1=1}$	$X_2 = 0$	$X_2 = 1$
Class = 0	0	1
Class = 1	2	0

$$\overset{N=3}{\Longrightarrow} \begin{cases} P(\boldsymbol{Class} = 0 \mid X_4 = 0, X_1 = 1) = \frac{1}{3} \\ P(\boldsymbol{Class} = 1 \mid X_4 = 0, X_1 = 1) = \frac{2}{3} \\ P(X_2 = 0 \mid X_4 = 0, X_1 = 1) = \frac{2}{3} \\ P(X_2 = 1 \mid X_4 = 0, X_1 = 1) = \frac{1}{3} \end{cases}$$

The rationale behind the computation of the expected number of observations

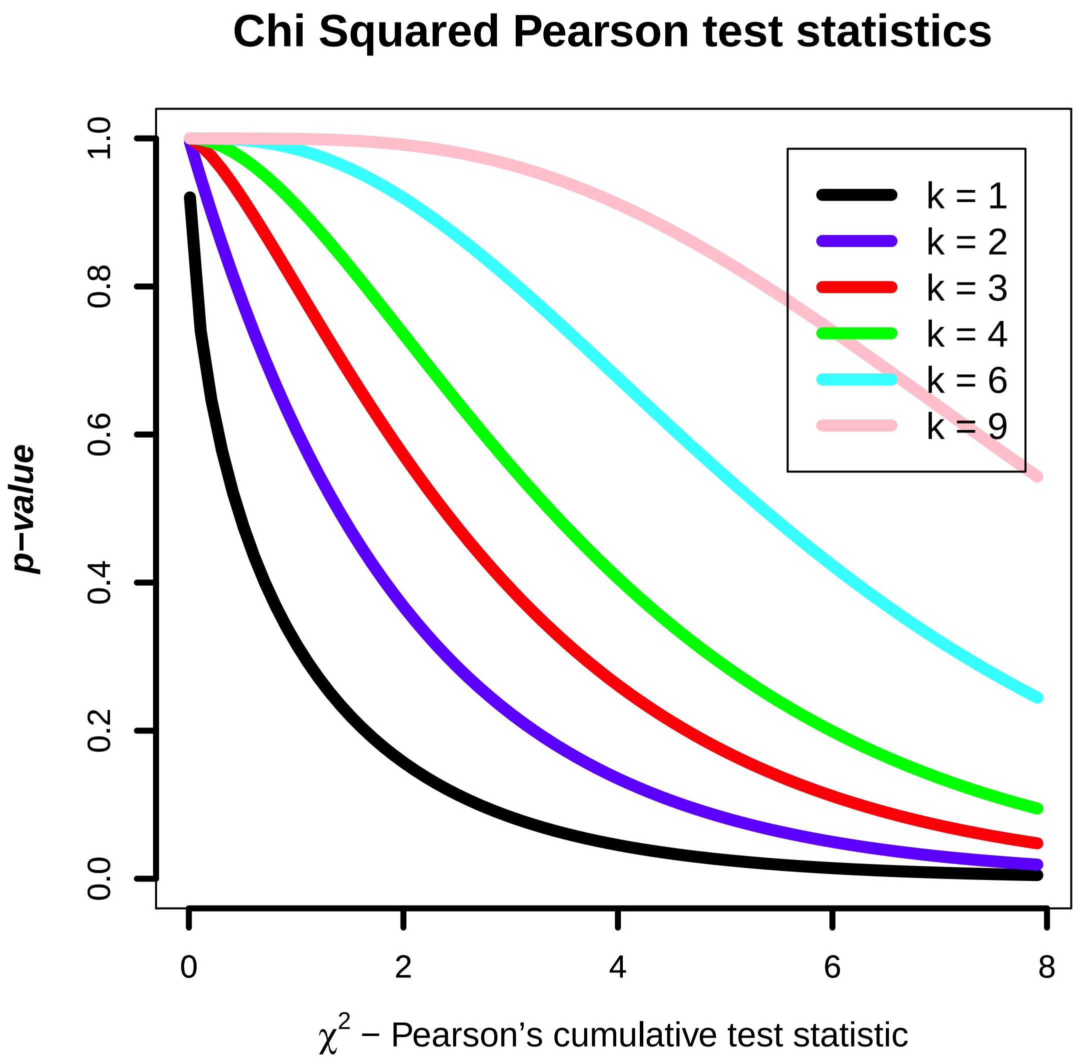
$$P(C=i, A=j) = P(C=i)\cdot P(A=j)$$

$$P(C=i) = \frac{\sum_{k=1}^{c} O_{i,k}}{N} \text{ and } P(A=j) = \frac{\sum_{k=1}^{r} O_{k,j}}{N}$$

$$P(C=i, A=j) \overset{\textit{indep.}}{=} \frac{\left(\sum_{k=1}^{c} O_{i,k}\right)\left(\sum_{k=1}^{r} O_{k,j}\right)}{N^2}$$

$$E[C=i, A=j] = N\cdot P(C=i, A=j)$$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$$


Chi Squared Pearson test statistics
p−value
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
χ^2 − Pearson's cumulative test statistic
k = 1
k = 2
k = 3
k = 4
k = 6
k = 9

Expected number of observations

E_{X_4}	$X_4 = 0$	$X_4 = 1$
Class = 0	2	2
Class = 1	4	4

$E_{X_1 \mid X_4}$	$X_1 = 0$	$X_1 = 1$
Class = 0	2	2
Class = 1	1	1

$E_{X_2 \mid X_4, X_1 = 1}$	$X_2 = 0$	$X_2 = 1$
Class = 0	$\frac{2}{3}$	$\frac{1}{3}$
Class = 1	$\frac{4}{3}$	$\frac{2}{3}$

$E_{X_4}(\boldsymbol{Class} = 0, X_4 = 0)$:

$N = 12, P(\boldsymbol{Class} = 0) = \frac{1}{3}$ **şi** $P(X_4 = 0) = \frac{1}{2} \Rightarrow$

$$N \cdot P(\boldsymbol{Class} = 0, X_4 = 0) = N \cdot P(\boldsymbol{Class} = 0) \cdot P(X_4 = 0) = 12 \cdot \frac{1}{3} \cdot \frac{1}{2} = 2$$

χ^2 Statistics

$$\chi^2{}_{X_4} = \frac{(4-2)^2}{2} + \frac{(0-2)^2}{2} + \frac{(2-4)^2}{4} + \frac{(6-4)^2}{4} = 2+2+1+1 = 6$$

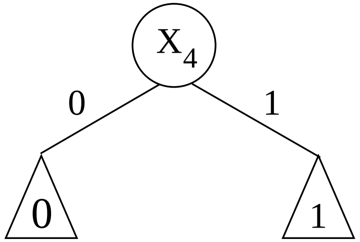
$$\chi^2{}_{X_1|X_4=0} = \frac{(3-2)^2}{2} + \frac{(1-2)^2}{2} + \frac{(0-1)^2}{1} + \frac{(2-1)^2}{1} = 3$$

$$\chi^2{}_{X_2|X_4=0,X_1=1} = \frac{\left(0-\frac{2}{3}\right)^2}{\frac{2}{3}} + \frac{\left(1-\frac{1}{3}\right)^2}{\frac{1}{3}} + \frac{\left(2-\frac{4}{3}\right)^2}{\frac{4}{3}} + \frac{\left(0-\frac{2}{3}\right)^2}{\frac{2}{3}} = \frac{4}{9}\cdot\frac{27}{4} = 3$$

***p*-values:** 0.0143, 0.0833, **and respectively** 0.0833.

The first one of these *p*-values is smaller than ε, therefore the root node (X_4) cannot be prunned.

Output (pruned tree) for 95% confidence level



Some exercises

Exemplifying

Text classification using the Naive Bayes algorithm

CMU, 2009 spring, Ziv Bar-Joseph, midterm, pr. 2

Training data:

‘study’	‘free’	‘money’	Category	*count*
1	0	0	Regular	1
0	0	1	Regular	1
1	0	0	Regular	1
1	1	0	Regular	1
0	1	0	Spam	4
0	1	1	Spam	4

Estimating the parameters, by MLE and applying Laplace’s rule (“add-one”):

$$P(\textbf{study}|\textbf{spam}) = \frac{0+1}{8+2} = \frac{1}{10} \qquad P(\textbf{study}|\textbf{regular}) = \frac{3+1}{4+2} = \frac{2}{3}$$

$$P(\textbf{free}|\textbf{spam}) = \frac{8+1}{8+2} = \frac{9}{10} \qquad P(\textbf{free}|\textbf{regular}) = \frac{1+1}{4+2} = \frac{1}{3}$$

$$P(\textbf{money}|\textbf{spam}) = \frac{4+1}{8+2} = \frac{1}{2} \qquad P(\textbf{money}|\textbf{regular}) = \frac{1+1}{4+2} = \frac{1}{3}$$

Classification of the message
$s =$ "money for psychology study",
using the a priori probability $P(\text{spam}) = 0.1$:

$$P(\text{spam} \mid s) = P(\text{spam} \mid \text{study}, \neg\text{free}, \text{money})$$

$$\overset{F.\ Bayes}{=} \frac{P(\text{study},\ \neg\text{free},\ \text{money} \mid \text{spam}) \cdot P(\text{spam})}{P(\text{study},\neg\text{free},\text{money} \mid \text{spam})P(\text{spam}) + P(\text{study},\neg\text{free},\text{money} \mid \text{reg})P(\text{reg})}$$

$$P(\text{study},\ \neg\text{free},\ \text{money}|\text{spam}) \overset{indep.\ cdt.}{=} P(\text{study}|\text{spam})\cdot P(\neg\text{free}|\text{spam})\cdot P(\text{money}|\text{spam})$$

$$= \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{2} = \frac{1}{200}$$

$$P(\text{study},\ \neg\text{free},\ \text{money}|\text{reg}) \overset{indep.\ cdt.}{=} P(\text{study}|\text{reg})\cdot P(\neg\text{free}|\text{reg})\cdot P(\text{money}|\text{reg})$$

$$= \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} = \frac{4}{27}$$

Therefore,

$$P(\text{spam}|\ s) = \frac{\frac{1}{200} \cdot \frac{1}{10}}{\frac{1}{200} \cdot \frac{1}{10} + \frac{4}{27} \cdot \frac{9}{10}} \approx 0.0037$$

Exemplifying

The computation of the *error rate* for the Naive Bayes algorithm

CMU, 2010 fall, Aarti Singh, HW1, pr. 4.2

Consider a simple learning problem of determining whether Alice and Bob from CA will go to hiking or not $Y : \mathit{Hike} \in \{T, F\}$ given the weather conditions $X_1 : \mathit{Sunny} \in \{T, F\}$ and $X_2 : \mathit{Windy} \in \{T, F\}$ by a Naive Bayes classifier.

Using training data, we estimated the parameters

$$\begin{array}{ll} P(\mathit{Hike}) = 0.5 & \\ P(\mathit{Sunny} \mid \mathit{Hike}) = 0.8, & P(\mathit{Sunny} \mid \neg \mathit{Hike}) = 0.7 \\ P(\mathit{Windy} \mid \mathit{Hike}) = 0.4, & P(\mathit{Windy} \mid \neg \mathit{Hike}) = 0.5 \end{array}$$

Assume that the true distribution of X_1, X_2, and Y satisfies the Naive Bayes assumption of conditional independence with the above parameters.

a. What is the joint probability that Alice and Bob go to hiking and the weather is sunny and windy, that is $P(\mathit{Sunny}, \mathit{Windy}, \mathit{Hike})$?

Solution:

$$P(\mathit{Sunny}, \mathit{Windy}, \mathit{Hike}) \overset{cdt.\ indep.}{=} P(\mathit{Sunny} | \mathit{Hike}) \cdot P(\mathit{Windy} | \mathit{Hike}) \cdot P(\mathit{Hike}) = 0.8 \cdot 0.4 \cdot 0.5 = 0.16.$$

b. What is the expected error rate of the Naive Bayes classifier?
(Informally, the expected error rate is the probability that an "observation"/instance randomly generated according to the *true* probabilistic distribution of data is incorrectly classified by the Naive Bayes algorithm.)

Solution:

X_1	X_2	Y	$P(X_1, X_2, Y) =$ $P(X_1 \vert Y) \cdot P(X_2 \vert Y) \cdot P(Y)$	$Y_{NB}(X_1, X_2)$	$P_{NB}(Y \vert X_1, X_2)$
F	F	F	$0.3 \cdot 0.5 \cdot 0.5 = 0.075$	F	0.555556
F	F	T	$0.2 \cdot 0.6 \cdot 0.5 = \mathbf{0.060}$	F	0.444444
F	T	F	$0.3 \cdot 0.5 \cdot 0.5 = 0.075$	F	0.652174
F	T	T	$0.2 \cdot 0.4 \cdot 0.5 = \mathbf{0.040}$	F	0.347826
T	F	F	$0.7 \cdot 0.5 \cdot 0.5 = \mathbf{0.175}$	T	0.421686
T	F	T	$0.8 \cdot 0.6 \cdot 0.5 = 0.240$	T	0.578314
T	T	F	$0.7 \cdot 0.5 \cdot 0.5 = 0.175$	F	0.522388
T	T	T	$0.8 \cdot 0.4 \cdot 0.5 = \mathbf{0.160}$	F	0.477612

***Note*:** Joint probabilities corresponding to incorrect predictions are shown in bold.

$$
\begin{aligned}
\textit{error} &\overset{def.}{=} E_P[I_{Y_{NB}(X_1,X_2) \neq Y}] \\
&= \sum_{X_1, X_2, Y} I[Y_{NB}(X_1, X_2) \neq Y] \cdot P(X_1, X_2, Y) \\
&= \mathbf{0.060 + 0.040 + 0.175 + 0.160 = 0.435}
\end{aligned}
$$

***Note*:** I is the *indicator* function; its value is 1 whenever the associated condition (in our case, $f_{NB}(X_1, X_2) \neq Y$) is true, and 0 otherwise.

Next, suppose that we gather more information about weather conditions and introduce a new feature denoting whether the weather is X_3 : *Rainy* or not. *Assume* that each day the weather in CA can be either *Rainy* or *Sunny*. That is, it can not be both *Sunny* and *Rainy*. (Similarly, it can not be $\neg$ *Sunny* and $\neg$*Rainy*).

c. In the above new case, are any of the Naive Bayes assumptions violated? Why (not)? What is the joint probability that Alice and Bob go to hiking and the weather is sunny, windy and not rainy, that is $P(Sunny, Windy, \neg Rainy, Hike)$?

Solution:

The conditional independence of variables given the class label assumption of Naive Bayes is violated. Indeed, knowing if the weather is *Sunny* completely determines whether it is *Rainy* or not. Therefore, *Sunny* and *Rainy* are clearly NOT conditionally independent given *Hike*.

$$\begin{aligned}
&P(Sunny, Windy, \neg Rainy, Hike)\\
&\quad = \underbrace{P(\neg Rainy | Hike, Sunny, Windy)}_{1} \cdot P(Sunny, Windy | Hike) \cdot P(Hike)\\
&\quad \overset{cond.\ indep.}{=} P(Sunny | Hike) \cdot P(Windy | Hike) \cdot P(Hike)\\
&\quad = 0.8 \cdot 0.4 \cdot 0.5 = 0.16.
\end{aligned}$$

d. What is the expected error rate when the Naive Bayes classifier uses all three attributes? Does the performance of Naive Bayes improve by observing the new attribute Rainy? Explain why.

Solution:

X_1	X_2	X_3	Y	$P(X_1, X_2, Y)$	$P_{NB}(X_1, X_2, X_3, Y) = P(X_3)\cdot$ $P(X_1\|Y)\cdot P(X_2\|Y)\cdot P(Y)$	$Y_{NB}(X_1, X_2, X_3)$	$P_{NB}(Y\|X_1, X_2, X_3)$
F	F	F	F	0	$0.075 \cdot 0.7 = 0.0525$	F	0.522388
F	F	F	T	**0**	$0.060 \cdot 0.8 = 0.0480$	F	0.477612
F	F	T	F	0.075	$0.075 \cdot 0.3 = 0.0225$	F	0.652174
F	F	T	T	**0.060**	$0.060 \cdot 0.2 = 0.0120$	F	0.347826
F	T	F	F	0	$0.075 \cdot 0.7 = 0.0525$	F	0.621302
F	T	F	T	**0**	$0.040 \cdot 0.8 = 0.0320$	F	0.378698
F	T	T	F	0.075	$0.075 \cdot 0.3 = 0.0225$	F	0.737705
F	T	T	T	**0.040**	$0.040 \cdot 0.2 = 0.0080$	F	0.262295
T	F	F	F	**0.175**	$0.175 \cdot 0.7 = 0.0525$	T	0.389507
T	F	F	T	0.240	$0.240 \cdot 0.8 = 0.1920$	T	0.610493
T	F	T	F	0	$0.175 \cdot 0.3 = 0.0525$	F	0.522388
T	F	T	T	**0**	$0.240 \cdot 0.2 = 0.0480$	F	0.477612
T	T	F	F	**0.175**	$0.175 \cdot 0.7 = 0.0525$	T	0.489022
T	T	F	T	0.160	$0.160 \cdot 0.8 = 0.1280$	T	0.510978
T	T	T	F	0	$0.175 \cdot 0.3 = 0.0225$	F	0.621302
T	T	T	T	**0**	$0.060 \cdot 0.2 = 0.0120$	F	0.378698

The new error rate is:
$0.060 + 0.040 + 0.175 + 0.175 = 0.45 > 0.435$ (see question b).

The Naive Bayes classifier performance drops because the conditional independence assumptions do not hold for the correlated features.

Computing

The *sample complexity* of the Naive Bayes and Joint Bayes Clssifiers

CMU, 2010 spring, Eric Xing, Tom Mitchell, Aarti Singh, HW2, pr. 1.1

A big reason we use the Naive Bayes classifier is that it requires less training data than the Joint Bayes Classifier. This exercise should give you a "feeling" for how great the disparity really is.

Imagine that each *instance* is an independent "*observation*" of the multivariate random variable $\bar{X} = X_1, ..., X_d$, where the X_i are i.i.d. and Bernoulli of parameter $p = 0.5$.

To train the Joint Bayes classifier, we need to see every value of $\bar{X}$ "enough" times; training the Naive Bayes classifier only requires seeing both values of X_i "enough" times.

Main Question: How many "observations"/instances are needed until, with probability $1 - \varepsilon$, we have seen every variable we need to see at least once?

Note: To train the classifiers well would require more than this, but for this problem we only require one observation.

Hint: You may want to use the following *inequalities*:

- **For any $k \geq 1$, $(1 - 1/k)^k \leq e^{-1}$**
- **For *any* events $E_1, ..., E_k$, $Pr(E_1 \cup \ldots \cup E_k) \leq \sum_{i=1}^{k} Pr(E_i)$.**
  (This is called the "union bounds" property.)

Consider the Naive Bayes classifier.

a. **Show that if N observations have been made, the probability that a given value of X_i (either 0 or 1) has *not* been seen is $\leq \dfrac{1}{2^{N-1}}$.**

b. **Show that if more than $N_{NB} = 1 + \log_2 \left(\dfrac{d}{\varepsilon}\right)$ observations have been made, then the probability that *any* X_i has not been observed in both states is $\leq \varepsilon$.**

Solution:

a. $P($**component** X_i **not seen in both states**$) = \left(\dfrac{1}{2}\right)^N + \left(\dfrac{1}{2}\right)^N = \dfrac{2}{2^N} = \dfrac{1}{2^{N-1}}$

b. $P($**any component not seen in both states**$)$

$\leq \sum_{i=1}^{d} P($**component** X_i **not seen in both states**$)$

$= \sum_{i=1}^{d} \dfrac{1}{2^{N_{NB}-1}} = d \cdot \dfrac{1}{2^{N_{NB}-1}} = d \cdot \dfrac{1}{2^{1+\log_2 \frac{d}{\varepsilon}-1}} = d \cdot \dfrac{1}{2^{\log_2 \frac{d}{\varepsilon}}} = d \cdot \dfrac{1}{\dfrac{d}{\varepsilon}} = d \cdot \dfrac{\varepsilon}{d} = \varepsilon$

Consider the Joint Bayes classifier.

c. **Let $\bar{x}$ be a particular value of $\bar{X}$. Show that after N observations, the probability that we have never seen $\bar{x}$ is $\leq e^{-N/2^d}$.**

d. **Using the "union bounds" property, show that if more than $N_{JB} = 2^d \ln\left(\frac{2^d}{\varepsilon}\right)$ observations have been made, then the probability that an arbitrarily chosen (but fixed) value of $\bar{X}$ has not been seen is $\leq \varepsilon$.**

Solution:

c. $P(\bar{x}$ **not seen in** N **observations**$)$

$$= \left(1 - \frac{1}{2^d}\right)^N = \left[\left(1 - \frac{1}{2^d}\right)^{2^d}\right]^{N/2^d} \leq \left(\frac{1}{e}\right)^{N/2^d} = e^{-N/2^d}$$

d. $P($**any** $\bar{x}$ **not seen in** N_{JB} **observations**$)$

$\leq \sum_{\bar{x}} P(\bar{x}$ **not seen in** N_{JB} **observations**$)$

$$= \sum_{\bar{x}} e^{-N_{JB}/2^d} = 2^d \cdot e^{-N_{JB}/2^d} = 2^d \cdot e^{-\ln \frac{2^d}{\varepsilon}} = 2^d \cdot \frac{1}{e^{\ln \frac{2^d}{\varepsilon}}} = \frac{2^d}{\frac{2^d}{\varepsilon}} = \varepsilon$$

e. Let $d = 2$ and $\varepsilon = 0.1$. What are the values of N_{NB} and N_{JB}?
What about $d = 5$?
And $d = 10$?

Solution:

$$\varepsilon = 0.1,\ d = 2 \quad \Rightarrow \quad \begin{cases} N_{NB} = & 1 + \log_2 \dfrac{2}{0.1} & = & 1 + \log_2 20 \approx & 5.32 \\ N_{JB} = & 2^2 \cdot \ln \dfrac{2^2}{0.1} & = & 4 \cdot \ln 40 \approx & 14.75 \end{cases}$$

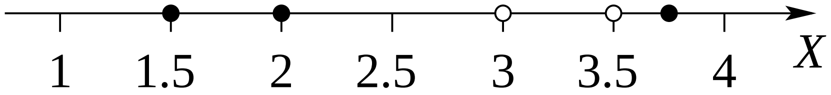
$$\varepsilon = 0.1,\ d = 5 \quad \Rightarrow \quad \begin{cases} N_{NB} = & 1 + \log_2 \dfrac{5}{0.1} & = & 1 + \log_2 50 \approx & 6.64 \\ N_{JB} = & 2^5 \cdot \ln \dfrac{2^5}{0.1} & = & 32 \cdot \ln 320 \approx & 184.58 \end{cases}$$

$$\varepsilon = 0.1,\ d = 10 \quad \Rightarrow \quad \begin{cases} N_{NB} = & 1 + \log_2 \dfrac{10}{0.1} & = & 1 + \log_2 100 \approx & 7.64 \\ N_{JB} = & 2^{10} \cdot \ln \dfrac{2^{10}}{0.1} & = & 1024 \cdot \ln 10240 \approx & 9455.67 \end{cases}$$

Exemplifying

ML hypotheses and *MAP hypotheses*

CMU, 2009 spring, Tom Mitchell, midterm, pr. 2.3-4



Let's consider the 1-dimensional data set shown above, based on the single real-valued attribute X. Notice there are two classes (values of Y), and five data points.

Consider a special type of *decision trees* where leaves have *probabilistic labels*. Each leaf node gives the probability of each possible label, where the probability is the fraction of points at that leaf node with that label.

For *example*, a decision tree learned from the data set above with zero splits would say $P(Y = 1) = 3/5$ and $P(Y = 0) = 2/5$. A decision tree with one split (at $X = 2.5$) would say $P(Y = 1) = 1$ if $X < 2.5$, and $P(Y = 1) = 1/3$ if $X \geq 2.5$.

a. For the above data set, draw a tree that maximizes the *likelihood* of the data.

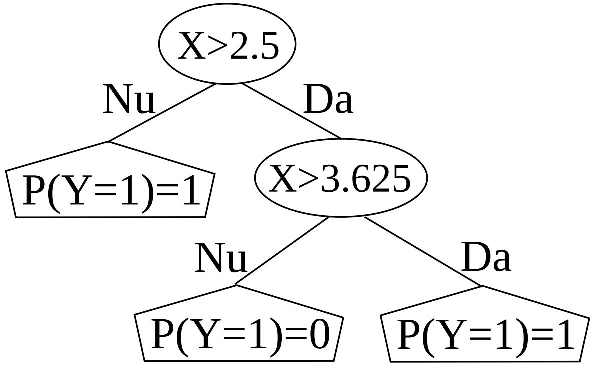
$T_{ML} = \text{argmax}_T \, P_T(D)$, **where**

$P_T(D) \overset{def.}{=} P(D|T) \overset{i.i.d.}{=} \prod_{i=1}^5 P(Y = y_i|X = x_i, T)$,

where y_i is the label/classs of the instance x_i

$(x_1 = 1.5,\ x_2 = 2,\ x_3 = 3,\ x_4 = 3.5,\ x_5 = 3.75.)$

Solution:



b. Consider a prior probability distribution $P(T)$ over trees that penalizes the number of splits in the tree.

$$P(T) \propto \left(\frac{1}{4}\right)^{splits(T)^2}$$

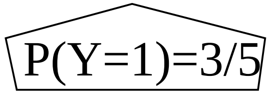
where T is a tree, $splits(T)$ is the number of splits in T, and $\propto$ means "is proportional to".

For the same data set, give the MAP tree when using this prior, $P(T)$, over trees.

Solution:

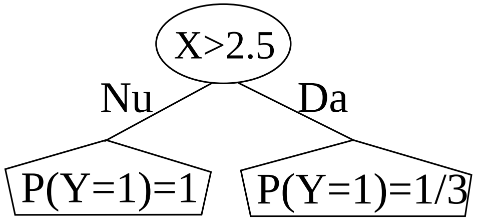
0 nodes:

$$P(T_0 \mid D) \propto \left(\frac{3}{5}\right)^3 \cdot \left(\frac{2}{5}\right)^2 \cdot \left(\frac{1}{4}\right)^0 = \frac{3^3 \cdot 2^2}{5^5} = \frac{108}{3125} = 0.0336$$



1 node:

$$P(T_1 \mid D) \propto 1^2 \cdot \left(\frac{2}{3}\right)^2 \cdot \frac{1}{3} \cdot \left(\frac{1}{4}\right)^1 = \frac{1}{27} = 0.037$$



2 nodes:

$$P(T_2) \propto \left(\frac{1}{4}\right)^4 \Rightarrow P(T_2 \mid D) \propto 1 \cdot \left(\frac{1}{4}\right)^4 = \frac{1}{256} = 0.0039 \Rightarrow \textbf{the MAP tree is } T_1.$$

The relationship between [the decision rules of] *Naive Bayes* and *Logistic Regression*

CMU, 2005 fall, Tom Mitchell, HW2, pr. 2

CMU, 2009 fall, Carlos Guestrin, HW1, pr. 4.1.2

CMU, 2009 fall, Geoff Gordon, HW4, pr. 1.2

CMU, 2012 fall, Tom Mitchell, Ziv Bar-Joseph, HW2, pr. 3.a

a. [Equivalence of NB and LR]

In Tom's draft chapter (*Generative and discriminative classifiers: Naive Bayes and logistic regression*) it has been proved that when Y is Boolean and $X = (X_1, \ldots, X_n)$ is a vector of continuous variables, then under certain assumptions the Gaussian Naive Bayes classifier implies that $P(Y|X)$ is given by the logistic function with appropriate parameters W. In particular:

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$

and

$$P(Y = 0|X) = \frac{\exp(w_0 + \sum_{i=1}^{n} w_i X_i)}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$

Consider instead the case where Y is Boolean and $X = (X_1, \ldots, X_n)$ is a vector of Boolean variables. Prove for this case also that $P(Y|X)$ follows this same form and hence that logistic regression is also the discriminative counterpart to a Naive Bayes generative classifier over Boolean features.

***Note*:**

***Discriminative classifiers* learn the parameters of $P(Y|X)$ directly, whereas *generative classifiers* instead learn the parameters of $P(X|Y)$ and $P(Y)$.**

***Hints*:**

1. Simple notation will help. Since the X_i are Boolean variables, you need only one parameter to define $P(X_i|Y = y_k)$. Define $\theta_{i1} = P(X_i = 1|Y = 1)$, in which case $P(X_i = 0|Y = 1) = 1 - \theta_{i1}$. Similarly, use θ_{i0} to denote $P(X_i = 1|Y = 0)$.

2. Notice with the above notation you can represent $P(X_i|Y = 1)$ as follows:

$$P(X_i = 1|Y = 1) = \theta_{i1}^{X_i}(1 - \theta_{i1})^{(1-X_i)}$$

Note when $X_i = 1$ the second term is equal to 1 because its exponent is zero. Similarly, when $X_i = 0$ the first term is equal to 1 because its exponent is zero.

Solution

b. [Relaxing the conditional independence assumption]

To capture interactions between features, the Logistic Regression model can be supplemented with extra terms. For example, a term can be added to capture a dependency between X_1 and X_2:

$$P(Y=1|X) = \frac{1}{1+\exp(w_0 + w_{1,2}X_1X_2 + \sum_{i=1}^{n} w_i X_i)}$$

Similarly, the conditional independence assumptions made by Naive Bayes can be relaxed so that X_1 and X_2 are not assumed to be conditionally independent. In this case, we can write:

$$P(Y|X) = \frac{P(Y)\, P(X_1, X_2|Y)\ \prod_{i=3}^{n} P(X_i|Y)}{P(X)}$$

Prove that for this case, that $P(Y|X)$ follows the same form as the logistic regression model supplemented with the extra term that captures the dependency between X_1 and X_2 (and hence that the supplemented Logistic Regression model is the discriminative counterpart to this generative classifier).

***Hints*:**

1. Using simple notation will help here as well. You need more parameters than before to define $P(X_1, X_2, Y)$. Define $\beta_{ijk} = P(X_1 = i, X_2 = j, Y = k)$.

2. The above notation can be used to represent $P(X_1, X_2 | Y = k)$ as follows:

$$P(X_1, X_2, Y = k) = (\beta_{11k})^{X_1 X_2} \, (\beta_{10k})^{X_1(1-X_2)} \, (\beta_{01k})^{(1-X_1)X_2} \, (\beta_{00k})^{(1-X_1)(1-X_2)}$$

Solution

Exemplifying the Gaussian [Naive] Bayes algorithm

CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW2, pr. 5.c

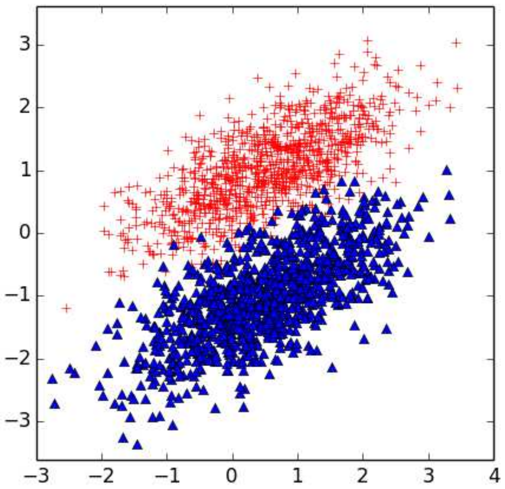
In a two dimensional case, we can visualize how Gaussian Naive Bayes behaves when input features are correlated. A data set is shown in Figure (A), where red points are in Class 0, blue points are in Class 1. The conditional distributions are two-dimensional Gaussians. In (B), (C) and (D), the ellipses represent conditional distributions for each class. The centers of ellipses show the means, and the contours show the boundary of two standard deviations.

a. Which of them is most likely to be the true conditional distribution?

