2018/03/27

STA414: Partial Problem Set Solutions.

P5, Q4

$\ell_x(\theta) = \log f(x|\theta)$

a) $f(x,z|\theta) = f(x|\theta) f(z|x,\theta)$

$\Rightarrow f(x|\theta) = f(x,z|\theta) / f(z|x,\theta)$ (Bayes')

$$\boxed{\Rightarrow \ell_x(\theta) = \log f(x|\theta) = \log f(x,z|\theta) - \log f(z|x,\theta)}$$

b) $Q(\theta, \theta^t) \equiv E_{z|x,\theta^t}\left[\log f(x,z|\theta)\right]$ as defined in class.

Note: θ^t, not θ Note: θ, not θ^t

$H(\theta,\theta^t) \equiv E_{z|x,\theta^t}\left(\log f(z|x,\theta)\right)$

c) i) The M-step maximizes $Q(\theta,\theta^t)$ w.r.t. θ. So $Q(\theta^{t+1},\theta^t) \geq Q(\theta,\theta^t)\ \forall\, \theta$; in particular, $Q(\theta^{t+1},\theta^t) \geq Q(\theta^t,\theta^t)$

ii) $H(\theta^{t+1},\theta^t) - H(\theta^t,\theta^t)$

$$= E_{z|x,\theta^t}\left(\log f(z|x,\theta^{t+1}) - \log f(z|x,\theta^t)\right)$$

$$= E_{z|x,\theta^t}\left[\log \frac{f(z|x,\theta^{t+1})}{f(z|x,\theta^t)}\right]$$

$$\leq \log E_{z|x,\theta^t}\left(\frac{f(z|x,\theta^{t+1})}{f(z|x,\theta^t)}\right) \quad \text{(Jensen)}$$

$$= \log \int \frac{f(z|x,\theta^{t+1})}{f(z|x,\theta^t)} f(z|x,\theta^t)\,dz$$

$$= \log \int f(z|x,\theta^{t+1})\,dz = \log(1) = 0$$

PS, Q5.

The notation here is very confusing.

$\boxed{\{\underline{x}_n\}_{n=1}^{N} \text{ - SAMPLE.}}$ Each $\underline{x}_n$ is one datapoint

$\boxed{\underline{x}_n = (x_{n1}, \ldots x_{nD})}$ Each $\underline{x}_n$ is composed of D ~~(possibly dependent)~~ Bernoulli trials. (INDEPENDENT)

nope →

- eg, I flip a coin D times, and get a binary vector. The i^{th} flip had $P(x_{ni} = 1) = \mu_i$

  The probability of getting any particular sequence of results - any particular binary sequence - is

$$\boxed{P(\underbrace{\underline{x}_n}_{\text{Observed data.}} | \underbrace{\mu_1 \ldots \mu_D}_{\text{parameters}}) = \prod_{i=1}^{D} \underbrace{\mu_i}_{P(i^{th}\text{ flip is heads})}^{\overset{i^{th}\text{ flip is heads}}{x_i}} \underbrace{(1-\mu_i)}_{P(i^{th}\text{ flip is tails})}^{\overset{i^{th}\text{ flip is tails.}}{1-x_i}}}$$

Now, we have $K > 1$ component distributions, each with their own set of parameters $\{\mu_k\}_{k=1}^{K}$, with

$$\boxed{\mu_k = (\mu_{k1}, \ldots \mu_{kD})}$$

Define, for each x_n, $z_n = (0, \ldots, 1, \ldots 0)$ as the length-K binary (latent) vector indicating which group x_n came from. If $P(x_n \text{ came from group } k) = \pi_k$, then

$$\boxed{P(z_n | \pi) = \prod_{k=1}^{K} \pi_k^{z_{nk}}}$$

Finally denote the actual ~~density~~ distribution for $\underline{x}_n$ coming from group k as $\cancel{P(\underline{x}_n|\underline{z}_n)}$ $P(\underline{x}_n|\mu_k)$. So the distribution of $\underline{x}_n$ conditional on $\underline{z}_n$ is

$$P(\underline{x}_n|\underline{z}_n) = \prod_{k=1}^{K} P(\underline{x}_n|\mu_k)^{z_{nk}}$$

which just equals $P(\underline{x}_n|\mu_k)$ for the one and only k for which $z_{nk}=1$.

