機械学習概論後半レポート

05242628 三田村彰大

概要　レポート問題の問 1・問 2・問 3 に取り組んだ。また、生成 AI についての質問の回答をレポート初めに付している。問 1・問 2 については Python を用いてプログラミングを行った。その際授業で配布されたノートブックファイルのコードを流用した。またライブラリとして sklearn を用いている。コードは kadai1.ipynb が問 1 に、kadai2.ipynb が問 2 に対応している。

※AI との会話ログ：　https://chatgpt.com/g/g-p-6a5f6f9bb77081918e92f9e1a0a86cf6-ji-jie-xue-xi-gai-lun-hou-ban-rehoto/project

※ 生成 AI の利用について

使用した AI

本課題においては chatGPT5.6-sol を利用した。(契約していたため。)

工夫

AI との会話リンクは冒頭に示した通りである。今回はプログラミングにおいて授業で配布されたノートブックファイルのコードを流用しているため、コーディング自体はそれほど行っていない。よって今回のレポートに関わるコーディングにおける AI の主な使途は、コードのデバッグおよび各種 API の利用方法についての相談程度である。

また、レポートの誤字・脱字および内容の確認に AI を利用している。ただしレポートの本文は一度すべて自分の言葉で書き、一度レポートが完成してから AI の査読を通すようにしている。

また、計算問題については一度自分で計算してみてから、AI に正誤確認を依頼している。[*1]

AI を利用した感想

前半のレポートに引き続き、やはり機械学習系のようなライブラリをふんだんに用いるプログラミングは AI が必須であるように感じた。特にライブラリの利用方法（どのような引数が必要であるかなど）をいちいち調べるのはめんどくさく、今現在書いているコードに沿って使用方法を提案してくれる AI は非常に助かると感じた。

問 1

以下によって生成されるデータに対する回帰を考える。

$$y = \sin(2\pi x) + \varepsilon \tag{1}$$
$$\varepsilon \sim \mathcal{N}\left(0, 0.2^2\right) \tag{2}$$

1.1

Gaussian Process と MLP による回帰を行う。ただし Gaussian Process におけるカーネルとしては次のような RBF+White kernel を用いることにする。(Figure1参照。)

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2l^2}\|\mathbf{x} - \mathbf{x}'\|^2\right) + \sigma_n^2 \delta_{\mathbf{x},\mathbf{x}'} \tag{3}$$

[*2]また optimizer の設定や MLP のハイパーパラメタの指定については問題文中の指示に従っている。[*3]

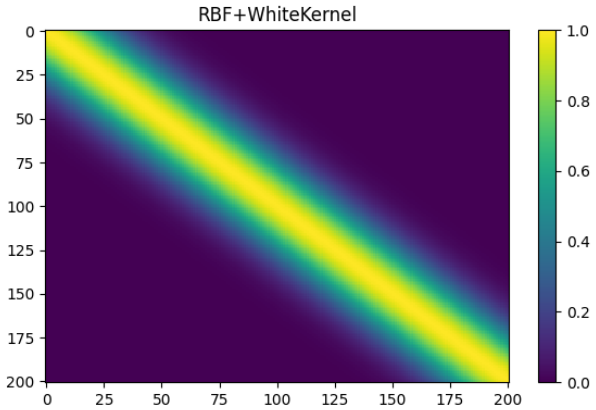

Figure1: カーネル $K(\mathbf{x}, \mathbf{x}')$ をヒートマップで表示している。

今、得られた学習結果を Figure2に示す。

1.2

Figure2で示された回帰の結果から次のことがわかる。

- MLP による回帰で得られる予測曲線（図中の緑線）は少ないデータ数においては安定しない。(真の関数から離れてしまう。)

[*1] 計算力が落ちるのが怖いので…

[*2] ただし l については授業で用いられていたサンプルコードの length_scale に従って $l = 0.2$ とした。また σ_n については今回与えているノイズに合わせて $\sigma_n = 0.2$ としている。

[*3] ただし GaussianProcessRegressor における観測ノイズの分散を決めるパラメタ α は $\alpha = 1 \times 10^{-4}$ としている。(ノイズは White kernel において最適化されるため、Regressor における α は数値安定化のみをしてくれれば良いと考えた。)