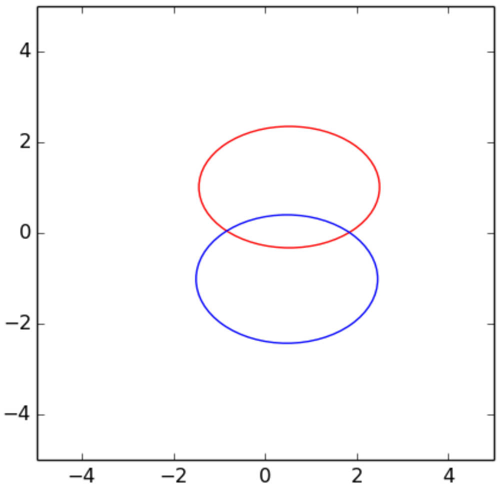
b. Which of them is most likely to be estimates by a Gaussian Naive Bayes model?

c. If we assume the prior probabilities for both classes are equal, which model will achieve a higher accuracy on the training data?

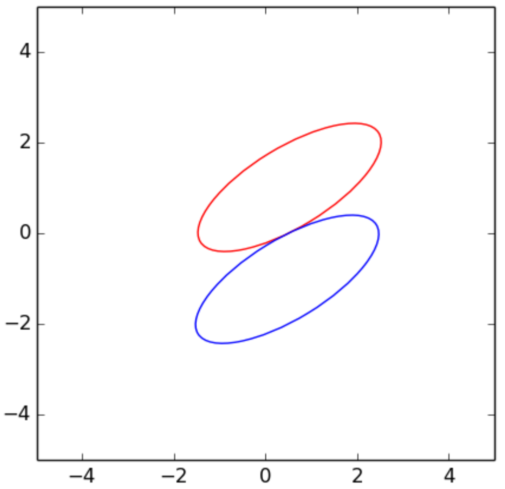
110.



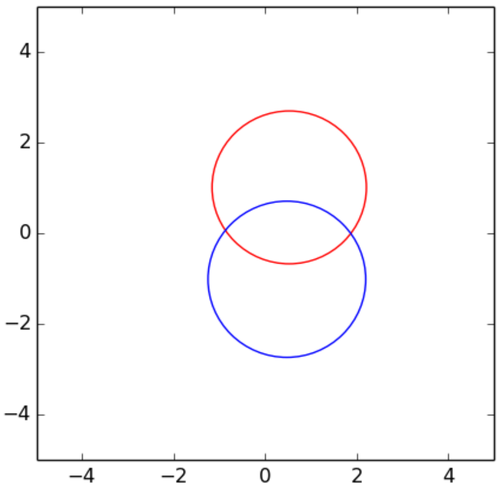
(A) Data



(B)



(C)



(D)

Solution:

a. (C) is the truth.

b. (B) corresponds to the Gaussian Naive Bayes estimates. [LC: Here follows the explanation:]
Because the Gaussian Naive Bayes model assume independence of the two features conditioned on the class label, the estimated model should be aligned with the axies. Both (B) and (D) satisfy this, but only in (B) the width and height of the oval, which are proportional to the standard deviation of each axis, matched the data.

c. (C) gives the lowest training error.

Proving

the relationship between the decision rules for *Gaussian Naive Bayes* and the *Logistic Regression* algorithm when the covariance matrices are diagonal and $\sigma_{i0}^2 = \sigma_{i1}^2$ for $i = 1, \ldots, d$

CMU, 2009 spring, Ziv Bar-Joseph, HW2, pr. 2

Assume a two-class ($Y \in \{0, 1\}$) Naive Bayes model over the d-dimensional real-valued input space $\mathbb{R}^d$, where the input variables $X|Y = 0 \in \mathbb{R}^d$ are distributed as

$$Gaussian(\mu_0 =< \mu_{01}, \ldots, \mu_{0d} >,\ \sigma =< \sigma_1, \ldots, \sigma_d >)$$

and $X|Y = 1 \in \mathbb{R}^d$ as

$$Gaussian(\mu_1 =< \mu_{11}, \ldots, \mu_{1d} >,\ \sigma =< \sigma_1, \ldots, \sigma_d >)$$

i.e., the inputs given the class have different means but identical variance for both classes.

Prove that, given the conditions stated above, the conditional probability $P(Y = 1|X = x)$, **where** $X = (X_1, \ldots, X_d)$ **and** $x = (x_1, \ldots, x_d)$ **can be written in a simiar form to Logistic Regression:**

$$\frac{1}{1 + \exp(w_0 + w \cdot x)}$$

with the parameters $w_0 \in \mathbb{R}$ **and** $w = (w_1, \ldots, w_d) \in \mathbb{R}^d$ **chosen in a suitable way.**

As a consequence, the decision rule for the Gaussean Bayes classifier supported by this model the desion rule has a linear form.

Solution

$$
\begin{aligned}
P(Y=1|X=x) &\overset{B.F.}{=} \frac{P(X=x|Y=1)\,P(Y=1)}{\sum_{y'\in\{0,1\}} P(X=x|Y=y')\,P(Y=y')} \\
&= \frac{1}{1+\dfrac{P(X=x|Y=0)P(Y=0)}{P(X=x|Y=1)P(Y=1)}} \\
&= \frac{1}{1+\exp\left(\ln\dfrac{P(X=x|Y=0)P(Y=0)}{P(X=x|Y=1)P(Y=1)}\right)} \\
&= \frac{1}{1+\exp\underbrace{\left(\ln\dfrac{P(X_1=x_1,\ldots,X_d=x_d|Y=0)P(Y=0)}{P(X_1=x_1,\ldots,X_d=x_d|Y=1)P(Y=1)}\right)}_{exponent}}
\end{aligned}
$$

$$\begin{aligned}
exponent \overset{cond.\ indep.}{=} \quad & \ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d} \ln\frac{P(X_i = x_i|Y=0)}{P(X_i = x_i|Y=1)} \\
= \quad & \ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d} \ln\left(\frac{\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(x_i-\mu_{i0})^2}{2\sigma_i^2}\right)}{\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(x_i-\mu_{i1})^2}{2\sigma_i^2}\right)}\right) \\
= \quad & \ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d}\left(\frac{(x_i-\mu_{i1})^2}{2\sigma_i^2} - \frac{(x_i-\mu_{i0})^2}{2\sigma_i^2}\right) \\
= \quad & \ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d}\frac{2x_i(\mu_{0i}-\mu_{1i}) + (\mu_{1i}^2-\mu_{0i}^2)}{2\sigma_i^2} \\
= \quad & \ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d}\left(\frac{x_i(\mu_{0i}-\mu_{1i})}{\sigma_i^2} + \frac{(\mu_{1i}^2-\mu_{0i}^2)}{2\sigma_i^2}\right) \\
= \quad & \underbrace{\ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d}\frac{(\mu_{1i}^2-\mu_{0i}^2)}{2\sigma_i^2}}_{w_0} + \sum_{i=1}^{d}\underbrace{\frac{\mu_{0i}-\mu_{1i}}{\sigma_i^2}}_{w_i} x_i
\end{aligned}$$

In conclusion,

$$P(Y=1|X=x) = \frac{1}{1+e^{(w\cdot x+w_0)}}$$

with

$$w_0 = \ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d}\frac{(\mu_{1i}^2-\mu_{0i}^2)}{2\sigma_i^2} \text{ **and** } w_i = \frac{\mu_{0i}-\mu_{1i}}{\sigma_i^2}, i=1,\ldots,d$$

Note that

$$P(Y=0|X=x) = \frac{e^{(w\cdot x+w_0)}}{1+e^{(w\cdot x+w_0)}}$$

and

$$P(Y=1|X=x) > P(Y=0|X=x) \Leftrightarrow w\cdot x + w_0 < 0$$

Since the coefficients w_i for $i = 1, \ldots, d$ do not depend on x_i, it follows that this *decison rule* of Gaussian Naive Bayes [in the conditions stated in the beginning of this problem] is a linear rule, like in Logistic Regression.

However, this relationship does not mean that there is a one-to-one correspondence between the parameters w_i of Gaussian Naive Bayes (GNB) and the parameters w_i of logistic regression (LR) because LR is discriminative and therefore doesn't model $P(X)$, while GNB does model $P(X)$.

To be more specific, note that the coefficients w_i in the GNB decision rules should be devided by $P(x_1, \ldots, x_d)$ in order to correspond to $P(Y = 1|X = x)$, which means that then they will not anymore be independent of x_i, like the LR coefficients.

Estimating the parameters for the Gaussian Naive Bayes algorithm

CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW2, pr. 5.ab

Consider a Gaussian Naive Bayes model, where the conditional distribution of each feature is a one-dimensional Gaussian, $X^{(j)}|Y \sim N(\mu_Y^{(j)}, (\sigma_Y^{(j)})^2)$**,** $j = 1, \cdots, d$**.**

a. Given n **independent training data points,** $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$**, give a maximum-likelihood estimate (MLE) of the conditional distribution of feature** $X^{(j)}, j = 1, \ldots, d$**.**

Solution:

The likelihood of the samples in Class 0 is

$$
\begin{aligned}
L(X_{i,0}^{(j)}|\mu_0^{(j)}, (\sigma\mu_0^{(j)})^2) &= \prod_{i=1}^{n_0} \frac{1}{\sqrt{2\pi}\sigma_0^{(j)}} \exp\left(-\frac{(X_{i,0}^{(j)} - \mu_0^{(j)})^2}{2(\sigma_0^{(j)})^2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma_0^{(j)}}\right)^{n_0} \exp\left(-\sum_{i=1}^{n_0} \frac{(X_{i,0}^{(j)} - \mu_0^{(j)})^2}{2(\sigma_0^{(j)})^2}\right)
\end{aligned}
$$

and the log-likelihood is

$$
\ln L = -n_0 \ln \sigma_0^{(j)} - \frac{1}{2(\sigma_0^{(j)})^2} \sum_{i=1}^{n_0} (X_{i,0}^{(j)} - \mu_0^{(j)})^2 + constant
$$

Taking the partial derivatives of the log-likelihood, we have

$$
\begin{aligned}
\frac{\partial \ln L}{\partial \mu_0^{(j)}} = 0 \;\Leftrightarrow\; \sum_{i=1}^{n_0} (X_{i,0}^{(j)} - \mu_0^{(j)}) = 0 \Leftrightarrow \mu_0^{(j)} = \frac{1}{n_0} \sum_{i=1}^{n_0} X_{i,0}^{(j)} \\
\frac{\partial \ln L}{\partial \sigma_0^{(j)}} = 0 \;\Leftrightarrow\; -\frac{n_0}{\sigma_0^{(j)}} + \frac{1}{(\sigma_0^{(j)})^3} \sum_{i=1}^{n_0} (X_{i,0}^{(j)} - \mu_0^{(j)})^2 = 0 \Leftrightarrow (\sigma\mu_0^{(j)})^2 = \frac{1}{n_0} \sum_{i=1}^{n_0} (X_{i,0}^{(j)} - \hat{\mu}_0^{(j)})^2
\end{aligned}
$$

Similarly, one can derive the MLE for the parameters in Class 1.

b. Suppose the prior of Y is already given. How many parameters do you need to estimate in Gaussian Naive Bayes model?

Solution:

For each class, there are 2 parameters (the mean and variance) for each feature, therefore there are $2 \cdot 2d = 4d$ parameters for all features in the two classes.

c. In a full/Joint Gaussian Bayes model, we assume that the conditional distribution $\Pr(X|Y)$ is a multidimensional Gaussian, $X|Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$, where $\mu \in \mathbb{R}^d$ is the mean vector and $\Sigma \in \mathbb{R}^{d \times d}$ is the covariance matrix.

Again, suppose the prior of Y is already given. How many parameters do you need to estimate in a full/Joint Gaussian Bayes model?

Solution:

For each class, there are d parameters for the mean, $d(d+1)/2$ parameters for the covariance matrix, because the covariance matrix is symmetric. Therefore, the number of parameters is $2 \cdot (d + d(d+1)/2) = d(d+3)$ in total for the two classes.

Proving

the relationship between

The full Gaussian Bayes algorithm and *Logistic Regression*

when $\Sigma_0 = \Sigma_1$

CMU, 2009 spring, A. Singh, T. Mitchell, HW2, pr. 2.2

Let's make the following *assumptions*:

1. Y **is a boolean variable following a Bernoulli distribution, with parameter** $\pi = P(Y = 1)$ **and thus** $P(Y = 0) = 1 - \pi$**.**

2. $X = < X_1, X_2, \ldots, X_n >$ **is a vector of random variables *not* conditionally independent given** Y**, and** $P(X|Y = k)$ **follows a *multivariate normal distribution*** $N(\mu_k, \Sigma)$**.**

Note that μ_k **is the** $n \times 1$ **mean vector depending on the value of** Y**, and** Σ **is the** $n \times n$ **covariance matrix, which does not depend on** Y**. We will write/use the density of the multivariate normal distribution in vector/matrix notation.**

Is the form of $P(Y|X)$ **implied by such this not-so-naive Gaussian Bayes classifier [LC: similar to] the form used by logistic regression?**
Derive the form of $P(Y|X)$ **to prove your answer.**

We start with:

$$\begin{aligned} P(Y=1|X) &= \frac{P(X|Y=1)\,P(Y=1)}{P(X|Y=1)\,P(Y=1)+P(X|Y=0)\,P(Y=0)} \\ &= \frac{1}{1+\dfrac{P(Y=0)\,P(X|Y=0)}{P(Y=1)\,P(X|Y=1)}} = \frac{1}{1+\exp\left(\ln\dfrac{P(Y=0)\,P(X|Y=0)}{P(Y=1)\,P(X|Y=1)}\right)} \\ &= \frac{1}{1+\exp\left(\ln\dfrac{P(Y=0)}{P(Y=1)}+\ln\dfrac{P(X|Y=0)}{P(X|Y=1)}\right)} \end{aligned}$$

Next we will focus on the term $\ln\dfrac{P(X|Y=0)}{P(X|Y=1)}$**:**

$$\ln\frac{P(X|Y=0)}{P(X|Y=1)} = \ln\frac{\dfrac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}}{\dfrac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}} + \ln\exp[(\star)] = \ln\exp[(\star)] = (\star)$$

where $(\star)$ **is the formulation obtained as the difference between the exponential parts of two multivariate Gaussian densities** $P(X|Y=0)$ **and** $P(X|Y=1)$**.**

$$\begin{aligned}(\star) &= \frac{1}{2}[(X-\mu_1)^\top \Sigma^{-1}(X-\mu_1) - (X-\mu_0)^\top \Sigma^{-1}(X-\mu_0)] \\ &= (\mu_0^\top - \mu_1^\top)\Sigma^{-1}X + \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0\end{aligned}$$

As a result, we have:

$$\begin{aligned}P(Y=1|X) &= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi} + \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0 + (\mu_0^\top - \mu_1^\top)\Sigma^{-1}X\right)} \\ &= \frac{1}{1+\exp(w_0 + w^\top X)}\end{aligned}$$

where $w_0 = \ln\frac{1-\pi}{\pi} + \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0$ **is a scalar,**
and $w = \Sigma^{-1}(\mu_0 - \mu_1)$ **is a** $d \times 1$ **a parameter vector.**

Note that $((\mu_0^\top - \mu_1^\top)\Sigma^{-1})^\top = ((\mu_0 - \mu_1)^\top \Sigma^{-1})^\top = (\Sigma^{-1})^\top((\mu_0 - \mu_1)^\top)^\top = \Sigma^{-1}(\mu_0 - \mu_1)$ **because** Σ^{-1} **is symmetric.**
(Σ **is symmetric because it is a covariance matrix, and therefore** Σ^{-1} **is also symmetric.)**

In conclusion, $P(Y|X)$ **has the form of the logistic regression (in vector and matrix notation).**

Some proofs

k-NN and the Curse of Dimensionality

Proving that the number of examples needed by k-NN grows exponentially with the number of features

CMU, 2010 fall, Aarti Singh, HW2, pr. 2.2

[Slides originally drawn by Diana Mînzat, MSc student, FII, 2015 spring]

Consider a set of n points $x_1, x_2, ..., x_n$ independently and uniformly drawn from a p-dimensional zero-centered unit ball

$$B = \{x \colon \|x\|^2 \leq 1\} \subset \mathbb{R}^p,$$

where $\|x\| = \sqrt{x \cdot x}$ and $\cdot$ is the inner product in $\mathbb{R}^p$.

In this problem we will study the size of the 1-nearest neighbourhood of the origin O and how it changes in relation to the dimension p, thereby gain intuition about the downside of k-NN in a high dimension space.

Formally, this size will be described as the distance from O to its nearest neighbour in the set $\{x_1, ..., x_n\}$, denoted by d^*:

$$d^* := \min_{1 \leq i \leq n} ||x_i||,$$

which is a random variable since the sample is random.

a. For $p = 1$, calculate $P(d^* \leq t)$, the *cumulative distribution function (c.d.f.)* of d^*, for $t \in [0, 1]$.

Solution:

In the one-dimensional space ($p = 1$), the unit ball is the interval $[-1, 1]$. The cumulative distribution function will have the following expression:

$$F_{n,1}(t) \overset{not.}{=} P(d^* \leq t) = 1 - P(d^* > t) = 1 - P(|x_i| > t,\ \text{for } i = 1, 2, ..., n)$$

Because the points $x_1, ..., x_n$ were generated independently, the c.d.f. can also be written as:

$$F_{n,1}(t) = 1 - \prod_{i=1}^{n} P(|x_i| > t) = 1 - (1 - t)^n$$

b. Find the formula of the *cumulative distribution function* of d^* for the general case, when $p \in \{1, 2, 3, ...\}$.

Hint: You may find the following fact useful: the volume of a p-dimensional ball with radius r is

$$V_p(r) = \frac{(r\sqrt{\pi})^p}{\Gamma\left(\dfrac{p}{2} + 1\right)},$$

where Γ is Euler's Gamma function, defined by

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi},\ \Gamma(1) = 1,\ \textbf{and } \Gamma(x+1) = x\Gamma(x) \textbf{ for any } x > 1.$$

Note: It can be easily shown that $\Gamma(n + 1) = n!$ for all $n \in \mathbb{N}^*$, therefore the Gamma function is a generalization of the factorial function.

Solution:

In the general case, i.e. considering a fixed $p \in \mathbb{N}^*$, it is obvious that the cumulative distribution function of d^* will have a similar form to the $p = 1$ case:

$$\begin{aligned} F_{n,p}(t) \overset{not.}{=} P(d^* \leq t) &= 1 - P(d^* > t) = 1 - P(||\,x_i\,|| > t,\ i = 1, 2, \ldots, n) \\ &= 1 - \prod_{i=1}^{n} P(||\,x_i\,|| > t). \end{aligned}$$

Denoting the volume of the sphere of radius t by $V_p(t)$, and knowing that the points $x_1, ..., x_n$ follow a uniform distribution, we can rewrite the above formula as follows:

$$F_{n,p}(t) = 1 - \left(\frac{V_p(1) - V_p(t)}{V_p(1)} \right)^n = 1 - \left(1 - \frac{V_p(t)}{V_p(1)} \right)^n .$$

Using the suggested formula for the volume of the sphere, it follows immediately that $F_{n,p} = 1 - (1 - t^p)^n$.

c. What is the *median* of the random variable d^* (i.e., the value of t for which $P(d^* \leq t) = 1/2$) ? The answer should be a *function of* both the sample size n and the dimension p.

Fix $n = 100$ and plot the values of the median function for $p = 1, 2, 3, ..., 100$ with the median values on the y-axis and the values of p on the x-axis. What do you see?

Solution:

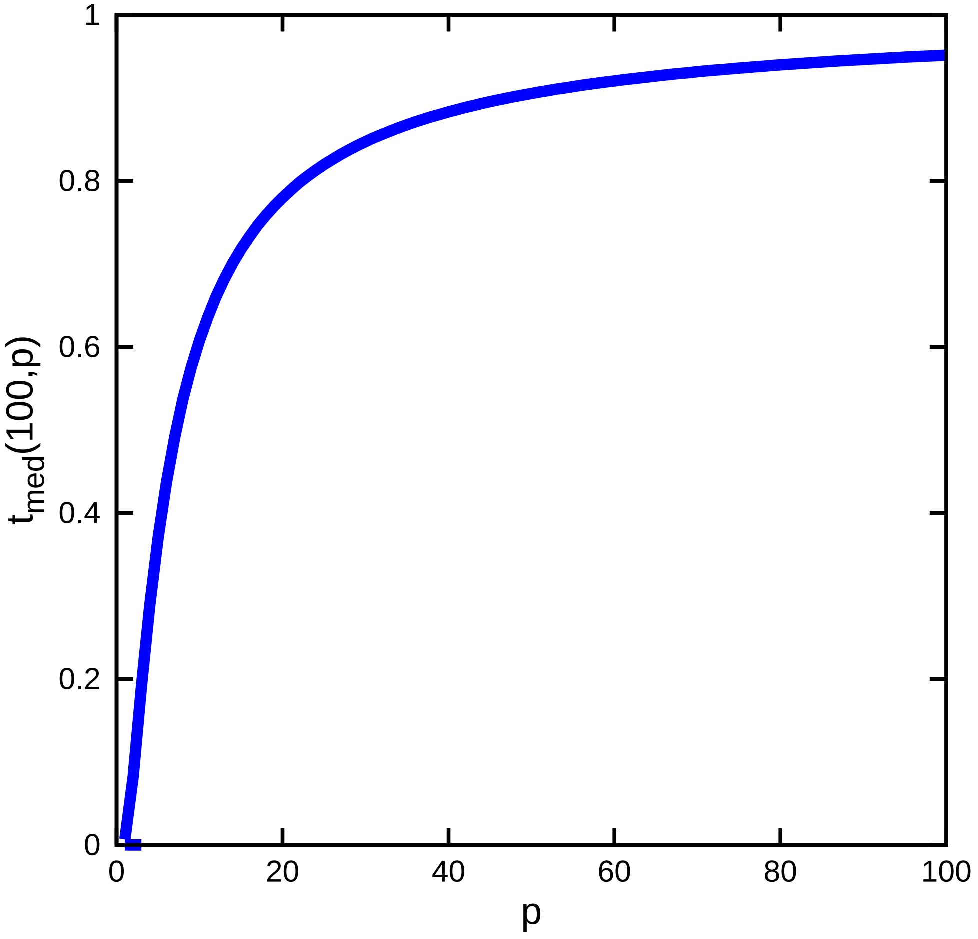
In order to find the median value of the random variable d^*, we will solve the equation $P(d^* \leq t) = 1/2$ of variable t:

$$P(d^* \leq t) = \frac{1}{2} \;\Leftrightarrow\; F_{n,p}(t) = \frac{1}{2} \Leftrightarrow 1 - (1 - t^p)^n = \frac{1}{2} \Leftrightarrow (1 - t^p)^n = \frac{1}{2}$$
$$\Leftrightarrow\; 1 - t^P = \frac{1}{2^{1/n}} \Leftrightarrow t^P = 1 - \frac{1}{2^{1/n}}$$

Therefore,

$$t_{med}(n, p) = \left(1 - \frac{1}{2^{1/n}}\right)^{1/p}.$$

The plot of the function $t_{med}(100, p)$ for $p = 1, 2, \ldots, 100$:



Remark:

The minimal sphere containing the nearest neighbour of the origin in the set $\{x_1, x_2, ..., x_n\}$ grows very fast as the value of p increases.

When p becomes greater than 10, most of the 100 training instances are closer to the surface of the unit ball than to the origin O.

d. Use the c.d.f. derived at point b to determine how large should the sample size n be such that with probability at least 0.9, the distance d^* from O to its nearest neighbor is less than $1/2$, i.e., half way from O to the boundary of the ball.

The answer should be a *function* of p.
Plot this function for $p = 1, 2, \ldots, 20$ with the function values on the y-axis and values of p on the x-axis. What do you see?

***Hint*: You may find useful the Taylor series expansion of $\ln(1-x)$:**

$$\ln(1-x) = -\sum_{i=1}^{\infty} \frac{x^i}{i} \textbf{ for } -1 \leq x < 1.$$

Solution:

$$P(d^* \leq 0.5) \geq 0.9 \quad \Leftrightarrow \quad F_{n,p}(0.5) \geq \frac{9}{10} \overset{b.}{\Leftrightarrow} 1 - \left(1 - \frac{1}{2^p}\right)^n \geq \frac{9}{10} \ \Leftrightarrow \ \left(1 - \frac{1}{2^p}\right)^n \leq \frac{1}{10}$$

$$\Leftrightarrow \quad n \cdot \ln\left(1 - \frac{1}{2^p}\right) \leq \ -\ln 10 \Leftrightarrow n \geq \ \frac{\ln 10}{-\ln\left(1 - \dfrac{1}{2^p}\right)}$$

Using the decomposition of $\ln(1 - 1/2^p)$ **into a Taylor series (with** $x = 1/2^p$**), we obtain:**

$$\begin{aligned} & P(d^* \leq 0.5) \geq 0.9 \\ & \Rightarrow \quad n \geq (\ln 10)\ 2^p \frac{1}{1 + \dfrac{1}{2} \cdot \dfrac{1}{2^p} + \dfrac{1}{3} \cdot \dfrac{1}{2^{2p}} + \ldots + \dfrac{1}{n} \dfrac{1}{2^{(n-1)p}} + \ldots} \\ & \Rightarrow \quad n \geq \ 2^{p-1} \ \ln 10. \end{aligned}$$

***Note*:**

In order to obtain the last inequality in the above calculations, we considered the following two facts:

i. $\frac{1}{3 \cdot 2^p} < \frac{1}{4}$ **holds for any** $p \geq 1$**, and**

ii. $\frac{1}{n \cdot 2^{(n-1)p}} \leq \frac{1}{2^n} \Leftrightarrow 2^n \leq n \cdot 2^{(n-1)p}$ **holds for any** $p \geq 1$ **and** $n \geq 2$**.**
(This can be proven by induction on p**).**

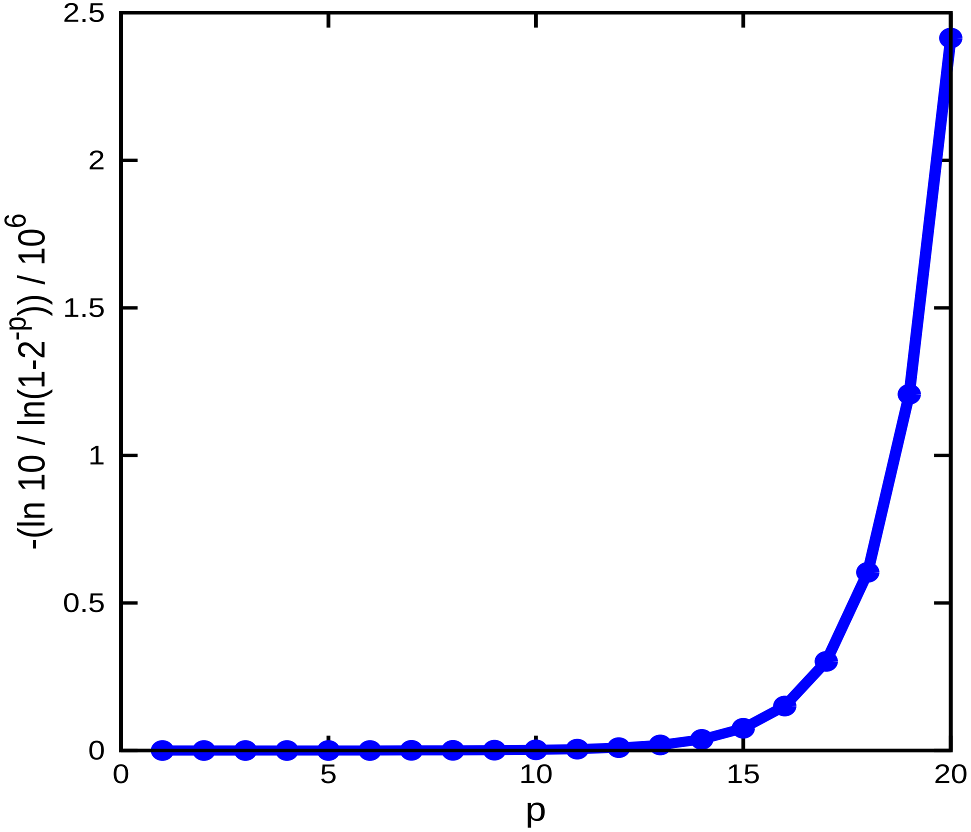
So, we got:

$$1 + \frac{1}{2} \cdot \frac{1}{2^p} + \frac{1}{3} \cdot \frac{1}{2^{2p}} + \ldots + \frac{1}{n} \frac{1}{2^{(n-1)p}} + \ldots <$$
$$1 + \frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots \to \frac{1}{1 - \frac{1}{2}} = 2.$$

The proven result

$$P(d^* \leq 0.5) \geq 0.9 \Rightarrow n \geq \; 2^{p-1} \; \ln 10$$

means that the sample size needed for the probability that $d^* < 0.5$ is large enough (9/10) grows exponentially with p.



e. Having solved the previous problems, what will you say about the downside of k-NN in terms of n and p?

Solution:

The k-NN classifier works well when a test instance has a “dense” neighborhood in the training data.

However, the analysis here suggests that in order to provide a dense neighborhood, the size of the training sample should be exponential in the dimension p, which is clearly infeasible for a large p.

(Remember that p is the dimension of the space we work in, i.e. the number of features of the training instances.)

An upper bound for the assimptotic error rate of 1-NN: twice the error rate of Joint Bayes

T. Cover and P. Hart (1967)

CMU, 2005 spring, Carlos Guestrin, HW3, pr. 1

Note: we will prove the *Covert & Hart' theorem* in the case of binary classification with real-values inputs.

Let $x_1, x_2, \ldots$ be the training examples in some fixed d-dimensional Euclidean space, and y_i be the corresponding binary class labels, $y_i \in \{0, 1\}$.

Let $p_y(x) \overset{not.}{=} P(X = x \mid Y = y)$ be the true conditional probability distribution for points in class y. We *assume* continuous and non-zero conditional probabilities: $0 < p_y(x) < 1$ for all x and y.

Let also $\theta \overset{not.}{=} P(Y = 1)$ be the probability that a random training example is in class 1. Again, *assume* $0 < \theta < 1$.

a. Calculate $q(x) \overset{not.}{=} p(Y = 1 \mid X = x)$, the true probability that a data point x belongs to class 1. Express $q(x)$ in terms of $p_0(x), p_1(x)$, and θ.

Solution:

$$\begin{aligned} q(x) &\overset{F.\ Bayes}{=} \frac{P(X = x|Y = 1)P(Y = 1)}{P(X = x)} \\ &= \frac{P(X = x|Y = 1)P(Y = 1)}{P(X = x|Y = 1)P(Y = 1) + P(X = x|Y = 0)P(Y = 0)} \\ &= \frac{p_1(x)\ \theta}{p_1(x)\ \theta + p_0(x)(1 - \theta)} \end{aligned}$$

b. The Joint Bayes classifier (usually called the Bayes Optimal classifier) always assigns a data point x the most probable class: $\text{argmax}_y\, P(Y = y \mid X = x)$.

Given some test data point x, what is the probability that example x will be misclassified using the Joint Bayes classifier, in terms of $q(x)$?

Solution:

The Joint Bayes classifier fails with probability $P(Y = 0|X = x)$ when $P(Y = 1|X = x) \geq P(Y = 0|X = x)$, and respectively with probability $P(Y = 1|X = x)$ when $P(Y = 0|X = x) \geq P(Y = 1|X = x)$. I.e.,

$$\begin{aligned} Error_{Bayes}(x) &= \min\{P(Y = 0|X = x), P(Y = 1|X = x)\} \\ &= \min\{1 - q(x),\, q(x)\} \\ &= \begin{cases} q(x) \textbf{ if } q(x) \in [0\,,\, 1/2] \\ 1 - q(x) \textbf{ if } q(x) \in (1/2\,,\, 1]. \end{cases} \end{aligned}$$

c. **The 1-nearest neighbor classifier assigns a test data point x the label of the closest training point x'.**

Given some test data point x and its nearest neighbor x', what is the *expected error* of the 1-nearest neighbor classifier, i.e., the probability that x will be misclassified, in terms of $q(x)$ and $q(x')$?

