STAT409 Final

2021320322 윤희서

1.

$P\{Yf(x)<0\}$

$= E[P(Y)\mathbb{1}\{Yf(X)<0\} \mid X=x]$ LOTP

$= E[P(Y=1)P(Y=1|X=x)\mathbb{1}\{f(x)<0\}$
$\quad + P(Y=-1)P(Y=-1|X=x)\mathbb{1}\{f(x)>0\}]$

$= E[(1-\pi)p(x)\mathbb{1}\{f(x)<0\} + \pi(1-p(x))\mathbb{1}\{f(x)>0\}]$

$= E[(1-\pi)p(x)\mathbb{1}\{f(x)<0\}] + E[\pi(1-p(x))\mathbb{1}\{f(x)>0\}]$

$= E[(1-\pi)p(x)\mathbb{1}\{f(x)\le 0\}] + E[\pi(1-p(x))\mathbb{1}\{f(x)\ge 0\}]$

위 식을 최소화하는 $f(x)$의 조건은

$$\begin{cases} f(x)<0 & \text{if } (1-\pi)p(x)<\pi(1-p(x)) \\ f(x)=0 & \text{if } (1-\pi)p(x)=\pi(1-p(x)) \\ f(x)>0 & \text{if } (1-\pi)p(x)>\pi(1-p(x)) \end{cases}$$

$$\Leftrightarrow \begin{cases} f(x)<0 & \text{if } p(x)<\pi \\ f(x)=0 & \text{if } p(x)=\pi \\ f(x)>0 & \text{if } p(x)>\pi \end{cases}$$

$$\therefore \operatorname{sign}\{f^*(x)\} = \begin{cases} -1 & \text{if } p(x)-\pi<0 \\ 0 & \text{if } p(x)-\pi=0 \\ 1 & \text{if } p(x)-\pi>0 \end{cases}$$

$$\Leftrightarrow \operatorname{sign}\{f^*(x)\} = \begin{cases} -1 & \text{if } \operatorname{sign}\{p(x)-\pi\}=-1 \\ 0 & \text{if } \operatorname{sign}\{p(x)-\pi\}=0 \\ 1 & \text{if } \operatorname{sign}\{p(x)-\pi\}=1 \end{cases}$$

$$\therefore \operatorname{sign}\{f^*(x)\} = \operatorname{sign}\{p(x)-\pi\}$$

2.

(a)

$$L(\mu|X_1,\cdots,X_n)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x_i-\mu)^2} \quad (\because X_i's \text{ are independent.})$$

$$=\left(\frac{1}{\sqrt{2\pi}}\right)^n\prod_{i=1}^{n}e^{-\frac{1}{2}(x_i-\mu)^2}$$

$$\log L=-\frac{n}{2}\log(2\pi)-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{\partial \log L}{\partial \mu}=\sum_{i=1}^{n}(x_i-\mu)\overset{\text{set}}{=}0 \Rightarrow \frac{1}{n}\sum_{i=1}^{n}x_i=\mu$$

$$\frac{\partial^2 \log L}{\partial \mu^2}=-n<0 \Rightarrow \log L : \text{concave}$$

$$\therefore \hat{\mu}^{MLE}=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}_n$$

(b)

$$P\left(-z_{0.025}\le\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\le z_{0.025}\right)=0.95 \text{ and } \overline{X}_n=12,\ \sigma=1,\ n=100$$

$$\Rightarrow P\left(\overline{X}_n-z_{0.025}\frac{\sigma}{\sqrt{n}}\le\mu\le\overline{X}_n+z_{0.025}\frac{\sigma}{\sqrt{n}}\right)$$

$$\approx P\left(12-1.96\times\frac{1}{10}\le\mu\le12+1.96\times\frac{1}{10}\right)$$

