Support vector machines for interval-valued data
区間をデータとするサポートベクターマシン

島根大学・自然科学研究科 益田大輝 (Taiki Masuda)
Graduate School of Natural Sciences and Technology, Shimane University
島根大学・自然科学研究科 森智也 (Tomoya Mori)
Graduate School of Natural Sciences and Technology, Shimane University
島根大学・総合理工学部 黒岩大史 (Daishi Kuroiwa)
Interdisciplinary Faculty of Science and Engineering, Shimane University

概 要

データには内在的な誤差が含まれることがある。不確実性のあるデータやパラメータに対処するために、数値を区間として表現し、数値計算や最適化の信頼性を向上させるための方法論（区間解析）は 1980 年代から行われてきた。先行研究では、区間をベクトル空間に埋め込む際や区間上の内積を扱う際に、ベクトル空間の公理や内積空間の性質が満たされていない場合がある。そのような中で、凸集合をベクトル空間に埋め込む研究 [Rådström, 1952] や、区間をヒルベルト空間に埋め込む研究 [Kuroiwa, M, 2024] が進められている。今回の発表では、区間上のサポートベクターマシンの理論について説明し、実際に区間データを用いてサポートベクターマシンを行った結果を紹介する。

1 サポートベクターマシンの理論

サポートベクターマシン (SVM : Support Vector Machine) は 2 値でラベル付けされたデータを分類するための手法である。ある最適化問題を解くことで、与えられたデータを分類する超平面を求めることが目的である。ここでは、X をユークリッド空間、すなわち $\mathbb{R}$ 上の有限次元ベクトル空間で内積が定義されているとし、説明変数を $x_i \in X$、ラベルを $y_i \in \{-1, 1\}$ とする以下のデータについて考える :

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

まず、データ $\{x_i \mid y_i = 1\}$ と $\{x_j \mid y_j = -1\}$ を強分離する超平面が存在すると仮定する。つまり、

$$\exists w \in X \backslash \{0\},\ \exists b \in \mathbb{R}\ s.t. \begin{cases} \langle w, x_i \rangle + b \geqq 1, & (i \in \{i \mid y_i = 1\}), \\ \langle w, x_j \rangle + b \leqq -1, & (j \in \{i \mid y_i = -1\}). \end{cases}$$

が成り立つとする。このように、データが強分離可能である場合に用いられる手法をハードマージン SVM という。上式について、$j \in \{i \mid y_i = -1\}$ のとき

$$-(\langle w, x_j \rangle + b) \geqq 1$$

と書けることから、$i \in \{i \mid y_i = 1\}$ のときと合わせて以下のように書き換えられる:

$$\exists w \in X \setminus \{0\},\ \exists b \in \mathbb{R},\ y_i(\langle w, x_i \rangle + b) \geqq 1, \quad i = 1, 2, \cdots, n.$$

i 番目のデータ x_i と超平面 $\langle w, x_i\rangle + b = 0$ との距離は、次のように表される :

$$\frac{y_i(\langle w, x_i\rangle + b)}{\|w\|}, \quad (\text{ただし } y_i(\langle w, x_i\rangle + b) \geqq 1)$$

$\{x_i \mid y_i = 1\}$ と $\{x_j \mid y_j = -1\}$ のそれぞれにおいて、超平面との距離が最小であるデータ (ベクトル) をサポートベクトルといい、サポートベクトルと超平面との距離をマージンという。次の最適化問題は、そのマージンを最大化するパラメータ w と b を解として得る。

$$\begin{array}{ll} \text{Maximize} & \min\limits_i \dfrac{y_i(\langle w, x_i\rangle + b)}{\|w\|} \\ \text{subject to} & y_i(\langle w, x_i\rangle + b) \geqq 1, \quad i = 1, 2, \cdots, n \\ & w \in X,\ b \in \mathbb{R}. \end{array}$$