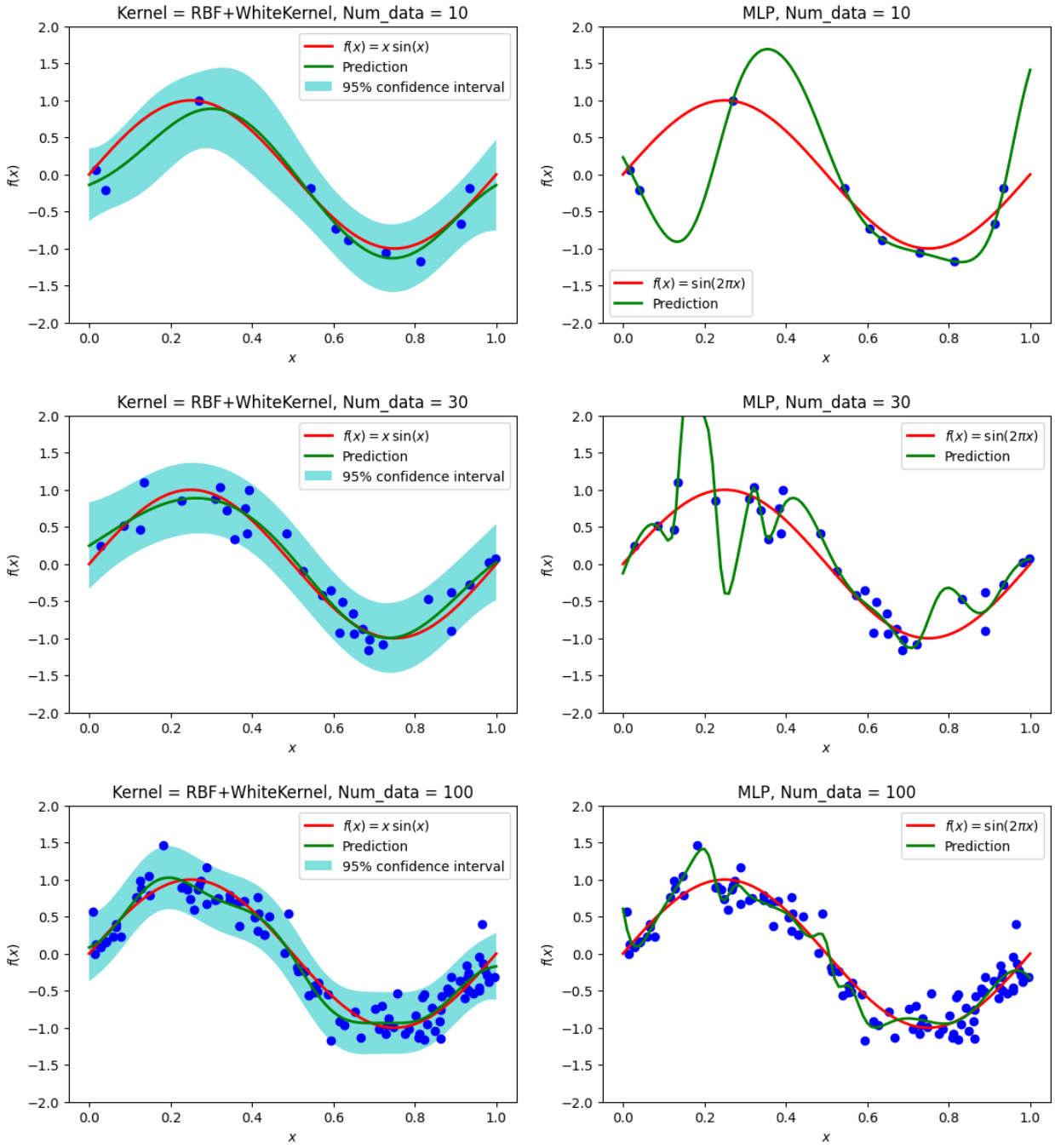

Figure2: Gaussian Process と MLP による回帰の結果。左列が Gaussian Process、右列が MLP による回帰を示しており、上から順にデータ数が $n = 10, 30, 100$ となっている。図中の赤線が真の関数 $y = \sin(2\pi x)$、青点が学習データ、緑線が予測曲線。また Gaussian Process については水色の領域で予測曲線の 95% 信頼区間を示している。

- MLP による回帰で得られる予測曲線はデータ数が多くなると真の関数に近くなるが、overfitting による局所的なノイズが見られる。
- Gaussian Process による回帰で得られる予測曲線は比較的少ないデータ数でも安定しているが、信頼区間が広くなってしまっている。
- Gaussian Process においてもデータ数が多くなると overfitting による予測曲線のノイズが見られるが、MLP 回帰と比べると安定している。

1.3

前章で述べた通り、データ数が少ない場合においては MLP による回帰が不安定になっていることがわかるが、これは MLP の自由度の高さに起因していると考えられる。今回隠れ層の次元数を 50 としていることを考えると MLP におけるパラメタの自由度は最低でも

$$\text{params} > 50 \times 50 = 2500 \tag{4}$$

となっているはずであり、データ数が 10 の場合を考えるとこれは強い劣決定問題となる。(比較として Gaussian Process の場合を考えるとこの場合におけるパラメタはせいぜい 2 つ程度であり[*4]、Gaussian Process と比べても MLP の劣決定性が非常に強いことがわかる。)

