Bayesian Approach of Fisher Discriminant Analysis

Machine Learning Algorithms

2021024966 김희성
2021031685 유성민

1 Description about FDA

FDA(Fisher Discriminant Analysis)는 classification의 일종으로, 두 개의 class를 가장 잘 나누는 벡터 $\mathbf{w}$를 찾는 방법이다. 두 class 간의 between-class covariance는 크게, class 내부의 within-class covariance는 작게 만드는 것을 목표로 한다.

1.1 Loss function

두 개의 mutually exclusive class를 각각 $\mathcal{C}_1$, $\mathcal{C}_2$라 하고, 그 class의 원소의 개수를 각각 N_1, N_2 이라 하자. 각 class의 mean $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$과 covariance Σ_1, Σ_2는 data $\mathbf{x}_n \in \mathbb{R}^D$에 대해 다음과 같다.

$$\begin{aligned}
\boldsymbol{\mu}_1 &= \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} \mathbf{x}_n \\
\boldsymbol{\mu}_2 &= \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} \mathbf{x}_n \\
\Sigma_1 &= \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^\top \\
\Sigma_2 &= \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^\top
\end{aligned} \tag{1}$$

전체 data의 개수를 N이라 하면, $N = N_1 + N_2$ 이므로 전체 data의 mean $\boldsymbol{\mu}$는 다음과 같다.

$$\begin{aligned}
\boldsymbol{\mu} &= \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \\
&= \frac{1}{N} (N_1 \boldsymbol{\mu}_1 + N_2 \boldsymbol{\mu}_2)
\end{aligned} \tag{2}$$

여기서 target t_n을 다음과 같이 정의한다.

$$t_n = \begin{cases} \dfrac{N}{N_1} & (n \in \mathcal{C}_1) \\ -\dfrac{N}{N_2} & (n \in \mathcal{C}_2) \end{cases} \tag{3}$$

따라서 $\sum_n t_n = 0$ 이며, 이는 class label에 가중치를 두어 평균을 0으로 설정한 것으로 생각할 수 있다.

그러면, Loss function을 다음과 같이 정의할 수 있다. (Duda and Hart, 1973)

$$L = \frac{1}{2} \sum_{n=1}^{N} (\mathbf{w}^\top \mathbf{x}_n + w_0 - t_n)^2 \tag{4}$$

여기서 data $\mathbf{x}_n$ 을 centralize 하면 ($\mathbf{x}_n \leftarrow \mathbf{x}_n - \boldsymbol{\mu}$) 새로운 mean은 $\boldsymbol{\mu} = 0$이고, bias $w_0 = -\mathbf{w}^\top \boldsymbol{\mu} = 0$ 이 되어 간단히 다음과 같이 쓸 수 있다.

$$L = \frac{1}{2} \sum_{n=1}^{N} (\mathbf{w}^\top \mathbf{x}_n - t_n)^2 \tag{5}$$

2 Bayesian Approach

2.1 Likelihood

일반성을 잃지 않고, 모든 data는 centralized 되어 있다고 가정한다. 그러면 식 (5) 를 사용하여 다음과 같은 likelihood를 구성할 수 있다.

$$\begin{aligned} p(t|\mathbf{x}, \mathbf{w}) &= \mathcal{N}(t|\mathbf{w}^\top \mathbf{x}, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{w}^\top \mathbf{x} - t)^2\right) \end{aligned} \tag{6}$$

2.2 Prior

parameter $\mathbf{w}$에 대한 prior은 다음과 같다.

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \sigma_{\mathbf{w}}^2 \mathbf{I}) \tag{7}$$

2.3 Posterior

posterior $p(\mathbf{w}|\mathcal{D})$ 는 Bayes' rule에 의해 다음과 같이 구할 수 있다.

$$\begin{aligned} p(\mathbf{w}|\mathcal{D}) &= p(\mathbf{w})p(\mathcal{D}|\mathbf{w}) \\ &= \prod_{n=1}^{N} p(\mathbf{w})p(t_n|\mathbf{x}_n, \mathbf{w}) \\ &= C \ \exp(①) \end{aligned} \tag{8}$$

likelihood와 prior 모두 gaussian 이므로, posterior 또한 gaussian 이다. 따라서, exp 내부 항만 전개하면 다음과 같다.

