Crowd Sourcing

True class	Image	w^1	w^2	w^3	w^4	...	w^m	Estimated Label
A	X_1	A	A		B			A
A	X_2	B	B	C				B
B	X_3	C	C		A			C
B	X_4	D	D		C			D
⋮	⋮							
	X_n							

Why do we believe we can beat MV?

workers may have low quality

can estimate quality (seeing how often they agree w/ majority)

Model of Worker Quality:

$$P(y^w = k \mid \text{true label is } c) = \pi^w_{kc}$$

all images in a class are equally hard

EM for crowd sourcing problem

Data $\{ y_i^{(r)}, w_i^{(r)} \}_{i=1\cdots n}$

r labels to img i r workers

n - # images

m - # raters

K - # classes

Model: True label $- Z_i$ comes from a dist $(q_1, q_2 \cdots q_k)$

For now, assume $Z_i \sim \mathrm{Unif}([K])$

Confusion matrix π^{a}_{kc} $\quad \mathbb{P}(y_{ij} = c \mid Z_i = k, w_{ij} = a)$

$$L(\pi; y, Z) = P(y \mid z, \pi)$$
$$= \prod_{i=1}^{n} P(y_i \mid Z_i, \pi)$$
$$= \prod_{i=1}^{n} \prod_{j=1}^{r} P(y_{ij} \mid Z_i, \pi^{w_{ij}})$$
$$= \prod_{i=1}^{n} \prod_{j=1}^{r} \prod_{c=1}^{K} \underbrace{P(y_{ij} = c \mid Z_i, \pi^{w_{ij}})}_{\pi^{w_{ij}}_{Z_i c}}{}^{\mathbb{1}_{y_{ij}=c}}$$

$$\log L(\pi; y, z) = \sum_{i=1}^{n} \sum_{j=1}^{r} \sum_{c=1}^{K} \mathbb{1}_{y_{ij}=c}\, \pi^{w_{ij}}_{z_i c}$$

Want to use EM

Find distribution of $Z \mid y, \hat{\pi}$

$$\mathbb{P}(z_i = k \mid y_i, \hat{\pi}) = \frac{P(y_i \mid z_i = k, \hat{\pi}) P(z_i = k \mid \hat{\pi})}{P(y_i \mid \hat{\pi})}$$

distribution over k, for each i — $q_{ik}(\hat{\pi}) \Big\{ = \dfrac{\prod_{j=1}^{r} \prod_{c=1}^{K} \left(\hat{\pi}^{w_{ij}}_{kc}\right)^{\mathbb{1}_{y_{ij}=c}} \cdot \frac{1}{K}}{\text{denominator}}$

Write $\mathbb{E}_{z \mid y, \hat{\pi}} \log L(\pi; y, z)$

E step:

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{r} \sum_{c=1}^{K} q_{ik} \mathbb{1}_{y_{ij}=c} \log \pi_{kc}^{w_{ij}} = Q(\pi, \hat{\pi})$$

M step:

$$\pi = \arg\max Q(\pi; \hat{\pi})$$

$$\sum_{c} \pi_{kc}^{a} = 1$$

Can get formula for $\pi_{kc}^{a} = \dfrac{\sum_i \sum_j q_{ik} \mathbb{1}_{y_{ij}=c} \mathbb{1}_{w_{ij}=a}}{\sum_{k'} \text{numerator}(k \to k')}$