実験データ解析演習　最終レポート
新潟県のお米の収量と天候の関連性について

2610180082 2 年 1 組 No. 39 水野 響
共同メンバー：藤波真尋, 山崎大知
グループ番号：1

平成 31 年 11 月 29 日

目 次

1　研究背景

一般的に, 農業はその土地の気候と密接に関係していることが知られており, その歴史は古い. 713 年の諸国への風土記の編纂命令には「産物」の記載が含まれており, その頃には既に特産品に相当する概念が既にあったことがうかがえる. しかし現在, 気候変動や異常気象などで伝統的にその土地で育てられている農作物が不作となってしまうことや, 育てる農作物が変化している現状がある.

その中で古くから親しまれている日本人の主食, 米は日本各地で生産されており, 特に生産量 1 位である新潟県は古くからの日本で有数のお米の生産地である. そこで我々は新潟県とその気候を調べ, どのような要因が生産量に影響を与えているかを調べ, 天候から豊作不作を判定したいと考えた. また, その結果から他の土地に稲作の可能性を見出せると考える.

2　テーマ

主成分分析で天候に大きく寄与する成分を見つける
サポートベクターマシーンで新潟のお米の収量とその気候の関連性を調べる

3　原理

3.1　サポートベクターマシン

サポートベクターマシンでは, 自明に 2 群に線形分離可能である学習データを用いてその 2 群を最もよく分離する関数を考える. 個体あるいは対象を特徴づける p 個の変数 $\boldsymbol{x} = (x_1, x_2, \cdots ., x_p)^T$ に関して観測された n 個の学習データ $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ とする. これらの学習データは 2 つの群 G_1, G_2 のどちらに属するものであるかが既に分かっているとする. ここで 2 群を分離する超平面

$$\boldsymbol{w}^T\boldsymbol{x} + b = 0$$

を考える. ここで, $\boldsymbol{w} = (w_1, w_2, \cdots, w_p)^T$ とする. 線形分離可能な学習データに対しては分離超平面によってデータは $\boldsymbol{w}^T\boldsymbol{x}_i + b > 0$ または $\boldsymbol{w}^T\boldsymbol{x}_i + b < 0$ のどちらかを満たす. ここで $(\boldsymbol{w}^T\boldsymbol{x} + b > 0 : G_1,\ \boldsymbol{w}^T\boldsymbol{x} + b < 0 : G_2)$ すなわち

$$\boldsymbol{x}_i \in G_1,\ \leftrightarrow \boldsymbol{w}^T\boldsymbol{x} + b > 0, \quad \boldsymbol{x}_i \in G_2,\ \leftrightarrow \boldsymbol{w}^T\boldsymbol{x} + b < 0,$$

$$y = \begin{cases} 1 & (\boldsymbol{x}_i \text{が } G_1 \text{に属するとき}) \\ -1 & (\boldsymbol{x}_i \text{が } G_2 \text{に属するとき}) \end{cases} \tag{1}$$

とすると

全てのデータに対して

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) > 0$$

が成り立つ.

では, 線形分離可能な学習データに対してどのように超平面を構成すれば良いのだろうか.

このための基準として用いるのが距離である. 一般に p 次元データ $\boldsymbol{x} = (x_1, x_2, \cdots ., x_p)^T$ から超平面 $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ 上への距離は

$$d_i = \frac{|w_1x_{i1} + w_2x_{i2} + \cdots + w_px_{ip} + b|}{\sqrt{w_1^2 + \cdots + w_p^2}} = \frac{|\boldsymbol{w}^T\boldsymbol{x}_i + b|}{||\boldsymbol{w}||}$$

である.

マージン最大化　線形分離可能な場合は二軍を分離する 1 つの超平面 H とそれを中間に挟む形で 2 つの超平面 H_+, H_- が存在する．ここで, 両端の超平面には少なくとも 1 つのデータが存在し, その間にはデータは存在しないとする．サポートベクターマシーンでは, この H_+ と H_- の中央にある H を最適な分離超平面と定義する．

このため, 超平面 H_+ 上のデータ $\boldsymbol{x}_+$ あるいは H_- 上のデータ $\boldsymbol{x}_-$ から超平面 H への距離

$$d = \frac{\boldsymbol{w}^T\boldsymbol{x}_+ + b}{||\boldsymbol{w}||} = \frac{-(\boldsymbol{w}^T\boldsymbol{x}_- + b)}{||\boldsymbol{w}||}$$