ここで、(w, b) がこの問題の実行可能解のとき、

$$\min_i y_i(\langle w, x_i\rangle + b) = m$$

とおくと $m \geqq 1$ であるが、w、b を m で割って改めて w、b と置き直すことで、$m = 1$ としてよいことがわかる。よって目的関数値は $\frac{1}{\|w\|}$ のみとなるが、$\frac{1}{\|w\|}$ の最大化は $\|w\|$ の最小化と同値であり、$\|w\|$ を $\frac{1}{2}\|w\|^2$ としても問題ないため、ハードマージン SVM の最適化問題は次のように表すことができる :

$$\begin{array}{ll} \text{Minimize} & \frac{1}{2}\|w\|^2 \\ \text{subject to} & y_i(\langle w, x_i\rangle + b) \geqq 1, \quad i = 1, 2, \cdots, n \\ & w \in X,\ b \in \mathbb{R}. \end{array}$$

ここでは、データを $\{x_i \mid y_i = 1\}$ と $\{x_j \mid y_j = -1\}$ に強分離する超平面が存在すると仮定したが、全てのデータ点が 2 つのグループに強分離されるとは限らない。そのため、超平面によって分離したグループとは異なるグループに含まれるデータが存在する場合を考える。このときに用いられる手法をソフトマージン SVM という。ハードマージン SVM では、超平面による分離を $w \in X \setminus \{0\}$ と $b \in \mathbb{R}$ を用いて、

$$y_i(\langle w, x_i\rangle + b) \geqq 1, \quad (i = 1, 2, \cdots, n)$$

と表していたが、異なるグループに含まれるデータ x_i に対して変数 $\xi_i \geqq 0$ を用いて調整をすると、上記の不等式は以下の不等式に書き換えられる :

$$y_i(\langle w, x_i\rangle + b) \geqq 1 - \xi_i, \quad (i = 1, 2, \cdots, n)$$

i 番目のデータが正しく分離されている場合は $\xi_i = 0$、異なるグループに含まれる場合は超平面の垂直方向に ξ_i で調整して引き戻すことで本来のグループに移動させる。このとき、マージンの最大化とともに ξ_i の最小化を行う最適化問題は、正の定数 C を定めることで次のように書ける :

$$\begin{array}{ll} \text{Minimize} & \dfrac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{n} \xi_i \\ \text{subject to} & y_i(\langle w, x_i\rangle + b) \geqq 1 - \xi_i,\ \xi_i \geqq 0, \quad i = 1, 2, \cdots, n \\ & w \in X,\ b \in \mathbb{R},\ \xi \in X. \end{array}$$

目的関数と制約条件を、

$$f(w,b,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$g_i(w,b,\xi) = 1 - \xi_i - y_i(\langle w, x_i\rangle + b)$$
$$h_i(w,b,\xi) = -\xi_i$$

と定義し、最適化問題を次のように書き換える :

$$\begin{array}{ll} \text{Minimize} & f(w,b,\xi) \\ \text{subject to} & g_i(w,b,\xi) \leqq 0,\ h_i(w,b,\xi) \leqq 0, \quad i = 1,2,\cdots,n \\ & w \in X,\ b \in \mathbb{R},\ \xi \in X. \end{array}$$

ソフトマージン最大化問題の双対問題を与えるため、一般論を示す。X を有限次元ユークリッド空間とし、関数 $f, g_1, g_2, \cdots, g_m : X \to \mathbb{R}$ に対して、最適化問題 (P) を次のように定める：

$$\begin{array}{lll} \text{(P)} & \text{Minimize} & f(x) \\ & \text{subject to} & g_i(x) \leqq 0, \quad i = 1,2,\cdots,m. \end{array}$$

問題 (P) を定めている関数が凸関数の場合に、以下の Lagrange 双対定理が成立している。

定理 1.1 (Goberna, Jeyakumar, López, 2008[1])**.** 関数 $g_1, g_2, \cdots, g_m$ が凸であるとき、次は同値：

(i) $\text{cone co} \bigcup_{i=1}^{m} \text{epi}\, g_i^* + \{0\} \times [0,\infty)$ は閉集合

(ii) 任意の凸関数 $f : X \to \mathbb{R}$ に対して、以下が成立する：

$$\inf_{x\in A} f(x) = \max_{\lambda_i \geqq 0} \inf_{x\in X} \left\{ f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \right\}$$

