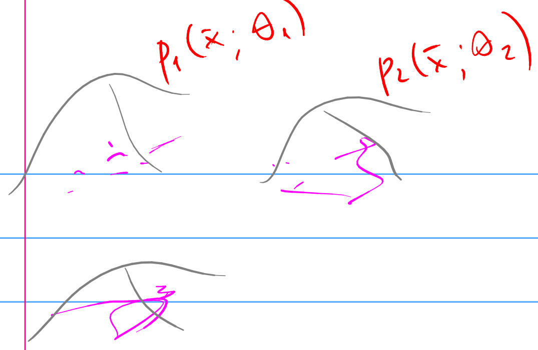


$$p(\bar x) = \sum_k \pi_k\, p_k(\bar x; \bar\theta_k)$$

$$p(D|\pi,\theta) = \prod_n \left(\sum_k \pi_k\, p_k(\bar x_n, \bar\theta_k)\right)$$

z_{nk}

$\to \max$

$\bar\pi, \bar\theta$

$$p(D, z|\pi,\theta) = \prod_n \prod_k (\pi_k p_k)$$

$X, \theta \qquad p(X|\theta) \to \max \qquad z \qquad p(X, z|\theta)$

$$\theta^{(0)} \to \theta^{(1)} \to \theta^{(2)} \to \_\_ \to \theta^{(m)} \to \theta^{(m+1)} \to \dots$$

$\log p(\theta|X) = \log p(X|\theta) + \log p(\theta)$

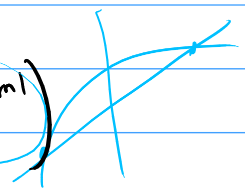
$\int f(z)\, q(z)\, dz$

$$\ell(\theta) = \log p(X|\theta) \to \max$$

$$\ell(\theta) - \ell(\theta^{(m)}) = \log p(X|\theta) - \log p(X|\theta^{(m)}) =$$

$$= \log \int p(X, z|\theta)\, dz - \log p(X|\theta^{(m)}) =$$

$$= \log \int p(X, z|\theta) \frac{p(z|X,\theta^{(m)})}{p(z|X,\theta^{(m)})}\, dz - \log p(X|\theta^{(m)}) \geq$$



$$\geq \int p(z|X,\theta^{(m)}) \log \frac{p(X,z|\theta)}{p(z|X,\theta^{(m)})}\, dz - \log p(X|\theta^{(m)})$$

$\theta^{(m)}$

$f(\alpha x + (1-\alpha) y) \geq \alpha f(x) + (1-\alpha) f(y)$

$$= \int p(z|X,\theta^{(m)}) \log \frac{p(X,z|\theta)}{p(z|X,\theta^{(m)})\, p(X|\theta^{(m)})}\, dz$$