[*4] (3) における l と σ_n

よって今回データ数が少ない場合に MLP による回帰が不安定となったのは、モデルの自由度の高さに対してデータ数が少なすぎたために真の関数以外にもデータにフィットするような予測曲線が多数考えられてしまったためであると思われる。[*5]

1.4

今回、Gaussian Process においては信頼区間を求めることができているが、これは Gaussian Process がモデルの出力として確率密度関数を返すためである。

今得られているデータ・ラベルを $\{\mathbf{x}_n, t_n\}_{n=1}^N$ とするとき、通常の MLP などのモデルにおいては

$$t = f(\mathbf{x}; \theta) \tag{5}$$

のような「値」を出力するようなモデルを考えるため、予測においても単一の値しか出力することができない。しかし、Gaussian Process においては

$$\begin{aligned} &p(t_1, \ldots, t_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N) \\ &\quad = \mathcal{N}(0, K(\mathbf{x}_1, \ldots, \mathbf{x}_N; \theta)) + \mathcal{N}\left(0, \beta^{-1}\right) \end{aligned} \tag{6}$$

のような「確率」を出力するようなモデルを考えるため、予測において確率分布を出力することができる。(ただし $K_{nm} = k_\theta(\mathbf{x}_n, \mathbf{x}_m)$ はカーネル行列。)

実際今、新たなデータ $\mathbf{x}_{N+1}$ が得られた時に t_{N+1} が得られる確率は

$$p(t_{N+1} \mid \mathbf{x}_{N+1}) \tag{7}$$

$$= p(t_{N+1} \mid t_1, \ldots t_N, \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}) \tag{8}$$

$$= \frac{p(t_1, \ldots, t_{N+1} \mid \mathbf{x}_1, \ldots, \mathbf{x}_{N+1})}{p(t_1, \ldots, t_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N)} \tag{9}$$

のように計算され、(6) と合わせればこれは

$$p(t_{N+1} \mid \mathbf{x}_{N+1}) \sim \mathcal{N}\left(m_{N+1}, \sigma_{N+1}^2\right) \tag{10}$$

$$\left(\begin{aligned} m_{N+1} &= \mathbf{k}^T C_N^{-1} (t_1, \ldots, t_N)^T \\ \sigma_{N+1}^2 &= \beta^{-1} + k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) - \mathbf{k}^T C_N^{-1} \mathbf{k} \end{aligned} \right) \tag{11}$$

のように解析的に求めることができる。[*6] これは t_{N+1} の予測を与える確率密度関数に他ならず、よってこれが Gaussian Process において信頼区間を求めることができる理由である。

また今回、Gaussian Process においてデータ数が小さくなると信頼区間の幅が広くなることが確認できた。これは授業で述べられた Gaussian Process における信頼性のデータ粗密依存性が表れていると考えられる。すなわちデータ数が少なくなるとデータが粗になってしまうため信頼区間が大きくなってしまっていると思われる。[*7]

[*5] ただ、データ数が 100 程度でそれなりに fit できるのはそれはそれで不思議ではある…

[*6] ただし

$$C_N = K(\mathbf{x}_1, \ldots, \mathbf{x}_N; \theta) + \beta^{-1} I_N \tag{12}$$

$$\mathbf{k} = (k(\mathbf{x}_1, \mathbf{x}_{N+1}), \ldots, k(\mathbf{x}_N, \mathbf{x}_{N+1})) \tag{13}$$

である。

[*7] より正確には、データ数によらずデータが密なところでは信頼区間が狭く粗なところでは信頼区間が広くなることが予想されたが、今回の実験においてはそのような傾向はあまり顕著には見られなかった。

問 2

scikit-learn の make_moons によって得られるデータの二値分類を行う。(Figure3参照。)

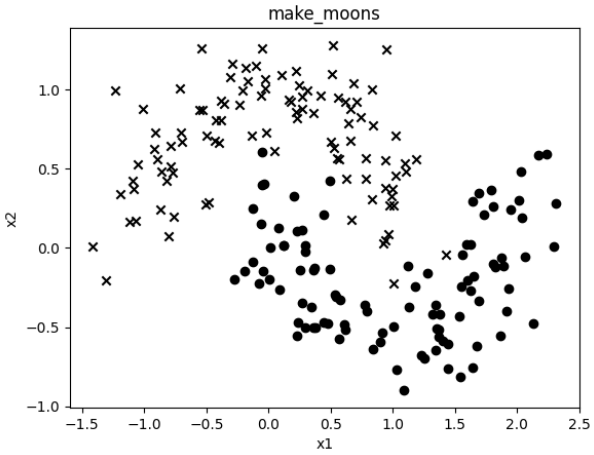

Figure3: scikit-learn の make_moons によって生成された学習データ。ただしデータ数は 200、noise は 0.2。

2.1

線形 SVM と RBF カーネル SVM を用いて二値分類を行う。ただしそれぞれのカーネルは次のよう。

$$\text{線形 SVM：} \quad k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}' \tag{14}$$

$$\text{RBF カーネル SVM：} \quad k(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2\right) \tag{15}$$