ただし、$A = \{x \mid g_i(x) \leqq 0,\ i = 1,\cdots,m\}$ は空でないとする。

注意 1.1. 定理 1.1 に関して注意を述べる。(i) の条件は FM（Farkas Minkowski）と呼ばれる制約想定である。また、g_i が全てアフィン関数、すなわち $g_i(x) = \langle a_i, x\rangle - b_i (i = 1,\cdots,m)$ のとき、FM が成立することが示される。実際、$g_i^* = \delta_{\{a_i\}} + b_i$ であり、$\text{epi}\, g_i^* = (a_i, b_i) + \{0\} \times [0,\infty)$ となる。よって、以下が示される。

$$\begin{aligned} \text{cone co} \bigcup_{i=1}^{m} \text{epi}\, g_i^* + \{0\} \times [0,\infty) &= \text{cone co} \bigcup_{i=1}^{m} \{(a_i,b_i) + \{0\} \times [0,\infty)\} + \{0\} \times [0,\infty) \\ &= \text{cone co}\{(a_1,b_1),\cdots,(a_m,b_m),(0,1)\} \end{aligned}$$

一般に有限集合 $\{x_1, x_2, \cdots, x_k\} \subset X$ の凸錐包 $\text{cone co}\{x_1, x_2, \cdots, x_k\}$ は閉集合であることが知られているため、FM が成立することが示された。

ここで、任意の $i = 1,2,\cdots,n$ に対して g_i と h_i はアフィン関数であり、定理 1.1 の条件 (i) をみたしているので、条件 (ii) の式が成り立つ:

$$\inf_{g_i\leqq 0, h_i \leqq 0} f(w,b,\xi) = \max_{u_i \geqq 0, v_i \geqq 0} \inf_{(w,b,\xi)} \left\{ f(w,b,\xi) + \sum_{i=1}^{n} u_i g_i(w,b,\xi) + \sum_{i=1}^{n} v_i h_i(w,b,\xi) \right\}$$

右辺の inf 以降の式は変数 (w,b,ξ) について微分可能な凸関数である。従って、右辺を偏微分して 0 になるときを観察する。

$$0 = \nabla f(w,b,\xi) + \sum_{i=1}^{n} u_i \nabla g_i(w,b,\xi) + \sum_{i=1}^{n} v_i \nabla h_i(w,b,\xi)$$

$$\therefore w = \sum_{i=1}^{n} u_i y_i x_i,\ 0 = \sum_{i=1}^{n} u_i y_i,\ 0 = (C,\cdots,C)^T - u - v$$

従ってこの右辺の inf 以降は次のように求められる：

$$\begin{aligned}
&\inf_{(w,b,\xi)} \left\{ f(w,b,\xi) + \sum_{i=1}^{n} u_i g_i(w,b,\xi) + \sum_{i=1}^{n} v_i h_i(w,b,\xi) \right\} \\
&= \inf_{(w,b,\xi)} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n} u_i \left\{1 - \xi_i - y_i(\langle w, x_i\rangle + b)\right\} - \sum_{i=1}^{n} v_i\xi_i \right\} \\
&= \inf_{b\in\mathbb{R},\xi\in X} \left\{ \frac{1}{2}\|\sum_{i=1}^{n} u_i y_i x_i\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n} u_i \left\{ 1 - \xi_i - y_i \left(\left\langle \sum_{i=1}^{n} u_i y_i x_i, x_i \right\rangle + b \right) \right\} - \sum_{i=1}^{n} v_i\xi_i \right\} \\
&= \inf_{\xi\in X} \left\{ \frac{1}{2}\|\sum_{i=1}^{n} u_i y_i x_i\|^2 + \sum_{i=1}^{n}\xi_i(C - u_i - v_i) + \sum_{i=1}^{n} u_i - \|\sum_{i=1}^{n} u_i y_i x_i\|^2 \right\} \\
&= -\frac{1}{2}\|\sum_{i=1}^{n} u_i y_i x_i\|^2 + \sum_{i=1}^{n} u_i \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} u_i u_j y_i y_j \langle x_i, x_j\rangle + \sum_{i=1}^{n} u_i
\end{aligned}$$