をマージンと呼び, 2 群を最もうまく分類する問題を, このマージンを最大とする分離超平面を求める問題へと帰着させる．

つまり

$$\max_{\boldsymbol{w},b} d \quad s.t. \quad \frac{y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)}{||\boldsymbol{w}||} \geq d$$

となる．しかし, このマージン最大化問題の解を求めることは計算上難しいことから, 2 次計画題とよばれる最適化問題へと式を変形する．

一般に超平面 $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ に任意の実数 $\boldsymbol{r} > \boldsymbol{0}$ をかけて, 係数のスケールを変えたとしても超平面は不変である．

ここでマージン最大化の条件 $\frac{|\boldsymbol{w}^T\boldsymbol{x}_i+b|}{||\boldsymbol{w}||} \geq d$ に注目する．この両辺をマージン d で割ると,

$$y_i\left(\frac{\boldsymbol{w}^T\boldsymbol{x}_i}{d||\boldsymbol{w}||} + \frac{b}{d||\boldsymbol{w}||}\right) \geq 1$$

となる．

そこで, $r = 1/d||\boldsymbol{w}||$ とおいて係数のスケールを変えると, 上の式は

$$y_i(\boldsymbol{w}^{*\boldsymbol{T}}\boldsymbol{x_i} + b^*) \geq 1$$

というようになる．ただし,

$$\boldsymbol{w}^* = \frac{1}{d||\boldsymbol{w}||}\boldsymbol{w}, \qquad b^* = \frac{1}{d||\boldsymbol{w}||}b$$

とおく．

これは特に超平面 H_+, H_- に対しては等号が成立する．ゆえに, 等号が成立する超平面上のデータから分離超平面 H 上への距離, すなわちマージン d^* は

$$d^* = \frac{y_i(\boldsymbol{w}^{*T}\boldsymbol{x}_i + b^*)}{||\boldsymbol{w}^*||} = \frac{1}{||\boldsymbol{w}^*||}$$

となることがわかる．

以上のことから, スケール変換した分離超平面 $\boldsymbol{w}^T\boldsymbol{x}_i + b^* = 0$ に対しては, 線形分離可能なすべての学習データは $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b^*) \geq 1$ を満たしており, それゆえマージン最大化問題はこの不等式制約条件のもとでマージン $d^* = 1/||\boldsymbol{w}^*||$ を最大にする $\boldsymbol{w}^*$ と b^* を求める問題に帰着する．

また, マージン d^* の最大化は分母の重みベクトルのノルムの 2 乗 $||\boldsymbol{w}||^{*2}$ の最小化と同等であることから, マージン最大化問題は次のように定式化することができる．ただし, d^*, $\boldsymbol{w}^*$, b^* はすべて, d, $\boldsymbol{w}$, b と表記する．

主問題：マージン最大化問題

$$\min_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w}||^2, \qquad \text{制約条件} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, \quad i = 1, 2, \cdots, n \qquad \cdots(1)$$

サポートベクターマーシンでは $(*)$ を主問題としてラグランジュ関数を導入して双対問題と呼ばれる最適化問題に帰着させて解く.

まず制約条件の個数 n に相当するラグランジュ乗数 $\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_n\ (\alpha_i \geq 0;\ i = 1, 2, \cdots, n)$ を用いて, ラグランジュ関数

$$L(\boldsymbol{w}, b, \alpha_1, \alpha_2, \cdots, \alpha_n) = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{n} \alpha_i \left\{ y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1 \right\} \qquad \cdots (*)$$

を導入する.

次に, 重みベクトル $\boldsymbol{w}$ と切片 b でラグランジュ関数を偏微分した式を 0 とおくと

$$\frac{\partial L(\boldsymbol{w}, b, \alpha_1, \alpha_2, \cdots, \alpha_n)}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_n y_i \boldsymbol{x}_i = \boldsymbol{0}$$

$$\frac{\partial L(\boldsymbol{w}, b, \alpha_1, \alpha_2, \cdots, \alpha_n)}{\partial b} = -\sum_{i=1}^{n} \alpha_n y_i = 0$$

となる. よって

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i, \qquad \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \cdots (2)$$

を得る.

$$L_D(\alpha_1, \alpha_2, \cdots, \alpha_n) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

が求まる.