a) $P(\underline{x}_n|\mu,\pi) = \cancel{\sum_{\underline{z}} P(\underline{x}_n, \underline{z}_n)}$

$= \sum_{\underline{z}} P(\underline{x}_n, \underline{z}_n)$ ← Sum is taken over all K possible $\underline{z}_n$ vectors.

$$= \sum_{\underline{z}} P(\underline{x}_n|\underline{z}_n)P(\underline{z}_n)$$

$$= \sum_{\underline{z}} \left[\prod_{k=1}^{K} P(\underline{x}_n|\mu_k)^{z_{nk}} \times \prod_{k=1}^{K} \pi_k^{z_{nk}} \right]$$

$$= \sum_{\underline{z}} \prod_{k=1}^{K} \left(\pi_k P(\underline{x}_n|\mu_k)\right)^{z_{nk}}$$

Key step!

$$= \underset{\uparrow \atop z=(1,0,\cdots 0)}{\pi_1 P(\underline{x}_n|\mu_1)} + \underset{\uparrow \atop z=(0,1,\cdots,0)}{\pi_2 P(\underline{x}_n|\mu_2)} + \cdots + \underset{\uparrow \atop z=(0,0,\cdots 1)}{\pi_K P(\underline{x}_n|\mu_K)}$$

$$\boxed{= \sum_{k=1}^{K} \pi_k P(\underline{x}_n|\mu_k).}$$

b) $P(\underline{X}) = \prod_{n=1}^{N} P(\underline{x}_n)$

$= \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k P(\underline{x}_n|\mu_k)$

$$\boxed{\Rightarrow \ell(\mu,\pi|X) = \log P(\underline{X}) = \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k P(\underline{x}_n|\mu_k)\right)}$$

c) As stated in the problem, the complete-data dist^n for each $\underline{x}_n$ (ie its dist^n if we knew $\underline{z}_n$) is

$$P(x_n|\mu_k) = \prod_{i=1}^{D} \mu_i^{x_{ni}} (1-\mu_i)^{1-x_{ni}}$$

so $$P(x_n, z_n) = P(x_n|z_n)P(z_n) = \prod_{k=1}^{K}\left(\prod_{i=1}^{D} \mu_i^{x_{ni}} (1-\mu_i)^{1-x_{ni}}\right)^{z_{nk}} \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

and $$P(\underline{X},\underline{Z}) = \prod_{n=1}^{N} P(x_n, z_n)$$
$$= \prod_{n=1}^{N}\prod_{k=1}^{K}\left(\pi_k \prod_{i=1}^{D} \mu_i^{x_{ni}} (1-\mu_i)^{1-x_{ni}}\right)^{z_{nk}}$$

which gives $l(\mu, \pi | \underline{X}, \underline{Z}) = \log P(\underline{X}, \underline{Z})$

$$\boxed{= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left(\log \pi_k + \sum_{i=1}^{D}\left(x_{ni}\log\mu_i + (1-x_{ni})\log(1-\mu_i)\right)\right)}$$

d) Dist^n of $z_{nk} | \underline{X}, \mu^t, \pi^t$ ...

$\cancel{P(z_{nk} = 1 | \underline{X}) = \frac{P(\underline{X}, 1)}{}}$ Notation is hard!

$$P(z_{nk} = 1 | x_n) = P(z_n = (0, \ldots \overset{k}{1} \ldots 0) | x_n)$$
$$= \frac{P(x_n, (0, \ldots, 1, \ldots 0))}{P(x_n)}$$
$$\boxed{= \frac{\pi_k P(x_n|\mu_k)}{\sum_{g=1}^{K} \pi_g P(x_n|\mu_g)}}$$

The question did ask for $P(z_{nk} | \underline{X})$, not $\underline{x}_n$ .. ~~big~~ but the x's are independent, so others don't affect the answer.