したがってソフトマージン SVM の最適化問題の双対問題は以下のようになる:

$$\begin{aligned}
&\text{Maximize} \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} u_i u_j y_i y_j \langle x_i, x_j\rangle + \sum_{i=1}^{n} u_i \\
&\text{subject to} \quad \sum_{i=1}^{n} u_i y_i = 0,\ 0 \leqq u_i \leqq C, \quad i = 1,2,\cdots,n.
\end{aligned}$$

2　区間を埋め込む空間 $\mathcal{N}$ 上の内積

一般的にデータは 1 つの値として収集されるが、身長や体重の微小な変化、四捨五入で切り捨てられる値などのように内在的な誤差を含むものも存在する。それらを区間データとして扱い、分類や推測を行うため、ここでは [3] の研究結果について紹介する。まず、$\mathbb{R}^n$ 上の二項関係 (半順序) と区間 $[a,b]$ を次のように定める。

定義 2.1. $\forall a = (a_1, a_2, \cdots, a_n), b = (b_1, b_2, \cdots, b_n) \in \mathbb{R}^n$ に対して、$\mathbb{R}^n$ 上の二項関係を次のように定義する。

$$a \leqq b \overset{\text{def}}{\iff} a_i \leqq b_i,\ \forall i \in \{1,2,\cdots,n\}$$

また、区間 $[a,b]$ を次のように定義する。

$$[a,b] := \{x \in \mathbb{R}^n \mid a \leqq x \leqq b\} = \prod_{i=1}^{n} [a_i, b_i]$$

上記で定義した区間に対して、$\mathbb{R}^n$ 上のすべての区間の族 $\mathcal{M}$ と $\mathcal{M}$ 上の和と非負スカラー積を次のように定義する。

$$\mathcal{M} = \{[a,b] \mid a \leqq b\}$$

$$[a,b] + [c,d] = [a+c, b+d], \quad t[a,b] = [ta, tb]\,(t \geqq 0)$$

また、$\mathcal{M}^2$ に対して $\sim$ で表記される同値関係を

$$([a,b],[c,d]) \sim \left(\left[a',b'\right],\left[c',d'\right]\right) \iff \begin{cases} a-c = a'-c' \\ b-d = b'-d', \end{cases}$$

と定義し、$([a,b],[c,d])$ が属する同値類を次のように定義する。

$$[a,b] \ominus [c,d] := \{\left(\left[a',b'\right],\left[c',d'\right]\right) \in \mathcal{M}^2 \mid ([a,b],[c,d]) \sim \left(\left[a',b'\right],\left[c',d'\right]\right)\}$$

この表記を用いて $\sim$ による $\mathcal{M}^2$ の商集合 $\mathcal{N}$ を次のように定義する。

$$\mathcal{N} = \mathcal{M}^2/\sim = \{[a,b] \ominus [c,b] \mid [a,b],[c,d] \in \mathcal{M}\}$$

$\mathcal{N}$ 上の和、スカラー積、次のように定義する。

$$[a,b] \ominus [c,d] + \left[a',b'\right] \ominus \left[c',d'\right] := \left[a+a', b+b'\right] \ominus \left[c+c', d+d'\right]$$

$$\lambda([a,b] \ominus [c,d]) := \begin{cases} [\lambda a, \lambda b] \ominus [\lambda c, \lambda d] & \lambda \geqq 0, \\ [-\lambda c, -\lambda d] \ominus [-\lambda a, -\lambda b] & \lambda < 0. \end{cases}$$

ここで、$\left\langle \begin{pmatrix} \cdot \\ \cdot \end{pmatrix}, \begin{pmatrix} \cdot \\ \cdot \end{pmatrix} \right\rangle$ を $\mathbb{R}^{2n}$ 上の内積とする。$[a,b] \ominus [c,d], [a',b'] \ominus [c',d'] \in \mathcal{N}$ に対して, $\mathcal{N}^2$ から $\mathbb{R}$ への関数を

$$\left\langle [a,b] \ominus [c,d], \left[a',b'\right] \ominus \left[c',d'\right] \right\rangle := \left\langle \begin{pmatrix} a-c \\ b-d \end{pmatrix}, \begin{pmatrix} a'-c' \\ b'-d' \end{pmatrix} \right\rangle,$$