Solution:

$$\begin{aligned} Error_{\text{1-NN}}(x) &= P(Y=1|X=x)P(Y=0|X=x') + \\ &\quad P(Y=0|X=x)P(Y=1|X=x') \\ &= q(x)(1-q(x')) + (1-q(x))q(x'). \end{aligned}$$

d. In the asymptotic case, i.e. when the number of training examples of each class goes to infinity, and the training data fills the space in a dense fashion, the nearest neighbor x' of x has $q(x')$ converging to $q(x)$, i.e. $P(Y=1|X=x') \to p(Y=1|X=x)$.

(This is true due to *i.* the result obtained at the above point a, and *ii.* the assumed continuity of the function $p_y(x) \overset{not.}{=} p(X=x|Y=y)$ w.r.t. x.)

By performing this substitution in the expression obtained at point c, give the *asymptotic error* for the 1-nearest neighbor classifier at point x, in terms of $q(x)$.

Solution:

$$\lim_{x' \to x} Error_{\textit{1-NN}}(x) = 2q(x)(1-q(x))$$

e. Show that the asymptotic error obtained at point d is less than twice the Bayes Optimal error obtained at point b and subsequently that this inequality leads to the corresponding relationship between the expected error rates:

$$E[\lim_{n\to\infty} Error_{1\text{-}NN}] \leq 2E[Error_{Bayes}].$$

Solution:

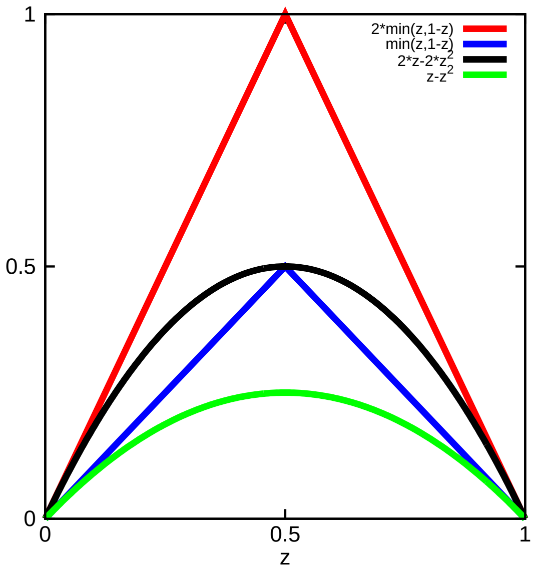
$z(1-z) \leq z$ **for all** z**, in particular for** $z \in [0\,,\,1/2]$**, and**
$z(1-z) \leq 1-z$ **for all** z**, in particular for** $z \in [1/2\,,\,1]$**.**
Therefore, for all x**,**

$$q(x)(1-q(x)) \leq \begin{cases} q(x) \text{ if } q(x) \in [0\,,\,1/2] \\ 1-q(x) \text{ if } q(x) \in (1/2\,,\,1]. \end{cases}$$

The results obtained at points b **and** d **lead to**

$$\lim_{n\to\infty} Error_{1\text{-}NN} = 2q(x)(1-q(x)) \leq 2Error_{Bayes}(x) \text{ for all } x.$$

By multiplication with $P(x)$ **and suming upon all values of** x**, we get:** $E[\lim_{n\to\infty} Error_{1\text{-}NN}] \leq 2E[Error_{Bayes}]$.



Remark

from: ***An Elementary Introduction to Statistical Learning Theory,***
S. Kulkarni, G. Harman, 2011, pp. 68-69

An even tighter bound exists for $E[\lim_{n \rightarrow \infty} Error_{\textit{1-NN}}]$: $2E[Error_{Bayes}](1 - E[Error_{Bayes}])$

Proof:

From $\lim_{x' \rightarrow x} Error_{\textit{1-NN}}(x) = 2q(x)(1 - q(x))$ **(see point d) and** $Error_{Bayes}(x) = \min\{1 - q(x),\ q(x)\}$ **(see point b),**

it follows that
$\lim_{x' \rightarrow x} Error_{\textit{1-NN}}(x) = 2Error_{Bayes}(x)(1 - Error_{Bayes}(x))$**.**

By multiplying this last equality with $P(x)$ and suming on all x — in fact, integrating upon x —, we get

$$E[\lim_{x' \rightarrow x} Error_{\textit{1-NN}}] = 2E[Error_{Bayes}(1 - Error_{Bayes})] = 2E[Error_{Bayes}] - 2E[(Error_{Bayes})^2].$$

Since $E[Z^2] \geq (E[Z])^2$ **for any** Z **(** $Var(Z) \overset{def.}{=} E[(Z - E[Z])^2] \overset{comp.}{=} E[Z^2] - (E[Z])^2 \geq 0$**),**
it follows that

$$E[\lim_{x' \rightarrow x} Error_{\textit{1-NN}}] \leq 2E[Error_{Bayes}] - 2(E[Error_{Bayes}])^2 = 2E[Error_{Bayes}](1 - E[Error_{Bayes}]).$$

Remarks

- $E[\lim_{n\to\infty} Error_{1\text{-}NN}] \geq E[Error_{Bayes}]$

  Proof:
  $2z - 2z^2 \geq z \ \forall z \in [0\,,\, 1/2]$ **and** $2z - 2z^2 \geq 1 - z \ \forall z \in [1/2\,,\, 1]$**.**
  Therefore,

  $$2q(x)(1 - q(x)) \geq Error_{Bayes}(x) \text{ for all } x,$$

  and

  $$\lim_{n\to\infty} Error_{1\text{-}NN}(x) = \lim_{x'\to x} Error_{1\text{-}NN}(x) \geq Error_{Bayes}(x) \text{ for all } x.$$

- **The Cover & Hart' upper bound for the asymptotic error rate of 1-NN doesn't hold in the non-asymptotic case (where the number of training examples is finite).**

Other Results

[from ***An Elementary Introduction to Statistical Learning Theory***,
S. Kulkarni, G. Harman, 2011, pp. 69-70]

- **When certain restrictions hold,**

$$E[\lim_{n\to\infty} Error_{k\text{-}NN}] \leq \left(1 + \frac{1}{k}\right) E[Error_{Bayes}].$$

○ **However, it can be shown that there are some distributions for which 1-NN outperforms k-NN for any fixed $k > 1$.**

- **If $\frac{k_n}{n} \to 0$ for $n \to \infty$ (for instance, $k_n = \sqrt{n}$), then**

$$E[\lim_{n\to\infty} Error_{k_n\text{-}NN}] = E[Error_{Bayes}].$$

Significance

The last result means that k_n-NN is

- a ***universally consistent learner*** (because when the amount of training data grows, its performance approaches that of Joint Bayes) and
- ***non-parametric*** (i.e., the underlying distribution of data can be arbitrary and we need no knowledge of its form).

Some other universally consistent learners exist.

However, the ***convergence rate*** is critical. For most learning methods, the convergence rate is very slow in high-dimensional spaces (due to “the curse of dimensionality”). It can be shown that ***there is no “universal” convergence rate,*** i.e. one can always find distributions for which the convergence rate is arbitrarily slow.

There is no one learning method which can universally beat out all other learning methods.

Conclusion

Such results make the ML field continue to be exciting, and makes the design of good learning algorithms and the understanding of their performance an important science and art!

Exemplifying the application of hierarchical agglomerative clustering (single- and complete-linkage)

CMU, 2012 fall, Tom Mitchell, Ziv Bar-Joseph, HW4, pr. 2.a

The table below is a distance matrix for 6 objects.

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

Show the final result of hierarchical clustering with single-linkage and complete-linkage by drawing the corresponding dendrograms.

Solution:

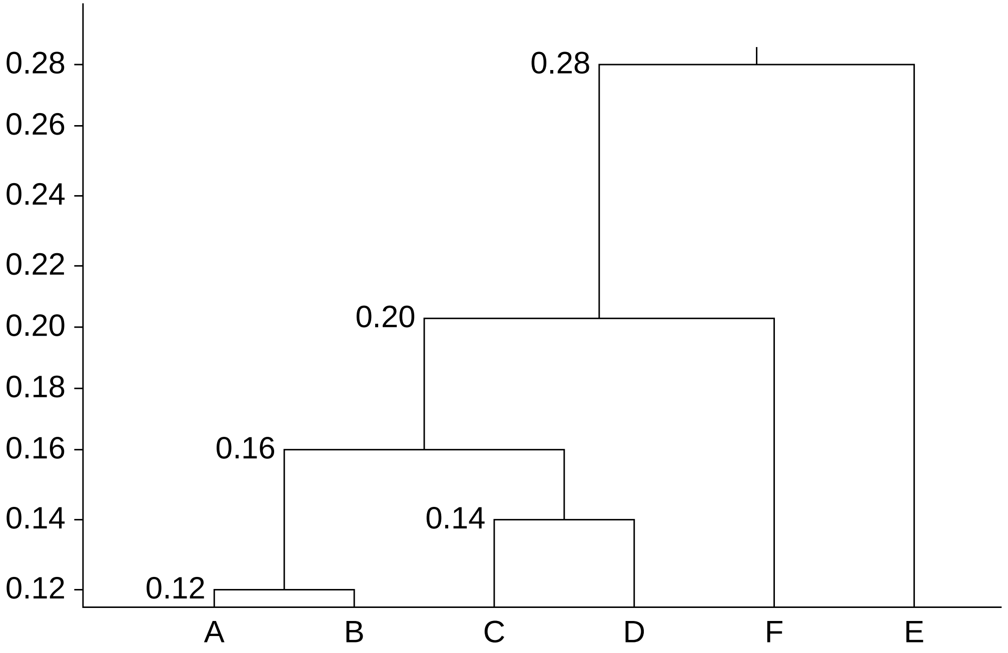
Single-linkage:

	AB	C	D	E	F
AB	0				
C	0.25	0			
D	0.16	**0.14**	0		
E	0.28	0.70	0.45	0	
F	0.34	0.93	0.20	0.67	0

	AB	CD	E	F
AB	0			
CD	**0.16**	0		
E	0.28	0.45	0	
F	0.34	0.20	0.67	0

	$ABCD$	E	F
$ABCD$	0		
E	0.28	0	
F	**0.20**	0.67	0

	$ABCDF$	E
$ABCDF$	0	
E	**0.28**	0



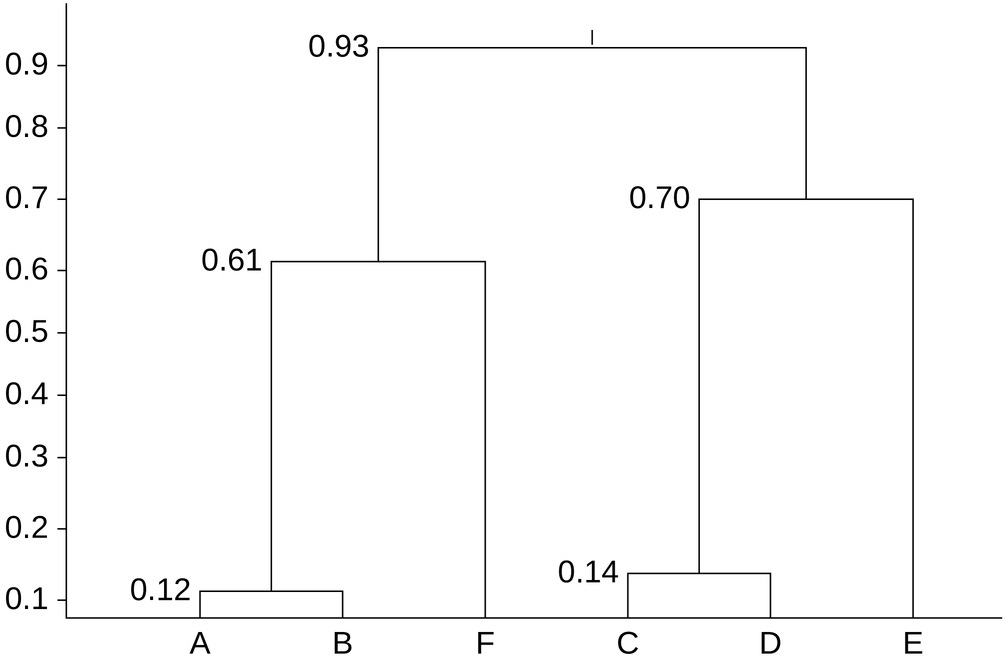
Complete-linkage:

	AB	C	D	E	F
AB	0				
C	0.51	0			
D	0.84	**0.14**	0		
E	0.77	0.70	0.45	0	
F	0.61	0.93	0.20	0.67	0

	AB	CD	E	F
AB	0			
CD	0.84	0		
E	0.77	0.70	0	
F	**0.61**	0.93	0.67	0

	ABF	CD	E
ABF	0		
CD	0.93	0	
E	0.77	**0.70**	0

	ABF	CDE
ABF	0	
CDE	**0.93**	0



Exemplifying

the application of hierarchical devisive clustering and the relationship between slingle-linkage hierarchies and Minimum Spanning Trees (MSTs)

CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 9.3

Hierarchical clustering may be bottom-up or top-down.
In this problem we will see whether a top-down clustering algorithm can be exactly analogous to a bottom-up clustering algorithm.

Consider the following *top-down clustering algorithm*:

1. Calculate the pairwise distance $d(P_i, P_j)$ between every two objects P_i and P_j in the set of objects to be clustered, and build a complete graph on the set of objects with edge weights being the corresponding distances.

2. Generate the Minimum Spanning Tree of the graph, i.e. choose the subset of edges E' with minimum sum of weights such that $G' = (P, E')$ is a single connected tree.

3. Throw out the edge with the heviest weight to generate two disconnected trees corresponding to top level clusters.

4. Repeat the previous step recursively on the lower level clusters to generate a top-down clustering on the set of n objects.

a. Apply this algoritm on the dataset given in the nearby table, using the Euclidian distance.

b. Does this top-down algorithm perform analogously to any bottom-up algorithm that you have encountered in class? Why?

Point	x	y
P_1	1	2
P_2	2	2
P_3	3	6
P_4	6	4
P_5	6	6
P_6	12	12

Solution:

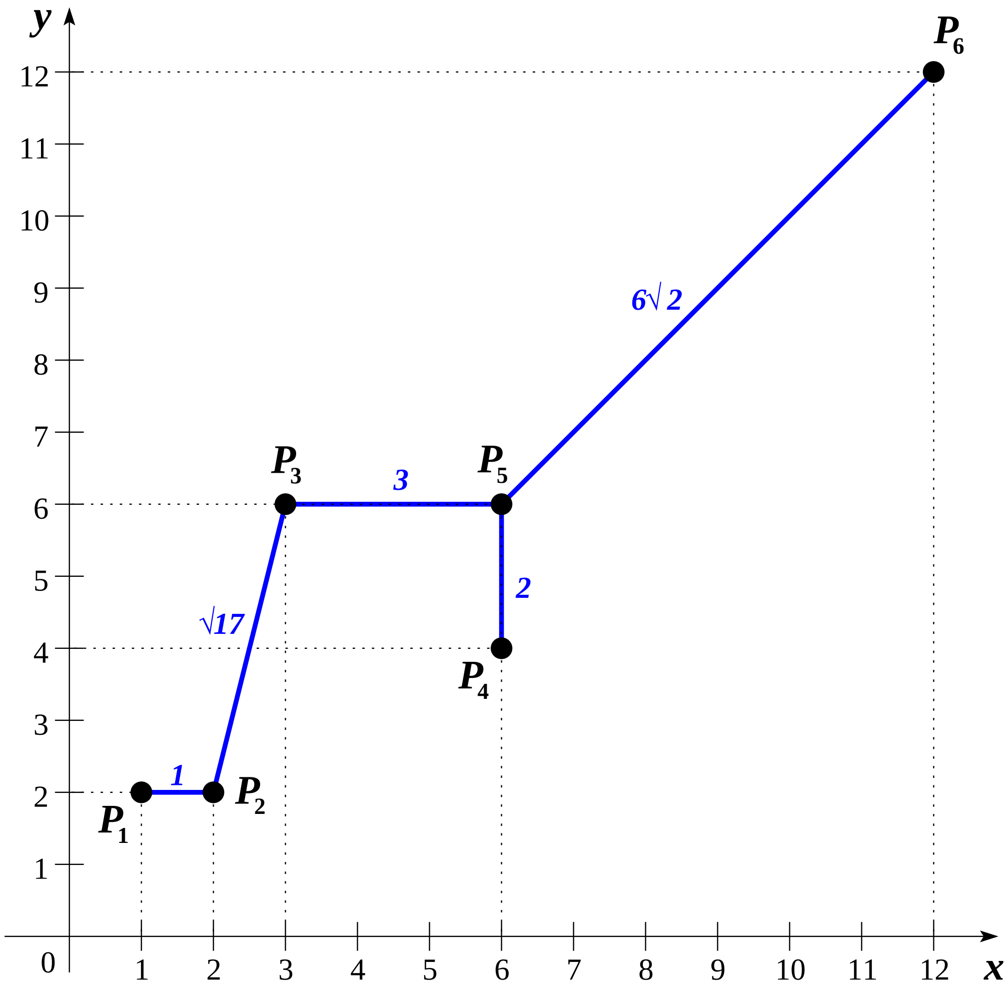
a.

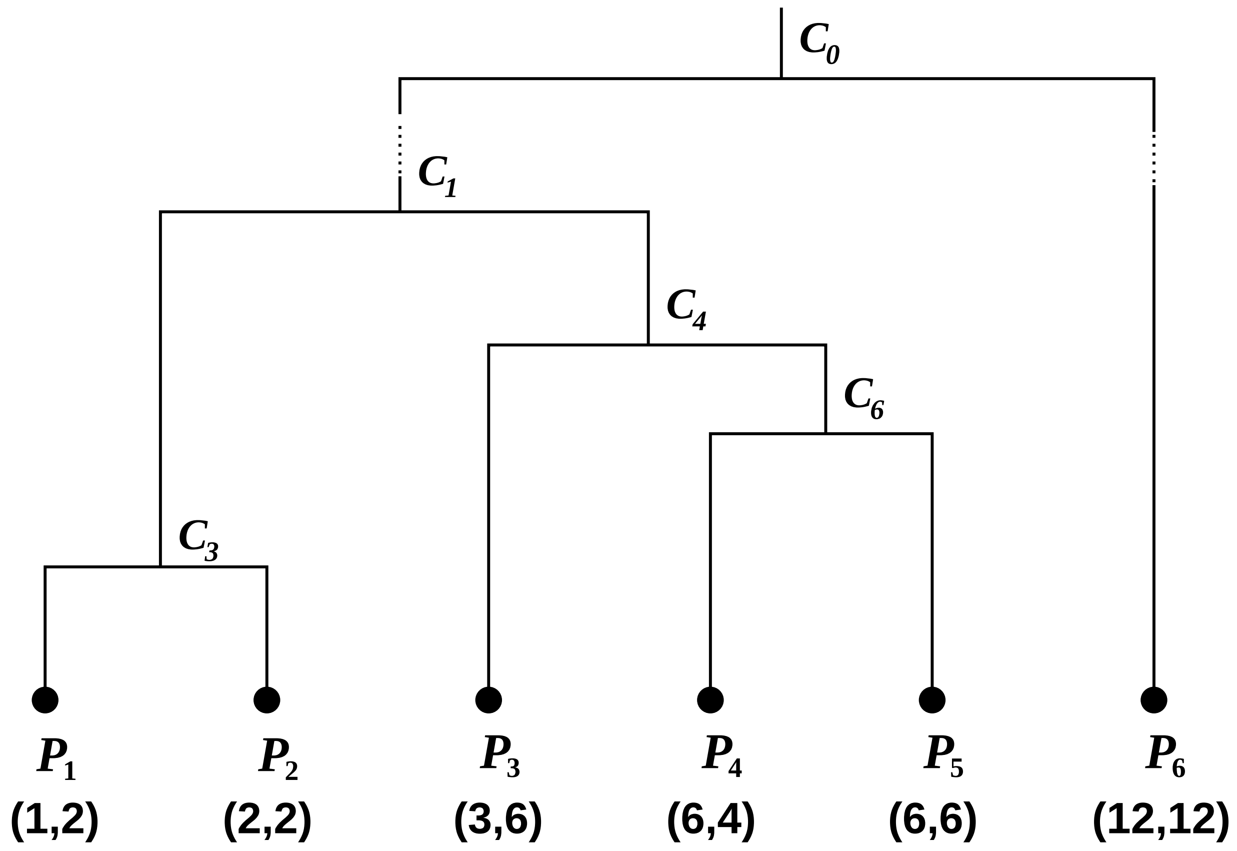
Kruskal algorithm:

1. (P_1, P_2), **cost** 1,
2. (P_4, P_5), **cost** 2,
3. (P_3, P_5), **cost** 3,
4. (P_2, P_3), **cost** $\sqrt{17}$
5. (P_5, P_6), **cost** $6\sqrt{2}$.

Prim algorithm:

1. (P_1, P_2), **cost** 1,
2. (P_2, P_3), **cost** $\sqrt{17}$
3. (P_3, P_5), **cost** 3,
4. (P_5, P_4), **cost** 2,
5. (P_5, P_6), **cost** $6\sqrt{2}$.




C_0
C_1
C_4
C_6
C_3
P_1
(1,2)
P_2
(2,2)
P_3
(3,6)
P_4
(6,4)
P_5
(6,6)
P_6
(12,12)

***Note*: If there is only one MST for the given dataset, then both Kruskal's and Prim's algorithm will find it. Otherwise, the two algorithms can produce differents results.**

One can see (both on this dataset and also in general) that Kruskal's algorithm is exactly analogous to the single-linkage bottom-up clustering algorithm.

Therefore, there is indeed a bottom-up equivalent to the top-down clustering algorithm presented in this exercise.

Exemplifying non-hierarchical clustering using the K-means algorithm

T.U. Dresden, 2006 summer, Steffen Höldobler, Axel Grossmann, HW3

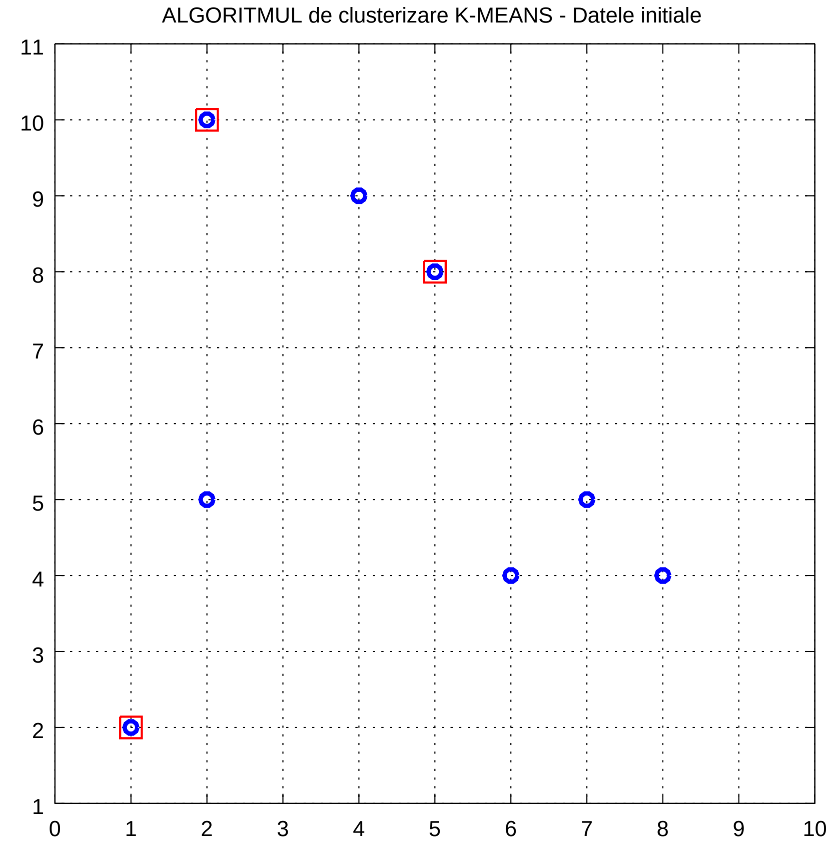
Folosiţi algoritmul K-means şi distanţa euclidiană pentru a grupa următoarele 8 instanţe din $\mathbb{R}^2$ în 3 clustere:

$$A(2,10),\ B(2,5),\ C(8,4),\ D(5,8),\ E(7,5),\ F(6,4),\ G(1,2),\ H(4,9).$$

Se vor lua drept centroizi iniţiali punctele A, D şi G.

a. Rulaţi prima iteraţie a algoritmului K-means. Pe un grid de valori 10×10 veţi marca instanţele date, poziţiile centroizilor la începutul primei iteraţii şi componenţa fiecărui cluster la finalul acestei iteraţii. (Trasaţi mediatoarele segmentelor determinate de centroizi, ca separatori ai clusterelor.)

b. Câte iteraţii sunt necesare pentru ca algoritmul K-means să conveargă? Desenaţi pe câte un grid rezultatul rulării fiecărei iteraţii.


ALGORITMUL de clusterizare K-MEANS - Datele initiale
11
10
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10

Solution:

***Iteration 0*:**

$$\left.\begin{array}{l}\mu_1^0=(2,\ 10)\\ \mu_2^0=(5,\ 8)\\ \mu_3^0=(1,\ 2)\end{array}\right\}\Rightarrow$$

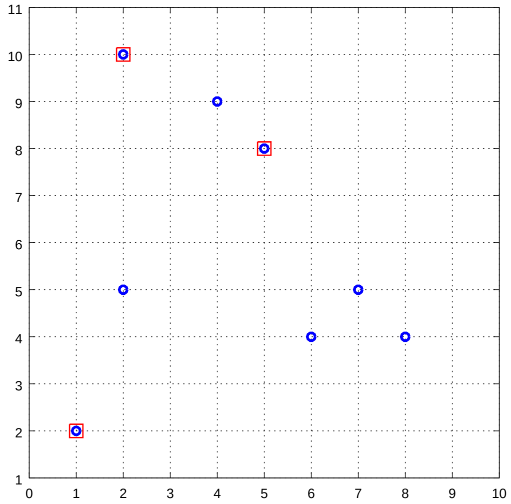
P_i	$d(\mu_1^0,P_i)$	$d(\mu_2^0,P_i)$	$d(\mu_3^0,P_i)$
$A(2,10)$	0	$\dots$	$\dots$
$B(2,5)$	5	$3\sqrt{2}$	$\sqrt{10}$
$C(8,4)$			
$D(5,8)$	$\dots$	0	$\dots$
$E(7,5)$			
$F(6,4)$			
$G(1,2)$	$\dots$	$\dots$	0
$H(4,9)$			

$$\Rightarrow\left\{\begin{array}{l}C_1^0=\{A\}\\ C_2^0=\{C,D,E,F,H\}\\ C_3^0=\{B,G\}\end{array}\right.$$

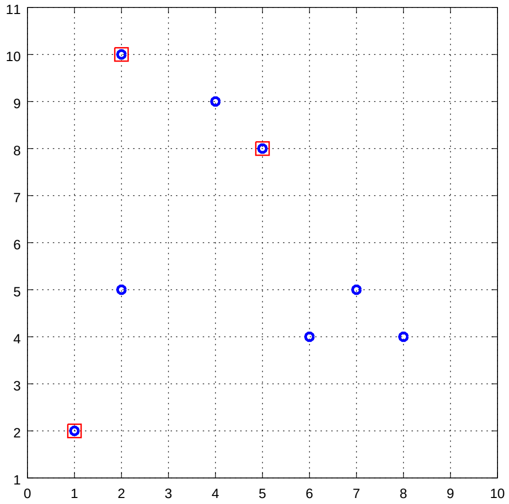
***Iteration 1*:**

$$\left.\begin{array}{l}\mu_1^1=\mu_1^0=(2,10)\\ \mu_2^1=\left(\dfrac{4+5+6+7+8}{5},\dfrac{4+4+5+8+9}{5}\right)=(6,\ 6)\\ \mu_3^1=\left(\dfrac{2+1}{2},\dfrac{5+2}{2}\right)=(1.5,\ 3.5)\end{array}\right\}\Rightarrow\dots\Rightarrow\left\{\begin{array}{l}C_1^0=\{A,H\}\\ C_2^0=\{C,D,E,F\}\\ C_3^0=\{B,G\}\end{array}\right.$$

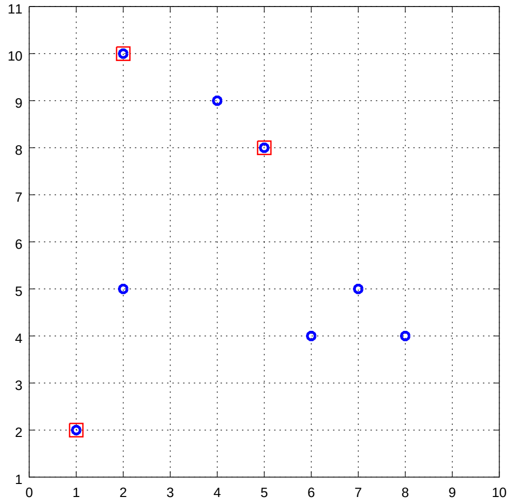
ALGORITMUL de clusterizare K-MEANS - Datele initiale



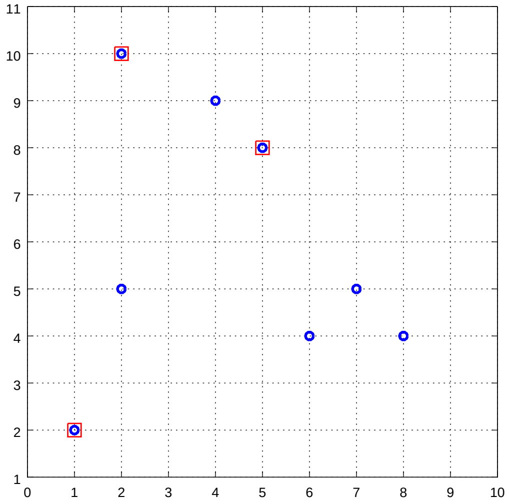
ALGORITMUL de clusterizare K-MEANS - Datele initiale



ALGORITMUL de clusterizare K-MEANS - Datele initiale



ALGORITMUL de clusterizare K-MEANS - Datele initiale

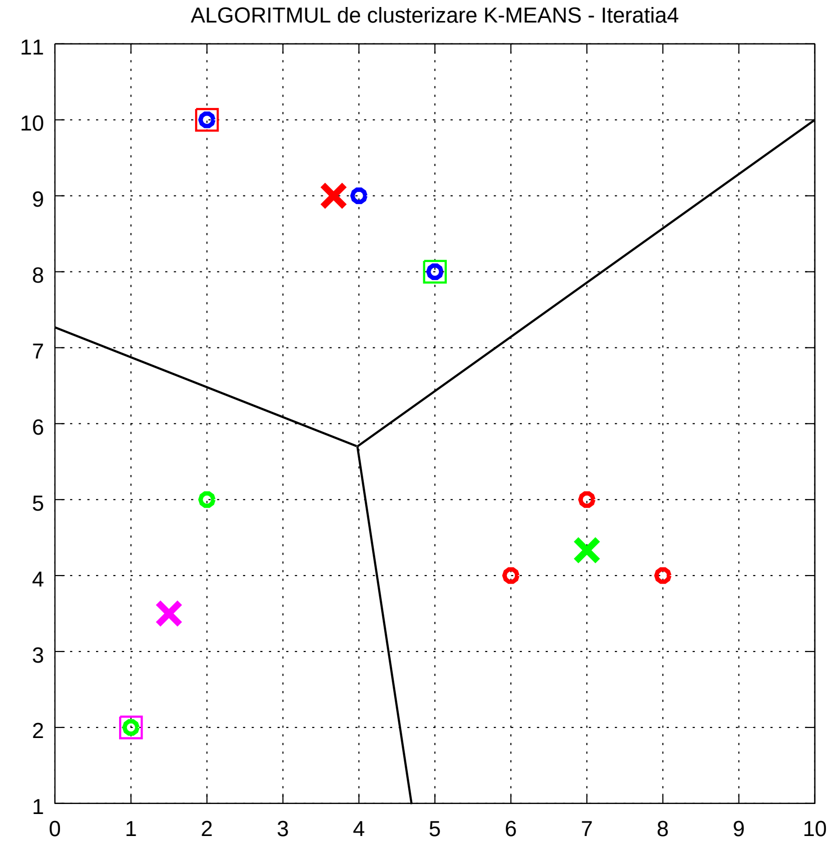


***Iteration 2*:**

$$\left.\begin{array}{l} \mu_1^2 = (3,\ 9.5) \\ \mu_2^2 = \left(\frac{26}{4}, \frac{21}{4}\right) = (6.5,\ 5.25) \\ \mu_3^2 = \mu_3^1 = (1.5,\ 3.5) \end{array}\right\} \Rightarrow \ldots \Rightarrow \left\{\begin{array}{l} C_1^2 = \{A, D, H\} \\ C_2^2 = \{C, E, F\} \\ C_3^2 = \{B, G\} \end{array}\right.$$

***Iteration 3*:**

$$\left.\begin{array}{l} \mu_1^3 = \left(\frac{2+4+5}{3}, \frac{8+9+10}{3}\right) = (11/3, 9) \\ \mu_2^3 = (7,\ 13/3) \\ \mu_3^3 = \mu_3^2 = (1.5,\ 3.5) \end{array}\right\} \Rightarrow \ldots \Rightarrow \left\{\begin{array}{l} C_1^3 = \{A, D, H\} = C_1^2 \\ C_2^3 = \{C, E, F\} = C_2^2 \\ C_3^3 = \{B, G\} = C_3^2 \end{array}\right\} \Rightarrow \textbf{\textit{Stop}}$$


ALGORITMUL de clusterizare K-MEANS - Iteratia4
11
10
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10

Some proofs

K-means as an optimisation algorithm: The monotonicity of the J_K criterion

[CMU, 2009 spring, Ziv Bar-Joseph, HW5, pr. 2.1]

Algoritmul K-means (S. P. Lloyd, 1957)

***Input*: $x_1, \ldots, x_n \in \mathbb{R}^d$, cu $n \geq K$.**

***Output*: o anumită K-partiţie pentru $\{x_1, \ldots, x_n\}$.**

***Procedură*:**

[***Iniţializare/Iteraţia 0:***] $t \leftarrow 0$;

se fixează în mod arbitrar $\mu_1^0, \ldots, \mu_K^0$, centroizii iniţiali ai clusterelor, şi se asignează fiecare instanţă x_i la centroidul cel mai apropiat, formând astfel clusterele $C_1^0, \ldots, C_K^0$.