$= p(X, z|\theta^{(m)})$

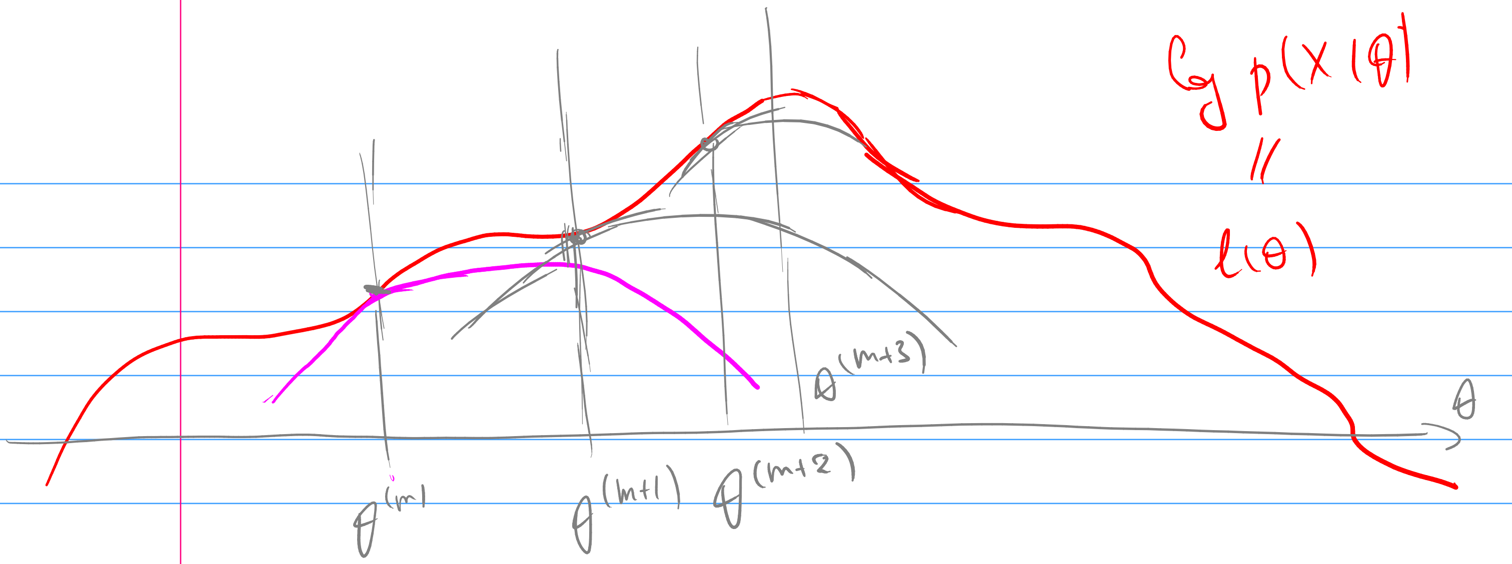
$f(\mathbb{E}_{p(x)}[x]) \geq \mathbb{E}_{p(x)}[f(x)]$

$$= h(\theta, \theta^{(m)})$$

$$\ell(\theta) \geq \ell(\theta^{(m)}) + h(\theta, \theta^{(m)})$$

$h(\theta, \theta^{(m)}) = \mathbb{E}_{z|}[\log p(x,z|\theta)] - \mathbb{E}_z[\log \dots]$

$$h(\theta^{(m)}, \theta^{(m)}) = 0$$



$$\boxed{\theta^{(m+1)} := \underset{\theta}{\arg\max}\, L(\theta, \theta^{(m)})} = \underset{\theta}{\arg\max}\, Q(\theta, \theta^{(m)})$$

$$\mathbb{E}_{q(z)}[\ldots] - c = \mathbb{E}_{q(z)}[\ldots - c]$$

$$Q(\theta, \theta^{(m)}) = \mathbb{E}_{p(z|X,\theta^{(m)})}\left[\log p(X,z|\theta)\right]$$

$$= \int p(z|X,\theta^{(m)}) \log p(X,z|\theta)\, dz$$

$$\theta = (\bar\theta_1, \bar\theta_2, \ldots \bar\theta_K, \bar\pi)$$

$$z = (\bar z_1, \bar z_2, \ldots, \bar z_N) \qquad \bar z_n = (0 \ldots 1 \ldots 0)$$

$$z_{nk} = 1 \text{ iff } \bar x_n \in C_k$$

$$p(X|\theta) = \prod_n \left(\sum_k \pi_k p_k(\bar x_n | \bar\theta_k) \right) \xrightarrow[\theta]{} \max$$

$$p(X,z|\theta) = \prod_n \prod_k \left(\pi_k p_k(x_n | \bar\theta_k) \right)^{z_{nk}}$$

$$\log p(X,z|\theta) = \sum_n \sum_k z_{nk} \left(\log \pi_k + \log p_k(\bar x_n | \bar\theta_k) \right)$$

$$p(z_{nk}\mid X,\theta^{(m)}) = p(\bar{x}_n \in C_k \mid \bar{x}_n, \theta^{(m)}) =$$

E-шаг

$$= \frac{p(C_k)\,p(\bar{x}_n\mid C_k)}{\sum_\ell p(C_\ell)\,p(\bar{x}_n\mid C_\ell)} = \frac{\pi_k^{(m)}\, p_k(\bar{x}_n\mid\theta_k^{(m)})}{\sum_\ell \pi_\ell^{(m)}\, p_k(\bar{x}_n\mid\theta_k^{(\ell)})}$$

$$Q(\theta,\theta^{(m)}) = \mathbb{E}_{p(Z\mid X,\theta^{(m)})}\left[\log p(X,Z\mid\theta)\right] =$$

$$= \sum_n \sum_k \left(\mathbb{E}_{p(Z\mid X,\theta^{(m)})}[z_{nk}]\right)\cdot\left(\log \pi_k + \log p_k(\bar{x}_n\mid\bar{\theta}_k)\right)$$