このようにして主問題 (1) をラグランジュ関数に関する最適化問題へと帰着させる.

$$\max_{\alpha_1, \alpha_2, \cdots, \alpha_n} L_D(\alpha_1, \alpha_2, \cdots, \alpha_n) = \max_{\alpha_1, \alpha_2, \cdots, \alpha_n} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \right\},$$

$$\text{制約条件：}\ \alpha_i \geq 0\ \ (i = 1, 2, \cdots, n),\ \ \sum_{i=1}^{n} \alpha_i y_i = 0$$

この双対問題の解を $\hat{\alpha_1}, \hat{\alpha_2}, \cdots, \hat{\alpha_n}$ とすると, 式 (2) の重みベクトル $\boldsymbol{w}$ に含まれるラグランジュ乗数を双対問題の解で置き換えることにより最適解

$$\hat{\boldsymbol{w}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x}_i$$

を得る. また切片 b については

$$H_+ : \hat{\boldsymbol{w}}^T \boldsymbol{x}_+ + b = 1, \qquad H_- : \hat{\boldsymbol{w}}^T \boldsymbol{x}_- + b = -1$$

が成り立つことを用いると,

$$\hat{b} = -\frac{1}{2}(\boldsymbol{w}^T \boldsymbol{x}_+ + \boldsymbol{w}^T \boldsymbol{x}_-)$$

が求まる. 以上より線形分離可能である学習データに対してマージンが最大となる 2 群を分離する超平面は

$$\hat{\boldsymbol{w}}^T \boldsymbol{x} + \hat{b} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + \hat{b} = 0$$

である. 従って最適な分離超平面を判別関数とする 2 群 $G_1,\ G_2$ の判別法は

$$\hat{\boldsymbol{w}}^T\boldsymbol{x}+\hat{b}=\sum_{i=1}^{n}\hat{\alpha}_i y_i\boldsymbol{x}_i^T\boldsymbol{x}+\hat{b}\begin{cases}\geq 0\Longrightarrow & G_1\\ <0\Longrightarrow & G_2\end{cases}\qquad\cdots(3)$$

で与えられる. サポートベクターマシンは実質的に判別関数を構築するデータは最適な超平面を挟む 2 つの超平面 H_-, H_+ 上のデータ, すなわち

$$y_i(\hat{\boldsymbol{w}}^T\boldsymbol{x}_i+\hat{b})=1$$

を満たすデータのみである. それらのデータはサポートベクターと呼ばれ, サポートベクターによって構成される識別機をサポートベクターマシンという. 従って, サポートベクターの添字集合を S とすると, (3) は

$$\hat{\boldsymbol{w}}^T\boldsymbol{x}+\hat{b}=\sum_{i\in S}\hat{\alpha}_i y_i\boldsymbol{x}_i^T\boldsymbol{x}+\hat{b}\begin{cases}\geq 0\Longrightarrow & G_1\\ <0\Longrightarrow & G_2\end{cases}\qquad\cdots(4)$$

と表せる.

サポートベクターマシンの大きな特徴は, 次の Karush-Kuhn-Tucker 条件から導かれることが知られている.

Karush-Kuhn-Tucker 条件　p 次元実数値関数を $f(w_1,\cdots w_p)=f(\boldsymbol{w})$ とする. この目的関数を最小とする解を, n 個の不等式制約条件 $g_i(\boldsymbol{w})\leq 0,\ i=1,2,\cdots,n$ のもとで求めるとする. すなわち,

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}),\qquad \text{制約条件}\ \ g_i(\boldsymbol{w})\leq 0,\quad i=1,2,\cdots,\ n\qquad\cdots(5)$$

次にラグランジュ乗数 $\boldsymbol{\alpha}=(\alpha_1,\cdots\alpha_n)^T$ を導入したラグランジュ関数を

$$L(\boldsymbol{w},\boldsymbol{\alpha})=f(\boldsymbol{w})+\sum_{i=1}^{n}\alpha_i g_i(\boldsymbol{w})\quad\cdots(6)$$

とする.　このとき, 不等式制約条件下での最適化問題は次の 4 つの条件を満足する重みベクトル $\boldsymbol{w}$ とラグランジュ乗数 $\boldsymbol{\alpha}$ を求める問題に帰着される.

