Statistical learning homework 5

Dai Yuehao (1800010660@pku.edu.cn)

December 31, 2020

1 ESL 14.2

1.1 1

The likelihood function for the data set $\{x_i\}_{i=1}^N$ is given by

$$l = \prod_{i=1}^{N} g(x_i) = \prod_{i=1}^{N} \left[\sum_{k=1}^{K} \pi_k g_k(x_i) \right] = \prod_{i=1}^{N} \left[\sum_{k=1}^{K} \frac{1}{\sqrt{(2\pi)^p \sigma^2 |L|}} \exp\left\{ -\frac{\pi_k}{2\sigma^2} (x_i - \mu_k)^T L^{-1} (x_i - \mu_k) \right\} \right].$$

1.2 2

Algorithm 1 EM Algorithm for k-component Gaussian Mixture.

1: Take initial guesses for the parameters $\hat{\mu}_k, \hat{\sigma}^2, \hat{\pi}$

2: **while** Not convergence **do**

3: Compute the responsibilities

$$\hat{\gamma}_{i,k} = \frac{\hat{\pi}_k \phi_{\hat{\theta}_k}(x_i)}{\sum_{k=1}^{K} \hat{\pi}_k \phi_{\hat{\theta}_k}(x_i)}, \quad i = 1, \cdots, N.$$

4: Compute the weighted means and variances

$$\hat{\mu}_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{i,k} y_k}{\hat{\gamma}_{i,k}}, \quad \hat{\sigma}^2 = \sum_{k=1}^{K} \hat{\pi}_k \frac{\sum_{i=1}^{N} \hat{\gamma}_{i,k} (y_i - \hat{\mu}_k)^2}{\sum_{i=1}^{N} \hat{\gamma}_{i,k}}, \quad \hat{\pi}_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{i,k}}{N}.$$

5: **end while**

1.3 3

If we take $\sigma \to 0$, then the density function of ϕ_{θ_k} will concentrate at μ_k, that means

$$\lim_{\sigma \to 0} \frac{\hat{\pi}_k \phi_{\hat{\theta}_k}(x_i)}{\sum_{k=1}^{K} \hat{\pi}_k \phi_{\hat{\theta}_k}(x_i)} = \hat{\pi}_k$$

if $\hat{\mu}_k$ is the nearest point of x_i among $\{\hat{\mu}_k\}_{k=1}^K$, other wise $\hat{\gamma}_{i,k} \to 0$, hence the weighted means

$$\hat{\mu}_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{i,k} y_k}{\hat{\gamma}_{i,k}}$$

is approximately the weighted means of all the points whose nearest neighbor is $\hat{\mu}_k$, and

$$\hat{\pi}_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{i,k}}{N}$$

is approximately the portion of each class. Thereby the EM algorithm coincides with K-means clustering.

2 ESL 14.7

The optimization problem is

$$\min_{\mu,\lambda,V_q} \quad \text{tr}\left[(x-\mu\cdot 1-V_q\lambda)^T(x-\mu\cdot 1-V_q\lambda)\right],$$

then we take the derivatives of all the variables and set them to zero we have

$$N\mu+V_q\left(\sum_{i=1}^N\lambda_i\right)-\sum_{i=1}^N x_i=0,\quad V_q^T\left(V_q\lambda_i+\mu-x_i\right)=0,$$

we can see that

$$\hat{\mu}=\bar{x},\quad \hat{\lambda}_i=V_q^T(x_i-\bar{x})$$

is obviously one of solution of the equations above. On the other hand, we must have

$$\lambda_i=V_q^T(x_i-\mu)$$

since $V_q^TV_q=I$, then we have

$$\mu=\bar{x}-\frac{1}{N}V_qV^T\left(N\bar{x}-N\mu\right),$$

hence

$$(I-V_qV_q^T)(\mu-\bar{x})=0,$$

all $\hat{\mu}$ that satisfies $(I-V_qV_q^T)(\mu-\bar{x})=0$ will be the solution, hence it is not unique.

3 ESL 14.21

Suppose for f we have $Lf=0$, then

$$f^TLf=\frac{1}{2}\sum_{k=1}^m\sum_{i\in A_k}\sum_{j\in A_k}w_{ij}(f_i-f_j)^2=0,$$

hence we have

$$f=\sum_{k=1}^m t_kI_{A_k}$$

where I_{A_k} is the indicator vectors of component A_k. On the other hand from the structure of L, f indeed satisfies $Lf=0$. Finally, if $k=1$, then $\text{rank}(L)=N-1$, hence when $k>1$, we have $\text{rank}(L)=N-k$, it is obvious that $\{I_{A_k},k=1,\cdots,m\}$ are linearly independent, thereby we have concluded the proof.