[***Recursivitate:***] **Se execută iteraţia $++\,t$:**

***Pasul 1*: se calculează noile poziţii ale centroizilor:**

$\mu_j^t = \dfrac{1}{|C_j^{t-1}|} \sum_{x_i \in C_j^{t-1}} x_i$ **pentru $j = \overline{1, K}$;**

***Pasul 2*:**

se reasignează fiecare x_i la [clusterul cu] centroidul cel mai apropiat, adică se stabileşte noua componenţă a clusterelor la iteraţia t: $C_1^t, \ldots, C_K^t$;

[***Terminare:***] **până când o anumită condiţie este îndeplinită (de exemplu: până când poziţiile centroizilor — sau: componenţa clusterelor — nu se mai modifică de la o iteraţie la alta).**

a. Demonstraţi că, de la o iteraţie la alta, algoritmul K-means măreşte *coeziunea de ansamblu* a clusterelor. I.e., considerând funcţia

$$J(C^t, \mu^t) \overset{def.}{=} \sum_{i=1}^{n} ||x_i - \mu^t_{C^t(x_i)}||^2 \overset{def.}{=} \sum_{i=1}^{n} \left(x_i - \mu^t_{C^t(x_i)}\right) \cdot \left(x_i - \mu^t_{C^t(x_i)}\right),$$

unde:

$C^t = (C^t_1, C^t_2, \ldots, C^t_K)$ **este colecţia de clustere (i.e., K-partiţia) la momentul t,**

$\mu^t = (\mu^t_1, \mu^t_2, \ldots, \mu^t_K)$ **este colecţia de centroizi ai clusterelor (K-configuraţia) la momentul t,**

$C^t(x_i)$ **desemnează clusterul la care este asignat elementul x_i la iteraţia t,**

operatorul $\cdot$ desemnează produsul scalar al vectorilor din $\mathbb{R}^d$,

arătaţi că $J(C^t, \mu^t) \geq J(C^{t+1}, \mu^{t+1})$ pentru orice t.

Ideea demonstraţiei

Inegalitatea de mai sus rezultă din două inegalităţi (care corespund paşilor 1 şi 2 de la iteraţia t):

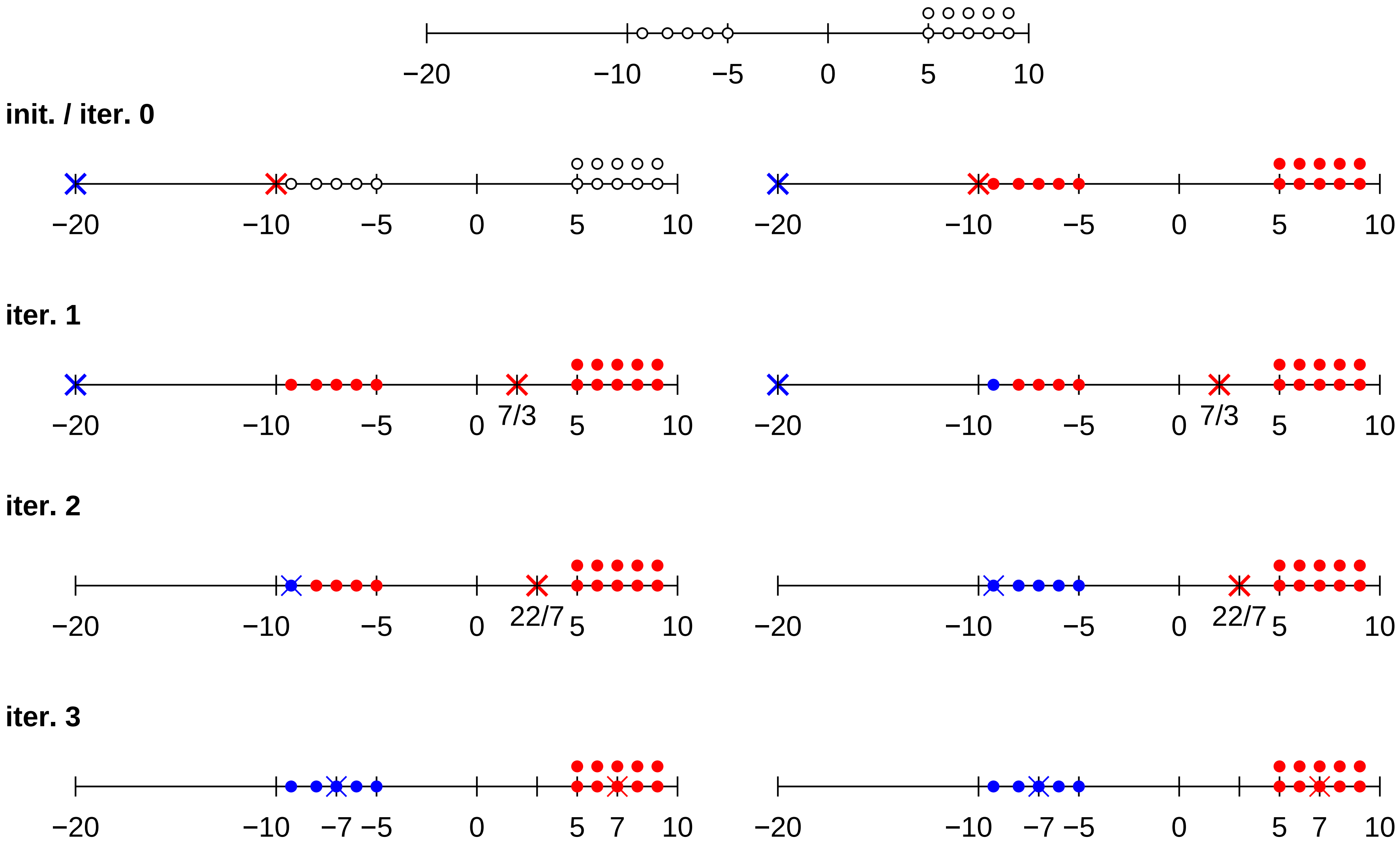
$$J(C^t, \mu^t) \overset{(1)}{\geq} J(C^t, \mu^{t+1}) \overset{(2)}{\geq} J(C^{t+1}, \mu^{t+1})$$

La prima inegalitate (cea corespunzătoare pasului 1) se poate considera că parametrul C^t este fixat iar μ este variabil, în vreme ce la a doua inegalitate (cea corespunzătoare pasului 2) se consideră μ^t fixat şi C variabil.

Prima inegalitate se poate obţine însumând o serie de inegalităţi, şi anume câte una pentru fiecare cluster C_j^t. A doua inegalitate se demonstrează imediat.

Ilustrarea acestei idei, pe un exemplu particular:

Vezi următoarele 3 slide-uri
[Edinburgh, 2009 fall, C. Williams, V. Lavrenko, HW4, pr. 3]


−20
−10
−5
0
5
10
init. / iter. 0
iter. 1
7/3
iter. 2
22/7
iter. 3
−7
7

Pentru acest exemplu de aplicare a algoritmului K-means, scriem expresiile numerice pentru valoarea criteriului $J_2(C^t, \mu^t)$ pentru fiecare iteraţie ($t = 0, 1, 2, 3$).

iter.	$J_2(C^t, \mu^t)$
0.	$0 + \{(-9-(-10))^2 + \ldots + (-5-(-10))^2$ $+2[(5-(-10))^2 + \ldots + (9-(-10))^2]\} \quad \geq$
1.	$(-9-(-20))^2 + \{(-8-7/3))^2 + \ldots (-5-7/3)^2$ $+2[(5-7/3)^2 + \ldots + (9-7/3)^2]\} \quad \geq$
2.	$(-9-(-9))^2 + \ldots (-5-(-9))^2$ $+2[(5-22/7)^2 + \ldots + (9-22/7)^2] \quad \geq$
3.	$(-9-(-7))^2 + \ldots (-5-(-7))^2$ $+2[(5-7)^2 + \ldots + (9-7)^2]$

***Observaţie*: La prima vedere, este greu să dovedim aceste inegalităţi ($J_2(C^{t-1}, \mu^{t-1}) \geq J_2(C^t, \mu^t)$, pentru $t = 1, 2, 3$) ...altfel decât calculând efectiv valoarea expresiilor care se compară. Însă, introducând nişte termeni intermediari, inegalităţile acestea se vor demonstra într-un mod foarte elegant...**

iter.	$J_2(C^{t-1}, \mu^t)$		$J_2(C^t, \mu^t)$	
0.			$0 + \{(-9-(-10))^2 + \ldots + (-5-(-10))^2$ $+2[(5-(-10))^2 + \ldots + (9-(-10))^2]\}$	$\geq$
1.	$0 + \{(-9-7/3)^2 + (-8-7/3)^2 + \ldots + (-5-7/3))^2$ $+2[(5-7/3)^2 + \ldots + (9-7/3)^2]\}$	$\geq$	$(-9-(-20))^2 + \{(-8-7/3))^2 + \ldots + (-5-7/3)^2$ $+2[(5-7/3)^2 + \ldots + (9-7/3)^2]\}$	$\geq$
2.	$(-9-(-9))^2 + \{(-8-22/7))^2 + \ldots + (-5-22/7)^2$ $+2[(5-22/7)^2 + \ldots + (9-22/7)^2]\}$	$\geq$	$(-9-(-9))^2 + (-8-(-9))^2 + \ldots + (-5-(-9))^2$ $+2[(5-22/7)^2 + \ldots + (9-22/7)^2]$	$\geq$
3.	$(-9-(-7))^2 + \ldots + (-5-(-7))^2$ $+2[(5-7)^2 + \ldots + (9-7)^2]$	$=$	$(-9-(-7))^2 + \ldots + (-5-(-7))^2$ $+2[(5-7)^2 + \ldots + (9-7)^2]$	

Explicaţii:

1. Inegalităţile pe orizontală ($J_2(C^{t-1}, \mu^t) \geq J_2(C^t, \mu^t)$, pentru $t = 1, 2, 3$) sunt uşor de demonstrat, pe baza corespondenţei termen cu termen. (Ele corespund eventualelor micşorări ale distanţelor atunci când se face reasignarea instanţelor la centroizi.)

2. Restul inegalităţilor ($J_2(C^t, \mu^t) \geq J_2(C^t, \mu^{t+1})$, pentru $t = 1, 2, 3$) se rezolvă printr-o metodă de optimizare simplă. De exemplu, pentru $t = 1$ este imediat că funcţia $(-9-x)^2 + (-8-x)^2 + \ldots + (-5-x)^2 + 2[(5-x)^2 + \ldots + (9-x)^2]$ îşi atinge minimul pentru $x = 7/3$, deci $J_2(C^1, \mu^2) \geq J_2(C^1, \mu^1)$.

Demonstraţie, pentru cazul general

Observaţie: **Pentru convenienţă, ne vom limita la cazul $d = 1$. Extinderea demonstraţei la cazul $d > 1$ nu comportă dificultăţi.**

Demonstrarea inegalităţii (1): $J(C^t, \mu^t) \geq J(C^t, \mu^{t+1})$

(Vezi pasul 1 al iteraţiei t.)

Fixăm $j \in \{1, \ldots, K\}$. Dacă notăm cu $C_j^t = \{x_{i_1}, x_{i_2}, \ldots, x_{i_l}\}$, unde $l \overset{not.}{=} |C_j^t|$, atunci

$$J(C_j^t, \mu_j^t) = \sum_{p=1}^{l} \left(x_{i_p} - \mu_j^t\right)^2, \text{ deci } J(C^t, \mu^t) = \sum_{j=1}^{K} J(C_j^t, \mu_j^t).$$

Dacă se consideră C_j^t fixat, iar μ_j^t variabil, atunci putem minimiza imediat funcţia

$$f(\mu) \overset{def.}{=} J(C_j^t, \mu) = l\mu^2 - 2\mu \sum_{p=1}^{l} x_{i_p} + \sum_{p=1}^{l} x_{i_p}^2 \Rightarrow \arg\min_{\mu} J(C_j^t, \mu) = \frac{1}{l} \sum_{p=1}^{l} x_{i_p} \overset{def.}{=} \mu_j^{t+1}.$$

Aşadar, $J(C_j^t, \mu) \geq J(C_j^t, \mu_j^{t+1})$, pentru $\forall \mu$. În particular, pentru $\mu = \mu_j^t$ vom avea: $J(C_j^t, \mu_j^t) \geq J(C_j^t, \mu_j^{t+1})$. Inegalitatea aceasta este valabilă pentru toate clusterele $j = 1, \ldots, K$. Dacă sumăm toate aceste inegalităţi, rezultă: $J(C^t, \mu^t) \geq J(C^t, \mu^{t+1})$.

Demonstrarea inegalităţii (2): $J(C^t, \mu^{t+1}) \geq J(C^{t+1}, \mu^{t+1})$

(Vezi pasul 2 al iteraţiei t.)

La acest pas, o instanţă oarecare x_i, unde $i \in \{1, \ldots, n\}$, este reasignată de la clusterul cu centroidul μ_j^{t+1}, la un alt centroid μ_q^{t+1}, dacă

$||x_i - \mu_{j'}^{t+1}||^2 \geq ||x_i - \mu_q^{t+1}||^2 \Leftrightarrow (x_i - \mu_{j'}^{t+1})^2 \geq (x_i - \mu_q^{t+1})^2,$ **pentru orice** $j' = 1, \ldots, K.$

În contextul iteraţiei t, acest lucru implică

$$\left(x_i - \mu_{C^t(x_i)}^{t+1}\right)^2 \geq \left(x_i - \mu_{C^{t+1}(x_i)}^{t+1}\right)^2.$$

Sumând membru cu membru inegalităţile de acest tip obţinute pentru $i = \overline{1, n}$, rezultă: $J(C^t, \mu^{t+1}) \geq J(C^{t+1}, \mu^{t+1})$, ceea ce era de demonstrat.

b. Ce puteţi spune despre oprirea algoritmului K-means? (Termină oare acest algoritm într-un număr finit de paşi, sau este posibil ca el să reviziteze o K-configuraţie anterioară $\mu = (\mu_1, \ldots, \mu_K)$?)

Răspuns:

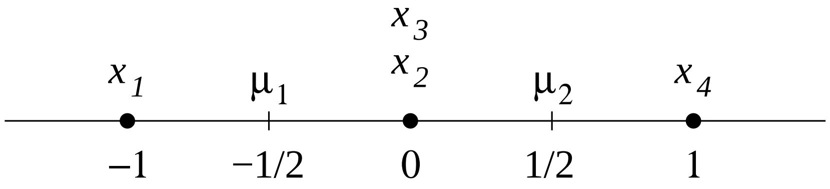
Dacă algoritmul revizitează o K-partiţie, atunci rezultă că pentru un anumit t avem $J(C^{t-1}, \mu^t) = J(C^t, \mu^{t+1})$. Este posibil ca acest fapt să se întâmple, şi anume atunci când:

– există instanţe multiple (i.e., $x_i = x_j$, deşi $i \neq j$),

– criteriul de oprire al algoritmului K-means este de forma "până când componenţa clusterelor nu se mai modifică",

– se presupune că, în cazul în care o instanţă x_i este situată la egală distanţă faţă de doi sau mai mulţi centroizi, ea poate fi asignată în mod aleatoriu la oricare dintre ei.

Aşa se întâmplă în *exemplul* din figura alăturată dacă se consideră că la o iteraţie t avem $x_2 = 0 \in C_1^t$ şi $x_3 = 0 \in C_2^t$, iar la iteraţia următoare alegem ca $x_3 = 0 \in C_1^{t+1}$ şi $x_2 = 0 \in C_2^{t+1}$ şi, din nou, invers la iteraţia $t + 2$.



Observaţii

- **Dacă se păstrează criteriul dat ca exemplu în enunţul problemei – adică se iterează până când centroizii "staţionează" – algoritmul se poate opri fără ca la ultima iteraţie $J(C, \mu)$ să fi atins minimul posibil. În cazul exemplului de mai sus, vom avea $\frac{1}{4} + 2 \cdot \frac{1}{4} + \frac{1}{4} = 1 > \frac{2}{3}$.**

- **Dacă nu există instanţe multiple care să fie situate la distanţe egale faţă de doi sau mai mulţi centroizi la o iteraţie oarecare a algoritmului K-means (precum sunt x_2 şi x_3 în *exemplul* de mai sus), sau dacă se impune *restricţia* ca în astfel de situaţii instanţele identice să fie asignate la un singur cluster, este evident că algoritmul K-means se opreşte într-un număr finit de paşi.**

Concluzii

- Algoritmul K-means explorează — pornind de la o anumită iniţializare a celor K centroizi —, doar un subset din totatul de K^n K-partiţii, asigurându-ne însă că are loc proprietatea $J(C^0, \mu^1) \geq J(C^1, \mu^2) \geq \ldots \geq J(C^{t-1}, \mu^t) \geq J(C^t, \mu^{t+1})$, conform punctului a al acestei probleme.

- Atingerea minimului global al funcţiei $J(C, \mu)$ — unde C este o variabilă care parcurge mulţimea tuturor K-partiţiilor care se pot forma cu instanţele $\{x_1, \ldots, x_n\}$ — nu este garantată pentru algoritmul K-means. Valoarea funcţiei J care se obţine la oprirea algoritmului K-means este dependentă de plasarea iniţială a centroizilor μ precum şi de modul concret în care sunt alcătuite clusterele în cazul în care o instanţă oarecare se află la distanţă egală de doi sau mai mulţi centroizi, după cum am arătat în exemplul de mai sus.

K-means algorithm: The "approximate" maximization of the "distance" between clusters

[CMU, 2010 fall, Aarti Singh, HW3, pr. 5.2]

Note: In this problem we will work with a version of the K-means algorithm which is slightly modified w.r.t. the one given in the problem CMU, 2009 spring, Ziv Bar-Joseph, HW5, pr. 2.1, where we have proved the monotonicity of the criterion J.

Let $X := \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be our sample points, and K denote the number of clusters to use. We represent the cluster assignments of the data points by an indicator matrix $\gamma \in \{0,1\}^{n \times K}$ such that $\gamma_{ij} = 1$ means $\mathbf{x}_i$ belongs to cluster j. We require that each point belongs to exactly one cluster, so $\sum_{j=1}^{K} \gamma_{ij} = 1$.

[We already know that] **the K-means algorithm "estimates" γ by minimizing the following "cohesion criterion" (or, "measure of distortion"):**

$$J(\gamma, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K) \; := \sum_{i=1}^{n} \sum_{j=1}^{K} \gamma_{ij} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2,$$

where $\|\cdot\|$ denotes the vector 2-norm.

K-means alternates between estimating γ and re-computing $\boldsymbol{\mu}_j$'s.

- **Initialize** $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K$**, and let** $C := \{1, \ldots, K\}$**.**
- **While the value of** J **is still decreasing, repeat the following:**

  1. **Determine** γ **by**

     $$\gamma_{ij} \leftarrow \begin{cases} 1, & \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \leq \|\mathbf{x}_i - \boldsymbol{\mu}_{j'}\|^2, \ \forall j' \in C, \\ 0, & \textbf{otherwise.} \end{cases}$$

     Break ties arbitrarily.

  2. **Recompute** $\boldsymbol{\mu}_j$ **using the updated** γ**:**
     For each $j \in C$**, if** $\sum_{i=1}^n \gamma_{ij} > 0$ **set**

     $$\boldsymbol{\mu}_j \leftarrow \frac{\sum_{i=1}^n \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^n \gamma_{ij}}.$$

     Otherwise, don't change μ_j**.**

Let $\bar{\mathbf{x}}$ denote the sample mean.
Consider the following three quantities:

$$\begin{aligned}
\textbf{Total variation:} \quad & V(X) = \frac{\sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2}{n}. \\
\textbf{Within-cluster variation:} \quad & V_j(X) = \frac{\sum_{i=1}^{n} \gamma_{ij} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2}{\sum_{i=1}^{n} \gamma_{ij}}. \\
\textbf{Between-cluster variation:} \quad & \widetilde{V}(X) = \sum_{j=1}^{K} \left(\frac{\sum_{i=1}^{n} \gamma_{ij}}{n} \right) \|\boldsymbol{\mu}_j - \bar{\mathbf{x}}\|^2.
\end{aligned}$$

What is the relation between these three quantities?

Based on this relation, show that K-means can be interpreted as minimizing a weighted average of within-cluster variations while approximately(!) maximizing the between-cluster variation. Note that the relation may contain an extra term that does not appear above.

Solution

To simplify the notation, we define $n_j = \sum_{i=1}^n \gamma_{ij}$**.**

We then have:

$$\begin{aligned}
V\left(X\right) &= \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K} \gamma_{ij}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 = \frac{1}{n}\sum_{j=1}^{K}\sum_{i=1}^{n} \gamma_{ij}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 \\
&= \frac{1}{n}\sum_{j=1}^{K}\sum_{i=1}^{n} \gamma_{ij}\|\mathbf{x}_i - \boldsymbol{\mu}_j + \boldsymbol{\mu}_j - \bar{\mathbf{x}}\|^2 \\
&= \frac{1}{n}\sum_{j=1}^{K}\sum_{i=1}^{n} \gamma_{ij}\left(\|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 + \|\boldsymbol{\mu}_j - \bar{\mathbf{x}}\|^2 + 2\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)\cdot\left(\boldsymbol{\mu}_j - \bar{\mathbf{x}}\right)\right) \\
&= \sum_{j=1}^{K} \frac{n_j}{n}\frac{\sum_{i=1}^{n}\gamma_{ij}\|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2}{n_j} + \sum_{j=1}^{K}\frac{n_j\|\boldsymbol{\mu}_j - \bar{\mathbf{x}}\|^2}{n} + \quad + \frac{2}{n}\sum_{j=1}^{K}\left(\boldsymbol{\mu}_j - \bar{\mathbf{x}}\right)\cdot\left(\sum_{i=1}^{n}\gamma_{ij}\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)\right) \\
&= \sum_{j=1}^{K}\frac{n_j}{n}V_j(X) + \widetilde{V}(X) - \frac{2}{n}\sum_{j=1}^{K} n_j\left(\boldsymbol{\mu}_j - \bar{\mathbf{x}}\right)\cdot\left(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}}_j\right), \ \textbf{where } \bar{\boldsymbol{\mu}}_j \overset{not.}{=} \frac{\sum_{i=1}^{n}\gamma_{ij}\mathbf{x}_i}{n_j}.
\end{aligned}$$

Note: The last equality on the previous slide holds because

$$\begin{aligned}\sum_{i=1}^{n} \gamma_{ij}(x_i - \mu_j) &= \left(\sum_{i=1}^{n} \gamma_{ij} x_i\right) - n_j \mu_j = n_j \frac{\sum_{i=1}^{n} \gamma_{ij} x_i}{n_j} - n_j \mu_j \\ &= n_j \left(\frac{\sum_{i=1}^{n} \gamma_{ij} x_i}{n_j} - \mu_j\right) = n_j \left(\bar{\mu}_j - \mu_j\right)\end{aligned}$$

We already know — see CMU, 2009 spring, Ziv Bar-Joseph, HW5, pr. 2.1 — that K-means aims to minimize J, and consequently $\frac{1}{n}J$, which coincides with the first term in the expression we obtained for $V(X)$, namely $\sum_{j=1}^{K} \frac{n_j}{n} V_j(X)$.

Since the total variation $V(X)$ is constant, minimizing the first term is equivalent to maximizing the sum of the other two terms, which is expected to be dominated by the between-cluster variation $\widetilde{V}(X)$ since a good $\boldsymbol{\mu}_j$ should be close to $\bar{\boldsymbol{\mu}}_j$, making the third term small in absolute value.

Exemplifying the application of a simple version of EM/GMM on data from $\mathbb{R}$

$(\sigma_1 = \sigma_2 = 1, \pi_1 = \pi_2 = 1/2)$

CMU, 2012 spring, Ziv Bar-Joseph, final exam, pr. 3.1

enhanced by Liviu Ciortuz

Suppose a GMM has two components with known variance and an equal prior distribution

$$\frac{1}{2}N(\mu_1, 1) + \frac{1}{2}N(\mu_2, 1)$$

The observed data are $x_1 = 0.5$ and $x_2 = 2$, and the current estimates of μ_1 and μ_2 are 1 and 2 respectively.

Execute the first iteration of the EM algorithm.

***Hint*: Normal densities for the standardized variable $y_{(\mu=0,\sigma=1)}$ at 0, 0.5, 1, 1.5, 2 are 0.4, 0.35, 0.24, 0.13, 0.05 respectively.**

Solution

The E-step:

$$E[Z_{i1}] = P(Z_{i1}=1|x_i,\mu) \overset{B.Th.}{=} \frac{P(x_i|Z_{i1}=1,\mu_1)P(Z_{i1}=1)}{P(x_i|Z_{i1}=1,\mu_1)P(Z_{i1}=1)+P(x_i|Z_{i2}=1,\mu_2)P(Z_{i2}=1)}$$

$$= \frac{P(x_i|Z_{i1}=1,\mu_1)\cdot\frac{1}{2}}{P(x_i|Z_{i1}=1,\mu_1)\cdot\frac{1}{2}+P(x_i|Z_{i2}=1,\mu_2)\cdot\frac{1}{2}}$$

$$= \frac{P(x_i|Z_{i1}=1,\mu_1)}{P(x_i|Z_{i1}=1,\mu_1)+P(x_i|Z_{i2}=1,\mu_2)} \quad \textbf{for } i\in\{1,2\}.$$

Therefore,

$$\begin{aligned}
P(Z_{11}=1|x_1,\mu) &= \frac{N(0.5;1,1)}{N(0.5;1,1)+N(0.5;2,1)} = \frac{N(0.5;0,1)}{N(0.5;0,1)+N(1.5;0,1)} = \frac{0.35}{0.35+0.13} = \frac{35}{48} \\
P(Z_{21}=1|x_2,\mu) &= \frac{N(2;1,1)}{N(2;1,1)+N(2;2,1)} = \frac{N(1;0,1)}{N(1;0,1)+N(0;0,1)} = \frac{0.24}{0.24+0.4} = \frac{0.24}{0.64} = \frac{3}{8}
\end{aligned}$$

Similarly,

$$\begin{aligned}
P(Z_{12}=1|x_1,\mu) &= P(Z_{11}=0|x_1,\mu) = 1-P(Z_{11}=1|x_1,\mu_1) = \frac{13}{48} \\
P(Z_{22}=1|x_2,\mu) &= P(Z_{21}=0|x_2,\mu) = 1-P(Z_{21}=1|x_2,\mu_1) = \frac{5}{8}
\end{aligned}$$

The M-step:

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{2} E[Z_{ij}]\, x_i}{\sum_{i=1}^{2} E[Z_{ij}]} = \frac{\sum_{i=1}^{2} P(Z_{ij} = 1 | x_i, \mu^{(t)})\, x_i}{\sum_{i=1}^{2} P(Z_{ij} = 1 | x_i, \mu^{(t)})}$$

Therefore,

$$\mu_1^{(1)} = \frac{\dfrac{35}{48} \cdot 0.5 + \dfrac{3}{8} \cdot 2}{\dfrac{35}{48} + \dfrac{3}{8}} = \frac{107}{106} \approx 1.009 \quad \textbf{and} \quad \mu_2^{(1)} = \frac{\dfrac{13}{48} \cdot 0.5 + \dfrac{5}{8} \cdot 2}{\dfrac{13}{48} + \dfrac{5}{8}} = \frac{133}{86} \approx 1.54$$

Derivation of the EM algorithm for a mixture of K uni-variate Gaussians: the general case (i.e., when all parameters π, μ, σ^2 are free)

following Dahua Lin,
An Introduction to Expectation-Maximization
(MIT, ML 6768 course, 2012 fall)

Note

We will first consider $K = 2$.
Generalization to $K > 2$ will be shown afterwards.