$$\begin{aligned}
&(1)\quad \frac{\partial L(\boldsymbol{w},\boldsymbol{\alpha})}{\partial w_i}=0 && i=1,2,\cdots,p\\
&(2)\quad \frac{\partial L(\boldsymbol{w},\boldsymbol{\alpha})}{\partial \alpha_i}=g_i(\boldsymbol{w})\leq 0 && i=1,2,\cdots,n\\
&(3)\quad \alpha_i\geq 0 && i=1,2,\cdots,n\\
&(4)\quad \alpha_i g_i(\boldsymbol{w})=0 && i=1,2,\cdots,n
\end{aligned}$$

サポートベクター　マージンを最大とするという意味で最適な分離超平面を求めるために用いた式 $(*)$ のラグランジュ関数は, 目的関数 $f(\boldsymbol{w})$ と制約条件式 $g_i(\boldsymbol{w})$ をそれぞれ

$$f(\boldsymbol{w}=\frac{1}{2}||\mathrm{w}||^2,\ \ g_i(\boldsymbol{w})=-\left\{y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)-1\right\}\quad i=1,2,\cdots,n$$

とおいたものである. このとき, Karush-Kuhn-Tucker 条件の一つである (4) に対して

$$\alpha_i\{y_i(\boldsymbol{w}^T\boldsymbol{x}+b)-1\}=0,\qquad i=1,2,\cdots,n$$

が成り立っていることを示している. データがサポートベクターでない場合は $y_i(\boldsymbol{w}^T\boldsymbol{x}+b)>1$ となり, 上式より $\alpha_i=0$ となる. 一方サポートベクターに対しては $y_i(\boldsymbol{w}^T\boldsymbol{x}+b)=1$ が成り立ち, $\alpha_i>0$ が成り立つ. 従って式 (4) の判別関数を構成するデータは, $\alpha_i>0$ に対応するサポートベクターのみであることが示される.

4　データ

使用したデータの種類と出典は表 1 の通りである.

表 1: データ

	データの種類	出典
米の収量	10a あたりの収量 (t)	eStat
天気	日照時間, 気温, 湿度, 降水量, 風速, 雲量	気象庁

5　操作

1. 6 次元の気候データを主成分分析により 2 次元に圧縮した.
2. 収量を 480 (t) 以上の年を豊作とし, ラベルを豊作の年を 1, 不作の年を 0 としてラベル付けをした.
3. 主成分分析した 2 次元データと固有ベクトルを月ごとにプロットした.
4. 第一主成分と第二主成分の配列を用意した.
5. 主成分を train データと test データを 7 : 3 の割合で分割した.
6. サポートベクターマシーンを学習させた.
7. サポートベクターマシーンの正答率を求めた.
8. 横軸を第一主成分, 縦軸を第二主成分としてデータをプロットし, サポートベクターマシーンの分類結果を可視化した.
9. 以上を 12ヶ月分同様に行った.
10. result 配列を用意し, 月ごとの第一主成分と第二主成分とサポートベクターマシーンの正答率を格納した.

6　結果

操作 3 による散布図が図 1 である. 操作 8 による散布図が図 2 である.

月ごとの第一主成分と第二主成分とサポートベクターマシーンの正答率が表 2 である. オレンジ色の方が豊作で, 青色の部分が不作を表している.