(ただしハイパーパラメタは問題文で指定されたものを用いることにする。)

二値分類の結果を Figure4に示す。また、テストデータに対する分類精度を Figure5に示す。

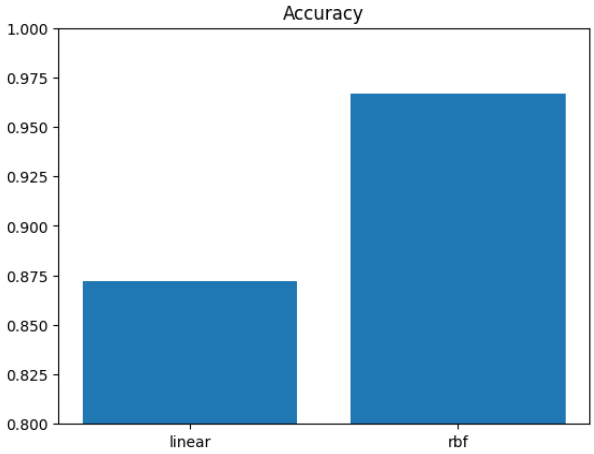

Figure5: テストデータに対する二値分類の精度の比較。左が線形 SVM、右が RBF カーネル SVM。

2.2

RBF カーネル SVM における γ の値 ((15) 式参照) の値を変えて二値分類を行い違いを観察する。結果を Figure7に示す。また、γ

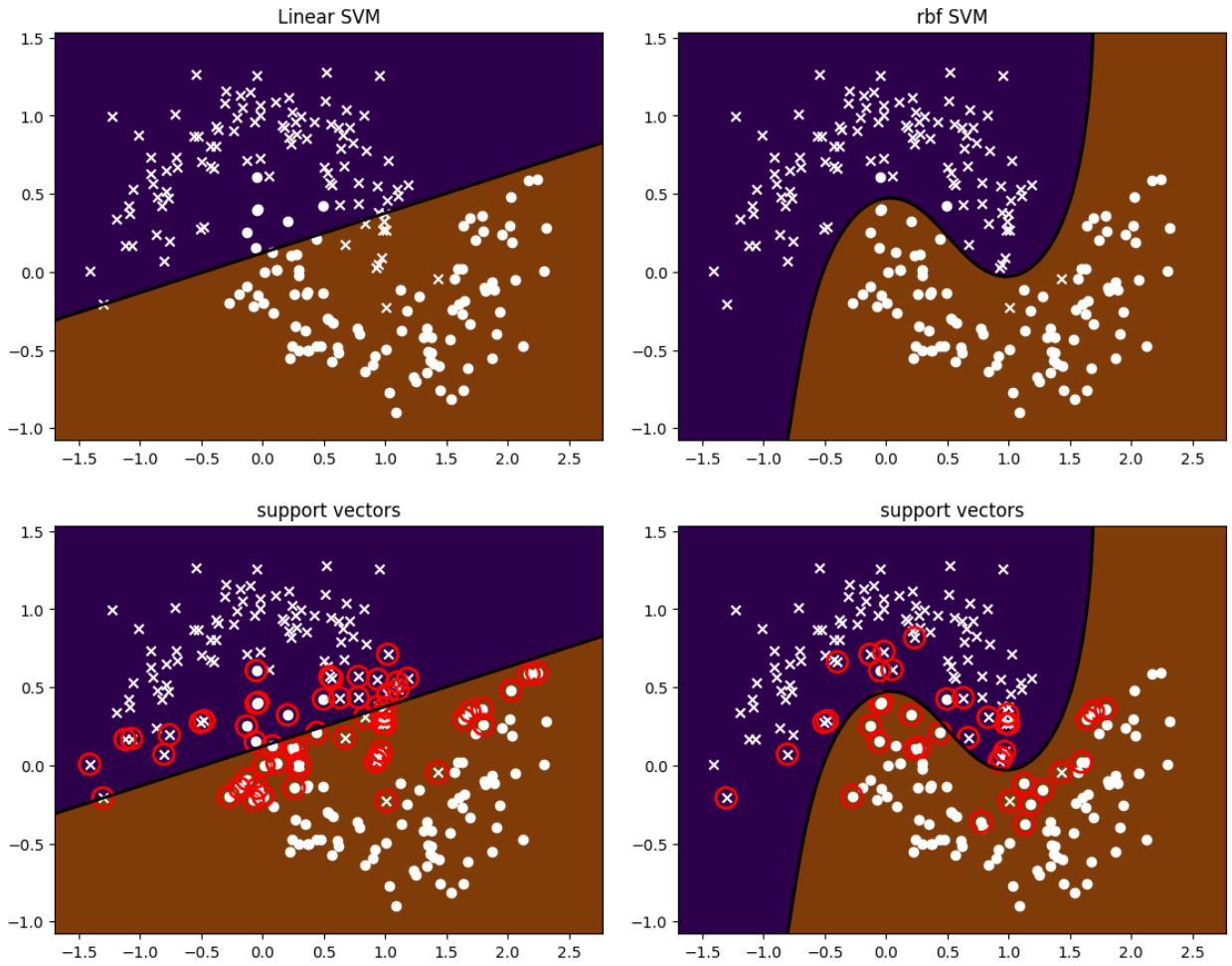

Figure4: 線形 SVM と RBF カーネル SVM による二値分類の結果。左列が線形 SVM、右列が RBF カーネル SVM による二値分類の結果を示している。黒線で分けられた二つの色の領域が決定領域を表し、下図の赤丸がサポートベクトルを表す。