Estimation (E) Step:

$$
\begin{aligned}
p_{ij} &\overset{not.}{=} P(Z_{ij} = 1 \mid X_i, \mu, \sigma, \pi) \overset{calcul}{=} E[Z_{ij} \mid X_i, \mu, \sigma, \pi] \\
&= \frac{P(X_i = x_i \mid Z_{ij} = 1, \mu, \sigma, \pi) \cdot P(Z_{ij} = 1 \mid \mu, \sigma, \pi)}{\sum_{j'=1}^{2} P(X_i = x_i \mid Z_{ij'} = 1, \mu, \sigma, \pi) \cdot P(Z_{ij'} = 1 \mid \mu, \sigma, \pi)} \\
&= \frac{\dfrac{1}{\sqrt{2\pi}\sigma_j} \cdot \exp\left(-\dfrac{(x_i - \mu_j)^2}{2\sigma_j^2}\right) \cdot \pi_j}{\dfrac{1}{\sqrt{2\pi}\sigma_1} \cdot \exp\left(-\dfrac{(x_i - \mu_1)^2}{2\sigma_1^2}\right) \cdot \pi_1 + \dfrac{1}{\sqrt{2\pi}\sigma_2} \cdot \exp\left(-\dfrac{(x_i - \mu_2)^2}{2\sigma_2^2}\right) \cdot \pi_2}
\end{aligned}
$$

Therefore, for $t > 0$ we will have:

$$
p_{ij}^{(t)} = \frac{\dfrac{\pi_j^{(t-1)}}{\sigma_j^{(t-1)}} \cdot \exp\left(-\dfrac{(x_i - \mu_j^{(t-1)})^2}{2(\sigma_j^{(t-1)})^2}\right)}{\dfrac{\pi_1^{(t-1)}}{\sigma_1^{(t-1)}} \cdot \exp\left(-\dfrac{(x_i - \mu_1^{(t-1)})^2}{2(\sigma_1^{(t-1)})^2}\right) + \dfrac{\pi_2^{(t-1)}}{\sigma_2^{(t-1)}} \cdot \exp\left(-\dfrac{(x_i - \mu_2^{(t-1)})^2}{2(\sigma_2^{(t-1)})^2}\right)}
$$

The likelihood of a "complete" instance (x_i, z_{i1}, z_{i2}):

$$P(X_i = x_i, Z_{i1} = z_{i1}, Z_{i2} = z_{i2} \mid \mu, \sigma, \pi)$$

$$= P(X_i = x_i \mid Z_{i1} = z_{i1}, Z_{i2} = z_{i2}, \mu_i, \sigma_i, \pi_i) \cdot P(Z_{i1} = z_{i1}, Z_{i2} = z_{i2} \mid \mu_i, \sigma_i, \pi_i)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_j} \cdot \exp\left(-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right) \cdot \pi_j, \ \textbf{ where } z_{ij} = 1 \textbf{ and } z_{ij'} = 0 \textbf{ for } j' \neq j$$

$$= \frac{1}{\sqrt{2\pi}\ \sigma_1^{z_{i1}}\ \sigma_2^{z_{i2}}} \cdot \exp\left(-\frac{1}{2}\sum_{j\in\{1,2\}} z_{ij}\frac{(x_i - \mu_j)^2}{\sigma_j^2}\right) \cdot \pi_1^{z_{i1}}\pi_2^{z_{i2}}$$

The log-likelihood of the same "complete" instance will be:

$$\ln P(X_i = x_i, Z_{i1} = z_{i1}, Z_{i2} = z_{i2} \mid \mu, \sigma, \pi)$$

$$= -\frac{1}{2}\ln(2\pi) - \sum_{j=1}^{2} z_{ij} \ln\sigma_j - \frac{1}{2}\sum_{j=1}^{2} z_{ij}\frac{(x_i - \mu_j)^2}{\sigma_j^2} + \sum_{j=1}^{2} z_{ij}\ \ln\pi_j$$

Given the dataset $X = \{x_1, \ldots, x_n\}$, the log-likelihood function will be:

$$l(\mu, \sigma, \pi) \overset{def.}{=} \ln P(X, Z_1, Z_2 \mid \mu, \sigma, \pi) \overset{i.i.d.}{=} \ln \prod_{i=1}^{n} P(X_i = x_i, Z_{i1}, Z_{i2} \mid \mu, \sigma, \pi)$$

$$= \sum_{i=1}^{n} \ln P(X_i = x_i, Z_{i1}, Z_{i2} \mid \mu, \sigma, \pi)$$

$$= -\frac{n}{2} \ln(2\pi) - \sum_{i=1}^{n} \sum_{j=1}^{2} Z_{ij} \ln \sigma_j - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{2} Z_{ij} \frac{(x_i - \mu_j)^2}{\sigma_j^2} + \sum_{i=1}^{n} \sum_{j=1}^{2} Z_{ij} \ln \pi_j$$

The expectation of the log-likelihood function:

$$E[\ln P(X, Z_1, Z_2 \mid \mu, \sigma, \pi)] =$$

$$-\frac{n}{2}\ln(2\pi) - \sum_{i=1}^{n}\sum_{j=1}^{2} E[Z_{ij}]\ln\sigma_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{2} E[Z_{ij}]\frac{(x_i-\mu_j)^2}{\sigma_j^2} + \sum_{i=1}^{n}\sum_{j=1}^{2} E[Z_{ij}]\,\ln\pi_j$$

Here above, the probability function w.r.t. which the expectation was computed was left unspecified. Now we will make it explicit:

$$\begin{aligned} Q(\mu, \sigma, \pi \mid \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) &\overset{not.}{=} E_{Z|X,\,\mu^{(t)},\sigma^{(t)},\pi^{(t)}}[\ln P(X, Z_1, Z_2 \mid X, \mu, \sigma, \pi)] \\ = \quad & -\frac{n}{2}\ln(2\pi) - \sum_{i=1}^{n}\sum_{j=1}^{2} E[Z_{ij} \mid X_i, \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}]\ln\sigma_j \\ & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{2} E[Z_{ij} \mid X_i, \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}]\frac{(x_i-\mu_j)^2}{\sigma_j^2} \\ & +\sum_{i=1}^{n}\sum_{j=1}^{2} E[Z_{ij} \mid X_i, \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}]\,\ln\pi_j \end{aligned}$$

$$p_{ij}^{(t)} \overset{not.}{=} E[Z_{ij} \mid X, \mu^{(t-1)}, \sigma^{(t-1)}, \pi^{(t-1)}] \Rightarrow$$

$$Q(\mu, \sigma, \pi \mid \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) =$$

$$-\frac{n}{2}\ln(2\pi) - \sum_{i=1}^{n}\sum_{j=1}^{2} p_{ij}^{(t)} \ln\sigma_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{2} p_{ij}^{(t)} \frac{(x_i-\mu_j)^2}{\sigma_j^2} + \sum_{i=1}^{n}\sum_{j=1}^{2} p_{ij}^{(t)} \ln\pi_j$$

Since $K = 2$ and $\pi_1 + \pi_2 = 1$, we get

$$Q(\mu, \sigma, \pi \mid \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) =$$

$$-\frac{n}{2}\ln 2\pi - \sum_{i=1}^{n}\sum_{j=1}^{2} p_{ij}^{(t)} \ln\sigma_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{2} p_{ij}^{(t)} \frac{(x_i-\mu_j)^2}{\sigma_j^2} + \sum_{i=1}^{n}(p_{i1}^{(t)} \ln\pi_1 + p_{i2}^{(t)} \ln(1-\pi_1))$$

Maximization (M) Step:

[For $K = 2$:]

$$\frac{\partial}{\partial \pi_1} Q(\mu, \sigma, \pi \mid \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) = 0 \Leftrightarrow \frac{1}{\pi_1} \sum_{i=1}^{n} p_{i1}^{(t)} = \frac{1}{1 - \pi_1} \sum_{i=1}^{n} p_{i2}^{(t)} \Leftrightarrow$$

$$\sum_{i=1}^{n} p_{i1}^{(t)} = \pi_1 (\sum_{i=1}^{n} p_{i1}^{(t)} + \sum_{i=1}^{n} p_{i2}^{(t)}) \Leftrightarrow \sum_{i=1}^{n} p_{i1}^{(t)} = \pi_1 \sum_{i=1}^{n} (\underbrace{p_{i1}^{(t)} + p_{i2}^{(t)}}_{1}) \Leftrightarrow \sum_{i=1}^{n} p_{i1}^{(t)} = n\pi_1$$

$$\Rightarrow \pi_1^{(t+1)} \leftarrow \frac{1}{n} \sum_{i=1}^{n} p_{i1}^{(t)}$$

Taking into account that $\pi_1^{(t+1)} + \pi_2^{(t+1)} = 1$ **and** $p_{i1}^{(t)} + p_{i2}^{(t)} = 1$ **for** $i = 1, \ldots, n$**,**

$$\Rightarrow \pi_2^{(t+1)} \leftarrow \frac{1}{n} \sum_{i=1}^{n} p_{i2}^{(t)}$$

$$\frac{\partial}{\partial \mu_1} \, Q(\mu, \sigma, \pi \mid \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) = 0 \Leftrightarrow \frac{1}{\sigma_1^2} \sum_{i=1}^{n} p_{i1}^{(t)}(x_i - \mu_1) = 0 \Leftrightarrow \sum_{i=1}^{n} p_{i1}^{(t)}(x_i - \mu_1) = 0$$

$$\Rightarrow \mu_1^{(t+1)} \leftarrow \frac{\sum_{i=1}^{n} p_{i1}^{(t)} x_i}{\sum_{i=1}^{n} p_{i1}^{(t)}}$$

Similarly, $\mu_2^{(t+1)} \leftarrow \frac{\sum_{i=1}^{n} p_{i2}^{(t)} x_i}{\sum_{i=1}^{n} p_{i2}^{(t)}}$

$$\frac{\partial}{\partial \sigma_1} Q(\mu, \sigma, \pi \mid \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) = 0 \Leftrightarrow -\frac{1}{\sigma_1} \sum_{i=1}^{n} p_{i1}^{(t)} + \frac{1}{\sigma_1^3} \sum_{i=1}^{n} p_{i1}^{(t)} (x_i - \mu_1)^2 = 0,$$

$$\Rightarrow \left(\sigma_1^{(t+1)}\right)^2 \leftarrow \frac{\sum_{i=1}^{n} p_{i1}^{(t)} (x_i - \mu_1^{(t+1)})^2}{\sum_{i=1}^{n} p_{i1}^{(t)}}$$

Similarly, $\left(\sigma_2^{(t+1)}\right)^2 \leftarrow \frac{\sum_{i=1}^{n} p_{i2}^{(t)} (x_i - \mu_2^{(t+1)})^2}{\sum_{i=1}^{n} p_{i2}^{(t)}}$

***Note*: One could relatively easy prove that these solutions (namely, $\pi^{(t+1)}, \mu^{(t+1)}, \sigma^{(t+1)}$) of the partial derivatives of the *auxiliary function* Q designate the values for which Q reaches its *maximum*.**

Generalization to $K > 2$

In this case, the Bernoulli distribution is replaced by a categorical one. The only one change needed in the above proof concerns updating the parameters of this distribution.
Since $\pi_1 + \ldots + \pi_K = 1$, we must solve the following *constraint optimization problem*:

$$\max_{\pi,\mu,\sigma} Q(\pi,\mu,\sigma|\pi^{(t)},\mu^{(t)},\sigma^{(t)})$$

$$\text{subject to } \sum_{i=1}^{K} \pi_j = 1 \textbf{ and } \pi_j \geq 0,\ \forall j = 1,\ldots,K.$$

By letting aside the $\geq$ constraints, and using the *Lagrangean multiplier* $\lambda \in \mathbb{R}$, this problem becomes:

$$\max_{\pi,\mu,\sigma} \left(Q(\pi,\mu,\sigma|\pi^{(t)},\mu^{(t)},\sigma^{(t)}) + \lambda(1 - \sum_{i=1}^{K} \pi_j) \right).$$

For $j = 1, \ldots, K$:

$$\frac{\partial}{\partial \pi_j} \, Q(\mu, \sigma, \pi \mid \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) = 0 \Leftrightarrow \sum_{i=1}^{n} p_{ij}^{(t)} \frac{1}{\pi_j} = \lambda \Leftrightarrow \pi_j^{(t+1)} = \frac{1}{\lambda} \sum_{i=1}^{n} p_{ij}^{(t)}.$$

Because $\sum_{j=1}^{K} \pi_j^{(t+1)} = 1$, it follows that

$$\lambda = \sum_{j=1}^{K} \sum_{i=1}^{n} \pi_j^{(t+1)} = \sum_{i=1}^{n} \underbrace{\sum_{j=1}^{K} \pi_j^{(t+1)}}_{1} = \sum_{i=1}^{n} 1 = n.$$

Therefore,

$$\pi_j^{(t+1)} \leftarrow \frac{1}{n} \sum_{i=1}^{n} p_{ij}^{(t)}.$$

***Note* that indeed $\pi_j^{(t+1)} \geq 0$, because the $p_{ij}^{(t)}$ terms designate some probabilities (see E-step).**

To summarize:

E Step:

$$p_{ij}^{(t)} \overset{not.}{=} P(z_{ij}=1 \mid x_i; \mu^{(t)}, (\sigma^2)^{(t)}, \pi^{(t)}) = \frac{N(x_i \mid \mu_j^{(t)}, (\sigma_j^2)^{(t)}) \cdot \pi_j^{(t)}}{\sum_{l=1}^{K} N(x_i \mid \mu_l^{(t)}, (\sigma_l^2)^{(t)}) \cdot \pi_l^{(t)}}$$

$$\textbf{where } N(x_i \mid \mu_j, \sigma_j^2) \overset{def.}{=} \frac{1}{\sqrt{2\pi}\sigma_j} \cdot \exp\left(-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}\right).$$

M Step:

$$\pi_j^{(t+1)} \leftarrow \frac{1}{n}\sum_{i=1}^{n} p_{ij}^{(t)}$$

$$\mu_j^{(t+1)} \leftarrow \frac{\sum_{i=1}^{n} p_{ij}^{(t)} x_i}{\sum_{i=1}^{n} p_{ij}^{(t)}}$$

$$\left(\sigma_j^{(t+1)}\right)^2 \leftarrow \frac{\sum_{i=1}^{n} p_{ij}^{(t)} (x_i - \mu_j^{(t+1)})^2}{\sum_{i=1}^{n} p_{ij}^{(t)}}$$

Exemplifying

some methodological issues regarding the application of the EM algorithmic schema

(using a simple EM/GMM algorithm on data from $\mathbb{R}$ $(\pi_1 = \pi_2 = 1/2)$)

CMU, 2007 spring, Eric Xing, final exam, pr. 1.8

A long time ago there was a village amidst hundreds of lakes. Two types of fish lived in the region, but only one type in each lake.

These types of fish both looked exactly the same, smelled exactly the same when cooked, and had the exact same delicious taste – except one was poisonous and would kill any villager who ate it. The only other difference between the fish was their effect on the pH (acidity) of the lake they occupy.

The pH for lakes occupied by the non-poisonous type of fish was distributed according to a Gaussian with unknown mean (μ_{safe}) and variance (σ^2_{safe}) and the pH for lakes occupied by the poisonous type was distributed according to a different Gaussian with unknown mean (μ_{deadly}) and variance (σ^2_{deadly}). (Poisonous fish tended to cause slightly more acidic conditions).

Naturally, the villagers turned to machine learning for help. However, there was much debate about the right way to apply EM to their problem. For each of the following procedures, indicate whether it is an accurate implementation of Expectation-Maximization and will provide a reasonable estimate for parameters μ and σ^2 for each class.

a.
Guess initial values of μ and σ^2 for each class.
(1) For each lake, find the most likely class of fish for the lake.
(2) Update the μ and σ^2 values using their maximum likelihood estimates based on these predictions.
Iterate (1) and (2) until convergence.

b.
For each lake, guess an initial probability that it is safe.
(1) Using these probabilities, find the maximum likelihood estimates for the μ and σ values for each class.
(2) Use these estimates of μ and σ to reestimate lake safety probabilities.
Iterate (1) and (2) until convergence.

c.
Compute the mean and variance of the pH levels across all lakes.
Use these values for the μ and σ^2 value of each class of fish.
(1) Use the μ and σ^2 values of each class to compute the belief that each lake contains poisonous fish.
(2) Find the maximum likelihood values for μ and σ^2.
Iterate (1) and (2) until convergence.

Solution

a. It'll do ok if we give sensible enough μ and σ^2 initial values.

b. Ok, this is the same as a after the first M-step.
(See the general EM algorithmic schema on the next slide.)

c. This will be stuck at the initial μ and σ^2:

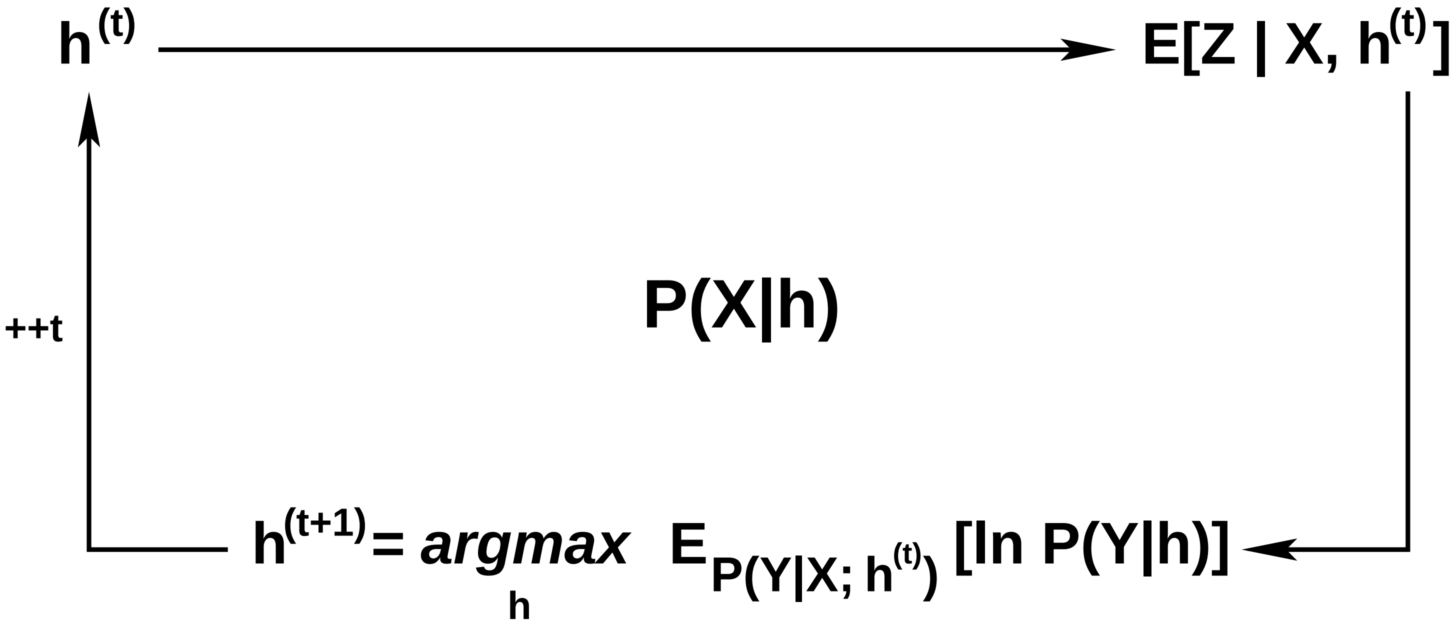
In the E-step we'll get:

$$P(safe|x) = \frac{P(x|safe) \cdot P(safe)}{P(x|safe) \cdot P(safe) + P(x|deadly) \cdot P(deadly)} = \frac{1}{2}$$

since on one side we assume $P(safe) = P(poison) = \dfrac{1}{2}$ **and on the other side** $P(x|safe) = P(x|poison)$ **because** $\mu_{safe} = \mu_{deadly}$ **and** $\sigma^2_{safe} = \sigma^2_{deadly}$**.**

In the M-step μ and σ^2 will not change since we are again letting them be calculated from all lakes (weighted equally).

The [general] EM algorithmic schema



$h^{(t)}$

$E[Z \mid X, h^{(t)}]$

$P(X|h)$

++t

$h^{(t+1)} = \underset{h}{\textit{argmax}}\ E_{P(Y|X;\, h^{(t)})}[\ln P(Y|h)]$

The EM algorithm for modeling mixtures of multi-variate Gaussians

Stanford University, Prof. Andrew Ng

ML course, 2009, lecture notes, parts VIII and IX

[adapted by Liviu Ciortuz]

Suppose that we are given the *instances* $x_1, \ldots, x_n \in \mathbb{R}^d$ (all seen as column-vectors). We wish to *model* these data by specifying a joint distribution $p(x_i, z_i) = p(x_i|z_i) \cdot p(z_i)$. Here,

$z_i \sim \text{Categorical}(\pi)$**,**

K **denotes the number of values that the z_i's can take on, namely** $\pi_j \overset{not.}{=} p(z_i = j)$ **for** $j = 1, \ldots, K$**, with** $\sum_{j=1}^{K} \pi_j = 1$**, and**

the [conditional] distribution $x_i|z_i = j$ is a Gaussian of mean vector μ_j and covariance matrix Σ_j.

Thus, our model posits that each x_i was *generated* by randomly choosing z_i from $\{1, \ldots, K\}$, and then x_i was drawn from one of the K Gaussians, depending on z_i. This is called *the mixture of* [*multi-variate*] *Gaussians* model. Remember that

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

***Note* that the z_i's are *latent* random variables, meaning that they're hidden/unobserved.**

[Use the EM general scheme (see Tom Mitchell's *Machine Learning* book, 1997, pag. 194-195) to] prove that the EM algorithm for estimating the parameters π, μ and Σ of our mixture of multi-variate Gaussian distributions has the following *update rules*:

E-step:

$$w_{ij} \overset{not.}{=} E[z_i = j | x_i; \pi', \mu', \Sigma'] = \frac{\mathcal{N}(x_i; \mu', \Sigma')\ \pi_j}{\sum_{l=1}^{K} \mathcal{N}(x_i; \mu', \Sigma')\ \pi_l} \tag{3}$$

M-step:

$$\pi_j = \frac{1}{n} \sum_{i=1}^{n} w_{ij}, \tag{4}$$

$$\mu_j = \frac{\sum_{i=1}^{n} w_{ij} x_i}{\sum_{i=1}^{n} wij}. \tag{5}$$

$$\Sigma_j = \frac{\sum_{i=1}^{n} w_{ij} (x_i - \mu_j)(x_i - \mu_j)^\top}{\sum_{i=1}^{n} w_{ij}}. \tag{6}$$

where π', μ' and Σ' represent the values of our parameters at initialization, and respectively the previous iteration of the EM algorithm.

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

(1e) $(A^{-1})^\top = (A^\top)^{-1}$

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

(4a) $\frac{\partial}{\partial X}|AXB| = |AXB|(X^{-1})^\top = |AXB|(X^\top)^{-1}$

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

(5a) $\frac{\partial}{\partial X}a^\top X = \frac{\partial}{\partial X}X^\top a = a$

(5b) $\frac{\partial}{\partial X}X^\top AX = (A + A^\top)X$

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

(5e) $\frac{\partial}{\partial X}a^\top Xa = \frac{\partial}{\partial X}a^\top X^\top a = aa^\top$

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

Solution

The *E-step* is easy (use Bayes rule):

$$\begin{aligned} w_{ij} &\overset{not.}{=} E[z_i = j | x_i; \pi', \mu', \Sigma'] = p(z_i = j | x_i; \pi', \mu', \Sigma') = \\ &= \frac{p(x_i | z_i = j; \mu', \Sigma')\, p(z_i = j; \pi')}{\sum_{l=1}^{K} p(x_i | z_i = l; \mu', \Sigma')\, p(z_i = l; \pi')} = \frac{\mathcal{N}(x_i; \mu', \Sigma')\, \pi'_j}{\sum_{l=1}^{K} \mathcal{N}(x_i; \mu', \Sigma')\, \pi_l} \end{aligned}$$

We will now concentrate on the *M-step*:

According to *the general EM scheme*, we need to maximize, with respect to our parameters π, μ, Σ, the "auxiliary" function

$$
\begin{aligned}
Q(\pi, \mu, \Sigma | \pi', \mu', \Sigma') &\overset{def.}{=} E_{p(z_i = j | x_i; \pi', \mu', \Sigma')} \ln p(x, z; \mu, \Sigma, \pi) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K} p(z_i = j | x_i; \pi', \mu', \Sigma') \ln p(x_i, z_i = j; \mu, \Sigma, \pi) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K} p(z_i = j | x_i; \pi', \mu', \Sigma') \ln (p(x_i | z_i = j; \mu, \Sigma)\, p(z_i = j; \pi)) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij} \ln \frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \cdot \exp(-\frac{1}{2}(x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j)) \cdot \pi_j \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij} \left[- \ln((2\pi)^{d/2} |\Sigma_j|^{1/2}) - \frac{1}{2}(x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j) + \ln \pi_j \right] \qquad (7)
\end{aligned}
$$

First, let's derive the M-step update rule for μ_l, with $l = 1, \ldots, K$.
We have to maximize (7) with respect to μ_l, so let's compute the corresponding derivative:

$$\frac{\partial}{\partial \mu_l} \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij} \left[-\ln((2\pi)^{d/2} |\Sigma_j|^{1/2}) - \frac{1}{2}(x_i - \mu_j)^\top \Sigma_j^{-1}(x_i - \mu_j) + \ln \pi_j \right]$$

$$= -\frac{\partial}{\partial \mu_l} \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij} \frac{1}{2}(x_i - \mu_j)^\top \Sigma_j^{-1}(x_i - \mu_j) = -\frac{1}{2} \sum_{i=1}^{n} w_{ij} \frac{\partial}{\partial \mu_l} \sum_{i=1}^{n} (x_i - \mu_j)^\top \Sigma_j^{-1}(x_i - \mu_j)$$

$$\overset{(5g)}{=} -\frac{1}{2} \sum_{i=1}^{n} w_{ij} \frac{\partial}{\partial \mu_l} \sum_{i=1}^{n} (\Sigma_j^{-1} + (\Sigma_j^{-1})^\top)\,(x_i - \mu_j)$$

$$\overset{(1e)}{=} \frac{1}{2} \sum_{i=1}^{n} w_{ij} \sum_{i=1}^{n} (\Sigma_j^{-1} + (\underbrace{\Sigma_j^\top}_{\Sigma_j})^{-1})\,(x_i - \mu_j) = \frac{1}{2} \sum_{i=1}^{n} w_{ij} \sum_{i=1}^{n} 2\Sigma_j^{-1}\,(x_i - \mu_j)$$

$$= \sum_{i=1}^{n} w_{il} \left(\Sigma_l^{-1} x_i - \Sigma_l^{-1} \mu_l \right) = \sum_{i=1}^{n} w_{il} \Sigma_l^{-1} x_i - \sum_{i=1}^{n} w_{il} \Sigma_l^{-1} \mu_l.$$

Setting this to zero and solving for μ_l therefore yields the update rule

$$\mu_l = \frac{\sum_{i=1}^{n} w_{il} x_i}{\sum_{i=1}^{n} w_{il}}.$$

Secondly, we'll derive the M-step updates to Σ_j, for $j = 1, \ldots, K$.
Grouping together only the terms that depend on Σ_j in (7), we find that we need to maximize

$$\sum_{i=1}^{n}\sum_{j=1}^{K} w_{ij}\left[\ln \frac{1}{|\Sigma_j|^{1/2}} - \frac{1}{2}(x_i-\mu_j)^\top \Sigma_j^{-1}(x_i-\mu_j)\right]$$

$$\overset{(2b)}{=} \sum_{i=1}^{n}\sum_{j=1}^{K} w_{ij}\left[\ln |\Sigma_j^{-1}|^{1/2} - \frac{1}{2}(x_i-\mu_j)^\top \Sigma_j^{-1}(x_i-\mu_j)\right].$$

We use the usual trick of working with the precision matrix $\Lambda_j \overset{not.}{=} \Sigma_j^{-1}$, where Σ_j is assumed invertible.

When maximizing the above quantity with respect to Λ_j by taking derivatives, we find:

$$
\begin{aligned}
&\frac{\partial}{\partial \Lambda_j} \sum_{i=1}^{n} w_{ij} \left[\ln |\Lambda_j|^{1/2} - \frac{1}{2}(x_i - \mu_j)^\top \Lambda_j (x_i - \mu_j) \right] \\
&\quad = \frac{1}{2} \sum_{i=1}^{n} w_{ij} \frac{\partial}{\partial \Lambda_j} \ln |\Lambda_j| - \frac{1}{2} \sum_{i=1}^{n} w_{ij} \frac{\partial}{\partial \Lambda_j} \left[(x_i - \mu_j)^\top \Lambda_j (x_i - \mu_j) \right] \\
&\quad \overset{(4b),(5c)}{=} \frac{1}{2} \sum_{i=1}^{n} w_{ij} \Lambda_j^{-1} - \frac{1}{2} \sum_{i=1}^{n} w_{ij} (x_i - \mu_j)(x_i - \mu_j)^\top \\
&\quad = \frac{1}{2} \Lambda_j^{-1} \sum_{i=1}^{n} w_{ij} - \frac{1}{2} \sum_{i=1}^{n} w_{ij} (x_i - \mu_j)(x_i - \mu_j)^\top .
\end{aligned}
$$

Setting this to zero and solving, we get:

$$
\Sigma_j = \Lambda_j^{-1} = \frac{\sum_{i=1}^{n} w_{ij} (x_i - \mu_j)(x_i - \mu_j)^\top}{\sum_{i=1}^{n} w_{ij}} .
$$

Finally, let's derive the M-step update for the parameters π_j.
Grouping together only the terms that depend on π_j in (7), we find that we need to maximize

$$\sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij} \ln \pi_j .$$

However, there is an additional constraint that the π_j's sum to 1, since they represent the probabilities $\pi_j = p(z_i = j; \pi)$. To deal with the constraint that $\sum_{j=1}^{K} \pi_j = 1$, we construct the Lagrangian

$$\mathcal{L}(\pi) = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij} \ln \pi_j + \beta \left(\sum_{j=1}^{K} \pi_j - 1 \right),$$

where β is the Lagrange multiplier.

***Note*: We don't need to worry about the constraint that $\pi_j \geq 0$, because as we'll shortly see, the solution we'll find from this derivation will automatically satisfy that anyway.**

Taking derivatives of $\mathcal{L}(\pi)$, we find:

$$\frac{\partial}{\partial \pi_j}\mathcal{L}(\pi) = \sum_{i=1}^{n} \frac{w_{ij}}{\pi_j} + \beta = \frac{1}{\pi_j}\sum_{i=1}^{n} w_{ij} + \beta.$$

Setting this to zero and solving, we get $\pi_j = \dfrac{\sum_{i=1}^{n} w_{ij}}{-\beta}$.

By using the constraint $\sum_j \pi_j = 1$**, and given the fact that** $\sum_j w_{ij} = 1$ **since** $w_{ij} \overset{not.}{=} p(z_i = j | x_i; \pi', \mu', \Sigma')$**, we easily find**

$$-\beta = \sum_{i=1}^{n}\sum_{j=1}^{K} w_{ij} = \sum_{i=1}^{n} 1 = n.$$

We therefore have our M-step derivation for the parameters π_j**:**

$$\pi_j = \frac{1}{n}\sum_{i=1}^{n} w_{ij},$$

and, obviously, $\pi_j \geq 0$**.**

Remarks

1. Let's contrast the update rules in the M-step with the formulas we would have when the z_is were known exactly (see the MLE of the parameters of a single multi-variate Gaussean distribution, CMU, 2010 fall, Aarti Singh, HW1, pr. 3.2.1):

$$\begin{aligned}
\pi_j &= \frac{1}{n}\sum_{i=1}^{n} 1\{z_i = j\}, \\
\mu_j &= \frac{\sum_{i=1}^{n} 1\{z_i = j\} x_i}{\sum_{i=1}^{n} 1\{z_i = j\}}, \\
\Sigma_j &= \frac{\sum_{i=1}^{n} 1\{z_i = j\}(x_i - \mu_j)(x_i - \mu_j)^\top}{\sum_{i=1}^{n} 1\{z_i = j\}},
\end{aligned}$$

with $1\{z_i = j\}$ ("indicator functions") indicating from which Gaussian each datapoint had come.

They are identical, except that instead of the indicator functions $1\{z_i = j\}$ indicating from which Gaussian each datapoint had come, we now have the w_{ij}s.

Remarks (cont'd)

2. The EM-algorithm is reminiscent of the K-means clustering algorithm, except that instead of the "hard" cluster assignments $c(i)$, we have the "soft" assignments w_{ij}.

3. Similar to K-means, the EM algorithm is also susceptible to local optima, so reinitializing at several different initial parameters may be a good idea.

4. It's clear that the EM algorithm has a very natural interpretation of repeatedly trying to guess the unknown z_i's.

A link between K-means and EM/GMM (the multi-variate case)

CMU, 2008 fall, Eric Xing, HW4, pr. 2.2

(see also CMU, 2010 fall, Aarti Singh, HW4, pr. 1.2)

Given N data points x_i, $(i = 1, \ldots, N)$, K-means will group them into K clusters by minimizing the *distortion* function

$$J = \sum_{i=1}^{N} \sum_{j=1}^{K} \gamma_{ij} \|x_i - \mu_j\|^2,$$

where μ_j is the centroid of the j-th cluster, and $\gamma_{ij} = 1$ if x_i belongs to the j-th cluster and $\gamma_{ij} = 0$ otherwise.