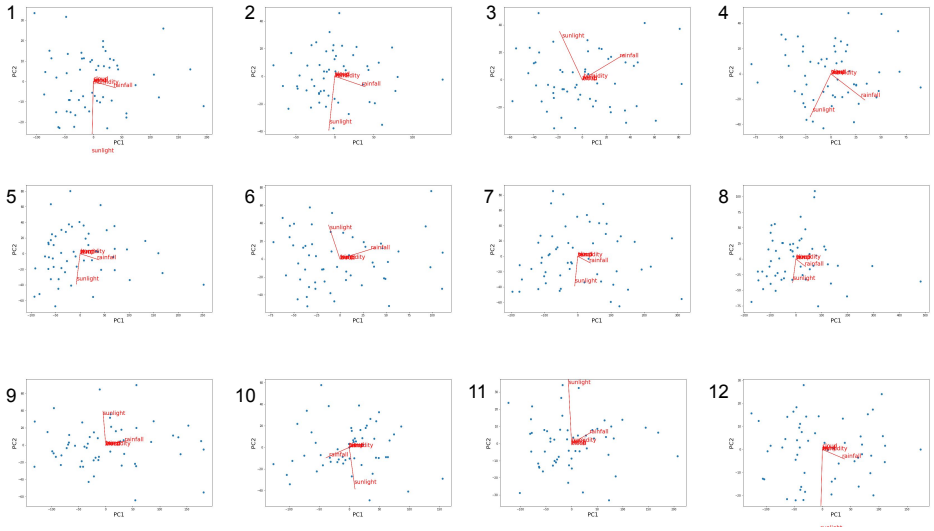

図 1: PCA による分析結果

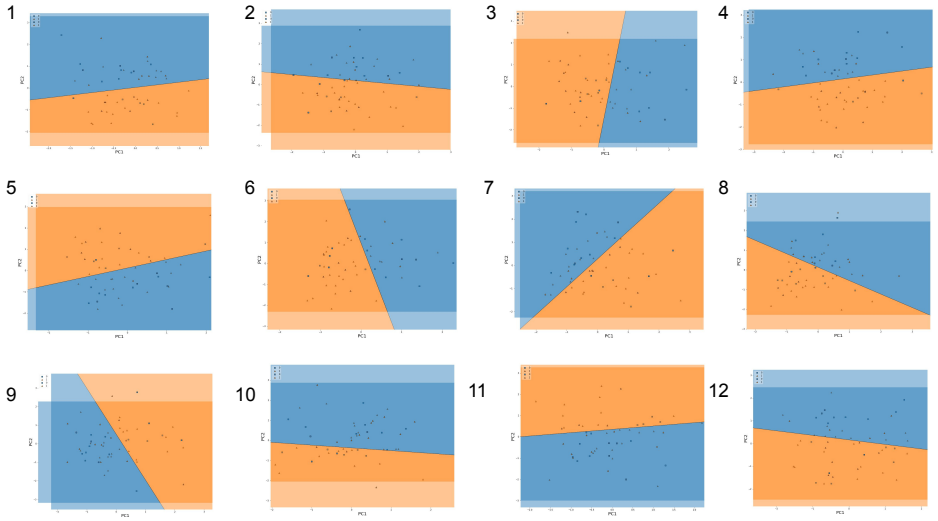

図 2: SVM による分析結果

表 2: 月ごとの固有ベクトルと SVM 正答率

month	PC1	PC2	SVM_score
1	0.948290	0.049088	0.529412
2	0.841302	0.151448	0.352941
3	0.726571	0.266117	0.588235
4	0.707992	0.285218	0.705882
5	0.718615	0.278711	0.882353
6	0.835200	0.163343	0.529412
7	0.907491	0.091873	0.588235
8	0.902660	0.096692	0.352941
9	0.906668	0.092054	0.352096
10	0.867753	0.129710	0.235294
11	0.953945	0.043966	0.470588
12	0.965803	0.031950	0.588200

7　考察

第二主成分の寄与率が高い月は SVM の正答率が高い. 特に 5 月は SVM の正答率が高い. それら以外はあてになる分析結果とはいえない. ここでは表 2 から SVM の正答率が 0.7 を超えている 4 月と 5 月について考える. 4 月は図 2 から PC1 の影響はほぼ関係がなく, PC2 の影響のみを考える. 図 1 の PCA の固有ベクトルと図 2 の SVM の結果から, 日照時間が長ければ豊作になり, 降水量は関係がない.

5 月は 4 月同様, 降水量は関係がなく, 日照時間は短い方が豊作になる.

ところで, 4 月は稲作では種籾を用意して発芽させてタネを育苗箱に蒔き, その後ハウス内で苗を育てる時期に相当する. すなわち 4 月の天候, 特に降水量と日照時間は稲作の様々な要因のうち, 田んぼの土のみに影響を与えると考えられる. 一般的に土に日光を当てる行為は, 日光消毒として知られている. そのため, 稲作においては 4 月における土の日光消毒が重要な役割を担っていると考えられる.

また. 5 月は田おこしをし, しろかき (水を引き入れてトラクターでよく混ぜる) を行った後, 田植えを行う時期である. 植えたばかりの稲は弱いので水を深めのまま管理する深水管理を行うことが知られていることから, 日照時間が短い方が良いということはこれに係る水の蒸発が稲作に良い影響を及ぼさないということが読み取れる.

すなわち以上をまとめると, 稲作は 4 月の田んぼの日光消毒と, 5 月の深水管理が最も需要な豊作の因子であることが考えられる.

8　次への課題

4 月 5 月以外の SVM による分析が使い物にならなかった. 次元拡張をすれば分離できたかもしれない.

参考文献

[1] 小西貞則「多変量解析入門——線形から非線形へ」p.193∼203