M-шаг

$$\xrightarrow[\pi,\,\theta]{} \max$$

$$\sum_k \left(\sum_n \mathbb{E} z_{nk}\right)\log \pi_k \xrightarrow[\pi_k]{} \max$$

$$\sum \pi_k = 1,\ \pi_k \geq 0$$

$$\pi_k^{(m+1)} = \frac{\sum_n \mathbb{E} z_{nk}}{\sum_\ell \sum_n \mathbb{E} z_{n\ell}} \qquad \sum_\ell \sum_n \mathbb{E} z_{n\ell} = N$$

1955

M N

MM – n_{MM}

MN – n_{MN}

NN – n_{NN}

$$p = \frac{2n_{MM} + n_{MN}}{2(n_\cdot + n_\cdot + n_\cdot)}$$

$$q = \frac{2n_{NN} + n_{MN}}{2(\cdot\ \text{—}\ )}$$

Закон Харди – Вайнберга

$p = p(M)$, $q = 1 - p = p(N)$

$$\Rightarrow\ p(MM, NN) = \frac{p^2}{p^2 + 2pq} \qquad \boxed{p(MN)} = \frac{2pq}{p^2+2pq}$$

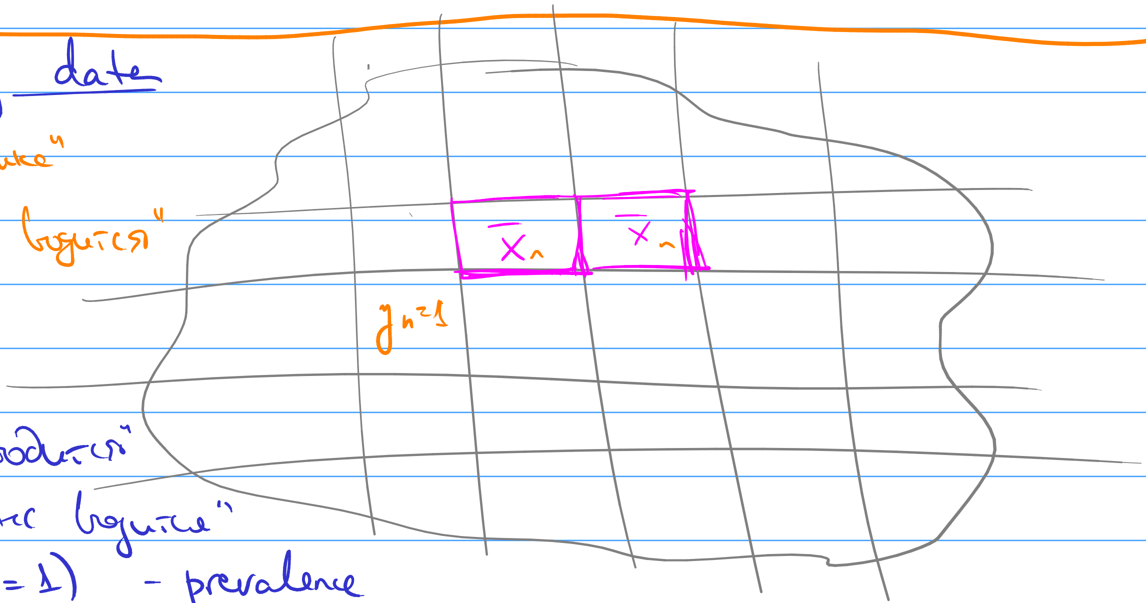
$$\boxed{p(MN)} = z$$

$$\mathbb{E}[z] = \frac{2p^{(m)}(1-p^{(m)})}{p^{(m)2} + 2p^{(m)}(1-p^{(m)})}$$

$$n_{NN},\quad \mathbb{E} n_{MN} = (\mathbb{E} z)\cdot(n_\cdot + n_\cdot + n_\cdot),\qquad \mathbb{E} n_{NM} = \text{―}\text{''}\text{―}\ \mathbb{E} n_{MN}$$

$$p^{(m+1)} = \frac{2\,\mathbb{E}\cdot n_{MM} + \mathbb{E}\, n_{MW}}{2\cdot(n_\cdot + n_\cdot + n_\cdot)}$$

Presence-only data



"видели суслика"

$y_n = 1$ => "суслик водится"

$y_n = 0$ => ?