In this exercise, we will use the following procedure for K-means:

- **Initialize [randomly] the cluster centroids μ_j, $j = 1, \ldots, K$;**
- **Iterate until *convergence*:**
  - **for every data point x_n, update its *cluster assignment*: $\gamma_{ij} = 1$ if $j = \arg\min_k \|x_i - \mu_k\|^2$, and $\gamma_{ij} = 0$ otherwise.**
  - **for each cluster j, update its centroid: $\mu_j = \dfrac{\sum_{i=1}^{N} \gamma_{ij} x_i}{\sum_{i=1}^{N} \gamma_{ij}}$**

Remember that in GMM, $p(x) = \sum_{j=1}^{K} \pi_k \, \mathcal{N}(x|\mu_j, \Sigma_j)$, where π_j is the prior [probability] for the j^{th} component, μ_j and Σ_j are the mean and covariance matrix for the j^{th} component respectively.
In the E-step of the EM algorithm, we will update

$$p(z_{ij} = 1|x_i) = \frac{\pi_j \, \mathcal{N}(x_i|\mu_j, \Sigma_j)}{\sum_{k=1}^{K} \pi_k \, \mathcal{N}(x_i|\mu_k, \Sigma_k)}$$

Now, suppose that

$i.$ $\Sigma_k = \sigma^2 I$, **for some** $\sigma > 0$, **and for all** $k = 1, \ldots, K$
$ii.$ $\pi_k \neq 0$ **for** $k = 1, \ldots, K$ **[LC: at any iteration of the EM algoritm], and**
$iii.$ $\|x_i - \mu_{k'}\| \neq \|x_i - \mu_k\|$ **for any** $k' \neq k$ **[at any iteration of the EM algoritm].**

Under the above assumptions, prove that when $\sigma \to 0_+$ we will get $p(z_{ij} = 1|x_i) \to \gamma_{ij}$, where γ_{ij} is the *cluster assignment* used in K-means.

Answer:

$$\begin{aligned} p(z_{ij}=1|x_i) &= \frac{\pi_j\,\mathcal{N}(x_i|\mu_j,\Sigma_j)}{\sum_{k=1}^K \pi_k\,\mathcal{N}(x_i|\mu_k,\Sigma_k)} = \frac{\pi_j\ \exp\left(-\dfrac{1}{2\sigma^2}\|x_i-\mu_j\|^2\right)}{\sum_{k=1}^K \pi_k \exp\left(-\dfrac{1}{2\sigma^2}\|x_i-\mu_k\|^2\right)} \\ &= \frac{1}{1+\sum_{k\neq j}\dfrac{\pi_k}{\pi_j}\exp\left(\dfrac{1}{2\sigma^2}(\|x_i-\mu_j\|^2-\|x_i-\mu_k\|^2)\right)} \end{aligned}$$

Case 1:
If $\|x_i-\mu_j\| = \min_k \|x_i-\mu_k\|$**, then for each** $k \neq j$ **we have** $\|x_i-\mu_j\|^2 - \|x_i-\mu_k\|^2 < 0$**. Since** $\sigma \to 0_+$**, it will follow that** $\exp\left(\dfrac{1}{2\sigma^2}(\|x_i-\mu_j\|^2-\|x_i-\mu_k\|^2)\right) \to 0$**. So,** $p(z_{ij}=1|x_i)\to 1$**.**

Case 2:
If $\|x_i-\mu_j\| \neq \min_k \|x_i-\mu_k\|$**, then**

- **for all** k **such that** $\|x_i-\mu_j\| < \|x_i-\mu_k\|$ **it will follow (exactly as above) that** $\exp\left(\dfrac{1}{2\sigma^2}(\|x_i-\mu_j\|^2-\|x_i-\mu_k\|^2)\right) \to 0$**;**
- **for all** k **such that** $\|x_i-\mu_j\| > \|x_i-\mu_k\|$ **we will have** $\exp\left(\dfrac{1}{2\sigma^2}(\|x_i-\mu_j\|^2-\|x_i-\mu_k\|^2)\right)$ $\to +\infty$**.**

Therefore, $p(z_{ij}=1|x_i) \to \dfrac{1}{1+\infty} = 0$**.**

Algoritmul EM, fundamentare teoretică: pasul E [şi pasul M]

CMU, 2008 fall, Eric Xing, HW4, pr. 1.1-3

Algoritmul EM (Expectation-Maximization) permite crearea unor modele probabiliste care pe de o parte depind de un set de parametri θ iar pe de altă parte includ pe lângă variabilele obişnuite ("observabile" sau "vizibile") x şi variabile necunoscute ("neobservabile", "ascunse" sau "latente") z.

În general, în astfel de situaţii/modele, nu se poate face în mod direct o estimare a parametrilor modelului (θ), în aşa fel încât să se garanteze atingerea maximului verosimilităţii datelor observabile x.

În schimb, algoritmul EM procedează în manieră iterativă, constituind astfel o modalitate foarte convenabilă de estimare a parametrilor θ.

Definim log-verosimilitatea datelor observabile (x) ca fiind $\log P(x \mid \theta)$, iar log-verosimilitatea datelor *complete* (observabile, x, şi neobservabile, z) ca fiind $\log P(x, z \mid \theta)$.

Observaţie: Pe tot parcursul acestui exerciţiu se va considera funcţia $\log$ ca având baza supraunitară, fixată.

a. Log-verosimilitatea datelor *observabile* (x) se poate exprima în funcţie de datele neobservabile (z), astfel:[a]

$$\ell(\theta) \overset{not.}{=} \log P(x \mid \theta) = \log \left(\sum_z P(x, z \mid \theta) \right)$$

În continuare vom nota cu q o funcţie / distribuţie de probabilitate definită peste variabilele ascunse/neobservabile z.

Folosiţi *inegalitatea lui Jensen* pentru a demonstra că are loc următoarea inegalitate:

$$\log P(x \mid \theta) \geq \sum_z q(z) \log \left(\frac{P(x, z \mid \theta)}{q(z)} \right) \tag{8}$$

pentru orice x (fixat), pentru orice valoare a parametrului θ şi pentru orice distribuţie probabilistă q definită peste variabilele neobservabile z.

[a] Reţineţi că x, vectorul de date observabile, este fixat (dat), în vreme ce z, vectorul de date neobservabile, este liber (variabil).

Observaţie (1)

Semnificaţia inegalităţii (8) este următoarea:

Funcţia (de fapt, orice funcţie de forma)

$$F(q,\theta) \overset{def.}{=} \sum_z q(z) \log \left(\frac{P(x,z \mid \theta)}{q(z)} \right)$$

constituie o margine inferioară pentru funcţia de log-verosimilitate a datelor incomplete / observabile, $\ell(\theta) \overset{not.}{=} \log P(x \mid \theta)$**.**

Remarcaţi faptul că F **este o funcţie de două variabile, iar prima variabilă nu este de tip numeric (cum este** θ**), ci este de tip funcţional.**

Mai mult, se observă că expresia funcţiei F **este de fapt o medie,**

$$E_{q(z)} \left[\log \frac{P(x,z \mid \theta)}{q(z)} \right],$$

atunci când x**,** q **şi parametrul** θ **se consideră fixaţi, iar** z **este lăsat să varieze.**

Soluţie

Inegalitatea lui Jensen, în contextul teoriei probabilităţilor:

considerând X este o variabilă aleatoare (unară),
dacă φ este o funcţie (reală) convexă, atunci $\varphi(E[X]) \leq E[\varphi(X)]$;

dacă φ este funcţie concavă, atunci $\varphi(E[X]) \geq E[\varphi(X)]$.

Aici vom folosi funcţia $\log$ cu bază supraunitară (funcţie concavă), deci aplicând inegalitatea lui Jensen vom obţine: $\log(E[X]) \geq E[\log(X)]$.

Log-verosimilitatea datelor observabile este:

$$\begin{aligned}
\ell(\theta) \overset{not.}{=} \log P(x \mid \theta) &= \log\left(\sum_z P(x, z \mid \theta)\right) = \log\left(\sum_z q(z) \frac{P(x, z \mid \theta)}{q(z)}\right) \\
&\overset{def.}{=} \log\left(E_{q(z)}\left[\frac{P(x, z \mid \theta)}{q(z)}\right]\right)
\end{aligned}$$

Conform inegalităţii lui Jensen (înlocuind X de mai sus cu $\dfrac{P(x, z \mid \theta)}{q(z)}$), rezultă:

$$\ell(\theta) \overset{not.}{=} \log P(x \mid \theta) \geq E_{q(z)}\left[\log \frac{P(x, z \mid \theta)}{q(z)}\right] \overset{def.}{=} \sum_z q(z) \log \frac{P(x, z \mid \theta)}{q(z)},$$

Notaţie

În continuare, pentru a vă aduce mereu aminte că distribuţia q se referă la datele neobservabile z, vom folosi notaţia $q(z)$ în loc de q.

În consecinţă, în cele ce urmează, în funcţie de context, $q(z)$ va desemna fie la distribuţia q, fie la valoarea acestei distribuţii pentru o valoare oarecare [a variabilei neobservabile] z.

(Este adevărat că această lejeră ambiguitate poate induce în eroare cititorul nexperimentat.)

b. Vă reamintim definiţia entropiei relative (numită şi divergenţa Kullback-Leibler):

$$KL(q(z) \,||\, P(z \mid x, \theta)) = -\sum_z q(z) \log \left(\frac{P(z \mid x, \theta)}{q(z)} \right)$$

Arătaţi că

$$\log P(x \mid \theta) = F(q(z), \theta) + KL(q(z) \,||\, P(z \mid x, \theta)).$$

Observaţie (2):

Semnificaţia egalităţii care trebuie demonstrată la acest punct este foarte interesantă: diferenţa dintre funcţia obiectiv $\ell(\theta) \overset{not.}{=} \log P(x \mid \theta)$ şi marginea sa inferioară $F(q(z), \theta)$ — a se vedea punctul a — este $KL(q(z) \,||\, P(z \mid x, \theta))$. Tocmai pe această chestiune se va "construi" punctul final, şi cel mai important, al problemei noastre.

Observaţie (3)

Ideile de bază ale algoritmului EM sunt două:

1. În loc să calculeze maximul funcţiei de log-verosimilitate $\log P(x \mid \theta)$ în raport cu θ, algoritmul EM va maximiza marginea sa inferioară, $F(q(z), \theta)$, în raport cu ambele argumente, $q(z)$ şi θ.

2. Pentru a căuta maximul (de fapt, un maxim local al) marginii inferioare $F(q(z), \theta)$, algoritmul EM aplică metoda *creşterii pe coordonate* (engl., coordinate ascent): după ce iniţial se fixează $\theta^{(0)}$ eventual aleatoriu, se maximizează *iterativ* funcţia $F(q(z), \theta)$, în mod *alternativ*: mai întâi în raport cu distribuţia $q(z)$ şi apoi în raport cu parametrul θ.

$$\textbf{Pasul E:} \qquad q^{(t)}(z) = \underset{q(z)}{\operatorname{argmax}}\, F(q(z), \theta^{(t)})$$

$$\textbf{Pasul M:} \qquad \theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, F(q^{(t)}(z), \theta)$$

Soluţie

$$
\begin{aligned}
F(q(z),\theta) &\overset{def.}{=} \sum_z q(z)\log\left(\frac{P(x,z\mid\theta)}{q(z)}\right)\\
&= \sum_z q(z)\log\left(\frac{P(z\mid x,\theta)\cdot P(x\mid\theta)}{q(z)}\right)\\
&= \sum_z q(z)\left[\log\frac{P(z\mid x,\theta)}{q(z)}+\log P(x\mid\theta)\right]\\
&= \sum_z q(z)\log\left(\frac{P(z\mid x,\theta)}{q(z)}\right)+\sum_z q(z)\log P(x\mid\theta)\\
&= -KL(q(z)\,||\,P(z\mid x,\theta))+\log P(x\mid\theta)\cdot\underbrace{\sum_z q(z)}_{=1}
\end{aligned}
$$

$$
\Rightarrow \qquad \log P(x\mid\theta)=F(q(z),\theta)+KL(q(z)\,||\,P(z\mid x,\theta)).
$$

Observaţie (4)

Conform proprietăţii $KL(p \mid\mid q) \geq 0$ **pentru** $\forall p, q$**, rezultă** $KL(q(z) \mid\mid P(z \mid x, \theta)) \geq 0$**.**

Aşadar, din egalitatea care tocmai a fost demonstrată la punctul b **obţinem (din nou!, după rezultatul de la punctul** a**) că** $F(q(z), \theta)$ **este o margine inferioară pentru log-verosimilitatea datelor observabile,** $\ell(\theta) \overset{not.}{=} \log P(x \mid \theta)$**.**

c. Fie $\theta^{(t)}$ valoarea obţinută pentru parametrul / parametrii θ la iteraţia t a algoritmului EM. Considerând această valoare fixată, arătaţi că maximul lui F în raport cu argumentul / distribuţia $q(z)$ este atins pentru distribuţia $P(z \mid x, \theta^{(t)})$, iar valoarea maximului este:

$$\max_{q(z)} F(q(z), \theta^{(t)}) = E_{P(z|x,\theta^{(t)})}[\log P(x, z \mid \theta^{(t)})] + H(P(z \mid x, \theta^{(t)}))$$

Soluţie

Trebuie să maximizăm $F(q(z), \theta^{(t)})$ — marginea inferioară a log-verosimilităţii datelor observabile x — în raport cu distribuţia $q(z)$.

Pe de o parte, rezultatul de la punctul a ne spune că $F(q(z), \theta) \leq \log P(x \mid \theta)$, pentru orice valoare a lui θ; în particular, pentru $\theta^{(t)}$ avem

$$\log P(x \mid \theta^{(t)}) \geq F(q(z), \theta^{(t)})$$

Pe de altă parte, dacă în egalitatea demonstrată la punctul b se înlocuieşte θ cu $\theta^{(t)}$, rezultă:

$$\log P(x \mid \theta^{(t)}) = F(q(z), \theta^{(t)}) + KL(q(z) || P(z \mid x, \theta^{(t)}))$$

În fine, dacă alegem $q(z) = P(z \mid x, \theta^{(t)})$, atunci termenul $KL(q(z)||P(z \mid x, \theta^{(t)}))$ din dreapta egalităţii de mai sus devine zero (vezi exerciţiul [PS-31] de la capitolul de Probabilităţi şi statistică).

Aşadar, valoarea $\max_{q(z)} F(q(z), \theta^{(t)})$ se obţine pentru distribuţia $q(z) = P(z \mid x, \theta^{(t)})$.

Acum vom calcula această valoare maximă:

$$\begin{aligned}
\max_{q(z)} F(q(z), \theta^{(t)}) &= \log P(x \mid \theta^{(t)}) \overset{def.\ F}{=} \sum_z P(z \mid x, \theta^{(t)}) \log \left(\frac{P(x, z \mid \theta^{(t)})}{P(z \mid x, \theta^{(t)})} \right) \\
&= E_{P(z|x,\theta^{(t)})} \left[\log \frac{P(x, z \mid \theta^{(t)})}{P(z \mid x, \theta^{(t)})} \right] \\
&= E_{P(z|x,\theta^{(t)})} \left[\log P(x, z|\theta^{(t)}) - \log P(z|x, \theta^{(t)}) \right] \\
&= E_{P(z|x,\theta^{(t)})} \left[\log P(x, z|\theta^{(t)}) \right] - E_{P(z|x,\theta^{(t)})} \left[\log P(z|x, \theta^{(t)}) \right] \\
&= E_{P(z|x,\theta^{(t)})} \left[\log P(x, z|\theta^{(t)}) \right] + H[P(z|x, \theta^{(t)})] \\
&= Q(\theta^{(t)} \mid \theta^{(t)}) + H[P(z \mid x, \theta^{(t)})]
\end{aligned}$$

unde $Q(\theta \mid \theta^{(t)}) \overset{not.}{=} E_{P(z|x,\theta^{(t)})}[\log P(x, z \mid \theta)]$**.**

Observaţie (5)

Notând $G_t(\theta) \overset{def.}{=} F(P(z \mid x, \theta^{(t)}), \theta)$**, din calculul de mai sus rezultă că** $G_t(\theta^{(t)}) = \log P(x \mid \theta^{(t)}) = Q(\theta^{(t)} \mid \theta^{(t)}) + H(P(z \mid x, \theta^{(t)}))$**. Se poate demonstra uşor — procedând similar cu calculul de mai sus — egalitatea**

$$G_t(\theta) = Q(\theta \mid \theta^{(t)}) + H[P(z \mid x, \theta^{(t)})]$$

Observând că termenul $H[P(z \mid x, \theta^{(t)})]$ **din această ultimă egalitate nu depinde de** θ**, rezultă imediat că**

$$\operatorname*{argmax}_{\theta} G_t(\theta) = \operatorname*{argmax}_{\theta} Q(\theta \mid \theta^{(t)})$$

În consecinţă,

$$\theta^{(t+1)} \overset{def.}{=} \operatorname*{argmax}_{\theta} F(P(z \mid x, \theta^{(t)}), \theta) = \operatorname*{argmax}_{\theta} G_t(\theta) = \operatorname*{argmax}_{\theta} Q(\theta \mid \theta^{(t)})$$

Egalitatea precedentă este responsabilă pentru următoarea reformulare (cea uzuală!) a algoritmului EM:

Pasul E′: **calculează** $Q(\theta \mid \theta^{(t)}) = E_{P(z|x,\theta^{(t)})}[\log P(x, z \mid \theta)]$

Pasul M′: **calculează** $\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta \mid \theta^{(t)})$

Algoritmul EM:
corectitudine / convergenţă

prelucrare de Liviu Ciortuz, după

en.wikipedia.org/wiki/Expectation-maximization

Pentru a maximiza funcţia de log-verosimilitate a datelor observabile, i.e. $\ell(\theta) \overset{def.}{=} \log P(x \mid \theta)$, unde baza logaritmului (nespecificată) este considerată supraunitară, algoritmul EM procedează în mod iterativ, optimizând la pasul M al fiecărei iteraţii (t) o funcţie "auxiliară"

$$Q(\theta \mid \theta^{(t)}) \overset{def.}{=} E_{P(z|x,\theta^{(t)})}[\log P(x, z \mid \theta)],$$

reprezentând media log-verosimilităţii datelor complete (observabile şi neobservabile) în raport cu distribuţia condiţională $P(z \mid x, \theta^{(t)})$.

Vom considera iteraţiile $t = 0, 1, \ldots$ şi $\theta^{(t+1)} = \operatorname{argmax}_\theta Q(\theta \mid \theta^{(t)})$, cu $\theta^{(0)}$ ales în mod arbitrar.

Demonstraţi că pentru orice t fixat (arbitrar) şi pentru orice θ astfel încât $Q(\theta \mid \theta^{(t)}) \geq Q(\theta^{(t)} \mid \theta^{(t)})$ are loc inegalitatea:

$$\log P(x \mid \theta) - \log P(x \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)}) \qquad (9)$$

Observaţii

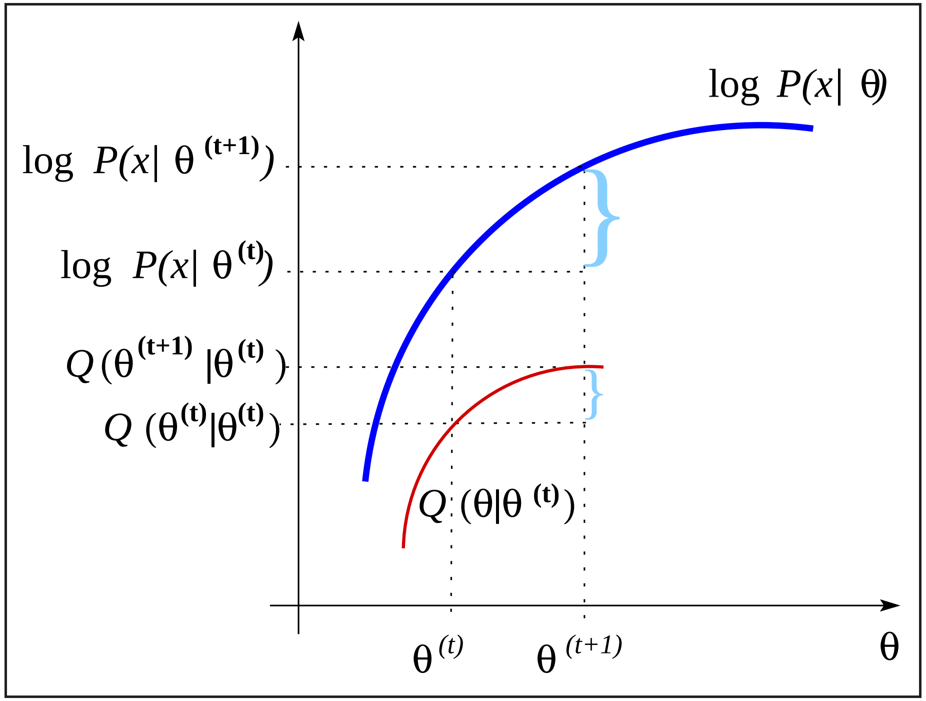
1. Semnificaţia imediată a relaţiei (9):
Orice îmbunătăţire a valorii funcţiei auxiliare $Q(\theta \mid \theta^{(t)})$ conduce la o îmbunătăţire cel puţin la fel de mare a valorii funcţiei obiectiv, $\ell(\theta)$.

2. Dacă în inegalitatea (9) se înlocuieşte θ cu $\theta^{(t+1)} = \operatorname{argmax}_\theta Q(\theta \mid \theta^{(t)})$, va rezulta

$$\log P(x \mid \theta^{(t+1)}) \geq \log P(x \mid \theta^{(t)}).$$

În final, vom avea
$\ell(\theta^{(0)}) \leq \ell(\theta^{(t)}) \leq \ell(\theta^{(t+1)}) \leq \ldots$.

Şirul acesta (monoton) este mărginit superior de 0 (vezi definiţia lui ℓ), deci converge la o anumită valoare ℓ^*. În anumite cazuri / condiţii, această valoare este un maxim (în general, local) al funcţiei de log-verosimilitate.



Observaţii (cont.)

3. Conform aceleiaşi inegalităţi (9), la pasul M de la iteraţia t a algoritmului EM, este suficient ca în loc să se ia $\theta^{(t+1)} = \operatorname{argmax}_\theta Q(\theta \mid \theta^{(t)})$, să se aleagă $\theta^{(t+1)}$ astfel încât $Q(\theta^{(t+1)} \mid \theta^{(t)}) > Q(\theta^{(t)} \mid \theta^{(t)})$. Aceasta constituie *versiunea "generalizată"* a algoritmului EM.

Demonstraţia relaţiei (9)

$$P(x, z \mid \theta) = P(z \mid x, \theta) \cdot P(x \mid \theta) \Rightarrow \log P(x \mid \theta) = \log P(x, z \mid \theta) - \log P(z \mid x, \theta) \Rightarrow$$

$$\underbrace{\sum_z P(z \mid x, \theta^{(t)})}_{1} \cdot \log P(x \mid \theta) =$$

$$\sum_z P(z \mid x, \theta^{(t)}) \cdot \log P(x, z \mid \theta) - \sum_z P(z \mid x, \theta^{(t)}) \cdot \log P(z \mid x, \theta) \Rightarrow$$

$$\log P(x \mid \theta) = Q(\theta \mid \theta^{(t)}) - \sum_z P(z \mid x, \theta^{(t)}) \cdot \log P(z \mid x, \theta)$$

Ultimul termen din egalitatea aceasta reprezintă o cross-entropie, pe care o vom nota cu $CH(\theta \mid \theta^{(t)})$. Aşadar,

$$\log P(x \mid \theta) = Q(\theta \mid \theta^{(t)}) + CH(\theta \mid \theta^{(t)})$$

Această egalitate este valabilă pentru toate valorile posibile ale parametrului θ. În particular pentru $\theta = \theta^{(t)}$, vom avea:

$$\log P(x \mid \theta^{(t)}) = Q(\theta^{(t)} \mid \theta^{(t)}) + CH(\theta^{(t)} \mid \theta^{(t)})$$

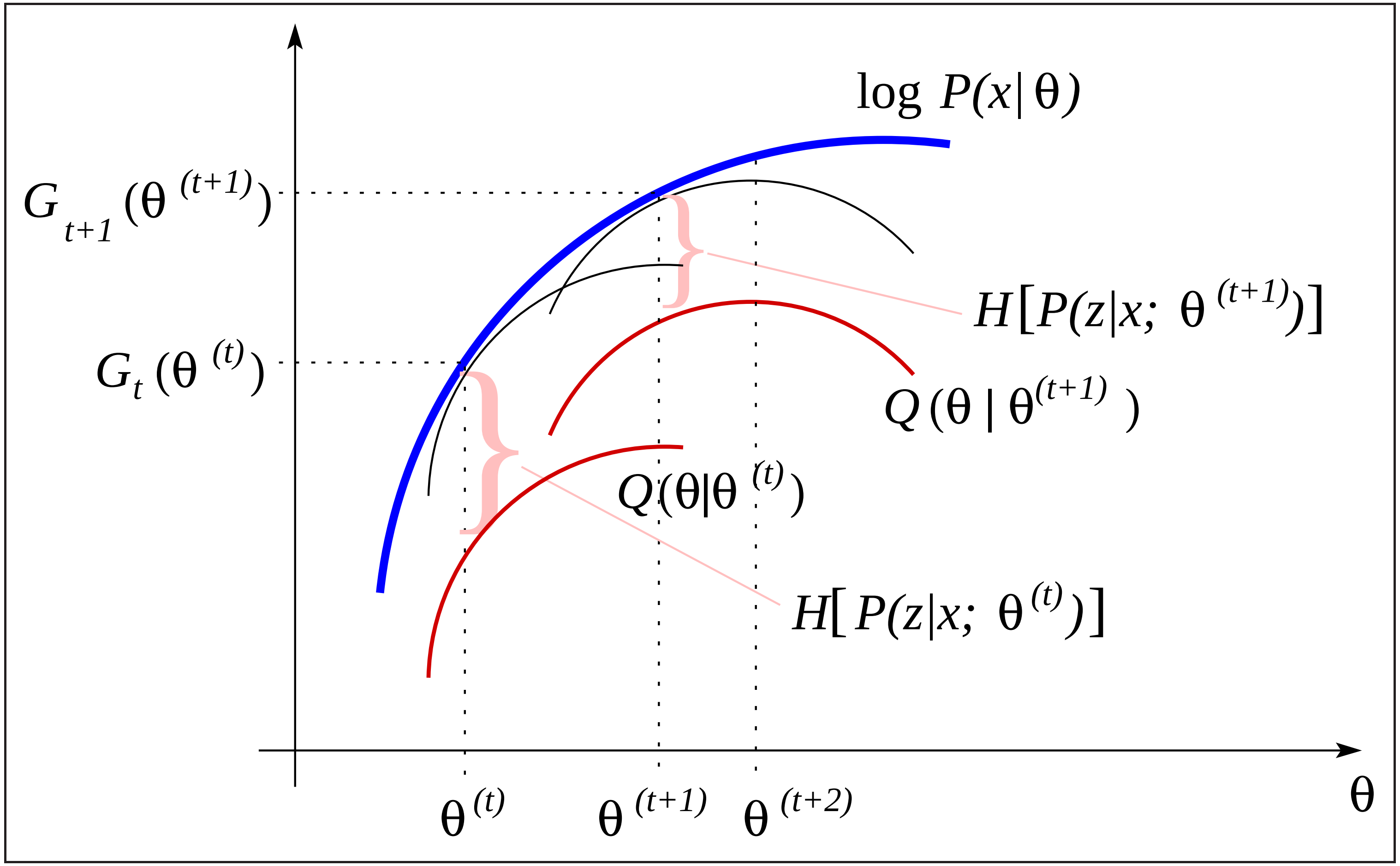
Demonstraţia relaţiei (9), cont.

Scăzând membru cu membru ultimele două egalităţi, obţinem:

$$\log P(x \mid \theta) - \log P(x \mid \theta^{(t)}) = Q(\theta \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)}) + CH(\theta \mid \theta^{(t)}) - CH(\theta^{(t)} \mid \theta^{(t)})$$

Conform inegalităţii lui Gibbs, avem $CH(\theta \mid \theta^{(t)}) \geq CH(\theta^{(t)} \mid \theta^{(t)})$**, deci în final rezultă:**

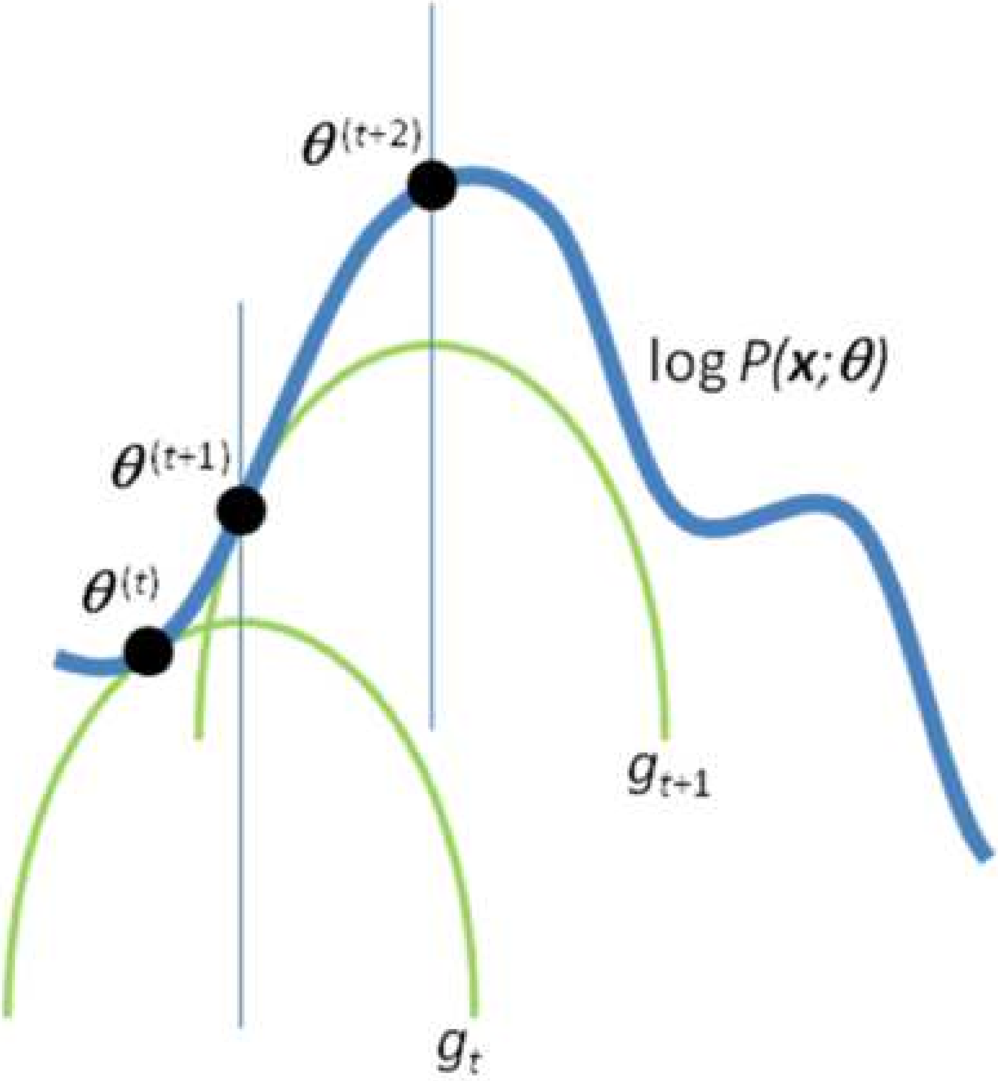
$$\log P(x \mid \theta) - \log P(x \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)})$$


$\log P(x|\theta)$
$G_{t+1}(\theta^{(t+1)})$
$H[P(z|x;\theta^{(t+1)})]$
$G_t(\theta^{(t)})$
$Q(\theta\,|\,\theta^{(t+1)})$
$Q(\theta|\theta^{(t)})$
$H[P(z|x;\theta^{(t)})]$
$\theta^{(t)}$
$\theta^{(t+1)}$
$\theta^{(t+2)}$
θ

From:

What is the expectation maximization algorithm?

