从 MLE 到 EM algo

- prob models: ① Bayes classifiers, ② Bayesian linear regression, ③ Logistic regression, ④ LS and RR. (ML and MAP)
  non-prob models: ① perceptron ② SVM ③ Decision trees ④ k means

- MLE: Seeks the θ that maximize the likelihood:
  $\theta_{ML} = \arg\max_\theta p(x_1, \dots x_n | \theta) = \arg\max_\theta p(x_1|\theta) \cdots p(x_n|\theta)$
  $\Leftrightarrow \arg\max_\theta \sum_{i=1}^n \ln p(x_i|\theta)$
  1) follows from i.i.d assumption. 2) $f(y) > f(x) \Rightarrow \ln f(y) > \ln f(x)$

What we have:
1. model parameters θ. 2. $\{x_1, \dots x_n\}$ data.
3. a prob dist $p(x|\theta)$ 4. an i.i.d assumption $x_i \overset{iid}{\sim} p(x|\theta)$

Expectation – Maximization algorithm.

1. x_i^o — observed portion (the sub-vector of x_i that's measured)
2. x_i^m — missing portion
3. The missing dimension can be different for diff x_i

we assume $x_i \overset{iid}{\sim} N(\mu, \Sigma)$, we want to solve

$\mu_{ML}, \Sigma_{ML} = \arg\max_{\mu,\Sigma} \sum_{i=1}^n \ln p(x_i^o | \mu, \Sigma)$

if we knew x_i^m, then $\mu_{ML}, \Sigma_{ML} = \arg\max_{\mu,\Sigma} \sum_{i=1}^n \ln p(x_i^o, x_i^m | \mu, \Sigma)$

goal: imputing missing values $\{x_1^m, \dots x_n^m\}$ $= \sum_{i=1}^n \ln p(x_i | \mu, \Sigma)$

idea: break θ into θ_1, θ_2.

$p(X|\theta_1) = \int p(X, \theta_2 | \theta_1) d\theta_2$, i.e. $p(x_i^o | \mu, \Sigma) = \int p(x_i^o, x_i^m | \mu, \Sigma) dx_i^m$
$= N(\mu_i^o, \Sigma_i^o)$

define a general objective function

① let us optimize marginal $p(X|\theta_1)$ over θ_1

② uses $p(X, \theta_2 | \theta_1)$ in doing so

The EM objective function:

$$\ln p(X|\theta_1) = \int q(\theta_2) \ln\left(\frac{p(X,\theta_2|\theta_1)}{q(\theta_2)}\right) d\theta_2 + \int q(\theta_2) \ln \frac{q(\theta_2)}{p(\theta_2|X,\theta_1)} d\theta_2$$

1) $q(\theta_2)$ = any prob dist

2) we assume we know $p(\theta_2|X,\theta_1)$

- derive em objective function.

$$\ln p(X|\theta_1) = \int q(\theta_2) \ln \cdot \frac{p(X,\theta_2|\theta_1)}{q(\theta_2)} d\theta_2 + \int q(\theta_2) \ln \frac{q(\theta_2)}{p(\theta_2|X,\theta_1)} d\theta_2$$

$$= \int q(\theta_2) \ln \frac{p(X,\theta_2|\theta_1)}{q(\theta_2)} \frac{q(\theta_2)}{p(\theta_2|X,\theta_1)} d\theta_2 \quad \rightarrow \; p(\theta_2|X,\theta_1) \cdot p(X|\theta_1)$$

$$= \int q(\theta_2) \ln p(X|\theta_1) d\theta_2 = \ln p(X|\theta_1)$$ 等式成立!

$$\therefore \text{EM objective} = \underbrace{\int q(\theta_2) \ln\left(\frac{p(X,\theta_2|\theta_1)}{q(\theta_2)}\right) d\theta_2}_{\text{a function only of } \theta_1} + \underbrace{\int q(\theta_2) \cdot \ln \frac{q(\theta_2)}{p(\theta_2|X,\theta_1)} d\theta_2}_{\text{KL divergence} \geq 0}$$

Given $\theta_1^{(t)}$, find value $\theta_1^{(t+1)}$:

E-step: set $q_t(\theta_2) = p(\theta_2|X,\theta_1^{(t)})$ and 计算.

$$\mathcal{L}_{q_t}(X,\theta_1) = \int p(\theta_2|X,\theta_1^{(t)}) \cdot \ln p(X,\theta_2|\theta_1) d\theta_2 - \underbrace{\int q_t(\theta_2) \cdot \ln q_t(\theta_2) d\theta_2}_{\text{ignore.}}$$

M-step: $\theta_1^{(t+1)} = \arg\max_{\theta_1} \mathcal{L}_{q_t}(X,\theta_1)$