$\therefore$ 95% CI for μ: $(11.804, 12.196)$

(c)

$$\text{Prior } \pi(\mu) = \frac{1}{100\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{\mu^2}{100^2}}$$

$$\propto e^{-\frac{1}{2\cdot 100^2}\cdot\mu^2}$$

$$\text{Likelihood } L(x|\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{1}{2}(x_i-\mu)^2}$$

$$\propto e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2} = e^{-\frac{1}{2}\left(\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + n\mu^2\right)}$$

$$\propto e^{-\frac{1}{2}(-2\mu n\bar{x}_n + n\mu^2)}$$

$$\text{Posterior } f(\mu|x_1,\cdots,x_n) \propto L(x|\mu)\pi(\mu)$$

$$\propto e^{-\frac{1}{2\cdot 100^2}\mu^2}\cdot e^{-\frac{1}{2}(-2\mu n\bar{x}+n\mu^2)}$$

$$= e^{-\frac{1}{2}\left(\frac{1}{100^2}+n\right)\mu^2 + n\bar{x}_n\mu}$$

$$\propto e^{-\frac{1}{2}\left(\frac{1}{100^2}+n\right)\left\{\mu^2 - \frac{2n\bar{x}_n}{\left(\frac{1}{100^2}+n\right)}\mu + \frac{n^2\bar{x}^2}{\left(\frac{1}{100^2}+n\right)^2}\right\}}$$

$$= e^{-\frac{1}{2}\left(\frac{1}{100^2}+n\right)\left(\mu - \frac{n\bar{x}_n}{\frac{1}{100^2}+n}\right)^2}$$

$$\Longleftrightarrow \mu|X_1,\cdots,X_n \sim N\left(\frac{n\bar{x}_n}{\frac{1}{100^2}+n}, \left(\frac{1}{100^2}+n\right)^{-1}\right)$$

(d)

$$P\left(-z_{0.025} \le \frac{\mu - \frac{n\bar{X}_n}{\frac{1}{100^2}+n}}{\left(\frac{1}{100^2}+n\right)^{-\frac{1}{2}}/\sqrt{n}} \le z_{0.025} \,\middle|\, X_1,\cdots,X_n\right) = 0.95$$ 에서 $\bar{X}_n = 12,\ n = 100$

$$\Rightarrow P\left(-1.96 \le \frac{\mu - \frac{100\times 12}{\frac{1}{100^2}+100}}{\left(\frac{1}{100^2}+100\right)^{-\frac{1}{2}}/\sqrt{100}} \le 1.96\right) \simeq P\left(-1.96 \le \frac{\mu-12}{0.01} \le 1.96\right)$$

$$= P(12-1.96\times 0.01 \le \mu \le 12+1.96\times 0.01)$$

$\therefore$ 95% Credible interval for μ: $(11.9804, 12.0196)$

3.