Chuong B. Do, Serafim Batzoglou,
Nature Biotechnology,
vol. 26, no. 8, 2008, pp. 897-899



The EM algorithm for solving a Bernoulli mixture model

CMU, 2008 fall, Eric Xing, HW4, pr. 1.4-7

Suppose I have two unfair coins. The first lands on heads with probability p, and the second lands on heads with probability q.

Imagine n tosses, where for each toss I choose to use the first coin with probability π and choose to use the second with probability $1 - \pi$. The outcome of each toss i is $x_i \in \{0, 1\}$.

Suppose I tell you the outcomes of the n tosses, $x \overset{not.}{=} \{x_1, x_2, \ldots, x_n\}$, but I don't tell you which coins I used on which toss.

Given only the outcomes, x, your job is to compute estimates for θ which is the set of all parameters, $\theta = \{p, q, \pi\}$ using the EM algorithm.

To compute these estimates, we will create a latent variable Z, where $z_i \in \{0, 1\}$ indicates the coin used for the n^{th} toss. For example $z_2 = 1$ indicates the first coin was used on the second toss.

We define the "incomplete" data log-likelihood as $\log P(x|\theta)$ and the "complete" data log-likelihood as log $P(x, z|\theta)$.

a. Show that $E[z_i|x_i,\theta] = P(z_i = 1|x_i,\theta)$**.**

b. Use Bayes rule to compute $P(z_i = 1|x_i,\theta^{(t)})$**, where** $\theta^{(t)}$ **denotes the parameters at iteration** t**.**

c. Write down the complete log-likelihood, $\log P(x,z|\theta)$**.**

d. ***E-Step*****: Show that the expected log-likelihood of the complete data** $Q(\theta \mid \theta^{(t)}) \overset{not.}{=} E_{P(z|x,\theta^{(t)})}[\log P(x,z \mid \theta)]$ **is given by**

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= \sum_{i=1}^{n} E[z_i \mid x_i,\theta^{(t)}] \cdot (\log \pi + x_i \log p + (1-x_i)\log(1-p)) + \\ &\quad + \left(1 - E[z_i \mid x_i,\theta^{(t)}]\right) \cdot (\log(1-\pi) + x_i \log q + (1-x_i)\log(1-q)) \end{aligned}$$

e. ***M-Step*****: Describe the process you would use to obtain the update equations for** $p^{(t+1)}$**,** $q^{(t+1)}$ **şi** $\pi^{(t+1)}$**.**

Solution

a.

$$E[z_i \mid x_i, \theta] = \sum_{z \in \{0,1\}} z_i P(z_i \mid x_i, \theta) = 0 \cdot P(z_i = 0 \mid x_i, \theta) + 1 \cdot P(z_i = 1 \mid x_i, \theta)$$

$$\Rightarrow E[z_i \mid x_i, \theta] = P(z_i = 1 \mid x_i, \theta)$$

b.

$$\begin{aligned} & P(z_i = 1 \mid x_i, \theta) \\ = & \frac{P(x_i \mid z_i = 1, \theta) P(z_i = 1 \mid \theta)}{P(x_i \mid z_i = 1, \theta) P(z_i = 1 \mid \theta) + P(x_i \mid z_i = 0, \theta) P(z_i = 0 \mid \theta)} \\ = & \frac{p^{x_i} \cdot (1-p)^{1-x_i} \cdot \pi}{p^{x_i} \cdot (1-p)^{1-x_i} \cdot \pi + q^{x_i} \cdot (1-q)^{1-x_i} \cdot (1-\pi)} \end{aligned}$$

c.

$$
\begin{aligned}
\log P(x, z \mid \theta) &\overset{i.i.d.}{=} \log \prod_{i=1}^{n} P(x_i, z_i \mid \theta) = \log \prod_{i=1}^{n} P(x_i \mid z_i, \theta) \cdot P(z_i \mid \theta) \\
&= \log \prod_{i=1}^{n} \left(p^{x_i}(1-p)^{1-x_i}\pi\right)^{z_i} \left(q^{x_i}(1-q)^{1-x_i}(1-\pi)\right)^{1-z_i} \\
&= \sum_{i=1}^{n} \log \left(\left(p^{x_i}(1-p)^{1-x_i}\pi\right)^{z_i} \left(q^{x_i}(1-q)^{1-x_i}(1-\pi)\right)^{1-z_i} \right) \\
&= \sum_{i=1}^{n} \left[z_i \log \left(p^{x_i}(1-p)^{1-x_i}\pi\right) + (1-z_i) \log \left(q^{x_i}(1-q)^{1-x_i}(1-\pi)\right) \right]
\end{aligned}
$$

d.

$$
\begin{aligned}
Q(\theta \mid \theta^{(t)}) &\overset{not.}{=} E_{P(z|x,\theta^{(t)})}[\log P(x, z \mid \theta)] \\
&= E_{P(z|x,\theta^{(t)})}[\sum_{i=1}^{n} \left[z_i \log \left(p^{x_i}(1-p)^{1-x_i}\pi\right) + \right. \\
&\qquad\qquad\qquad \left. (1-z_i) \log \left(q^{x_i}(1-q)^{1-x_i}(1-\pi)\right)\right]] \\
&= \sum_{i=1}^{n} \left[E[z_i \mid x_i, \theta^{(t)}] \cdot \log p^{x_i}(1-p)^{1-x_i}\pi + \right. \\
&\qquad \left. + \left(1 - E[z_i \mid x_i, \theta^{(t)}]\right) \cdot q^{x_i}(1-q)^{1-x_i}(1-\pi)\right] \\
&= \sum_{i=1}^{n} \left[E[z_i \mid x_i, \theta^{(t)}] \cdot (\log \pi + x_i \log p + (1-x_i)\log(1-p)) + \right. \\
&\qquad \left. + \left(1 - E[z_i \mid x_i, \theta^{(t)}]\right) \cdot (\log(1-\pi) + x_i \log q + (1-x_i)\log(1-q))\right]
\end{aligned}
$$

e.

$$\mu_i^{(t)} \overset{not.}{=} E[z_i \mid x_i, \theta^{(t)}] = \frac{(p^{(t)})^{x_i} \cdot (1-p^{(t)})^{1-x_i} \cdot \pi^{(t)}}{(p^{(t)})^{x_i} \cdot (1-p^{(t)})^{1-x_i} \cdot \pi^{(t)} + (q^{(t)})^{x_i} \cdot (1-q^{(t)})^{1-x_i} \cdot (1-\pi^{(t)})}$$

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) \;&=\; \sum_{i=1}^{n} \Big[\mu_i^{(t)} \left(\log \pi + x_i \log p + (1-x_i)\log(1-p)\right) + \\ &\quad + \left(1-\mu_i^{(t)}\right) \cdot \left(\log(1-\pi) + x_i \log q + (1-x_i)\log(1-q)\right)\Big] \end{aligned}$$

$$\frac{\partial Q(\theta \mid \theta^{(t)})}{\partial p} = 0 \Leftrightarrow \sum_{i=1}^{n} \mu_i^{(t)} \left(\frac{x_i}{p} - \frac{1-x_i}{1-p}\right) = 0 \Leftrightarrow \frac{1}{p}\sum_{i=1}^{n} \mu_i^{(t)} x_i = \frac{1}{1-p}\sum_{i=1}^{n} \mu_i^{(t)}(1-x_i)$$

$$\Leftrightarrow \quad (1-p)\sum_{i=1}^{n} \mu_i^{(t)} x_i = p\sum_{i=1}^{n} \mu_i^{(t)}(1-x_i) \Leftrightarrow \sum_{i=1}^{n} \mu_i^{(t)} x_i = p\left(\sum_{i=1}^{n} \mu_i^{(t)}(1-x_i) + \sum_{i=1}^{n} \mu_i^{(t)} x_i\right)$$

$$\Leftrightarrow \quad \sum_{i=1}^{n} \mu_i^{(t)} x_i = p\sum_{i=1}^{n} \mu_i^{(t)} \qquad \Rightarrow \qquad p^{(t+1)} = \frac{\sum_{i=1}^{n} \mu_i^{(t)} x_i}{\sum_{i=1}^{n} \mu_i^{(t)}}$$

$$\frac{\partial Q(\theta \mid \theta^{(t)})}{\partial q} = 0 \Leftrightarrow \sum_{i=1}^{n} (1-\mu_i^{(t)}) \left(\frac{x_i}{q} - \frac{1-x_i}{1-q} \right) = 0 \quad \Rightarrow \quad q^{(t+1)} = \frac{\sum_{i=1}^{n} (1-\mu_i^{(t)}) x_i}{\sum_{i=1}^{n} (1-\mu_i^{(t)})}$$

$$\begin{aligned}
\frac{\partial Q(\theta \mid \theta^{(t)})}{\partial \pi} = 0 \Leftrightarrow \sum_{i=1}^{n} \left(\frac{\mu_i^{(t)}}{\pi} - \frac{1-\mu_i^{(t)}}{1-\pi} \right) = 0 \Leftrightarrow \frac{1}{\pi} \sum_{i=1}^{n} \mu_i^{(t)} = \frac{1}{1-\pi} \sum_{i=1}^{n} (1-\mu_i^{(t)}) \\
\Leftrightarrow \quad (1-\pi) \sum_{i=1}^{n} \mu_i^{(t)} = \pi \sum_{i=1}^{n} (1-\mu_i^{(t)}) \Leftrightarrow \sum_{i=1}^{n} \mu_i^{(t)} = \pi \left(\sum_{i=1}^{n} (1-\mu_i^{(t)}) + \sum_{i=1}^{n} \mu_i^{(t)} \right) \\
\Leftrightarrow \quad \sum_{i=1}^{n} \mu_i^{(t)} = \pi \sum_{i=1}^{n} 1 \Leftrightarrow \sum_{i=1}^{n} \mu_i^{(t)} = n\pi \qquad \Rightarrow \qquad \pi^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \mu_i^{(t)}
\end{aligned}$$

Estimarea parametrilor unei mixturi de distribuţii binomiale

A. când toate variabilele sunt observabile: MLE;
B. când unele variabile sunt neobservabile: algoritmul EM

prelucrare de Liviu Ciortuz, după

***"What is the expectation maximization algorithm?"*,**
Chuong B. Do, Serafim Batzoglou,
Nature Biotechnology, vol. 26, no. 8, 2008, pag. 897-899

Fie următorul *experiment probabilist*:

Dispunem de două monede, A şi B.
Efectuăm 5 serii de operaţiuni de tipul următor:

- **Alegem în mod aleatoriu una dintre monedele A şi B, cu probabilitate egală (1/2);**
- **Aruncăm de 10 ori moneda care tocmai a fost aleasă (Z) şi notăm rezultatul (X), ca număr de feţe 'head' (ro. 'stemă') obţinute în urma aruncării.**

A. La acest punct vom considera că s-a obţinut următorul rezultat pentru experimentul nostru:

i	Z_i	X_i
1	B	$5H \quad (5T)$
2	A	$9H \quad (1T)$
3	A	$8H \quad (2T)$
4	B	$4H \quad (6T)$
5	A	$7H \quad (3T)$

Semnificaţia variabilelor aleatoare Z_i şi X_i pentru $i = 1, \ldots, 5$ din tabelul de mai sus este imediată.

i. **Calculaţi $\hat{\theta}_A$ şi $\hat{\theta}_B$, probabilităţile de apariţie a feţei 'head' pentru cele două monede, folosind *definiţia clasică a probabilităţilor*, şi anume raportul dintre numărul de cazuri favorabile şi numărul de cazuri posibile, relativ la întregul experiment.**

Răspuns:

Analizând datele din tabelul din enunţ, rezultă imediat

$$\hat{\theta}_A = \frac{24}{24+6} = 0.8 \text{ şi } \hat{\theta}_B = \frac{9}{9+11} = 0.45.$$

De observat că termenii 6 **şi respectiv** 11 **de la numitorii acestor fracţii reprezintă numărul de feţe 'tail' care au fost obţinute la aruncarea monedei** A **şi respectiv** B**:** $6T = 1T + 2T + 3T$, $11T = 5T + 6T$.

Observaţie

Dacă în locul variabilelor binare $Z_i \in \{A, B\}$ pentru $i = 1, \ldots, 5$ introducem în mod natural variabilele-indicator $Z_{i,A} \in \{0, 1\}$ şi $Z_{i,B} \in \{0, 1\}$ tot pentru $i = 1, \ldots, 5$, definite prin $Z_{i,A} = 1$ iff $Z_i = A$, şi $Z_{i,A} = 0$ iff $Z_{i,B} = 0$, atunci procesările necesare pentru calculul probabilităţilor/parametrilor $\hat{\theta}_A$ şi $\hat{\theta}_B$ pot fi prezentate în mod sintetizat ca în tabelul de mai jos.[a]

i	$Z_{i,A}$	$Z_{i,B}$	X_i
1	0	1	$5H$
2	1	0	$9H$
3	1	0	$8H$
4	0	1	$4H$
5	1	0	$7H$

$\Rightarrow$

[a] Prezentăm acest “artificiu” ca pregătire pentru rezolvarea (ulterioară a) punctului B al prezentei probleme.

$\Rightarrow$

$X_i \cdot Z_{i,A}$	$X_i \cdot Z_{i,B}$
$0H \quad (0T)$	$5H \quad (5T)$
$9H \quad (1T)$	$0H \quad (0T)$
$8H \quad (2T)$	$0H \quad (0T)$
$0H \quad (0T)$	$4H \quad (6T)$
$7H \quad (3T)$	$0H \quad (0T)$
$\sum_{i=1}^{5} X_i \cdot Z_{i,A} = 24H$	$\sum_{i=1}^{5} X_i \cdot Z_{i,B} = 9H$
$\sum_{i=1}^{5} (10 - X_i) \cdot Z_{i,A} = 6T$	$\sum_{i=1}^{5} (10 - X_i) \cdot Z_{i,B} = 11T$

$\Rightarrow$

$$\Rightarrow \quad \begin{cases} \hat{\theta}_A = \dfrac{24}{24+6} = 0.8 \\[2ex] \hat{\theta}_B = \dfrac{9}{9+11} = 0.45 \end{cases}$$

ii. **Calculaţi $l_1(\theta_A, \theta_B) \overset{not.}{=} P(X, Z \mid \theta_A, \theta_B)$, funcţia de verosimilitate a datelor "complete" — unde $X \overset{not.}{=} < X_1, \ldots, X_5 >$ sunt datele "observabile", iar $Z \overset{not.}{=} < Z_1, \ldots, Z_5 >$ sunt datele "neobservabile" —, în raport cu parametrii θ_A şi respectiv θ_B ai distribuţiilor binomiale care modelează aruncarea celor două monede.**

Răspuns:

Calculul verosimilităţii datelor complete:

$$
\begin{aligned}
l_1(\theta_A, \theta_B) &\overset{def.}{=} P(X, Z_A, Z_B \mid \theta_A, \theta_B) \overset{indep.\ cdt.}{=} \prod_{i=1}^{5} P(X_i, Z_{i,A}, Z_{i,B} \mid \theta_A, \theta_B) \\
&= \prod_{i=1}^{5} P(X_i \mid Z_{i,A}, Z_{i,B}; \theta_A, \theta_B) \cdot P(Z_{i,A}, Z_{i,B} \mid \theta_A, \theta_B) \\
&= P(X_1 \mid Z_{B,1} = 1, \theta_B) \cdot 1/2 \cdot \\
&\quad\ P(X_2 \mid Z_{A,2} = 1, \theta_A) \cdot 1/2 \cdot \\
&\quad\ P(X_3 \mid Z_{A,3} = 1, \theta_A) \cdot 1/2 \cdot \\
&\quad\ P(X_4 \mid Z_{B,4} = 1, \theta_B) \cdot 1/2 \cdot \\
&\quad\ P(X_5 \mid Z_{A,5} = 1, \theta_A) \cdot 1/2 \\
&= \theta_B^5 (1-\theta_B)^5 \cdot \theta_A^9 (1-\theta_A) \cdot \theta_A^8 (1-\theta_A)^2 \cdot \theta_B^4 (1-\theta_B)^6 \cdot \theta_A^7 (1-\theta_A)^3 \cdot \frac{1}{2^5} \\
&= \frac{1}{2^5}\, \theta_A^{24} (1-\theta_A)^6\, \theta_B^9 (1-\theta_B)^{11}
\end{aligned}
$$

iii. **Calculaţi** $\hat{\theta}_A \stackrel{not.}{=} \arg\max_{\theta_A} \log l_1(\theta_A, \theta_B)$ **şi** $\hat{\theta}_B \stackrel{not.}{=} \arg\max_{\theta_B} \log l_1(\theta_A, \theta_B)$ **folosind derivatele parţiale de ordinul întâi.**

Observaţii:

1. Baza logaritmului, fixată dar lăsată mai sus nespecificată, se va considera supraunitară (de exemplu 2, e **sau 10).**

2. Lucrând corect, veţi obţine acelaşi rezultat ca la punctul *i.*

Răspuns:

Funcţia de log-verosimilitate a datelor complete se exprimă astfel:

$$\log l_1(\theta_A, \theta_B) \;\; = \;\; -5\log 2 + 24\log\theta_A + 6\log(1-\theta_A) + 9\log\theta_B + 11\log(1-\theta_B)$$

Prin urmare, maximul acestei funcţii în raport cu parametrul θ_A se calculează astfel:

$$\frac{\partial \log l_1(\theta_A, \theta_B)}{\partial\theta_A} = 0 \;\; \Leftrightarrow \;\; \frac{24}{\theta_A} - \frac{6}{1-\theta_A} = 0 \Leftrightarrow \frac{4}{\theta_A} = \frac{1}{1-\theta_A} \Leftrightarrow 4 - 4\,\theta_A = \theta_A \Leftrightarrow \hat{\theta}_A = 0.8$$

Similar, se face calculul şi pentru $\dfrac{\partial \log l_1(\theta_A, \theta_B)}{\partial\theta_B}$ şi se obţine $\hat{\theta}_B = 0.45$.[a]

Cele două valori obţinute, $\hat{\theta}_A$ şi $\hat{\theta}_B$, reprezintă estimarea de verosimilitate maximă (MLE) a probabilităţilor de apariţie a feţei 'head' ('stema') pentru moneda A şi respectiv moneda B.

[a] Se verifică uşor faptul că într-adevăr rădăcinile derivatelor parţiale de ordinul întâi pentru funcţia de log-verosimilitate reprezintă puncte de maxim. Pentru aceasta se studiază semnele acestor derivate.

Observaţii

1. Am arătat pe acest caz particular că metoda de calculare a probabilităţilor ($\hat{\theta}_A$ şi $\hat{\theta}_B$) direct din datele observate (aşa cum o ştim din liceu) corespunde de fapt metodei de estimare în sensul verosimităţii maxime (MLE).

2. La punctul B vom arăta cum anume se poate face estimarea aceloraşi parametri θ_A şi θ_B în cazul in care o parte din variabile, şi anume $Z_{i,A}$ şi $Z_{i,B}$, sunt neobservabile.

B. La acest punct se va relua experimentul de la punctul A, însă de data aceasta vom considera că valorile variabilelor Z_i nu sunt cunoscute.

i	Z_i	X_i
1	?	$5H \quad (5T)$
2	?	$9H \quad (1T)$
3	?	$8H \quad (2T)$
4	?	$4H \quad (6T)$
5	?	$7H \quad (3T)$

iv. **Pentru convenienţă, pentru** $i = 1, \ldots, 5$ **vom considera variabilele-indicator “neobservabile”** $Z_{i,A}, Z_{i,B} \in \{0, 1\}$**, cu** $Z_{i,A} = 1$ **iff** $Z_{i,B} = 0$ **şi** $Z_{i,B} = 1$ **iff** $Z_{i,A} = 0$**.**

Folosind teorema lui Bayes, calculaţi mediile variabilelor neobservabile $Z_{i,A}$ **şi** $Z_{i,B}$ **condiţionate de variabilele observabile** X_i**. Veţi considera că parametrii acestor distribuţii binomiale care modelează aruncarea monedelor** A **şi** B **au valorile** $\theta_A^{(0)} = 0.6$ **şi respectiv** $\theta_B^{(0)} = 0.5$**.**

Aşadar, se cer: $E[Z_{i,A} \mid X_i, \theta_A^{(0)}, \theta_B^{(0)}]$ **şi** $E[Z_{i,B} \mid X_i, \theta_A^{(0)}, \theta_B^{(0)}]$ **pentru** $i = 1, \ldots, 5$**. Ca şi mai înainte, probabilităţile a priori** $P(Z_{i,A} = 1)$ **şi** $P(Z_{i,B} = 1)$ **se vor considera 1/2.**

Răspuns:

Algoritmul EM ne permite să facem în mod iterativ estimarea parametrilor θ_A şi θ_B în funcţie de valorile variabilelor observabile, X_i, şi de valorile inţiale atribuite parametrilor (în cazul nostru, $\theta_A^{(0)} = 0.6$ şi $\theta_B^{(0)} = 0.5$).

Vom face o sinteză a calculelor de la prima iteraţie a algoritmului EM — detaliate la punctele *iv*, v şi *vi* de mai jos — sub forma următoare, care seamănă într-o anumită măsură cu tabelele de la punctul A:

i	$Z_{i,A}$	$Z_{i,B}$	X_i
1	–	–	$5H$
2	–	–	$9H$
3	–	–	$8H$
4	–	–	$4H$
5	–	–	$7H$

$\overset{E}{\Rightarrow}$

$E[Z_{i,A}]$	$E[Z_{i,B}]$
$0.45H$	$0.55H$
$0.80H$	$0.20H$
$0.73H$	$0.27H$
$0.35H$	$0.65H$
$0.65H$	$0.35H$

$\overset{M}{\Rightarrow}$

$\overset{M}{\Rightarrow}$

$X_i \cdot E[Z_{i,A}]$	$X_i \cdot E[Z_{i,B}]$
$2.2H \quad (2.2T)$	$2.8H \quad (2.8T)$
$7.2H \quad (0.8T)$	$1.8H \quad (0.2T)$
$5.9H \quad (1.5T)$	$2.1H \quad (0.5T)$
$1.4H \quad (2.1T)$	$2.6H \quad (3.9T)$
$4.5H \quad (1.9T)$	$2.5H \quad (1.1T)$
$\sum_{i=1}^{5} X_i \cdot E[Z_{i,A}] = 21.3H$	$\sum_{i=1}^{5} X_i \cdot E[Z_{i,B}] = 11.7H$
$\sum_{i=1}^{5}(10 - X_i) \cdot E[Z_{i,A}] = 8.3T$	$\sum_{i=1}^{5}(10 - X_i) \cdot E[Z_{i,B}] = 8.3T$

$\Rightarrow$

$$\Rightarrow \quad \left\{ \begin{array}{l} \hat{\theta}_A^{(1)} = \dfrac{21.3}{21.3 + 8.7} \approx 0.71 \\[2ex] \hat{\theta}_B^{(1)} = \dfrac{11.7}{11.7 + 8.3} \approx 0.58 \end{array} \right.$$

Observaţii:

- este posibilă calcularea mediilor variabilelor neobservabile $Z_{i,A}$ şi $Z_{i,B}$, condiţionate de variabilele observabile X_i şi în funcţie de valorile asignate iniţial (0.6 şi 0.5 în enunţ, dar în general ele se pot asigna în mod aleatoriu) pentru parametrii θ_A şi θ_B;
- faţă de tabloul de sinteză de la punctul precedent, când toate variabilele erau observabile şi se calculau produsele $X_i \cdot Z_{i,A}$ şi $X_i \cdot Z_{i,B}$, aici se înlocuiesc variabilele $Z_{i,A}$ şi $Z_{i,B}$ cu mediile $E[Z_{i,A}]$ şi $E[Z_{i,B}]$ în produsele respective. De fapt, în loc să se calculeze $\sum_i X_i \cdot Z_{i,A}$ se calculează media $E[\sum_i X_i \cdot Z_{i,A}]$, şi similar pentru B.

Justificarea concretă pentru cele două observaţii de mai sus va fi dezvoltată la punctele mai jos.

Notaţie: Pentru simplitate, în cele de mai sus (inclusiv în tabelele precedente), prin $E[Z_{i,A}]$ am notat $E[Z_{i,A} \mid X_i, \theta^{(0)}]$, iar prin $E[Z_{i,B}]$ am notat $E[Z_{i,B} \mid X_i, \theta^{(0)}]$, unde $\theta^{(0)} \overset{not.}{=} (\theta_A^{(0)}, \theta_B^{(0)})$.

Întrucât variabilele $Z_{i,A}$ au valori boolene (0 sau 1), rezultă că

$$\begin{array}{rcl} E[Z_{i,A}\,|\,X_i,\theta^{(0)}] & = & 0\cdot P(Z_{i,A}=0\,|\,X_i,\theta^{(0)})+1\cdot P(Z_{i,A}=1\,|\,X_i,\theta^{(0)}) \\ & = & P(Z_{i,A}=1\,|\,X_i,\theta^{(0)}) \end{array}$$

Probabilităţile $P(Z_{i,A}=1\mid X,\theta^{(0)})=P(Z_{i,A}=1\mid X_i,\theta^{(0)})$**, pentru** $i=1,\ldots,5$**, se pot calcula folosind teorema lui Bayes:**

$$\begin{array}{rcl} P(Z_{i,A}=1\mid X_i,\theta^{(0)}) & = & \dfrac{P(X_i\,|\,Z_{i,A}=1,\theta^{(0)})\cdot P(Z_{i,A}=1\,|\,\theta^{(0)})}{\sum_{j\in\{0,1\}}P(X_i\,|\,Z_{i,A}=j,\theta^{(0)})\cdot P(Z_{i,A}=j\,|\,\theta^{(0)})} \\ & = & \dfrac{P(X_i\mid Z_{i,A}=1,\theta_A^{(0)})}{P(X_i\mid Z_{i,A}=1,\theta_A^{(0)})+P(X_i\mid Z_{i,B}=1,\theta_B^{(0)})} \end{array}$$

S-a ţinut cont că $P(Z_{i,A}=1\mid\theta^{(0)})=P(Z_{i,B}=1\mid\theta^{(0)})=1/2$ **(a se vedea enunţul).**

De exemplu, pentru $i = 1$ vom avea:

$$E[Z_{A,1} \mid X_1, \theta^{(0)}] \ = \ \frac{0.6^5(1-0.6)^5}{0.6^5(1-0.6)^5 + 0.5^5(1-0.5)^5} = \frac{1}{1+\left(\frac{0.25}{0.24}\right)^5} \approx 0.45$$

Similar cu $E[Z_{A,1} \mid X_1, \theta^{(0)}]$ se calculează şi celelalte medii $E[Z_{i,A} \mid X_i, \theta^{(0)}]$ pentru $i = 2, \ldots, 5$ şi $E[Z_{i,B} \mid X_i, \theta^{(0)}]$ pentru $i = 1, \ldots, 5$.

***Observaţie*: Se poate ţine cont că, de îndată ce s-a calculat $E[Z_{i,A} \mid X_i, \theta^{(0)}]$, se poate obţine imediat şi $E[Z_{i,B} \mid X_i, \theta^{(0)}] = 1 - E[Z_{i,A} \mid X_i, \theta^{(0)}]$, fiindcă $Z_{i,A} + Z_{i,B} = 1$.**

v. **Calculaţi media funcţiei de log-verosimilitate a datelor complete, X (observabile) şi Z (neobservabile):**

$$l_2(\theta_A, \theta_B) \overset{def.}{=} E_{P(Z|X,\theta^{(0)})}[\log P(X, Z \mid \theta)],$$

unde $\theta \overset{not.}{=} (\theta_A, \theta_B)$ şi $\theta^{(0)} \overset{not.}{=} (\theta_A^{(0)}, \theta_B^{(0)})$.

Semnificaţia **notaţiei de mai sus este următoarea:**

Funcţia $l_2(\theta_A, \theta_B)$ este o medie a variabilei aleatoare reprezentată de log-verosimilitatea datelor complete (observabile şi, respectiv, neobservabile), iar această medie se calculează în raport cu distribuţia probabilistă condiţională a datelor neobservabile, $P(Z \mid X, \theta^{(0)})$.

Observaţie: **La elaborarea calculului, veţi folosi mai întâi proprietatea de liniaritate a mediilor variabilelor aleatoare, şi apoi rezultatele de la punctul *iv.***

Răspuns:

Media funcţiei de log-verosimilitate a datelor complete, $l_2(\theta_A, \theta_B)$, se calculează astfel:

$$\begin{aligned}
l_2(\theta_A, \theta_B) &\overset{def.}{=} E_{P(Z|X,\theta^{(0)})}[\log P(X, Z \mid \theta)] \\
&\overset{indep.\ cdt.}{=} E_{P(Z|X,\theta^{(0)})}[\log \prod_{i=1}^{5} P(X_i, Z_{i,A}, Z_{i,B} \mid \theta_A, \theta_B)] \\
&= E_{P(Z|X,\theta^{(0)})}[\log \prod_{i=1}^{5} P(X_i \mid Z_{i,A}, Z_{i,B}; \theta_A, \theta_B) \cdot P(Z_{i,A}, Z_{i,B} \mid \theta_A, \theta_B)]
\end{aligned}$$

În continuare, omiţând din nou distribuţia probabilistă în raport cu care se calculează media aceasta întrucât ea poate fi subînţeleasă, vom scrie:

$$
\begin{aligned}
& l_2(\theta_A, \theta_B) \\
= \quad & E[\log \prod_{i=1}^{5} \cdot (\theta_A^{Z_{i,A}})^{X_i} \cdot [(1-\theta_A)^{Z_{i,A}}]^{10-X_i} \cdot (\theta_B^{Z_{i,B}})^{X_i} \cdot [(1-\theta_B)^{Z_{i,B}}]^{10-X_i} \cdot \frac{1}{2}\,] \\
= \quad & E[\sum_{i=1}^{5} [X_i \cdot Z_{i,A} \cdot \log\theta_A + (10-X_i) \cdot Z_{i,A} \cdot \log(1-\theta_A) + \\
& \qquad X_i \cdot Z_{i,B} \cdot \log\theta_B + (10-X_i) \cdot Z_{i,B} \cdot \log(1-\theta_B) - \log 2]\,] \\
= \quad & \sum_{i=1}^{5} [X_i \cdot E[Z_{i,A}] \cdot \log\theta_A + (10-X_i) \cdot E[Z_{i,A}] \cdot \log(1-\theta_A) + \\
& \qquad X_i \cdot E[Z_{i,B}] \cdot \log\theta_B + (10-X_i) \cdot E[Z_{i,B}] \cdot \log(1-\theta_B) - \log 2] \\
= \quad & \sum_{i=1}^{5} \log[\theta_A^{X_i \cdot E[Z_{i,A}]} \cdot (1-\theta_A)^{(10-X_i) \cdot E[Z_{i,A}]} \cdot \theta_B^{X_i \cdot E[Z_{i,B}]} \cdot (1-\theta_B)^{(10-X_i) \cdot E[Z_{i,B}]} \cdot \frac{1}{2}] \\
= \quad & \log(\theta_A^{2.2}(1-\theta_A)^{2.2}\theta_B^{2.8}(1-\theta_B)^{2.8} \cdot \ldots \cdot \theta_A^{4.5}(1-\theta_A)^{1.9}\theta_B^{2.5}(1-\theta_B)^{1.1} \cdot \frac{1}{2}).
\end{aligned}
$$

La ultima egalitate de mai sus, cantităţile fracţionare provin din calculele simple $X_1 \cdot E[Z_{1,A} \mid X_1, \theta] \approx 2.2$**,** $X_1 \cdot E[Z_{1,B} \mid X_1, \theta] \approx 2.8$**,** $\ldots$**,** $X_5 \cdot E[Z_{1,A} \mid X_5, \theta] \approx 4.5$**,** $X_5 \cdot E[Z_{1,B} \mid X_5, \theta] \approx 2.5$ **(a se vedea tabelele precedente).**

vi. **Calculaţi $\theta_A^{(1)} \overset{not.}{=} \arg\max_{\theta_A} l_2(\theta_A, \theta_B)$ şi $\theta_B^{(1)} \overset{not.}{=} \arg\max_{\theta_B} l_2(\theta_A, \theta_B)$.**

Răspuns:

Valorile parametrilor θ_A şi θ_B pentru care se atinge maximul mediei funcţiei de log-verosimilitate a datelor complete se obţin cu ajutorul derivatelor parţiale de ordinul întâi:

$$
\begin{aligned}
& \frac{\partial l_2(\theta_A, \theta_B)}{\partial \theta_A} = 0 \\
\Rightarrow\ & \frac{\partial}{\partial \theta_A}(2.2 \log \theta_A + 2.2 \log(1-\theta_A) + \ldots + 4.5 \log \theta_A + 1.9 \log(1-\theta_A)) = 0 \\
\Rightarrow\ & \frac{2.2}{\theta_A} - \frac{2.2}{1-\theta_A} + \ldots + \frac{4.5}{\theta_A} - \frac{1.9}{1-\theta_A} = 0 \Rightarrow \ldots \Rightarrow \theta_A^{(1)} \approx 0.71.
\end{aligned}
$$

Similar, vom obţine $\theta_B^{(1)} \approx 0.58$.