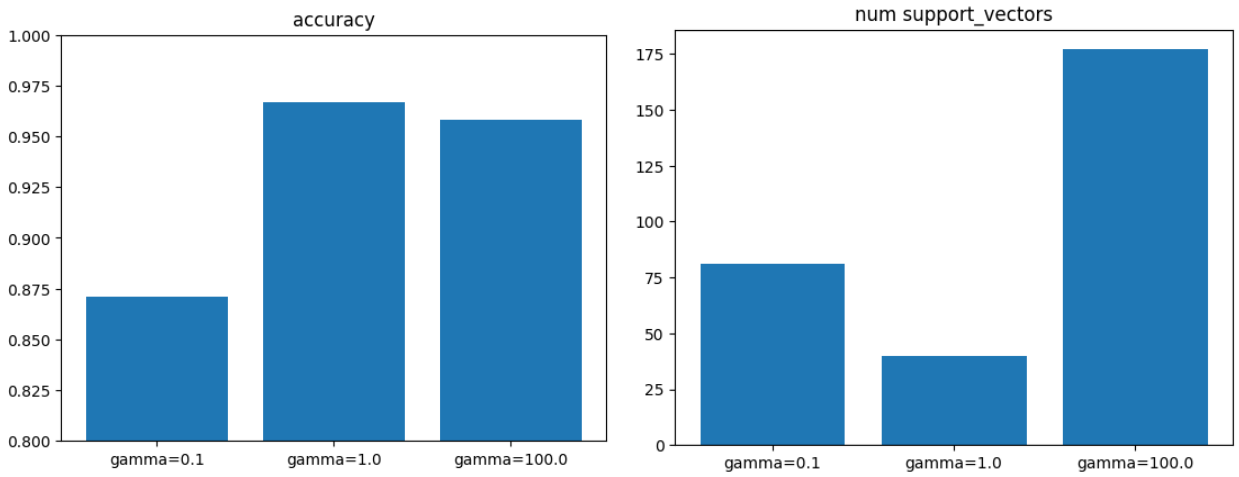

Figure6: RBF カーネル SVM において γ の値を様々に変えて二値分類を行った際の、テストデータに対する分類精度の変化（左図）とサポートベクトルの個数の変化（右図）。

の値を変えた時のテストデータに対する分類精度の変化とサポートベクトルの個数の変化を Figure6に示す。

2.3

Figure4〜Figure7の観察から、次のことがわかる。

- 線形 SVM では決定領域境界が直線に制限されるため、RBF カーネル SVM に比べて決定領域の柔軟性が低くテストデータに対する分類精度も悪くなっている。
- RBF カーネル SVM においては γ が大きいほど決定領域の柔軟性が高くなっている。ただしテストデータに対する分類精度は γ が小さい場合と大きい場合ノイズ入れにおいても（$\gamma = 1.0$ の場合に比べて）小さくなっている。
- RBF カーネル SVM における γ の値が極端に大きい場合