$$\begin{aligned} ① &= -\frac{1}{2\sigma^2}\left(\sum_{n=1}^{N}\left(\mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2t_n \mathbf{w}^\top \mathbf{x}_n + t_n^2\right) + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2}\mathbf{w}^\top \mathbf{w}\right) \\ &= -\frac{1}{2\sigma^2}\left(\sum_{n\in\mathcal{C}_1}\left(\mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} - 2\frac{N}{N_1}\mathbf{w}^\top \mathbf{x}_n + \frac{N^2}{N_1^2}\right) + \sum_{n\in\mathcal{C}_2}\left(\mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} + 2\frac{N}{N_2}\mathbf{w}^\top \mathbf{x}_n + \frac{N^2}{N_2^2}\right) + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2}\mathbf{w}^\top \mathbf{w}\right) \\ &= -\frac{1}{2\sigma^2}\left(\mathbf{w}^\top\left(\sum_{n\in\mathcal{C}_1}\mathbf{x}_n \mathbf{x}_n^\top + \sum_{n\in\mathcal{C}_2}\mathbf{x}_n \mathbf{x}_n^\top + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2}\mathbf{I}\right)\mathbf{w} - 2N\mathbf{w}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right) + C \end{aligned} \tag{9}$$

여기서 *within-class* covariance $\mathbf{S}_\mathrm{W}$와 *between-class* covariance $\mathbf{S}_\mathrm{B}$ 를 다음과 같이 정의하자.

$$\begin{aligned} \mathbf{S}_\mathrm{W} &= \sum_{n\in\mathcal{C}_1}(\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^\top + \sum_{n\in\mathcal{C}_2}(\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^\top \\ \mathbf{S}_\mathrm{B} &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \end{aligned} \tag{10}$$

그러면, 다음이 성립한다.

$$\begin{aligned}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top &= \sum_{n\in\mathcal{C}_1}\mathbf{x}_n\mathbf{x}_n^\top + \sum_{n\in\mathcal{C}_2}\mathbf{x}_n\mathbf{x}_n^\top \\ &= \mathbf{S}_\mathrm{W} + N_1\boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top + N_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top \\ &= \mathbf{S}_\mathrm{W} + \frac{N_1N_2}{N}\mathbf{S}_\mathrm{B}\end{aligned} \tag{11}$$

따라서, ① 은 다음과 같다.

$$\begin{aligned}① &= -\frac{1}{2\sigma^2}\left(\mathbf{w}^\top\left(\mathbf{S}_\mathrm{W} + \frac{N_1N_2}{N}\mathbf{S}_\mathrm{B} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)\mathbf{w} - 2N\mathbf{w}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right) + C \\ &= -\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}})^\top\Sigma_{\mathbf{w}|\mathcal{D}}^{-1}(\mathbf{w} - \boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}}) + C\end{aligned} \tag{12}$$

즉, posterior $p(\mathbf{w}|\mathcal{D})$ 는 다음과 같다.

$$p(\mathbf{w}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}}, \Sigma_{\mathbf{w}|\mathcal{D}}) \tag{13}$$

여기서 $\boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}}$와 $\Sigma_{\mathbf{w}|\mathcal{D}}$는 다음과 같다.

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}} &= N\left(\mathbf{S}_\mathrm{W} + \frac{N_1N_2}{N}\mathbf{S}_\mathrm{B} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ \Sigma_{\mathbf{w}|\mathcal{D}} &= \sigma^2\left(\mathbf{S}_\mathrm{W} + \frac{N_1N_2}{N}\mathbf{S}_\mathrm{B} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}\end{aligned} \tag{14}$$

2.4 Predictive Distribution

새로운 input data $\mathbf{x}$에 대하여 predictive distribution $p(t|\mathbf{x}, \mathcal{D})$ 는 likelihood와 posterior의 곱을 적분하여 구할 수 있다.

$$p(t|\mathbf{x}, \mathcal{D}) = \int_\mathcal{W} p(t|\mathbf{x}, \mathbf{w})p(\mathbf{w}|\mathcal{D})\,\mathrm{d}\mathbf{w} \tag{15}$$

이 또한 gaussian이므로, exp 내부 항만 보도록 하자. 또한, $\mathbf{w}$에 대해 적분한 후의 결과는 t에 대한 probability distribution 이므로, $p(t|\mathbf{x}, \mathbf{w})p(\mathbf{w}|\mathcal{D})$의 exp 내의 항을 다음과 같이 정리할 수 있다.