C. Formalizaţi paşii E şi M ai algoritmului EM pentru estimarea parametrilor θ_A şi θ_B în condiţiile de la punctul B.

Răspuns:

Formulele care se folosesc în cadrul algoritmului EM pentru rezolvarea problemei date — i.e. estimarea parametrilor θ_A şi θ_B atunci când variabilele Z_i sunt neobservabile —, se elaborează/deduc astfel:

Pasul E:

$$E[Z_{i,A} \mid X, \theta] = P(Z_{i,A} = 1 \mid X, \theta) = P(Z_{i,A} = 1 \mid X_i, \theta)$$

$$\overset{T.\ B.}{=} \frac{P(X_i \mid Z_{i,A} = 1; \theta) \cdot \overbrace{P(Z_{i,A} = 1 \mid \theta)}^{1/2}}{P(X_i \mid Z_{i,A} = 1; \theta) \cdot P(Z_{i,A} = 1 \mid \theta) + P(X_i \mid Z_{i,B} = 1; \theta) \cdot \underbrace{P(Z_{i,B} = 1 \mid \theta)}_{1/2}}$$

$$= \frac{P(X_i \mid Z_{i,A} = 1; \theta)}{P(X_i \mid Z_{i,A} = 1; \theta) + P(X_i \mid Z_{i,B} = 1; \theta)}$$

$$= \frac{\theta_A^{X_i}(1 - \theta_A)^{10 - X_i}}{\theta_A^{X_i}(1 - \theta_A)^{10 - X_i} + \theta_B^{X_i}(1 - \theta_B)^{10 - X_i}}$$

Similar, vom obţine:

$$E[Z_{i,B} \mid X, \theta] = \frac{\theta_B^{X_i}(1 - \theta_B)^{10 - X_i}}{\theta_A^{X_i}(1 - \theta_A)^{10 - X_i} + \theta_B^{X_i}(1 - \theta_B)^{10 - X_i}}$$

Notând cu

- x_i **valoarea variabilei** X_i,
- $\theta_A^{(t)}$ **şi respectiv** $\theta_B^{(t)}$ **estimările parametrilor** θ_A **şi** θ_B **la iteraţia** t **a algoritmului EM,**
- $\mu_{i,A}^{(t+1)}$ **şi respectiv** $\mu_{i,B}^{(t+1)}$**, mediile** $E[Z_{i,A} \mid X_i, \theta_A^{(t)}]$ **şi** $E[Z_{i,B} \mid X_i, \theta_B^{(t)}]$**,**

vom avea:

$$
\begin{aligned}
\mu_{i,A}^{(t+1)} &= \frac{(\theta_A^{(t)})^{x_i}(10-\theta_A^{(t)})^{10-x_i}}{(\theta_A^{(t)})^{x_i}(10-\theta_A^{(t)})^{10-x_i}+(\theta_B^{(t)})^{x_i}(10-\theta_B^{(t)})^{10-x_i}}\\
\mu_{i,B}^{(t+1)} &= \frac{(\theta_B^{(t)})^{x_i}(10-\theta_B^{(t)})^{10-x_i}}{(\theta_A^{(t)})^{x_i}(10-\theta_A^{(t)})^{10-x_i}+(\theta_B^{(t)})^{x_i}(10-\theta_B^{(t)})^{10-x_i}}
\end{aligned}
$$

Pasul M:

Ca şi mai înainte, în formulele de mai jos vom folosi notaţiile simplificate $E[Z_{i,A}] \overset{not.}{=} E[Z_{i,A} \mid X_i, \theta^{(t)}]$ **şi** $E[Z_{i,B}] \overset{not.}{=} E[Z_{i,B} \mid X_i, \theta^{(t)}]$**.**

Cu aceste notaţii, procedând similar cu calculul de la partea B, punctul v**, vom avea:**

$$l_2(\theta_A, \theta_B) = \log \prod_{i=1}^{5} \theta_A^{x_i E[Z_{i,A}]} (1-\theta_A)^{(10-x_i)E[Z_{i,A}]} \; \theta_B^{x_i E[Z_{i,B}]} (1-\theta_B)^{(10-x_i)E[Z_{i,B}]}$$

Prin urmare,

$$\frac{\partial}{\partial \theta_A} l_2(\theta_A, \theta_B) = 0 \Rightarrow \frac{1}{\theta_A} \sum_{i=1}^{5} x_i E[Z_{i,A}] = \frac{1}{1-\theta_A} \sum_{i=1}^{5} (10 - x_i) E[Z_{i,A}]$$

$$\Rightarrow \quad (1-\theta_A) \sum_{i=1}^{5} x_i E[Z_{i,A}] = \theta_A \sum_{i=1}^{5} (10 - x_i) E[Z_{i,A}]$$

$$\Rightarrow \quad \sum_{i=1}^{5} x_i E[Z_{i,A}] = 10\, \theta_A \sum_{i=1}^{5} E[Z_{i,A}]$$

$$\Rightarrow \quad \theta_A = \frac{\sum_{i=1}^{5} x_i \; E[Z_{i,A}]}{10 \sum_{i=1}^{5} E[Z_{i,A}]} \quad \textbf{şi, similar,} \;\; \theta_B = \frac{\sum_{i=1}^{5} x_i \; E[Z_{i,B}]}{10 \sum_{i=1}^{5} E[Z_{i,B}]}$$

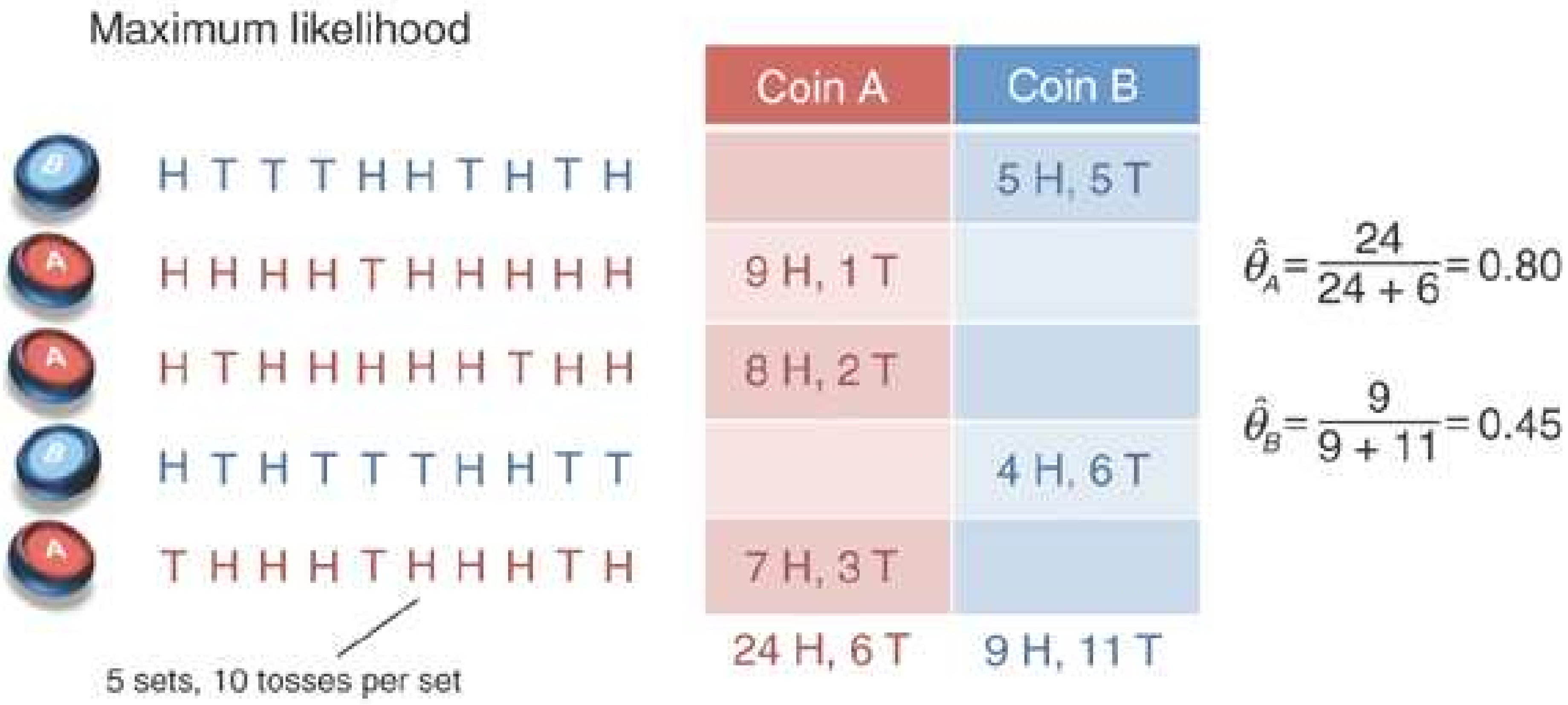
Aşadar, la pasul M al algoritmului EM vom avea:

$$\theta_A^{(t+1)} = \frac{\sum_{i=1}^{5} x_i \; \mu_{i,A}^{(t+1)}}{10 \sum_{i=1}^{5} \mu_{i,A}^{(t+1)}} \quad \textbf{şi} \quad \theta_B^{(t+1)} = \frac{\sum_{i=1}^{5} x_i \; \mu_{i,B}^{(t+1)}}{10 \sum_{i=1}^{5} \mu_{i,B}^{(t+1)}}$$

Observaţie

Implementând algoritmul EM cu relaţiile obţinute pentru pasul E şi pasul M, după execuţia a 10 iteraţii se vor obţine valorile $\theta_A^{(10)} \approx 0.80$ şi $\theta_B^{(10)} \approx 0.52$.

Este interesant de observat că estimarea obţinută pentru parametrul θ_A este acum la acelaşi nivel cu cea obţinută prin metoda verosimilităţii maxime (MLE, vezi punctul A) în cazul observării tuturor variabilelor (0.80, vezi rezolvarea de la partea A, punctul i), iar estimarea obţinută pentru parametrul θ_B a coborât de la valorea 0.58 care a fost obţinută la prima iteraţie a algoritmului EM la o valoare (0.52) care este considerabil mai apropiată de estimarea prin metoda MLE (0.45).



Maximum likelihood

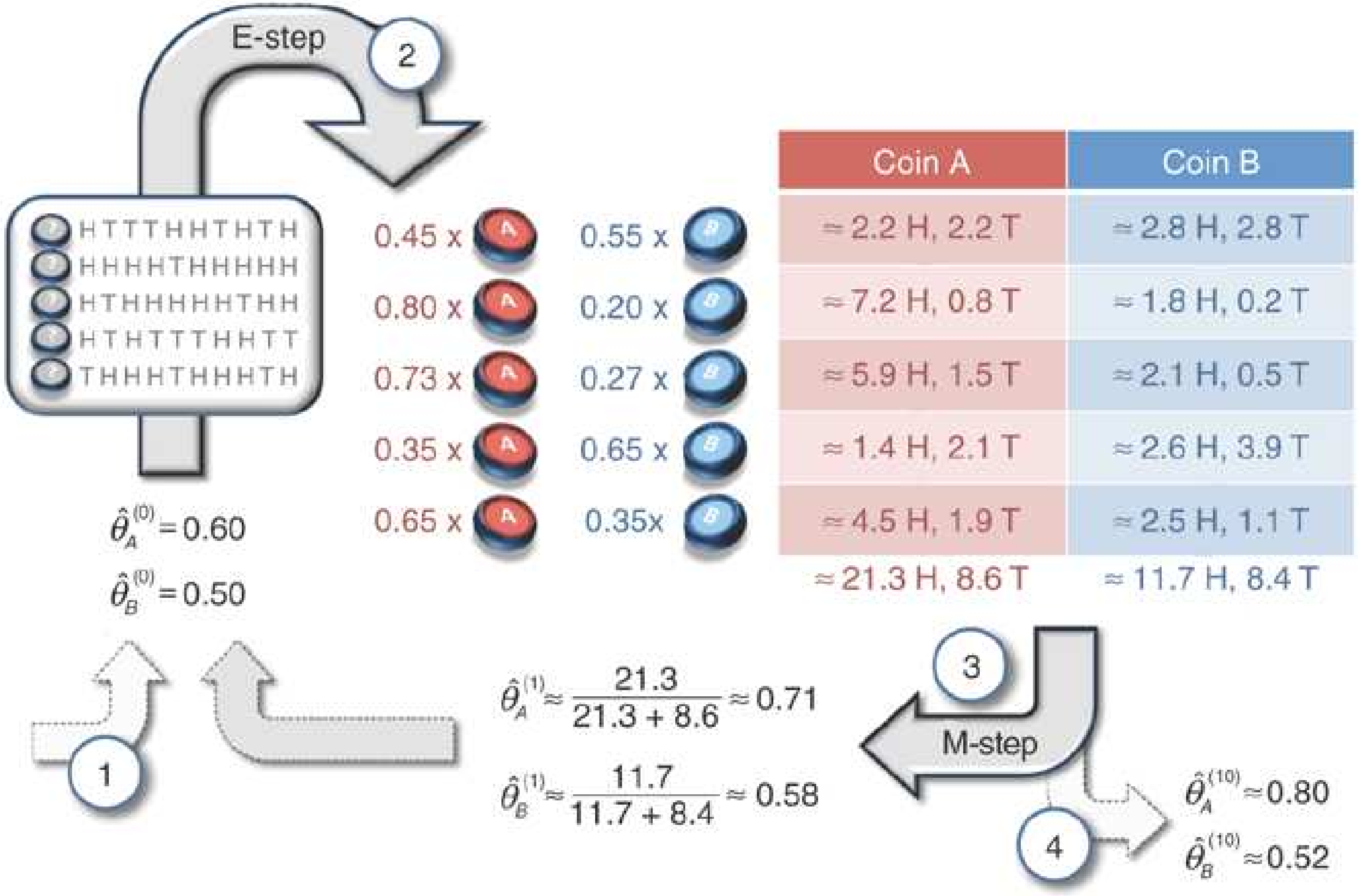
		Coin A	Coin B
B	H T T T H H T H T H		5 H, 5 T
A	H H H H T H H H H H	9 H, 1 T	
A	H T H H H H H T H H	8 H, 2 T	
B	H T H T T T H H T T		4 H, 6 T
A	T H H H T H H H T H	7 H, 3 T	
		24 H, 6 T	9 H, 11 T

5 sets, 10 tosses per set

$$\hat{\theta}_A = \frac{24}{24 + 6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9 + 11} = 0.45$$

Expectation maximization



			Coin A	Coin B
HTTTHHTHTH	0.45 x A	0.55 x B	≈ 2.2 H, 2.2 T	≈ 2.8 H, 2.8 T
HHHHTHHHHH	0.80 x A	0.20 x B	≈ 7.2 H, 0.8 T	≈ 1.8 H, 0.2 T
HTHHHHHTHH	0.73 x A	0.27 x B	≈ 5.9 H, 1.5 T	≈ 2.1 H, 0.5 T
HTHTTTHHTT	0.35 x A	0.65 x B	≈ 1.4 H, 2.1 T	≈ 2.6 H, 3.9 T
THHHTHHHTH	0.65 x A	0.35x B	≈ 4.5 H, 1.9 T	≈ 2.5 H, 1.1 T
			≈ 21.3 H, 8.6 T	≈ 11.7 H, 8.4 T

$$\hat{\theta}_A^{(0)} = 0.60$$

$$\hat{\theta}_B^{(0)} = 0.50$$

$$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3 + 8.6} \approx 0.71$$

$$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7 + 8.4} \approx 0.58$$

$$\hat{\theta}_A^{(10)} \approx 0.80$$

$$\hat{\theta}_B^{(10)} \approx 0.52$$

The EM algorithm for solving a mixture of K categorical distributions, applied to the problem of Word Sense Disambiguation,

i.e. identifying the semantic domains associated to words in a text document

CMU, 2012 fall, Eric Xing, Aarti Singh, HW3, pr. 3

The objective of this exercise is to derive the update equations of the EM algorithm for optimizing the latent variables [designating the semantic domains] involved in generating a text document.

Each *word* will be seen as a random variable w that can take values $1, \ldots, V$ from the *vocabulary* of words. In fact, we will denote each w by an array of V components such that $w(i) = 1$ if w takes the value of the i-th word in the vocabulary. Hence, $\sum_{i=1}^{V} w(i) = 1$.

Given a *document* containing words w_j, $j = 1, \ldots, N$, where N is the length of the document, we will assume that these words are generated from a mixture of K discrete *topics*:

$$P(w) \;=\; \sum_{m=1}^{K} \pi_m \, P(w|\mu_m) \text{ and } P(w|\mu_m) = \prod_{i=1}^{V} \mu_m(i)^{w(i)},$$

where

π_m denotes the prior [probability] for the latent topic variable $t = m$,

$\mu_k \overset{not.}{=} (\mu_k(1), \ldots, \mu_k(i), \ldots, \mu_k(V))$, with $\mu_k(i) \geq 0$ for $i = 1, \ldots, V$ and $\sum_{i=1}^{V} \mu_m(i) = 1$ for each $k = 1, \ldots, K$, and

$\mu_m(i) \overset{not.}{=} P(w(i) = 1|t = m)$.

a. In the expectation step, for each word w_j, compute

$$F_j(t) \overset{not.}{=} P(t|w_j;\theta),$$

the probability that w_j belongs to each of the K topics, where θ is the set of parameters of this mixture model.

Answer:

$$\begin{aligned} F_j(t_j = m) &\overset{not.}{=} P(t_j = m|w_j;\theta) \overset{Bayes\ T.}{=} \frac{P(w_j|t_j = m;\theta)\ P(t_j = m|\theta)}{P(w_j|\theta)} \\ &= \frac{\pi_m\ P(w_j|\mu_m)}{\sum_{m'=1}^{K} \pi_{m'}\ P(w_j|\mu_{m'})} = \frac{\pi_m \prod_{l=1}^{V} \mu_m(l)^{w_j(l)}}{\sum_{m'=1}^{K} \pi_{m'} \prod_{l=1}^{V} \mu_{m'}(l)^{w_j(l)}} \end{aligned}$$

b. In the maximization step, compute θ that maximizes the log-likelihood of the data

$$l(w;\theta) \overset{not.}{=} \log \prod_{j=1}^{N} P(w_j;\theta)$$

***Hint*: Suming over the latent topic variable, we can write $l(w;\theta)$ as**

$$l(w;\theta) = \sum_{j=1}^{N} \log \sum_{t} P(w_j, t; \theta) = \sum_{j=1}^{N} \log F_j(t) \frac{\sum_t P(w_j, t; \theta)}{F_j(t)}$$

Further on, using Jensen's inequality (see pr. EM-2) we get:

$$l(w;\theta) \quad \geq \quad \sum_{j=1}^{N} \sum_{t} F_j(t) \log \frac{P(w_j, t; \theta)}{F_j(t)} = \sum_{j=1}^{N} \sum_{t} F_j(t) \log P(w_j;\theta) - H(F_j)$$

Hence compute θ as:

$$\operatorname*{argmax}_{\theta} \sum_{j=1}^{N} \sum_{t} F_j(t) \log P(w_j, t; \theta)$$

Answer:

$$
\begin{aligned}
\theta &= \operatorname*{argmax}_{\theta} \sum_{j=1}^{N} \sum_{m=1}^{K} F_j(t_j = m) \log P(w_j, t_j = m | \theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{j=1}^{N} \sum_{m=1}^{K} F_j(t_j = m) \log P(w_j | t_j = m; \theta)\ P(t_j = m | \theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{j=1}^{N} \sum_{m=1}^{K} F_j(t_j = m) \log \left(\pi_m \prod_{l=1}^{V} \mu_m(l)^{w_j(l)} \right) \\
&= \operatorname*{argmax}_{\theta} \sum_{j=1}^{N} \sum_{m=1}^{K} [F_j(t_j = m) \log \pi_m + F_j(t_j = m) \sum_{l=1}^{V} \log \mu_m(l)^{w_j(l)}] \\
&= \operatorname*{argmax}_{\theta} \sum_{j=1}^{N} \sum_{m=1}^{K} [F_j(t_j = m) \log \pi_m + F_j(t_j = m) \sum_{l=1}^{V} w_j(l) \log \mu_m(l)] \quad (10)
\end{aligned}
$$

To optimize $\mu_m(l)$:

After eliminating from (10) the terms which are constant with respect to μ_m we get:

$$\sum_{j=1}^{N} F_j(t_j = m) \sum_{l=1}^{V} w_j(l) \log \mu_m(l)$$

We will use a Lagrangian to constrain μ_m to be a probability distribution:

$$\mathcal{L}(\mu_m(l)) = \sum_{j=1}^{N} F_j(t_j = m) \sum_{l=1}^{V} w_j(l) \log \mu_m(l) + \beta \left(\sum_{l=1}^{V} \mu_m(l) - 1 \right)$$

Then, solving for $\mu_m(l)$:

$$\begin{aligned} \frac{\partial}{\partial \mu_m(l)} \mathcal{L}(\mu_m(l)) = 0 \Leftrightarrow & \sum_{j=1}^{N} F_j(t_j = m) \frac{w_j(l)}{\mu_m(l)} + \beta = 0 \\ \Leftrightarrow \quad & \frac{1}{\mu_m(l)} \sum_{j=1}^{N} F_j(t_j = m)\; w_j(l) + \beta = 0 \Leftrightarrow \frac{1}{\mu_m(l)} = \frac{-\beta}{\sum_{j=1}^{N} F_j(t_j = m)\; w_j(l)} \\ \Leftrightarrow \quad & \mu_m(l) = \frac{\sum_{j=1}^{N} F_j(t_j = m)\; w_j(l)}{-\beta} \qquad (11) \end{aligned}$$

Knowing that $\sum_{l=1}^{V} \mu_m(l) = 1$, we have:

$$\sum_{l=1}^{V} \frac{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}{-\beta} = 1 \;\;\Leftrightarrow\;\; -\beta = \sum_{l=1}^{V} \sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)$$

Hence, substituting for $-\beta$ in (11), we get:

$$\begin{aligned}
\mu_m(l) &= \frac{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}{\sum_{l=1}^{V} \sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)} = \frac{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}{\sum_{j=1}^{N} \sum_{l=1}^{V} F_j(t_j = m)\, w_j(l)} \\
&= \frac{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}{\sum_{j=1}^{N} F_j(t_j = m) \underbrace{\sum_{l=1}^{V} w_j(l)}_{1}} = \frac{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}{\sum_{j=1}^{N} F_j(t_j = m)}
\end{aligned}$$

Note: Intuitively the last expression can be interpreted as the portion [of word[occurrence]s] that had $w(l) = 1$ among all words [in the given document] which are deemed to belong to cluster m.

To optimize π_m, we proceed similarly:

We begin by removing from (10) the terms that are constant with respect to π_m, thus getting:

$$\sum_{j=1}^{N} F_j(t_j = m) \log \pi_m$$

So, using the Lagrangian with the constraint that $\sum_{m=1}^{K} \pi_m = 1$

$$\mathcal{L}(\pi_m) = \sum_{j=1}^{N} F_j(t_j = m) \log \pi_m + \beta \left(\sum_{m=1}^{K} \pi_m - 1 \right),$$

and solving for π_m, we have:

$$\frac{\partial}{\partial \pi_m} \mathcal{L}(\pi_m) = 0 \Leftrightarrow \sum_{j=1}^{N} \frac{F_j(t_j = m)}{\pi_m} + \beta = 0 \Leftrightarrow \frac{1}{\pi_m} \sum_{j=1}^{N} F_j(t_j = m) = -\beta$$

$$\Leftrightarrow \pi_m = \frac{\sum_{j=1}^{N} F_j(t_j = m)}{-\beta} \tag{12}$$

Since $\sum_{m=1}^{K} \pi_m = 1$, it gives us:

$$\sum_{m=1}^{K} \frac{\sum_{j=1}^{N} F_j(t_j = m)}{-\beta} = 1 \;\Leftrightarrow\; \frac{1}{-\beta} \sum_{m=1}^{K} \sum_{j=1}^{N} F_j(t_j = m) = 1$$

$$\Leftrightarrow\; -\beta = \sum_{m=1}^{K} \sum_{j=1}^{N} F_j(t_j = m)$$

Substituting for $-\beta$ in (12), we get:

$$\pi_m = \frac{\sum_{j=1}^{N} F_j(t_j = m)}{\sum_{m=1}^{K} \sum_{j=1}^{N} F_j(t_j = m)} = \frac{\sum_{j=1}^{N} F_j(t_j = m)}{\sum_{j=1}^{N} \underbrace{\sum_{m=1}^{K} F_j(t_j = m)}_{1}} = \frac{\sum_{j=1}^{N} F_j(t_j = m)}{N}$$

Note: Intuitively the last expression can be interpreted as the portion [of word[occurrence]s] that belongs to cluster m among the total of N words.

To summarize:

E step:

$$F_j(t_j = m) = \frac{\pi_m \prod_{l=1}^{V} \mu_m(l)^{w_j(l)}}{\sum_{m'=1}^{K} \pi_{m'} \prod_{l=1}^{V} \mu_{m'}(l)^{w_j(l)}} \text{ for } j = 1, \ldots, N \text{ and } m = 1, \ldots, K$$

M step:

$$\mu_m(l) = \frac{\sum_{j=1}^{N} F_j(t_j = m)\, w_j(l)}{\sum_{j=1}^{N} F_j(t_j = m)} \text{ for } m = 1, \ldots, K \text{ and } l = 1, \ldots, V$$

$$\pi_m = \frac{\sum_{j=1}^{N} F_j(t_j = m)}{N} \text{ for } m = 1, \ldots, K$$

Using the EM algorithm for estimating the *selection probability* for a mixture of two (arbitrary) distributions

CMU, 2006 spring, ?, final exam, pr. 8

CMU, 2004 fall, Carlos Guestrin, HW2, pr. 2.1

We want to derive an EM algorithm for estimating the mixing parameter for a mixture of arbitrary probability densities f_1 and f_2.

For *example*, $f_1(x)$ could be a standard normal distribution centered at 0, and $f_2(x)$ could be the uniform distribution between $[0, 1]$. You can think about such mixtures in the following way: First, you flip a coin. With probability λ (i.e., the coin comes up *heads*), you will sample x from density f_1, and with probability $(1 - \lambda)$ you sample from density f_2.

More formally, let $f_\lambda(x) = \lambda f_1(x) + (1 - \lambda) f_2(x)$, where f_1 and f_2 are arbitrary probability density functions on $\mathbb{R}$, and $\lambda \in [0, 1]$ is an unknown mixture parameter.

a.
Given a data point x, and a value for the mixture parameter λ, compute the probability that x was generated from density f_1.

b.
Now, suppose you are given a data set $\{x_1, \ldots, x_n\}$ drawn i.i.d. from the mixture density, and a set of coin flips $\{z_1, z_2, \ldots, z_n\}$, such that $z_i = 1$ means that x_i is a sample from f_1, and $z_i = 0$ means that x_i was generated from density f_2.
For a fixed parameter λ, compute the complete log-likelihood of the data, i.e., $\ln P(x_1, z_1, x_2, z_2, \ldots, x_n, z_n | \lambda)$.

c.
Now, suppose you are given only a sample $\{x_1, \ldots, x_n\}$ drawn i.i.d. from the mixture density, without the knowledge about which component the samples were drawn from (i.e., the z_i are unknown).
Using your derivations from part a and b, derive the E- and M-steps for an EM algorithm to compute the maximum likelihood estimate (MLE) of the mixture parameter λ.

Solution

a. $P(Y=1|X=x) = \dfrac{P(X=x|Y=1)\cdot P(Y=1)}{P(X=x)} = \dfrac{\lambda f_1(x)}{f_\lambda(x)}$.

b. $P(x_1, z_1, x_2, z_2, \ldots, x_n, z_n|\lambda) = \prod_{i=1}^n P(x_i, z_i|\lambda) = \prod_{i=1}^n P(x_i|z_i, \lambda)\cdot P(z_i|\lambda)$.

$$P(x_i|z_i,\lambda)\cdot P(z_i|\lambda) = \begin{cases} \lambda f_1(x_i) & \textbf{if } z_i = 1 \\ (1-\lambda) f_2(x_i) & \textbf{if } z_i = 0 \end{cases} = f_1(x_i)^{z_i}(1-\lambda)^{1-z_i} f_2(x_i)^{1-z_i}$$

Therefore,

$$\begin{aligned}\ln P(x_1, z_1, x_2, z_2, \ldots, x_n, z_n|\lambda) &= \sum_{i=1}^n \ln P(x_i, z_i|\lambda) \\ &= \sum_{i=1}^n \ln\left(\lambda^{z_i} f_1(x_i)^{z_i}(1-\lambda)^{1-z_i} f_2(x_i)^{1-z_i}\right) \\ &= \sum_{i=1}^n z_i(\ln\lambda + \ln f_1(x_i)) + (1-z_i)(\ln(1-\lambda) + \ln f_2(x_i))\end{aligned}$$

c.

E-step: $q(z_i) \overset{not.}{=} P(z_i = 1 | x_i, \lambda^{(t)}) \overset{B.Th.}{=} \dfrac{\lambda^{(t)} f_1(x_i)}{f_{\lambda^{(t)}}(x_i)}$

M-step:

$$\begin{aligned}
\lambda^{(t+1)} &= \operatorname*{argmax}_{\lambda} E_q[\ln \prod_{i=1}^{n} P(x_1, z_1, x_2, z_2, \ldots, x_n, z_n | \lambda)] \\
&= \operatorname*{argmax}_{\lambda} (\ln \lambda \cdot \underbrace{\sum_{i=1}^{n} q(z_i)}_{c} + \ln(1-\lambda) \cdot \underbrace{\sum_{i=1}^{n} (1 - q(z_i))}_{n-c})
\end{aligned}$$

$$\frac{\partial}{\partial \lambda}(c \ln \lambda + (n-c)\ln(1-\lambda)) = 0 \Leftrightarrow \frac{c}{\lambda} = \frac{n-c}{1-\lambda}$$

$$\Leftrightarrow \quad c(1-\lambda) = \lambda(n-c) \Leftrightarrow c = n\lambda \Leftrightarrow \lambda = \frac{c}{n} = \frac{\sum_{i=1}^{n} q(z_i)}{n}$$

$$\Rightarrow \qquad \lambda^{(t+1)} = \frac{\lambda^{(t)}}{n} \sum_{i=1}^{n} \frac{f_1(x_i)}{f_{\lambda^{(t)}}(x_i)}$$