と定義する。このとき、次の定理が成り立つ。

定理 2.1 ([3])**.** 上で定義した関数は well-defined であり, $\mathcal{N}$ 上の内積となる。

定理 2.2 ([3])**.** $\mathcal{N}$ の次元は $2n$ であり, $\mathcal{N}$ はヒルベルト空間である。

上記の定理が成り立つことから、$\mathcal{N}$ 上で内積を用いた分析を行うことができる。

3　区間データのサポートベクターマシン

ここでは [3] と同様にして $\mathcal{N}$ 上で SVM を行うことができることを確認する。任意の $i = 1, 2, \cdots, n$ に対して $[\underline{x}_i, \overline{x}_i] \ominus [0,0] \in \mathcal{N}$ 、 $y_i \in \{-1, 1\}$ とすると $\mathcal{N}$ 上のデータセットは以下のようになる:

$$([\underline{x}_1, \overline{x}_1] \ominus [0,0], y_1), ([\underline{x}_2, \overline{x}_2] \ominus [0,0], y_2), \cdots, ([\underline{x}_n, \overline{x}_n] \ominus [0,0], y_n).$$

$\mathcal{N}$ はヒルベルト空間であるので、ソフトマージン SVM における最適化問題は以下のようになる。ただし、$[\underline{x}_i, \overline{x}_i] \ominus [0,0] = x_i \ominus 0$ と表記する:

$$
\begin{aligned}
&\text{Minimize} && \frac{1}{2}\|w \ominus w'\|^2 + C\sum_{i=1}^{n}\xi_i \\
&\text{subject to} && y_i(\langle w \ominus w', x_i \ominus 0\rangle - b) \geqq 1 - \xi_i,\ \xi_i \geqq 0,\ i = 1, 2, \cdots, n, \\
& && w \ominus w' \in \mathcal{N},\ b \in \mathbb{R},\ \xi \in \mathbb{R}^n
\end{aligned}
$$

制約条件について、

$$
\langle w \ominus w', x \ominus x'\rangle := \left\langle \begin{pmatrix} \underline{w} - \underline{w'} \\ \overline{w} - \overline{w'} \end{pmatrix}, \begin{pmatrix} \underline{x} - \underline{x'} \\ \overline{x} - \overline{x'} \end{pmatrix} \right\rangle
$$

であり、$\left\langle \begin{pmatrix} \cdot \\ \cdot \end{pmatrix}, \begin{pmatrix} \cdot \\ \cdot \end{pmatrix} \right\rangle$ は $\mathbb{R}^{2n}$ の任意の内積である。

目的関数と制約条件を、

$$
\begin{aligned}
f(w \ominus w', b, \xi) &= \frac{1}{2}\|w \ominus w'\|^2 + C\sum_{i=1}^{n}\xi_i \\
g_i(w \ominus w', b, \xi) &= 1 - \xi_i - y_i(\langle w \ominus w', x_i \ominus 0\rangle - b) \\
h_i(w \ominus w', b, \xi) &= -\xi_i
\end{aligned}
$$

と定義し、最適化問題を次のように書き換える :

$$
\begin{aligned}
&\text{Minimize} && f(w \ominus w', b, \xi) \\
&\text{subject to} && g_i(w \ominus w', b, \xi) \leqq 0,\ h_i(w \ominus w', b, \xi) \leqq 0, \quad i = 1, 2, \cdots, n \\
& && w \ominus w' \in \mathcal{N},\ b \in \mathbb{R},\ \xi \in \mathbb{R}^n.
\end{aligned}
$$

ここで、任意の $i = 1, 2, \cdots, n$ に対して g_i と h_i はアフィン関数であり、定理 1.1 の条件 (i) をみたしているので、条件 (ii) の式が成り立つ:

$$
\begin{aligned}
&\inf_{g_i \leqq 0, h_i \leqq 0} f(w \ominus w', b, \xi) \\
&= \max_{u_i \geqq 0, v_i \geqq 0} \inf_{(w \ominus w', b, \xi)} \left\{ f(w \ominus w', b, \xi) + \sum_{i=1}^{n} u_i g_i(w \ominus w', b, \xi) + \sum_{i=1}^{n} v_i h_i(w \ominus w', b, \xi) \right\}
\end{aligned}
$$