4 ESL 14.23

4.1 a

Since $\log(x)$ is concave, then in terms of Jensen's Inequality we have

$$\sum_{k=1}^r c_k\log\left(\frac{y_k}{c_k}\right)\le\log\left(\sum_{k=1}^r c_k\cdot\frac{y_k}{c_k}\right)=\log\left(\sum_{k=1}^r y_k\right).$$

Hence let

$$c_k=\frac{a_{ikj}^s}{b_{ij}^s}$$

we have

$$\log\left(\sum_{k=1}^r w_{ik}h_{kj}\right)\ge\sum_{k=1}^r\frac{a_{ikj}^s}{b_{ij}^s}\log\left(\frac{b_{ij}^s}{a_{ikj}^s}w_{ik}h_{kj}\right).$$

4.2 b

Since $\log(x)$ is concave, then in terms of Jensen's Inequality we have

$$\sum_{k=1}^{r} c_k \log\left(\frac{y_k}{c_k}\right) \leq \log\left(\sum_{k=1}^{r} c_k \cdot \frac{y_k}{c_k}\right) = \log\left(\sum_{k=1}^{r} y_k\right).$$

Hence let

$$c_k = \frac{a_{ikj}^s}{b_{ij}^s}$$

we have

$$\log\left(\sum_{k=1}^{r} w_{ik}h_{kj}\right) \geq \sum_{k=1}^{r} \frac{a_{ikj}^s}{b_{ij}^s} \log\left(\frac{b_{ij}^s}{a_{ikj}^s} w_{ik}h_{kj}\right).$$

We only need to show that

$$\begin{aligned}
g(W,H|W^s,H^s) &= \sum_{i=1}^{N}\sum_{j=1}^{p}\sum_{k=1}^{r} x_{ij}\frac{w_{ik}^s h_{kj}^s}{b_{ij}^s}\left[\log\left(\frac{b_{ij}^s}{a_{ikj}^s} w_{ik}h_{kj}\right) - \log\left(\frac{b_{ij}^s}{a_{ikj}^s}\right)\right] - \sum_{i=1}^{N}\sum_{j=1}^{p}\sum_{k=1}^{r} w_{ik}h_{kj} \\
&\leq \sum_{i=1}^{N}\sum_{j=1}^{p}\left[x_{ij}\log\left(\sum_{k=1}^{r} w_{ik}h_{kj}\right) - \sum_{k=1}^{r} w_{ik}h_{kj} - \sum_{k=1}^{r} x_{ij}\frac{a_{ikj}^s}{b_{ij}^s}\log\left(\frac{b_{ij}^s}{a_{ikj}^s}\right)\right] \\
&= L(W,H) - \sum_{i=1}^{N}\sum_{j=1}^{p}\sum_{k=1}^{r} x_{ij}\frac{a_{ikj}^s}{b_{ij}^s}\log\left(\frac{b_{ij}^s}{a_{ikj}^s}\right),
\end{aligned}$$

the equation holds if and only if $W = W^s, H = H^s$. Hence we conclude the proof.

4.3 c

Take the derivative of w_{ik} and set it to zero we have

$$\frac{\partial g}{\partial w_{ik}} = \sum_{j=1}^{p} x_{ij}\frac{a_{ikj}^s}{b_{ij}^s w_{ik}} - \sum_{j=1}^{p} h_{kj} = 0,$$

hence

$$w_{ik} = \frac{\sum_{j=1}^{p} x_{ij}\frac{a_{ikj}^s}{b_{ij}^s}}{\sum_{j=1}^{p} h_{kj}},$$

now to see

$$x_{ij}\frac{a_{ikj}^s}{b_{ij}^s} = x_{ij}w_{ik}^s h_{kj}^s/(WH)_{ij}$$

we have

$$w_{ik} = \frac{\sum_{j=1}^{p} x_{ij}\frac{a_{ikj}^s}{b_{ij}^s}}{\sum_{j=1}^{p} h_{kj}} = w_{ik}^s \frac{\sum_{j=1}^{p} x_{ij}h_{kj}^s/(WH)_{ij}}{\sum_{j=1}^{p} h_{kj}}.$$

The procedure of h_{ik} is exactly the same as w_{ik}.

5 ESL 15.2

We see that for a single sample x_i, the probability of it is not contained by a bootstrap sample is about 1/3, for the rest $N-1$ samples, as the number of bootstrap samples B gets large, the trees that use bootstrap samples not containing x_i will almost cover the rest $N-1$ samples, hence they form a forest trained with the rest $N-1$ samples. Hence the OOB error estimate for a random forest approaches its N-fold CV error estimate, and that in the limit, the identity is exact.