$z_n = 1$ — "суслик водится"

$z_n = 0$ — " — " — не водится"

$\pi = p(z_n = 1)$ — prevalence

Если $y_n = 1$, то $z_n = 1$ с вер-1.

$$p(z = 1 \mid \bar{x}) = \sigma(\eta(\bar{x})) = \frac{1}{1 + e^{-\eta(\bar{x})}}$$

	$d = 0$	$d = 1$
$X = 0$	π_{00}	π_{01}
$X = 1$	π_{10}	π_{11}
	$p(d=0)$	$p(d=1)$

$p(x=0)$

$p(x=1)$

Prospective studies

π_{00} π_{01}

$$\frac{\sigma(\eta(x))}{\frac{1}{1+e^{-\eta(x)}}} = p(d=1 \mid x=0) = \frac{\pi_{01}}{\pi_{00}+\pi_{01}}$$

$$p(d=1 \mid x=1) = \frac{\pi_{11}}{\pi_{10}+\pi_{11}}$$

Retrospective studies

$\pi_0 = p(s=1 \mid d=0)$

$\pi_1 = p(s=1 \mid d=1)$

$p(d=1 \mid x=0) = ?$

$$\frac{\pi_{01}}{\pi_{00}+\pi_{01}} = p(d=1 \mid x=0, s=1) =$$

$$= \frac{p(d=1 \mid x=0)\, \overbrace{p(s=1 \mid d=1, x=0)}^{\pi_1}}{p(d=1 \mid x=0)\, p(s=1 \mid d=1, x=0) + p(d=0 \mid x=0)\, \underbrace{p(s=1 \mid d=0, x=0)}_{\pi_0}}$$

$p(d=0 \mid x=0) = 1 - p(d=1 \mid x=0)$

$$\frac{\pi_{01}}{\pi_{00}+\pi_{01}} = \frac{\pi_1\, p(d=1|x=0)}{\pi_1\, p(d=1|x=0) + \pi_0\big(1-\underbrace{p(d=1|x=0)}_{=\sigma(\eta(x))}\big)}$$

$$\pi_1(\pi_{00}+\cancel{\pi_{01}})\,p = \pi_{01}(\cancel{\pi_1}-\pi_0)\,p + \pi_0\,\pi_{01}$$

$$\sigma(\eta(x)) = p = \frac{\pi_0\pi_{01}}{\pi_{00}\pi_1+\pi_{01}\pi_0}$$

$$\sigma(\eta(x)) = \frac{1}{1+e^{-\eta(x)}}$$

$$\frac{\pi_{01}}{\pi_{00}+\pi_{01}} = \frac{\pi_1\,\sigma(\eta(x))}{\pi_0+(\pi_1-\pi_0)\,\sigma(\eta(x))}$$

$$= \frac{\pi_1}{\cancel{1+e^{-\eta(x)}}} \cdot \frac{\cancel{1+e^{-\eta(x)}}}{\pi_0(\cancel{1}+e^{-\eta(x)}) + (\pi_1-\cancel{\pi_0})} =$$

$$= \frac{\pi_1}{\pi_1+\pi_0 e^{-\eta(x)}} = \frac{1}{1+\frac{\pi_0}{\pi_1}e^{-\eta(x)}} =$$

$$= \frac{1}{1+e^{-\eta(x)+\ln\frac{\pi_0}{\pi_1}}} = \sigma\Big(\overbrace{\eta(x)-\ln\frac{\pi_0}{\pi_1}}^{\eta^*(x)}\Big)$$

$$\eta(x) = w_0 + w_1 x$$

$$\eta^*(x) = \eta(x) - \ln\frac{\pi_0}{\pi_1}$$

$$w_1, \quad w_0 - \ln \pi_0/\pi_1$$

$$\sigma(\eta(\bar{x})) = p(z=1|\bar{x})$$

$s=1$ – выборка

$\pi = p(z=1)$ – встречаемость

n_p – positive

n_u – unknown

$$p(y=1|\bar{x}, s=1) = \sigma(\eta_{naive}(x))$$

[illegible]