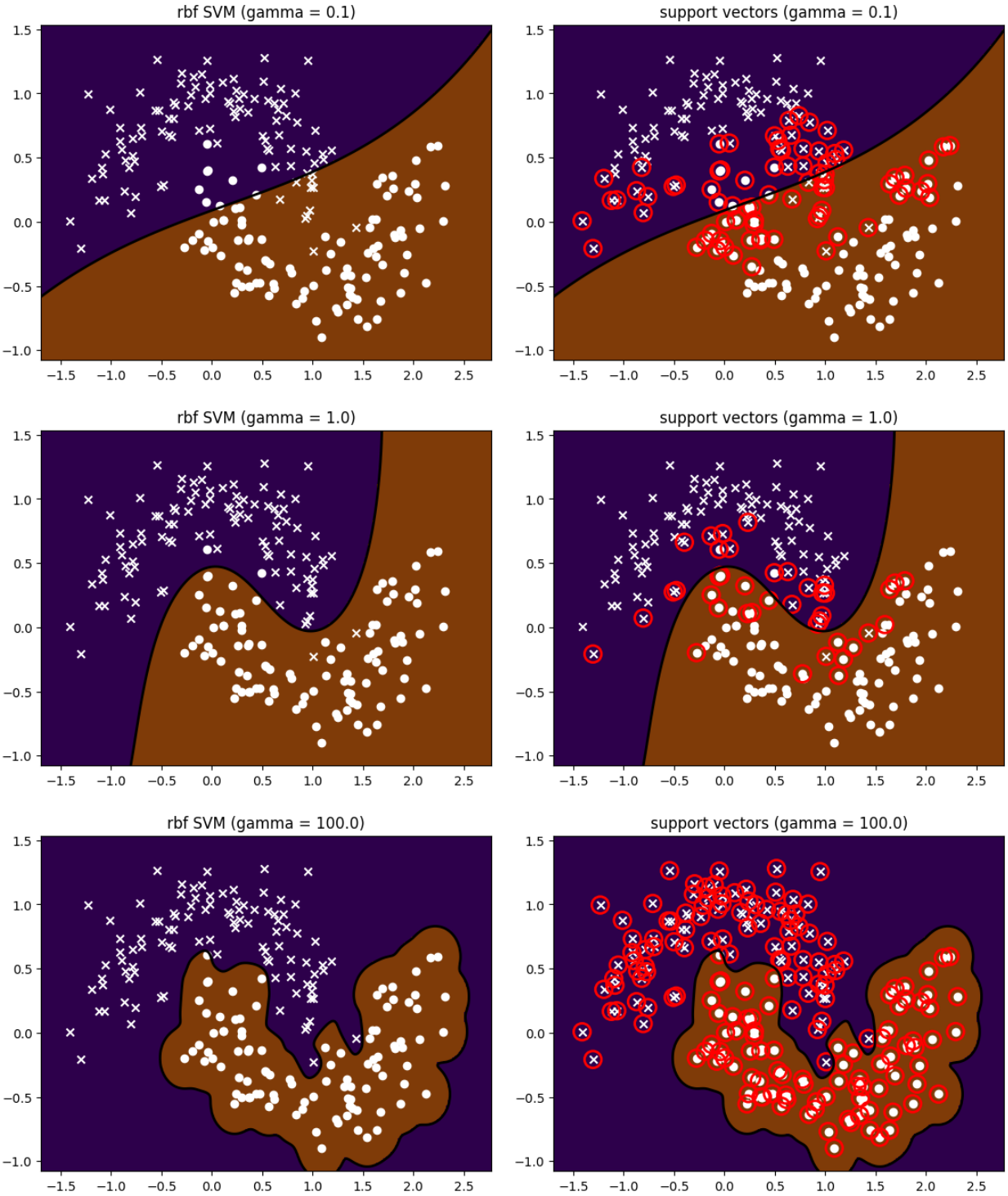

Figure7: RBF カーネル SVM において γ の値を様々に変えた時の二値分類の結果（左列）。右列はサーポートベクトルを明示した図。上から $\gamma = 0.1, 1, 100$。

($\gamma = 100$) では、すべてのデータ点がサポートベクトルとなっている。それ以外の場合においては基本的に決定領域境界近傍のデータ点がサポートベクトルとなっている。

2.4

先の章で見た通り線形 SVM に比べて RBF カーネル SVM の決定領域はより柔軟となっているが、これは特徴空間の次元から説明できる。

今、データを $\mathbf{x}$ で表し特徴空間への写像を $\phi(\mathbf{x})$ で表すとき、一般にカーネル $K(\mathbf{x}, \mathbf{x}')$ に対して $K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}')$ を満たす。よってまず線形 SVM について、(14) 式から

$$\phi(\mathbf{x}) = \mathbf{x} \tag{16}$$

と求まり、よってこの場合の特徴空間の次元は $\dim(\phi) = \dim(\mathbf{x})$ となっていることがわかる。

次に RBF カーネル SVM における特徴空間の次元を考察する。今

(15) 式から、カーネルを次のように変形することができる。

$$
\begin{aligned}
k(\mathbf{x},\mathbf{x}') &= \exp\left(-\gamma\|\mathbf{x}-\mathbf{x}'\|^2\right) \\
&= \exp\left(-\gamma\left(\mathbf{x}^2+\mathbf{x}'^2-2\mathbf{x}\cdot\mathbf{x}'\right)\right) \\
&= \exp\left(-\gamma\mathbf{x}^2\right)\exp\left(-\gamma\mathbf{x}'^2\right)\prod_{i=1}^{d}\exp\left(2\gamma x_i x_i'\right) \\
&= \exp\left(-\gamma\mathbf{x}^2\right)\exp\left(-\gamma\mathbf{x}'^2\right)\prod_{i=1}^{d}\left(\sum_{n=0}^{\infty}\frac{(2\gamma)^n}{n!}x_i^n x_i'^n\right) \\
&= \exp\left(-\gamma\mathbf{x}^2\right)\exp\left(-\gamma\mathbf{x}'^2\right) \\
&\qquad\times\sum_{(n_1,\ldots,n_d)\in\mathbb{N}^d}\left(\prod_{i=1}^{d}\frac{(2\gamma)^{n_i}}{n_i!}x_i^{n_i}x_i'^{n_i}\right) \\
&= \sum_{(n_1,\ldots,n_d)\in\mathbb{N}^d}\left\{\exp\left(-\gamma\mathbf{x}^2\right)\prod_{i=1}^{d}\sqrt{\frac{(2\gamma)^{n_i}}{n_i!}}x_i^{n_i}\right\} \\
&\qquad\times\left\{\exp\left(-\gamma\mathbf{x}^2\right)\prod_{i=1}^{d}\sqrt{\frac{(2\gamma)^{n_i}}{n_i!}}x_i'^{n_i}\right\}
\end{aligned}
$$