Now because complete-data log-likelihood is linear in z_{nk}, $Q(\theta, \theta^t)$ is obtained by plugging in $\hat{z}$ for z.

e) These are just the weighted MLE's for the Bernoulli dist^n - you may attempt the calculus yourselves.

P6Q2

c) Recall that a covariance matrix for a vector-valued random variable satisfies

$$\Sigma_{ij} = Cov(X_i, X_j)$$

so

$$\Sigma_{ii} = Var(X_i)$$

Hence S_{ii} is a sample estimate of $Var(X_i)$. The sum of the variances of all the X_i then is

$$\sum_{i=1}^{P} S_{ii} = tr(S) \quad \text{(trace)}$$

But it is known for any matrix A with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D$ that

$$tr(A) = \sum_{d=1}^{D} \lambda_d.$$

It is also true that the variance of the d^{th} principal component is λ_d, the corresponding eigenvalue of S. Putting it all together gives

$$\sum_{d=1}^{D} \lambda_d / tr(S)$$

as the prop^n of total variance explained by the

P7 Q5

a) A saturated model has $\hat{p}_n = t_n$.

For a binomial* distribution, the distribution function of t_n is

*Binom(1,p) is Bern(p)

$$P(t_n = x) = p_n^x (1-p_n)^{1-x}$$

so the likelihood for p is

$$L(p) = \prod_{n=1}^{N} P(t_n | p) = \prod_{n=1}^{N} p_n^{t_n} (1-p_n)^{1-t_n}$$

and the log-likelihood,

$$l(p) = \sum_{n=1}^{N} t_n \log p_n + (1-t_n)\log(1-p_n)$$

For the saturated model, this gives

$$l_{sat}(p) = \sum_{n=1}^{N} t_n \log t_n + (1-t_n)\log(1-t_n)$$

But $t_n \in \{0,1\}$, so the above is not defined unless we fudge it and say $0 \log 0 = 0$. In that case, every term in the log-likelihood is

$$0 \log 0 + 1 \log 1 = 0.$$

b) Just plug $\hat{t}$ into the above formula. (see below)

c) $$D = 2(\underbrace{l_{sat}(p)}_{=0} - l_{model}(p))$$

$$= -2 l_{model}(p)$$
$$= -2 \sum_{n=1}^{N} t_n \log \hat{p}_n + (1-t_n)\log(1-\hat{p}_n)$$

d) $AIC = -2 l_{model}(p) + 2d$, $d = \#$ parameters

$BIC = -2 l_{model}(p) + d \log N$.

Forward Stepwise selection:

i) Start with the null model, or the smallest model you would be willing to accept.

ii) Do:
- Add each available feature separately into the model, and calculate the AIC/BIC for the resulting (more complex) model

- Choose the feature that yields the largest decrease in AIC/BIC, and add it to the model permenently

- repeat.

UNTIL:
- AIC/BIC stops decreasing OR
- AIC/BIC " " "too much" OR
- All of the features have been added.

P7Q3

Let $y = g(x)$. Note that $G_i(y) = P(Y_i < y)$

$$= P(g(T_i) < g(x))$$
$$= P(T_i < x)$$
$$= F_i(x)$$

although this isn't quite enough to prove the statement; it is suggestive though.

We can write $F_1(x) - F_0(x) = \int_{-\infty}^{x} (f_1(s) - f_0(s))\,ds$.

Now,

$$G_1(x) - G_0(x) = \int_{-\infty}^{x} (g_1(s) - g_0(s))\,ds$$

$$= \int_{-\infty}^{g^{-1}(x)} (f_1(g^{-1}(s)) - f_0(g^{-1}(s)))\, g^{-1\prime}(s)\,ds.$$

$$= \int_{-\infty}^{u} (f_1(u) - f_0(u))\,du.$$

$$= F_1(u) - F_0(u)$$

$$\Rightarrow \max_x |G_1(x) - G_0(x)| = \max_u |F_1(u) - F_0(u)|,$$

So $KS = KS^+$.