$$\begin{aligned}&-\frac{1}{2\sigma^2}\left(\mathbf{w}^\top\left(\mathbf{S}_\mathrm{W} + \frac{N_1N_2}{N}\mathbf{S}_\mathrm{B} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)\mathbf{w} - 2N\mathbf{w}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + (\mathbf{w}^\top\mathbf{x} - t)^2\right) + \cdots \\ =&-\frac{1}{2\sigma^2}\left(\mathbf{w}^\top\underbrace{\left(\mathbf{x}\mathbf{x}^\top + \mathbf{S}_\mathrm{W} + \frac{N_1N_2}{N}\mathbf{S}_\mathrm{B} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)}_{\Delta}\mathbf{w} - 2N\mathbf{w}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 + t\mathbf{x}/N) + t^2\right) + \cdots\end{aligned} \tag{16}$$

여기서 $\mathbf{w}$를 marginalize out 시키면 t에 대한 gaussian distribution 을 얻을 수 있고, 그것의 exp 내부 항은 다음과 같다.

$$\begin{aligned}&-\frac{1}{2\sigma^2}\left((1 - \mathbf{x}^\top\Delta^{-1}\mathbf{x})t^2 - 2N\mathbf{x}^\top\Delta^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)t\right) + \cdots \\ =&\frac{(1 - \mathbf{x}^\top\Delta^{-1}\mathbf{x})}{2\sigma^2}\left(t - \frac{N\mathbf{x}^\top\Delta^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{(1 - \mathbf{x}^\top\Delta^{-1}\mathbf{x})}\right)^2\end{aligned} \tag{17}$$

따라서, prediction의 mean과 variance는 다음과 같다.

$$\mathbb{E}[t|\mathbf{x},\mathcal{D}] = \frac{N\mathbf{x}^\top \left(\mathbf{x}\mathbf{x}^\top + \mathbf{S}_\mathrm{W} + \frac{N_1 N_2}{N}\mathbf{S}_\mathrm{B} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{1 - \mathbf{x}^\top \left(\mathbf{x}\mathbf{x}^\top + \mathbf{S}_\mathrm{W} + \frac{N_1 N_2}{N}\mathbf{S}_\mathrm{B} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}\mathbf{x}} \tag{18}$$

$$\mathrm{Var}[t|\mathbf{x},\mathcal{D}] = \frac{\sigma^2}{1 - \mathbf{x}^\top \left(\mathbf{x}\mathbf{x}^\top + \mathbf{S}_\mathrm{W} + \frac{N_1 N_2}{N}\mathbf{S}_\mathrm{B} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}\mathbf{x}}$$

여기서 다음과 같은 행렬 $\mathbf{X}$를 도입하자.

$$\mathbf{X} = \begin{pmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_N \\ | & & | \end{pmatrix} \in \mathbb{R}^{D\times N} \tag{19}$$

그러면 다음이 성립한다.

$$\begin{aligned} \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top &= \mathbf{S}_\mathrm{W} + \frac{N_1 N_2}{N}\mathbf{S}_\mathrm{B} \\ &= \mathbf{X}\mathbf{X}^\top \end{aligned} \tag{20}$$

따라서, 식 (18) 은 다음과 같이 간단히 쓸 수 있다.

$$\mathbb{E}[t|\mathbf{x},\mathcal{D}] = \frac{N\mathbf{x}^\top \left(\mathbf{x}\mathbf{x}^\top + \mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{1 - \mathbf{x}^\top \left(\mathbf{x}\mathbf{x}^\top + \mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}\mathbf{x}} \tag{21}$$

$$\mathrm{Var}[t|\mathbf{x},\mathcal{D}] = \frac{\sigma^2}{1 - \mathbf{x}^\top \left(\mathbf{x}\mathbf{x}^\top + \mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}\mathbf{x}}$$

식 (21) 의 분모를 비교해보면 같다는 것을 알 수 있다. 이를 Woodbury identity를 이용하여 정리하면 다음과 같다.

