

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：藤浪真尋, 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	2
2	テーマ	2
3	原理	2
3.1	サポートベクターマシン	2
4	データ	4
5	操作	4
6	結果	5
7	考察	6
8	次への課題	7

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 $\mathbf{x} = (x_1, x_2, \dots, x_p)^T$ に関して観測された n 個の学習データ $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ 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

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

を考える。ここで、 $\mathbf{w} = (w_1, w_2, \dots, w_p)^T$ 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 $\mathbf{w}^T \mathbf{x}_i + b > 0$ または $\mathbf{w}^T \mathbf{x}_i + b < 0$ のどちらかを満たす。ここで $(\mathbf{w}^T \mathbf{x} + b > 0 : G_1, \mathbf{w}^T \mathbf{x} + b < 0 : G_2)$ すなわち

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

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

とすると

全てのデータに対して

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

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 $\mathbf{x} = (x_1, x_2, \dots, x_p)^T$ から超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 上への距離は

$$d_i = \frac{|w_1 x_{i1} + w_2 x_{i2} + \dots + w_p x_{ip} + b|}{\sqrt{w_1^2 + \dots + w_p^2}} = \frac{|\mathbf{w}^T \mathbf{x}_i + b|}{\|\mathbf{w}\|}$$

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 \mathbf{x}_+ あるいは H_- 上のデータ \mathbf{x}_- から超平面 H への距離

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

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

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

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 $\mathbf{w}^T \mathbf{x} + b = 0$ に任意の実数 $r > 0$ 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 $\frac{|\mathbf{w}^T \mathbf{x}_i + b|}{\|\mathbf{w}\|} \geq d$ に注目する. この両辺をマージン d で割ると,

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

となる.

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

$$y_i(\mathbf{w}^{*T} \mathbf{x}_i + b^*) \geq 1$$

というようになる. ただし,

$$\mathbf{w}^* = \frac{1}{d\|\mathbf{w}\|} \mathbf{w}, \quad b^* = \frac{1}{d\|\mathbf{w}\|} b$$

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

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

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 $\mathbf{w}^T \mathbf{x}_i + b^* = 0$ に対しては, 線形分離可能なすべての学習データは $y_i(\mathbf{w}^T \mathbf{x}_i + b^*) \geq 1$ 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 $d^* = 1/\|\mathbf{w}^*\|$ を最大にする \mathbf{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 $\|\mathbf{w}\|^{*2}$ 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^* , \mathbf{w}^* , b^* はすべて, d , \mathbf{w} , b と表記する。

マージン最大化問題

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

4 データ

使用したデータの種別と出典は表 1 の通りである。

表 1: データ

	データの種別	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ンの学習させた。
7. サポートベクターマシ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。

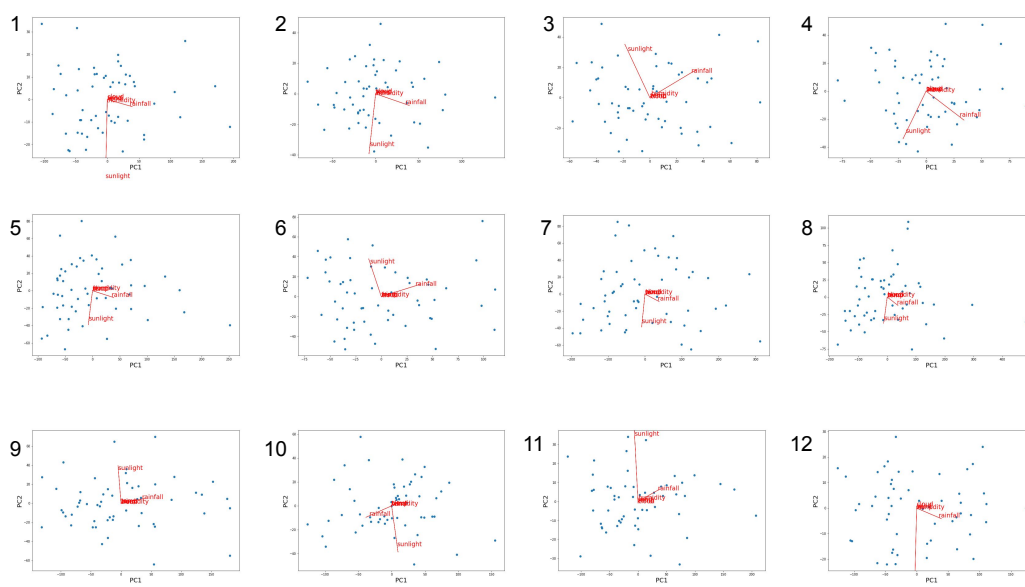


図 1: PCA による分析結果

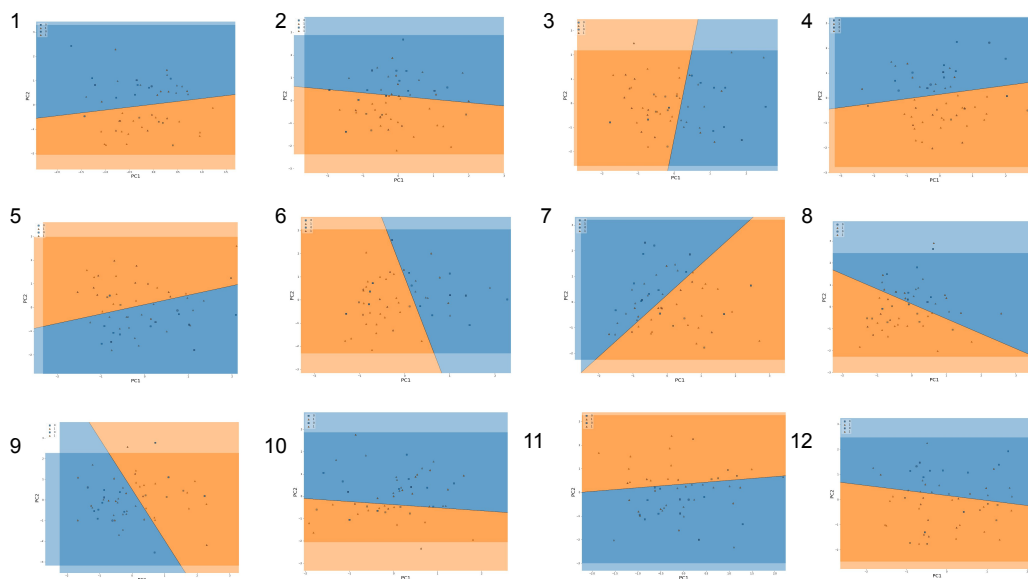


図 2: SVM による分析結果

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	0.3520961
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。

8 次への課題

データをもっとよく観察して、ある程度仮説を立てるべきだった。