右辺の inf 以降の式は変数 $(w \ominus w', b, \xi)$ について微分可能な凸関数である。従って、右辺を偏微分して 0 になるときを観察する。

$$
0 = \nabla f(w \ominus w', b, \xi) + \sum_{i=1}^{n} u_i \nabla g_i(w \ominus w', b, \xi) + \sum_{i=1}^{n} v_i \nabla h_i(w \ominus w', b, \xi)
$$

このとき、

$$
\begin{aligned}
\nabla f(w \ominus w', b, \xi) &= (w \ominus w', 0, (C, \cdots, C)^T)^T, \\
\nabla g_i(w \ominus w', b, \xi) &= (-y_i x_i \ominus 0, -y_i, -1)^T, \\
\nabla h_i(w \ominus w', b, \xi) &= (0, 0, -1)^T
\end{aligned}
$$

$$
\therefore w \ominus w' = \sum_{i=1}^{n} u_i y_i x_i \ominus 0,\ 0 = \sum_{i=1}^{n} u_i y_i,\ 0 = (C, \cdots, C)^T - u - v
$$

したがって、

$$
\begin{aligned}
&\inf_{(w\ominus w',b,\xi)}\left\{f(w\ominus w',b,\xi)+\sum_{i=1}^{n}u_ig_i(w\ominus w',b,\xi)+\sum_{i=1}^{n}v_ih_i(w\ominus w',b,\xi)\right\}\\
&=\inf_{(w\ominus w',b,\xi)}\left\{\frac{1}{2}\|w\ominus w'\|^2+C\sum_{i=1}^{n}\xi_i+\sum_{i=1}^{n}u_i\left\{1-\xi_i-y_i(\langle w\ominus w',x_i\ominus 0\rangle+b)\right\}-\sum_{i=1}^{n}v_i\xi_i\right\}\\
&=\inf_{\xi\in\mathbb{R}^n}\left\{\frac{1}{2}\|\sum_{i=1}^{n}u_iy_ix_i\ominus 0\|^2+\sum_{i=1}^{n}\xi_i(C-u_i-v_i)+\sum_{i=1}^{n}u_i-\|\sum_{i=1}^{n}u_iy_ix_i\ominus 0\|^2\right\}\\
&=-\frac{1}{2}\|\sum_{i=1}^{n}u_iy_ix_i\ominus 0\|^2+\sum_{i=1}^{n}u_i\\
&=-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}u_iu_jy_iy_j\left\langle\begin{pmatrix}\underline{x}_i\\ \overline{x}_i\end{pmatrix},\begin{pmatrix}\underline{x}_j\\ \overline{x}_j\end{pmatrix}\right\rangle+\sum_{i=1}^{n}u_i
\end{aligned}
$$

以上より、ソフトマージン SVM の最適化問題の双対問題は以下のようになる:

$$
\begin{aligned}
\text{Maximize}\quad & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}u_iu_jy_iy_j\left\langle\begin{pmatrix}\underline{x}_i\\ \overline{x}_i\end{pmatrix},\begin{pmatrix}\underline{x}_j\\ \overline{x}_j\end{pmatrix}\right\rangle+\sum_{i=1}^{n}u_i\\
\text{subject to}\quad & \sum_{i=1}^{n}u_iy_i=0,\ 0\leqq u_i\leqq C,\quad i=1,2,\cdots,n.
\end{aligned}
$$