$$(\text{denom}) = \left(1 + \mathbf{x}^\top \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}\mathbf{x}\right)^{-1} \tag{22}$$

$$= \left(1 + \frac{\sigma_\mathbf{w}^2}{\sigma^2}\left(\mathbf{x}^\top\mathbf{x} - \mathbf{x}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\right)\mathbf{X}^\top\mathbf{x}\right)\right) \tag{23}$$

식 (23) 를 variance에 plugin 하면 다음과 같다.

$$\begin{aligned} \mathrm{Var}[t|\mathbf{x},\mathcal{D}] &= \sigma^2 + \sigma_\mathbf{w}^2\mathbf{x}^\top\mathbf{x} - \sigma_\mathbf{w}^2\mathbf{x}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{x} \\ &\geq \sigma^2 \end{aligned} \tag{24}$$

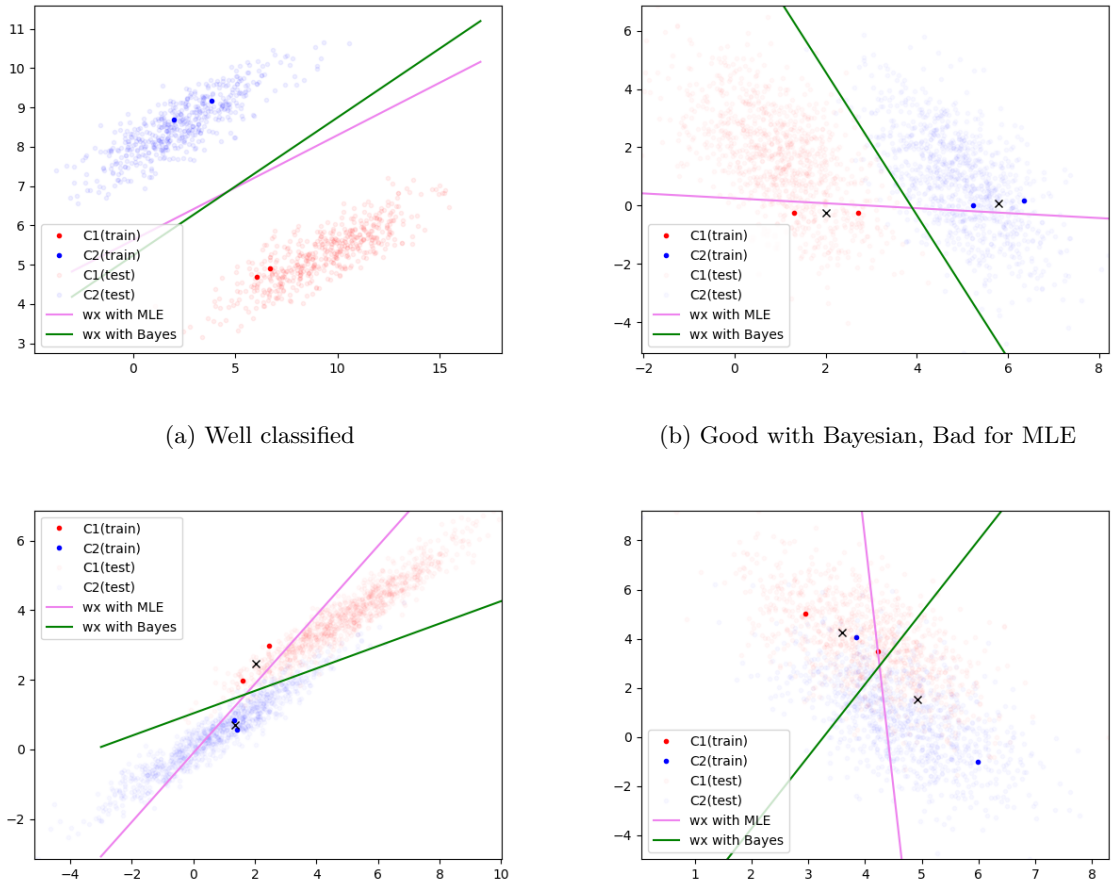
mean 의 분자에 Woodbury identity를 적용하면 다음과 같다.

$$(\text{num}) = N\mathbf{x}^\top \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2}\mathbf{I}\right)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{N\mathbf{x}^\top \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2}\mathbf{I}\right)^{-1}\mathbf{x}\mathbf{x}^\top \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2}\mathbf{I}\right)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{1 + \mathbf{x}^\top \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2}\mathbf{I}\right)^{-1}\mathbf{x}} \tag{25}$$