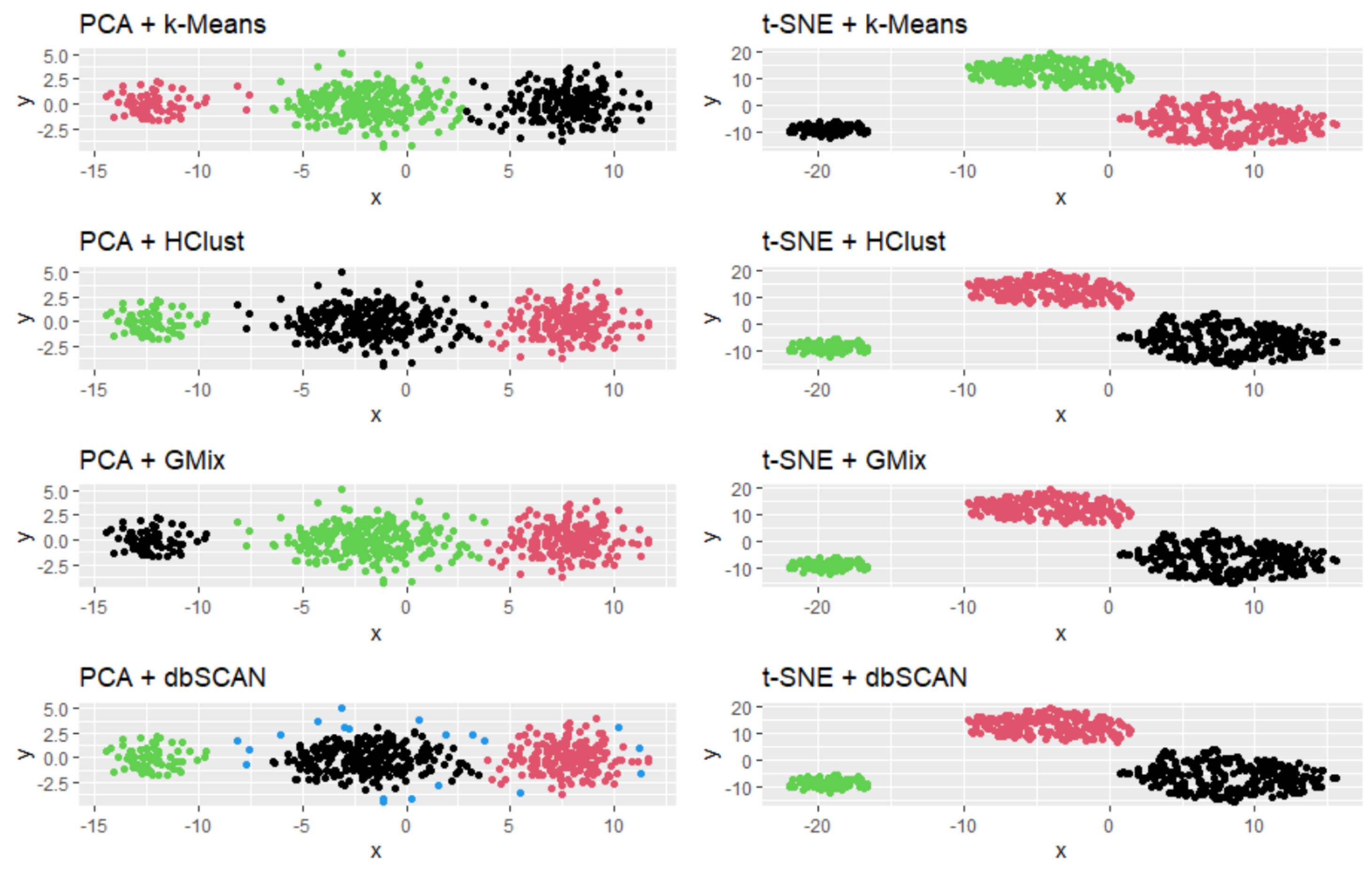
	Accuracy
Logistic Regression	0.8655
LASSO-penalized Logistic Regression	0.8930
Linear SVM	0.8870
Gaussian Kernel SVM	0.8715
Classification Tree	0.8715
Random Forest	0.8845
Logit Boosting	0.8685

LASSO-penalized Logistic Regression의 Accuracy가 가장 높다.
실험 이전에는 앙상블 모델의 성능이 더 높을 것으로 예상하였으나 실제로 보고된 실험 결과는 그렇지 않았다. 이러한 결과가 나온 이유는 Training set의 크기가 500으로, 앙상블 모델의 성능이 높게 나오기에는 충분한 크기의 Training set이 아니었기 때문이라고 추측할 수 있다.
앙상블 모델은 여러 개의 개별 모델의 에측을 결합하여 더 좋은 예측을 하기 위한 모델인데, 여기에서 얻을 수 있는 다양성을 활용하기 위해서는 각각의 개별 모델이 다른 샘플과 특징을 학습할 수 있도록 충분한 양의 데이터가 필요하다.
이러한 이유로 Training set이 작은 경우에는 단일 모델을 사용했을 때 더 좋은 성능이 보고될 수 있다.

Tuning을 진행한 Hyperparameter는 다음과 같다.
Logistic Regression: 없음
LASSO-penalized Logistic Regression: lambda
Linear SVM: C
Gaussian Kernel SVM: C, sigma
Classification Tree: size
Random Forest: mtry, ntree
Logit Boosting: n.trees

(STAT409_Final_2021320322.R에 과정 첨부)

4.



(STAT409_Final_2021320322.R에 과정 첨부)