выборка:

$z=1$: $n_p + \pi \cdot n_u$

$z=0$: $(1-\pi)n_u$

$$p(y=1\mid\bar{x},s=1) = \underbrace{p(y=1\mid z=1,s=1,\cancel{\bar{x}})}\cdot p(z=1\mid s=1,\bar{x}) +$$

$$+\underbrace{\cancel{p(y=1\mid z=0,s=1,\cancel{\bar{x}})}}_{=0}\, p(z=0\mid s=1,\bar{x})$$

$$= \underbrace{p(y=1\mid z=1,s=1)}\cdot p(z=1\mid s=1,\bar{x})$$

$$\eta(\bar{x}) = w_1x_1+\dots+w_dx_d$$

$$\sigma(\eta(x)) = p(z=1\mid x)$$

$$\| \qquad \frac{p(y=1,z=1\mid s=1)}{p(z=1\mid s=1)} = \frac{n_p}{n_p+\pi\cdot n_u}$$

$$\cancel{\sigma(\eta_{naive}(\bar{x}))}$$

$$\|$$

$$\underbrace{p(y=1\mid\bar{x},s=1)} = \frac{n_p}{n_p+\pi n_u}\cdot \underbrace{p(z=1\mid s=1,\bar{x})}$$

$$\| \qquad \sigma\left(\eta(x) - \log\frac{\pi_0}{\pi_1}\right)$$

$$p(z=1\mid s=1)p(s=1) \qquad \frac{n_p+\pi n_u}{n_p+n_u} = p(z=1\mid s=1)$$

$$\|$$

$$\pi_1 = p(s=1\mid z=1) = \frac{p(z=1,s=1)}{p(z=1)} = \frac{n_p+\pi n_u}{\pi(n_p+n_u)}\cdot p(s=1)$$

$$p(z=1) = \pi$$

$$\pi_0 = p(s=1\mid z=0) = \frac{p(z=0,s=1)}{p(z=0)} = \frac{(1-\pi)n_u}{(1-\pi)(n_p+n_u)}\, p(s=1)$$

$$= \alpha$$

$$\sigma(\eta_{naive}(\bar{x})) = \frac{n_p}{n_p+\pi n_u}\cdot\sigma\left(\eta(x) - \log\frac{\pi n_u}{n_p+\pi n_u}\right)$$

$$L(\bar{\eta}\mid y,X) = p(y\mid X,\eta) = \prod_{i=1}^{n_p+n_u} p(y_i\mid s_i=1,\bar{x}_i) =$$

$$= \prod_i \underline{p(y_i=1\mid s_i=1,\bar{x}_i)}^{y_i}\left(1-p(y_i=1\mid s_i=1,\bar{x}_i)\right)^{1-y_i}$$

$$= \prod_i \left(\frac{n_p}{n_p+\pi n_u} \cdot \sigma\left(\eta(x) - \log \frac{\pi n_u}{n_p+\pi n_u}\right)\right)^{y_i} \times$$

$$\times \left(1 - \frac{n_p}{n_p+\pi n_u} \sigma(- - -)\right)^{1-y_i} \longrightarrow \max$$

$$\eta^{(0)} \to \eta^{(1)} \to \dots \to \eta^{(m)} \to \eta^{(m+1)} \to \dots$$

z_i: $\quad \sigma\left(\underline{\eta(\bar{x}_i)} - \log \frac{\pi_0}{\pi_1}\right) \approx \underline{\underline{z_i}}$ $\qquad \eta^{(0)} = \eta_{naive}$

<u>М-шаг</u>: $\quad \eta^{(m+1)}(\bar{x}) = \eta^*(\bar{x}) + \log \frac{\pi_0}{\pi_1}$, где η^* обуч на $z_i^{(m+1)}$

<u>E-шаг</u>: $\quad z_i^{(m+1)} = \underset{\dots}{E}[z_i] = \sigma\left(\eta^{(m)}(\bar{x})\right)$ для $y_i = 0$

$\qquad z_i^{(m+1)} = 1$ для $y_i = 1$

<u>data imputation</u> $\equiv$ pseudolabels