4　区間データを用いた分析

ここで、実データを用いて区間データの SVM を行った結果を紹介する。今回は、実際に区間データとして収集されているものが少なく、適当なものを見つけることができなかったため、[最低気温, 最高気温] を区間データとみなして SVM を行う。データは 1964 年〜2022 年の、「北海道の水稲の作況指数（e-stat より）」と「北海道（岩見沢）の気象データ（気象庁より）」を使用する。上記のデータより、説明変数として最低気温、最高気温、降水量、最低湿度、最高湿度、またラベルとして作況指数 (平年並みを 100 とし、作況指数 $\geqq 104 \to 1$、作況指数 $< 104 \to -1$) を抽出し、計 57 個のデータに対して SVM を行う。作況指数について、各ラベルのデータ数ができるだけ等しくなるようにするため、今回は 104 を基準にラベルを付けた。水稲は田植えが 5 月頃、収穫が 10 月頃となるので、期間を 5〜10 月とすると 1 年分のデータは以下の通り :

$$
\left[\begin{pmatrix}\text{最低気温 5 月}\\ \vdots\\ \text{最低気温 10 月}\\ \text{降水量 5 月}\\ \vdots\\ \text{降水量 10 月}\\ \text{最低湿度 5 月}\\ \vdots\\ \text{最低湿度 10 月}\end{pmatrix},\begin{pmatrix}\text{最高気温 5 月}\\ \vdots\\ \text{最高気温 10 月}\\ \text{降水量 5 月}\\ \vdots\\ \text{降水量 10 月}\\ \text{最高湿度 5 月}\\ \vdots\\ \text{最高湿度 10 月}\end{pmatrix}\right]\ominus\left[\begin{pmatrix}0\\ \vdots\\ 0\end{pmatrix},\begin{pmatrix}0\\ \vdots\\ 0\end{pmatrix}\right]
$$

ここで、SVM の精度を評価するための指標として、以下で定義する正答率（Accuracy）を用いた。

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{Total samples}}$$

た、TP(True positives) は実際のラベルが 1 で、1 だと予測されたサンプル数を表し、TN(True negatives) は実際のラベルが -1 で、-1 だと予測されたサンプル数を表す。今回、サンプル数が少なかったため、ランダムに並べられたデータを分割して分析を行うことでトレーニングデータとテストデータのラベルの偏りを解消した。テストの手順は以下の通り:

(i) データ 1 とデータ 2 をトレーニングデータ、データ 3 をテストデータ

(ii) データ 2 とデータ 3 をトレーニングデータ、データ 1 をテストデータ

(iii) データ 3 とデータ 1 をトレーニングデータ、データ 2 をテストデータ

データをランダムに並び変えるための指標を変えるごとに、3 回分の SVM を行うことになる。この三回分の正答率の平均値をそれぞれのデータの並べ方に対して導出した。その結果、50 回分の平均を取ると 61.66% だった。適切な結果が得られたと判断できる基準は約 70% 以上とされている場合が多く、今回の結果はその基準に満たなかった。改善点として、以下の 2 つが挙げられる:

(i) 区間データの取得方法

(ii) $\mathbb{R}^{2n}$ 上で定義された内積

(i) について、本来、区間データに対して分析を行うことは一般的ではなく、データが区間のまま収集されることはほとんどない。今回はもともと 1 つの値として収集されたデータを区間データに加工したが、最初から区間データとして収集されたデータに対して SVM を行うことでよりよい結果が得られると考える。また、$\mathcal{N}$ 上の内積については今回採用した $\mathbb{R}^{2n}$ 上で定義された内積だけでなく、いくつか存在する。扱うデータに対して適切な内積を選択することができれば、より精度を向上させることができると考える。

参考文献

[1] M .A .Goberna, V. Jeyakumar,M. A. López; Necessary and sufficient constraint qualifications for solvability of systems of infinite convex inequalities. Nonlinear Anal. 68 (2008), no.5, 1184–1194.

[2] F. J. Gould, Jon W. Tolle; A Necessary and Sufficient Qualification for Constrained Optimization. SIAM Journal on Applied Mathematics, Vol.20, No.2 (Mar., 1971), pp.164-172.

[3] D. Kuroiwa, T. Mori; Inner products on intervals in $\mathbb{R}^n$, Thai Journal of Mathematics. Vol.22, no.3 (2024),509-518.

[4] 気象庁, 北海道岩見沢の気象データ (1964〜2022 年), `https://www.jma.go.jp` より取得, [2024/11/14].

[5] 農林水産省, 作物統計調査, e-Stat, `https://www.e-stat.go.jp` より取得, [2024/11/27].