ただし $\dim(\mathbf{x})=d$ とした。[*8]

よって、$(n_1,\ldots,n_d)\in\mathbb{N}^d$ によって特徴量ベクトルの次元の引数をとり、その具体的な形を

$$
\phi_{n_1,\ldots,n_d}(\mathbf{x}) = \exp\left(-\gamma\mathbf{x}^2\right)\prod_{i=1}^{d}\sqrt{\frac{(2\gamma)^{n_i}}{n_i!}}x_i^{n_i} \tag{17}
$$

によって定めれば、ϕ は $k(\mathbf{x},\mathbf{x}')=\phi(\mathbf{x})\cdot\phi(\mathbf{x}')$ を満たす。このことから、RBF カーネル SVM における特徴空間の次元が（概念的には）$\dim(\phi)=\mathbb{N}^d=\infty$ となっていることがわかる。

よって無限次元の特徴空間を考えている RBF カーネル SVM と特徴空間がデータ次元と同じ次元をもつ線形 SVM ではその特徴空間の次元数の違いによって特徴量の表現能力に差が出ることが予想される。具体的には、RBF カーネル SVM の方が特徴空間の次元数が大きいためにより柔軟に決定領域を選び取ることが可能になると思われ、今回見られた決定領域の柔軟性の違いを説明することができると考えられる。

次に、γ と分類精度の関係について述べる。今回 RBF カーネル SVM で γ を大きくしていったところより決定領域が柔軟になることがわかったが、テストデータに対する分類精度は $\gamma=0.1$ から $\gamma=1.0$ への変化では向上するものの $\gamma=100$ では逆に悪化する。今回これについては、決定領域が柔軟になりすぎたことによって過学習が生じていると考えることができる。

問 3

3.1

潜在変数 $Z\in\mathbb{R}^N$ とそれに対して導入した任意の確率分布 $q(Z|X)$ に対し

$$
\log p(X|\theta) = \mathcal{L}(q,\theta) + D_{\mathrm{KL}}\left(q(Z|X)\,||\,p(Z|X,\theta)\right) \tag{18}
$$

が成立することは次のように示すことができる。

$$
\begin{aligned}
\log p(X|\theta) &= \int dz\; q(Z|X)\log p(X|\theta) && (19)\\
&= \int dz\; q(Z|X)\log\frac{p(X,Z|\theta)}{p(Z|X,\theta)} && (20)\\
&= \int dz\; q(Z|X)\log\frac{p(X,Z|\theta)}{q(Z|X)}\frac{q(Z|X)}{p(Z|X,\theta)} && (21)\\
&= \left[\int dz\; q(Z|X)\log p(X,Z|\theta)\right. \\
&\qquad\left.-\int dz\; q(Z|X)\log q(Z|X)\right] \\
&\qquad+\int dz\; q(Z|X)\log\frac{q(Z|X)}{p(Z|X,\theta)} && (22)\\
&= \mathcal{L}(q,\theta) + D_{\mathrm{KL}}\left(q(Z|X)\,||\,p(Z|X,\theta)\right) && (23)
\end{aligned}
$$

ただし（19）から（20）への変形は

$$
p(X,Z|\theta) = p(X|\theta)\,p(Z|X,\theta) \tag{24}
$$

を用いている。よってこの結果と KL ダイバージェンスが非負となることを用いれば

$$
\log p(X|\theta) \geq \mathcal{L}(q,\theta) \tag{25}
$$

となっていることが直ちにわかり、$\mathcal{L}(q,\theta)$ が $\log p(X|\theta)$ の下界となっていることがわかる。

3.2

確率分布を次のように与えることを考える。

$$
q_\phi(Z|X) = \mathcal{N}\left(\mu,\mathrm{diag}\left(\sigma_i^2\right)\right) \tag{26}
$$

$$
p(Z|X,\theta) = p(Z) = \mathcal{N}(0,I) \tag{27}
$$

この時、$D_{\mathrm{KL}}\left(q_\phi(Z|X)\,||\,p(Z)\right)$ は次のように計算することができる。

$$
\begin{aligned}
&D_{\mathrm{KL}}\left(q_\phi(Z|X)\,||\,p(Z)\right) \\
&= \int dz\; q_\phi(Z|X)\log q_\phi(Z|X) - \int dz\; q_\phi(Z|X)\log p(Z) \\
&= \mathbb{E}_{q_\phi}\left[\log q_\phi(Z|X)\right] - \mathbb{E}_{q_\phi}\left[\log p(z)\right]
\end{aligned}
$$

