Introduction to Machine Learning 2021 Term Project Final Report

Group 168 Kai Hartzell

Predicting npf events

In this project, I used NB classifier to predict event and nonevent days. The training data, *npf_train.csv*, included 104 columns and 458 observations in total. The test data, *npf_test_hidden.csv*, included 965 unclassified observations. There were columns `id`, `date`, `class4`, `partlybad` and 100 other variables measured.

I dropped out columns `id`, `date` and `partlybad` of both train and test data, because `id` and `date` did not have any impact on the results, and the value of `partlybad` was always false. The `class4` column indicated the observed class, and it was one of these: `II`, `Ia`, `Ib`, `nonevent`.

The binary classification task was to identify nonevent and event classes, ie. `II`, `Ia`, `Ib`. The hidden test data did not contain the class values, but I replaced the `NA` values with a placeholder `nonevent`. I factored the training data classes as `II`, `Ia`, `Ib`, `nonevent`.

Next, I constructed a data frame and table of the class sd and mean values:

	class_non.mean	class_non.sd	classII.mean	classII.sd	classIa.mean	classIa.sd	classIb.mean	classIb.sd
CO2168.mean	383.3257592	12.0163656	379.7816558	9.6374969	377.3983620	7.9319611	377.5385938	8.5728401
CO2168.std	3.8076389	3.7136048	3.4647404	3.0638298	1.8811666	3.3126275	3.3173097	3.0186016
CO2336.mean	383.3010772	12.0127173	379.8355082	9.6230363	377.4714719	7.8794670	377.6037099	8.5602995
CO2336.std	3.5851381	3.4501180	3.2415884	2.8516257	1.7707188	3.1835412	3.1055142	2.7644834
CO242.mean	384.3454936	11.4029920	380.5672238	9.4285409	378.0545534	7.3849728	378.4111268	8.3496778
CO242.std	4.5978001	4.4815486	4.2714589	3.9259490	2.1576477	3.4830427	4.2127625	4.3486898
CO2504.mean	383.1770183	12.0503130	379.7501748	9.6367019	377.4236330	7.8790896	377.5267348	8.5898615
CO2504.std	3.4192060	3.1583781	2.9971388	2.5841796	1.6631054	3.0425304	2.8382270	2.4519342
Glob.mean	124.1843709	108.2884333	265.4114777	96.6022100	217.4514936	136.5090434	273.8362652	94.1315775
Glob.std	100.3933153	89.5181144	196.3972508	70.2540049	148.3179284	97.8803925	197.6149565	70.4300363
H2O168.mean	8.3952240	4.2909264	6.6839081	3.2319905	4.7256178	3.0865356	6.1144060	2.7852484
H2O168.std	0.5305522	0.4666248	0.6812579	0.4348018	0.3949500	0.3097822	0.6062125	0.3944818
H2O336.mean	8.3234504	4.2414113	6.6003523	3.1875941	4.6750551	3.0275964	6.0381327	2.7308122
H2O336.std	0.5293356	0.4655487	0.6725026	0.4262756	0.3950264	0.3053176	0.6076353	0.3960382
H2O42.mean	8.5241697	4.3895086	6.8414829	3.3113565	4.8100257	3.1670608	6.2500283	2.8805575
H2O42.std	0.5447998	0.4825249	0.7035039	0.4504993	0.4007320	0.3314933	0.6189761	0.3851627
H2O504.mean	8.2824671	4.2084802	6.5496678	3.1622404	4.6486063	2.9945831	5.9924726	2.7011730
H2O504.std	0.5274781	0.4679204	0.6699628	0.4251228	0.3928798	0.3059966	0.6094654	0.3996900
H2O672.mean	8.2507405	4.1830761	6.5114863	3.1356407	4.6267247	2.9575198	5.9550852	2.6803969
H2O672.std	0.5299801	0.4680479	0.6695066	0.4254028	0.3879299	0.3077983	0.6088883	0.4012817
H2O84.mean	8.4704343	4.3504865	6.7679388	3.2783437	4.7685026	3.1332267	6.1851300	2.8420338
H2O84.std	0.5398556	0.4769061	0.6954969	0.4453180	0.3943370	0.3208854	0.6118577	0.3918308
NET.mean	80.5261547	74.3785508	167.8337368	73.0851478	128.5413076	93.6546789	171.6901507	73.6059595
NET.std	87.6830282	78.6126033	174.5705225	62.2764092	133.0591203	82.5527469	172.4932287	64.8169662
NO168.mean	0.0832628	0.1178991	0.0486565	0.0551810	0.0840824	0.1708901	0.0618202	0.0845130
NO168.std	0.0868361	0.0701328	0.0762156	0.0385910	0.0756751	0.0743735	0.1103664	0.1480846
NO336.mean	0.0891754	0.1240921	0.0505879	0.0589467	0.0906982	0.1825832	0.0630157	0.0884501
NO336.std	0.0903346	0.0772049	0.0800398	0.0558533	0.0763280	0.0787067	0.0865659	0.0772434