분모에 식 (22) 을 plugin 하여 정리하면 mean은 다음과 같다.

$$\mathbb{E}[t|\mathbf{x}, \mathcal{D}] = N\mathbf{x}^\top \left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2}\mathbf{I}\right)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{26}$$

2.5 Result



(a) Well classified

(b) Good with Bayesian, Bad for MLE

(c) Good with Bayesian, Bad for MLE

(d) Slightly outlying training data

그림 1: Result of FDA with MLE & Bayesian ($\sigma^2/\sigma_{\mathbf{w}}^2 = 0.1$)

그림 1 은 MLE를 이용해 구한 $\mathbf{w}$와 Bayesian approach를 이용해 구한 $\mathbf{w}$를 비교한 결과이다. 진한 점은 training data이고, 연한 점은 test data이다. 또한, 분홍색 실선은 MLE로 구한 boundary 이고 녹색 실선은 Bayesian approach로 구한 boundary 이다.

극단적인 성능 비교를 위하여, training data의 개수를 각 class 당 2개[1]로 설정하였다. 그림 1a 과 같이 MLE와 Bayesian classifier가 모두 잘 분류하는 경우도 있는 반면, 대부분의 상황에서 그림 1b, 1c 과 같이

[1] 1개로 하면 MLE에서 singular matrix가 생기기 때문에 불가능하다.

Bayesian boundary가 더 잘 분류함을 확인할 수 있다. 자세히는, MLE로 생성된 boundary는 over-fitted 된 것을 확인할 수 있는 반면, Bayesian으로 생성된 boundary는 비교적 덜 그렇다는 것을 확인할 수 있다. 이는 Bayesian approach 에서 사용된 prior variance 가 regularization term 역할을 했기 때문으로 사료된다.

또한, 그림 1d 와 같이 training data가 약간 outlying 하여도 MLE에 비해 Bayesian이 더 좋은 classification 을 하는 것을 확인할 수 있었다.

3 Kernel Extension

다음과 같은 mapping function ϕ 에 대해:

$$\phi : \mathbb{R}^D \to \mathbb{R}^M \tag{27}$$

mapping 된 data $\phi(\mathbf{x}) \in \mathbb{R}^M$를 생각하자. 그러면, 이렇게 mapping 된 data의 mean을 다음과 같이 정의한다.

$$\begin{aligned} \boldsymbol{\mu}_1^\phi &= \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} \phi(\mathbf{x}_n) \\ \boldsymbol{\mu}_2^\phi &= \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} \phi(\mathbf{x}_n) \end{aligned} \tag{28}$$

여기서 mapped data를 centralized 해주면, loss function은 다음과 같다.

$$\frac{1}{2} \sum_{n=1}^{N} (\mathbf{w}^\top \phi(\mathbf{x}_n) - t_n)^2 \tag{29}$$

다음과 같은 행렬을 도입하자.

$$\boldsymbol{\Phi} = \begin{pmatrix} | & & | \\ \phi(\mathbf{x}_1) & \cdots & \phi(\mathbf{x}_N) \\ | & & | \end{pmatrix} \in \mathbb{R}^{M \times N} \tag{30}$$

$$\mathbf{t} = (t_1, \cdots, t_N)^\top \in \mathbb{R}^N$$

또한, positive definite function k에 대하여, 다음 행렬을 정의하자.

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$$

$$\mathbf{K} = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix} \tag{31}$$

$$\mathbf{k} = (k(\mathbf{x}, \mathbf{x}_1), \cdots, k(\mathbf{x}, \mathbf{x}_N))^\top$$

여기서, 식 (26) 은 다음과 같이 변형할 수 있으므로:

$$\mathbb{E}[t|\mathbf{x}, \mathcal{D}] = \mathbf{x}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X} + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2} \mathbf{I} \right)^{-1} \mathbf{t} \tag{32}$$