6 Wainwrigth 11.2

The log-likelihood function can be written as

$$l(\Theta|x) = \frac{n}{2}\log\det(\Theta) - \frac{pN}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n} x_i^T \Theta x_i,$$

then Θ should be invertible in order to maximize l, then we take the derivative of Θ and set it to zero we have

$$\frac{\partial l}{\partial \Theta} = \frac{n}{2}\Theta^{-1} - \frac{1}{2}\sum_{i=1}^{n} x_i x_i^T = 0 \quad \Longrightarrow \Theta^{-1} = \hat{\Sigma},$$

if $\hat{\Sigma} \succ 0$ then $\hat{\Theta}_{\text{MLE}} = \hat{\Sigma}^{-1}$. If $\hat{\Sigma}$ is singular, there must exist a semidefinite matrix Θ_0 that satisfies $\sum_{i=1}^{n} x_i^T \Theta_0 x_i = 0$, since

$$\frac{1}{n}\sum_{i=1}^{n} x_i^T \Theta_0 x_i = \text{tr}(\hat{\Sigma}\Theta_0),$$

we simply choose θ_0 such that $\hat{\Sigma}\theta = 0$, then let $\Theta_0 = \theta\theta^T$ we have $\text{tr}(\hat{\Sigma}\Theta_0) = 0$. Now consider

$$\Theta_0 = Q\text{diag}\{1, 0, \cdots, 0\}Q^T$$

where Q is orthogonal, then let

$$\Theta_n = n\Theta_0 + QQ^T,$$

then $\text{tr}(\Theta_n\hat{\Sigma}) = \text{tr}(\Theta_1\hat{\Sigma})$, whereas $\det(\Theta_n) = n$, hence the log-likelihood can diverge to infinity.

7 Wainwrigth 11.7

7.1 a

We have

$$\Phi(\theta) = \log\left(\sum_{X\in\{-1,1\}^d} \exp\left\{\sum_{(j,k)\in E} \theta_{jk}x_j x_k\right\}\right),$$

hence

$$\frac{\partial\Phi(\theta)}{\partial\theta_{jk}} = \frac{1}{\Phi(\theta)}\sum_{X\in\{-1,1\}^d} x_j x_k \exp\left\{\sum_{(j,k)\in E} \theta_{jk}x_j x_k\right\} = \mathbb{E}_\theta[X_j X_k].$$

7.2 b

We have

$$p_\theta\left(x_j|X_{\setminus\{j\}}\right) = \frac{p_\theta(X)}{p_\theta\left(X_{\setminus\{j\}}\right)} = \frac{\exp\left\{\sum_{(i,k)\in E}\theta_{ik}x_i x_k\right\}}{\sum_{X_j\in\{-1,1\}}\exp\left\{\sum_{(i,k)\in E}\theta_{ik}x_i x_k\right\}},$$

which can also be written as

$$p_\theta\left(x_j|X_{\setminus\{j\}}\right) = \frac{\exp\left\{2\sum_{(j,k)\in E}\theta_{jk}x_j x_k\right\}}{1+\exp\left\{2\sum_{(j,k)\in E}\theta_{jk}x_j x_k\right\}},$$

hence

$$p_\theta\left(x_j|X_{\setminus\{j\}}\right) = \left.\frac{\partial f}{\partial t}\right|_{t=\exp\left\{2\sum_{(j,k)\in E}\theta_{jk}x_j x_k\right\}}.$$

8 Partial Correlations

Since changing mean and variance will not change the correlation coefficient of random variables, we can simply assume that i, j, k have zero mean and variance 1. Assume that

$$\rho_{i,h|K\setminus\{h\}} = a, \quad \rho_{j,h|K\setminus\{h\}} = b,$$

then there exists x, y such that

$$i|K\setminus\{h\} = ah|K\setminus\{h\} + \sqrt{1-a^2}x, \quad j|K\setminus\{h\} = bh|K\setminus\{h\} + \sqrt{1-b^2}y,$$

where

$$\mathbb{E}[x] = \mathbb{E}[y] = 0, \quad \mathrm{Var}[x] = \mathrm{Var}[y] = 1, \quad \rho_{x,h|K\setminus\{h\}} = \rho_{y,h|K\setminus\{h\}} = 0,$$

then

$$\rho_{i,j|K\setminus\{h\}} = ab + \sqrt{(1-a^2)(1-b^2)}\rho_{x,y}|K, \quad \rho_{i,j|} = \rho_{x,y}|K,$$

hence

$$\rho_{i,j|K\setminus\{h\}} - \rho_{i,h|K\setminus\{h\}} - \rho_{j,h|K\setminus\{h\}} = \sqrt{(1-a^2)(1-b^2)}\rho_{x,y}|K,$$

then we have concluded our proof.