特に（26）（27）から、

$$
\begin{aligned}
&\mathbb{E}_{q_\phi}\left[\log q_\phi(Z|X)\right] \\
&= \mathbb{E}_{q_\phi}\left[\frac{1}{(2\pi)^{N/2}\left\{\det\left(\mathrm{diag}\left(\sigma_i^2\right)\right)\right\}^{1/2}}\right. \\
&\qquad\left.\exp\left(-\frac{1}{2}(\mathbf{z}-\mu)^T\,\mathrm{diag}\left(\sigma_i^2\right)^{-1}(\mathbf{z}-\mu)\right)\right] \\
&= \mathbb{E}_{q_\phi}\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{(z_i-\mu_i)^2}{\sigma_i^2} - \log(2\pi)^{N/2}\left(\prod_{i=1}^{N}\sigma_i^2\right)^{1/2}\right] \\
&= \underbrace{-\frac{1}{2}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\mathbb{E}_{q_\phi}\left[(z_i-\mu_i)^2\right]}_{=-\frac{N}{2}} - \log(2\pi)^{N/2} - \sum_{i=1}^{N}\log\sigma_i
\end{aligned}
$$

[*8] このカーネルの展開の計算は AI の助けを借りている。

及び

$$\begin{aligned}\mathbb{E}_{q_\phi}\left[\log p\left(z\right)\right] &= \mathbb{E}_{q_\phi}\left[\frac{1}{\left(2\pi\right)^{N/2}}\exp\left(-\frac{1}{2}\mathbf{z}^2\right)\right] \\ &= \mathbb{E}_{q_\phi}\left[-\frac{1}{2}\sum_{i=1}^{N} z_i^2 - \log\left(2\pi\right)^{N/2}\right] \\ &= -\frac{1}{2}\sum_{i=1}^{N}\left(\mu_i^2+\sigma_i^2\right) - \log\left(2\pi\right)^{N/2}\end{aligned}$$

が従うので、最終的に求めたい値は

$$\begin{aligned}&D_{\mathrm{KL}}\left(q_\phi\left(Z|X\right)||\,p\left(Z\right)\right) \\ &\qquad = -\frac{N}{2} + \frac{1}{2}\sum_{i=1}^{N}\mu_i^2 + \sum_{i=1}^{N}\left(\frac{1}{2}\sigma_i^2 - \log\sigma_i\right) \qquad (28)\end{aligned}$$

となる。

3.3

今損失関数が

$$\mathrm{Loss}\left(\phi,\theta\right) = -\mathbb{E}_{q_\phi}\left[\log p\left(X|\theta\right)\right] + D_{\mathrm{KL}}\left(q_\phi\left(Z|X\right)||\,p\left(Z\right)\right) \quad (29)$$

のように定義されるとき、停留点条件を計算すると

$$\mu_i = 0 \qquad (30)$$

$$\sigma_i - \frac{1}{\sigma_i} = 0 \;\rightsquigarrow\; \sigma_i = 1 \qquad (31)$$

が得られる。よって損失関数における KL 項は $q_\phi\left(Z|X\right)$ の平均と分散 μ, σ^2 を $p\left(z\right)$ の平均と分散 $0, I$ に近づけるような制約を与える。

3.4

今 VAE における損失関数は

$$\mathrm{Loss}\left(\phi,\theta\right) = \underbrace{-\mathbb{E}_{q_\phi}\left[\log p\left(X|\theta\right)\right]}_{\text{再構成項目}} + \underbrace{D_{\mathrm{KL}}\left(q_\phi\left(Z|X\right)||\,p\left(Z\right)\right)}_{\text{KL 項}} \quad (32)$$

のように再構成項と KL 項からなっている。この時前章での考察から、KL 項は $q_\phi\left(Z|X\right)$ を $p\left(Z\right)$ に近づけるような役割を持っており、これは VAE において潜在変数に関する事後分布 $p\left(Z|X,\theta\right)$ を任意に与えた確率分布 $q\left(Z|X\right)$ で近似する際の近似精度を良くするような役割を持つ。

また再構成項は尤度の最大化を要請する項目であり、これは VAE をエンコーダ＋デコーダの結合モデルと考えた際に、$\mathbf{x} \to \mathbf{z} \to \mathbf{x}$ と $\mathbf{x}$ を再構成する際の精度が高くなることを要請していることに相当する。

以上