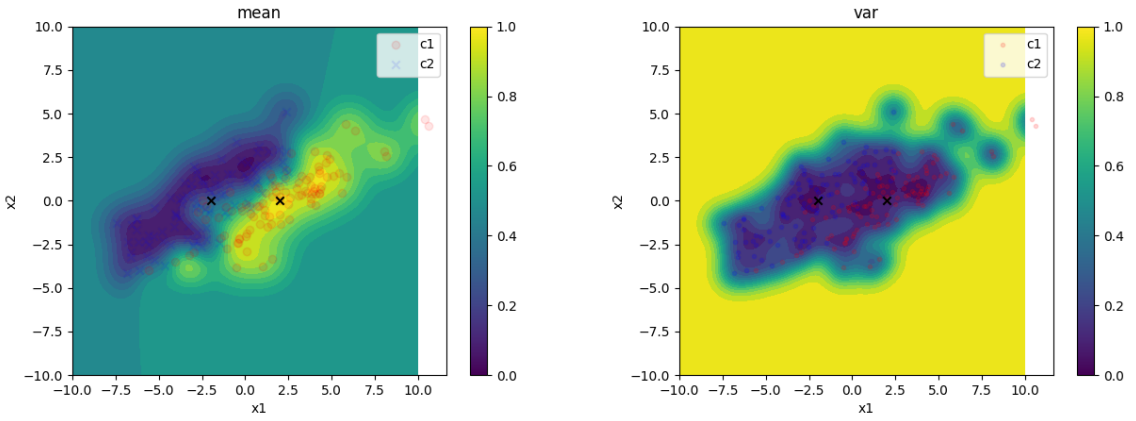
mapping 된 data의 predictive distribution의 mean과 variance는 각각 다음과 같다.

$$
\begin{aligned}
\mathbb{E}[t|\mathbf{x},\mathcal{D}] &= \phi(\mathbf{x})^\top \mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Phi} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}\mathbf{t} \\
&= \mathbf{k}^\top\left(\mathbf{K} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}\mathbf{t} \\
\mathrm{Var}[t|\mathbf{x},\mathcal{D}] &= \sigma^2 + \sigma_\mathbf{w}^2\phi(\mathbf{x})^\top\phi(\mathbf{x}) - \sigma_\mathbf{w}^2\phi(\mathbf{x})^\top\mathbf{\Phi}\left(\mathbf{\Phi}^\top\mathbf{\Phi} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}\mathbf{\Phi}^\top\phi(\mathbf{x}) \\
&= \sigma^2 + \sigma_\mathbf{w}^2 k(\mathbf{x},\mathbf{x}) - \sigma_\mathbf{w}^2\mathbf{k}^\top\left(\mathbf{K} + \frac{\sigma^2}{\sigma_\mathbf{w}^2}\mathbf{I}\right)^{-1}\mathbf{k}
\end{aligned}
\tag{33}
$$

이는 kernel vector $\mathbf{k}'$과 kernel matrix $\mathbf{K}'$ 이 다음과 같은 **Gaussian process** 라고 생각할 수 있다.

$$
\begin{aligned}
\mathbf{k}'_i &= \sigma_\mathbf{w}^2\phi(\mathbf{x})^\top\phi(\mathbf{x}_i) \\
\mathbf{K}'_{ij} &= \sigma_\mathbf{w}^2\phi(\mathbf{x}_i)^\top\phi(\mathbf{x}_j) + \sigma^2\delta_{ij}
\end{aligned}
\tag{34}
$$

3.1 Result



(a) Mean of classifier

(b) Variance of classifier

그림 2: Result of Kernel-extended Bayesian FDA

그림 2 는 kernel-extended Bayesian FDA의 결과를 나타낸 것으로, Gaussian process를 이용한 classification으로 볼 수 있다. Kernel function으로는 Gaussian kernel에 Bayesian approach로 얻어진 regularization term 을 추가한 것을 사용하였다.

그림 2a 를 보면 각 class의 data가 잘 분류되었음을 확인할 수 있다. 단순히 하나의 linear boundary를 제공하는 FDA에 비해 훨씬 정교하게 분류됨을 확인할 수 있다. 또한, regularization term이 자연스럽게 추가되어서 이상치나 적은 data에 강한 것을 확인할 수 있다.

그림 2b 는 classifier의 variance를 나타낸 것으로, data가 있는 곳의 variance는 작고, 그것으로부터 멀어질수록 증가하는 gaussian process 의 특징을 확인할 수 있다. 또한, regularization term 덕분에 극단적으로 variance가 낮은 점이 없다는 것도 확인할 수 있다.

4 References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York :Springer.

Duda, R. O. and P. E. Hart (1973). Pattern Classification and Scene Analysis. Wiley.