Johns Hopkins Engineering

625.464 Computational Statistics

An Expectation Maximization Example

Module 3 Lecture 3C

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

The Peppered Moth Example

C I T alleles

observe n moths

$n = n_C + n_I + n_T$

carbonia CC CI CT

insularia II IT

typica TT

want to know prob P_C, P_I, P_T

observed data

$X = (n_C, n_I, n_T)$

$P_C + P_I + P_T = 1$

complete data

CC	CI	CT	II	IT	TT
P_C^2	$P_C P_I$	$P_C P_T$	P_I^2	$P_I P_T$	P_T^2

$Y = (n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$

$X = M(Y) = (n_{CC} + n_{CI} + n_{CT}, n_{II} + n_{IT}, n_{TT})$

Peppered Moth MLE Problem

Want to know p_C, p_I, p_T ← parameters we need to find

$P=(p_C, p_I, p_T)$

Multinomial R.V. $\quad Y=(n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$

$$f_Y(y|p) = \binom{n}{n_{CC}\cdots n_{TT}} (p_C^2)^{n_{CC}} \cdot (2p_Cp_I)^{n_{CI}} \cdot (2p_Cp_T)^{n_{CT}} \cdot (p_I^2)^{n_{II}} (2p_Ip_T)^{n_{IT}} \cdot (p_T^2)^{n_{TT}}$$

$$\log f_Y(y|p) = n_{CC}\log p_C^2 + n_{CI}\log 2p_Cp_I + n_{CT}\log 2p_Cp_T + n_{IT}\log p_I^2 + n_{IT}\log 2p_Ip_T + n_{TT}\log p_T^2 + \log\binom{n}{n_{CC}\cdots n_{TT}}$$

Peppered Moth EM Algorithm

$Y = (N_{CC}, N_{CI}, N_{CT}, N_{II}, N_{IT}, N_{TT})$

only $N_{TT} = n_{TT} = n_T$

Step ① fill in missing data.

Find $Q(p|p^{(t)}) = E\left[\log f_Y(y|p) \,\middle|\, x, p^{(t)}\right]$

want to find

$E\left[N_{??} \,\middle|\, \underline{n_C, n_I, n_T,}\, p^{(t)}\right]$ for all ?? pairs

Peppered Moth EM Algorithm

$$E\left[\overset{n_{CC}^{(t)}}{\cancel{N_{CC}}} \,\middle|\, x, p^{(t)}\right] = \frac{n_C \left(p_C^{(t)}\right)^2}{\left(p_C^{(t)}\right)^2 + 2 p_C^{(t)} p_I^{(t)} + 2 p_C^{(t)} p_T^{(t)}}$$

$$E\left[\overset{n_{CI}^{(t)}}{\cancel{N_{CI}}} \,\middle|\, x, p^{(t)}\right] = \frac{n_C \left(2 p_C^{(t)} p_I^{(t)}\right)}{\text{denom}}$$

$$E[N_{TT}] = n_{TT}$$

$$E\left[\overset{n_{CT}^{(t)}}{\cancel{N_{CT}}} \,\middle|\, x, p^{(t)}\right] = \frac{n_C \left(2 p_C^{(t)} p_T^{(t)}\right)}{\text{denom}}$$

$$E\left[\overset{n_{II}^{(t)}}{\cancel{N_{II}}}\right] = \frac{n_I \left(p_I^{(t)}\right)^2}{\left(p_I^{(t)}\right)^2 + 2 p_I^{(t)} p_T^{(t)}}$$

$$E\left[\overset{n_{IT}^{(t)}}{\cancel{N_{IT}}}\right] = \frac{n_I \left(2 p_I^{(t)} p_T^{(t)}\right)}{\text{denom}}$$

EM Algorithm

E Step: $Q(p|p^{(t)}) = n_{CC}^{(t)} \log p_C^2 + n_{CI}^{(t)} \log 2p_C p_I$

$+ n_{CT}^{(t)} \log 2p_C p_T + n_{II}^{(t)} \log p_I^2$

$+ n_{IT}^{(t)} \log p_I p_T + n_{TT} \log p_T^{(2)}$

M Step: Take partial derivatives w/ respect to $p = (p_C, p_I, p_T)$

$p_T = 1 - p_C - p_I$

M Step

$$\frac{dQ(p|p^{(t)})}{dp_C} = \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{p_C} - \frac{n_{CT}^{(t)} + n_{IT}^{(t)} + 2n_{TT}}{1-p_C-p_I}$$

$$\frac{dQ(p|p^{(t)})}{dp_I} = \frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{p_I} - \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{1-p_C-p_I}$$

Set $=0$ and solve.

$$p_C^{(t+1)} = \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{2n}$$

$$p_T^{(t+1)} = \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{2n}$$

$$p_I^{(t+1)} = \frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{2n}$$