二値分類により天気予測を行う機械学習モデルの作成およびアルゴリズム検討

CSVファイルをダウンロード

- 気象庁のサイトから気象データをダウンロード
  (https://www.data.jma.go.jp/risk/obsdl/index.php)
  観測期間: 2020年4月1日から2023年3月31日まで
  観測地域: 東京都

図. 気象庁サイト

	A	B	C	D	E	F	G	H	I	J
1	年月日	平均気温(℃)	天気概況(昼：	降水量の合計	降水量の合計	日照時間(時間	平均風速(m/s	平均蒸気圧(h	平均湿度(%)	平均雲量(10分比)
2	2020/4/1	11.4	雨	31	0	0	2.7	13.1	97	10
3	2020/4/2	13.9	晴	0	0	11.9	6.2	7	45	3.8
4	2020/4/3	12.8	晴後薄曇	0	1	9.9	3.3	7.6	50	4.8
5	2020/4/4	15.7	快晴	0	1	11.5	3.9	10.6	62	0.5
6	2020/4/5	10.2	曇後雨一時晴	0	0	1.2	2.6	9.2	73	6.3
7	2020/4/6	12.1	晴	0	1	11.3	3	6	43	4.5
8	2020/4/7	12.4	薄曇	0	1	8.9	2.7	7.2	50	6.5
9	2020/4/8	14.3	快晴	0	1	11.7	2.6	9.5	60	2
10	2020/4/9	13.3	晴後曇	0.5	0	8.3	3.3	9.1	62	5
11	2020/4/10	11.2	晴一時薄曇	0	1	9.9	3.8	6.2	49	6.8
12	2020/4/11	11.4	晴後時々曇	0.5	0	9	3.2	6.4	48	6.5
13	2020/4/12	10.3	曇一時雨	6.5	0	0.6	2.5	10.4	84	10
14	2020/4/13	8	大雨	132	0	0	3.9	10.5	97	10
15	2020/4/14	10.9	快晴	1.5	0	11.8	4.7	6.6	54	2.5
16	2020/4/15	13.8	薄曇一時晴	0	1	9.1	3.1	7.7	48	7.3
17	2020/4/16	11.7	曇	0.5	0	5.9	3.7	9.1	66	10
18	2020/4/17	12.4	曇時々晴	0.5	0	4.8	3	9	62	9.5
19	2020/4/18	12.9	大雨	89.5	0	0.7	3.1	14.3	96	8
20	2020/4/19	14.7	晴	0	1	11.3	3.5	10.1	64	4.3
21	2020/4/20	9.7	雨	15	0	0	2.5	11.3	94	9.8

図. 気象データのCSVファイル

データの整理

- 天気概況に関する文字情報を数値化
  晴→4, 曇→3, 雨→2, その他(みぞれなど)→1
- 降水量を二値で表現
  降水量=0→0
  降水量>0→1

	A	B	C	D	E	F	G	H	I	J
1	年月日	平均気温(℃)	天気概況(昼:(	降水量の合計	降水量(0or1)	日照時間(時間	平均風速(m/s	平均蒸気圧(h	平均湿度(%)	平均雲量(10分比)
2	2020/4/1	11.4	2	31	1	0	2.7	13.1	97	10
3	2020/4/2	13.9	4	0	0	11.9	6.2	7	45	3.8
4	2020/4/3	12.8	4	0	0	9.9	3.3	7.6	50	4.8
5	2020/4/4	15.7	4	0	0	11.5	3.9	10.6	62	0.5
6	2020/4/5	10.2	2	0	0	1.2	2.6	9.2	73	6.3
7	2020/4/6	12.1	4	0	0	11.3	3	6	43	4.5
8	2020/4/7	12.4	3	0	0	8.9	2.7	7.2	50	6.5
9	2020/4/8	14.3	4	0	0	11.7	2.6	9.5	60	2
10	2020/4/9	13.3	4	0.5	1	8.3	3.3	9.1	62	5
11	2020/4/10	11.2	4	0	0	9.9	3.8	6.2	49	6.8
12	2020/4/11	11.4	4	0.5	1	9	3.2	6.4	48	6.5
13	2020/4/12	10.3	2	6.5	1	0.6	2.5	10.4	84	10
14	2020/4/13	8	2	132	1	0	3.9	10.5	97	10
15	2020/4/14	10.9	4	1.5	1	11.8	4.7	6.6	54	2.5
16	2020/4/15	13.8	3	0	0	9.1	3.1	7.7	48	7.3
17	2020/4/16	11.7	3	0.5	1	5.9	3.7	9.1	66	10
18	2020/4/17	12.4	3	0.5	1	4.8	3	9	62	9.5
19	2020/4/18	12.9	2	89.5	1	0.7	3.1	14.3	96	8
20	2020/4/19	14.7	4	0	0	11.3	3.5	10.1	64	4.3
21	2020/4/20	9.7	2	15	1	0	2.5	11.3	94	9.8

図. 整理後の気象データ

線形回帰で実装

- 全データに対して訓練用:評価用=8:2となるようランダムに配分
- 予測精度の計算を10,000回繰り返し，中央値を採用

結果: 予測精度は**54.24%**

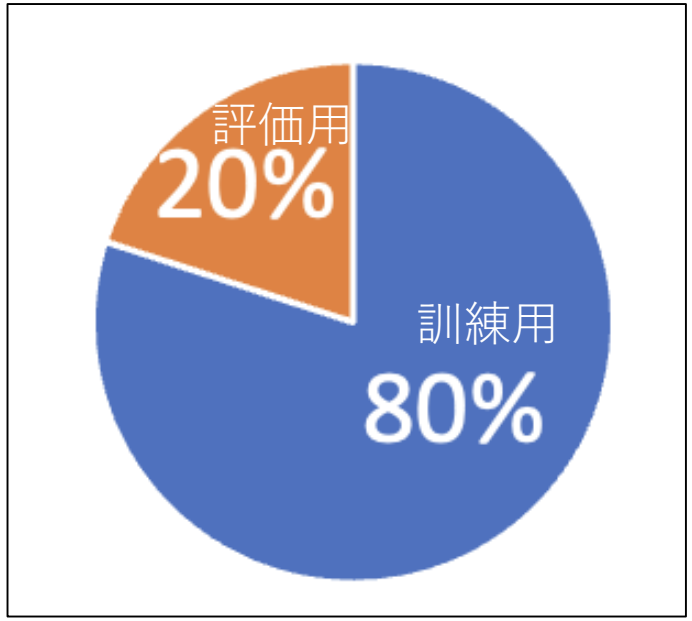

図. 訓練用/評価用データの配分

精度upするための着眼点

- 過学習の可能性

  →CSVファイルから抽出する気象データを絞る
  ①気温は日照時間に伴って上昇するため**日照時間**を除外
  ②**平均蒸気圧**: 気温が上がると増大するため除外

  →①, ②を実施するも、予測精度は**54.27%**とほぼ改善されず

- 別の機械学習アルゴリズムを検討
  →ロジスティック回帰, ナイーブベイズ(正規分布, ベルヌーイ分布)を適用

ロジスティック回帰で実装

結果: 予測精度は**87.61%**

ナイーブベイズで実装

結果: (正規分布)予測精度は**80.82%**
結果: (ベルヌーイ分布)予測精度は**65.30%**

結論

今回検討した

- 線形回帰
- ロジスティック回帰
- ナイーブベイズ(正規分布, ベルヌーイ分布)

のうち、二値分類による予測に最適なアルゴリズムは**ロジスティック回帰**

最後までお読み頂きありがとうございました。