	class_non.mean	class_non.sd	classII.mean	classII.sd	classIa.mean	classIa.sd	classIb.mean	classIb.sd
NO42.mean	0.0708654	0.0952515	0.0372286	0.0413754	0.0679048	0.1414973	0.0495214	0.0668548
NO42.std	0.0985503	0.1232974	0.0735775	0.0415377	0.0749556	0.0684068	0.1128484	0.1676579
NO504.mean	0.0886274	0.1226654	0.0490583	0.0586438	0.0893289	0.1906810	0.0615647	0.0863516
NO504.std	0.0905429	0.0839672	0.0763539	0.0406231	0.0748648	0.0805113	0.0839574	0.0689388
NO672.mean	0.0871272	0.1183312	0.0482930	0.0574997	0.0906560	0.1928020	0.0599156	0.0845935
NO672.std	0.0914093	0.0878955	0.0744356	0.0396351	0.0750297	0.0788534	0.0850076	0.0706539
NO84.mean	0.0696912	0.1038470	0.0393163	0.0464958	0.0749010	0.1535954	0.0525418	0.0760259
NO84.std	0.0797140	0.0603073	0.0729624	0.0419833	0.0751914	0.0719955	0.0944649	0.1160699
NOx168.mean	1.8006764	1.6301601	0.8619501	0.7170419	1.4469960	1.8247297	1.0933323	0.8587382
NOx168.std	0.5197397	0.4535324	0.4625667	0.7003889	0.3706119	0.3411748	0.4621600	0.3910069
NOx336.mean	1.7959607	1.6180526	0.8485011	0.7146270	1.4394845	1.8213682	1.0824423	0.8536414
NOx336.std	0.5499471	0.5687775	0.3930678	0.3140737	0.3676035	0.3367659	0.4216894	0.3590915
NOx42.mean	1.8093522	1.6224146	0.8625380	0.7014339	1.4401144	1.7962648	1.1101887	0.8502164
NOx42.std	0.6384147	0.7791902	0.4397016	0.3326902	0.3895916	0.3643103	0.6164194	0.8641324
NOx504.mean	1.7811794	1.5945210	0.8390183	0.7190962	1.4176531	1.8252514	1.0648497	0.8472283
NOx504.std	0.5897877	0.6577056	0.4328458	0.6036045	0.3526593	0.3312644	0.4132702	0.3432988
NOx672.mean	1.7631446	1.5703224	0.8313492	0.7164193	1.4099343	1.8261845	1.0594163	0.8493411
NOx672.std	0.5487984	0.5379051	0.3815342	0.3356013	0.3555551	0.3316057	0.4136651	0.3522673
NOx84.mean	1.7905142	1.6296561	0.8608148	0.7053100	1.4411304	1.8118328	1.0974465	0.8517585
NOx84.std	0.5127998	0.4434093	0.4490562	0.4547183	0.3744475	0.3504128	0.4846042	0.4074471
O3168.mean	28.9563955	8.7015444	37.1354305	7.9375591	35.5817172	9.5820739	38.1966494	7.6531242
O3168.std	3.5209114	2.3207539	4.0125807	2.2471876	3.1157460	2.1628869	4.0468967	2.3752102
O342.mean	27.7531718	8.5799124	35.9105184	8.1806544	34.7646199	9.6125137	37.1545134	8.0531688
O342.std	3.9320093	2.5715429	4.5629995	2.5121122	3.4300530	2.2983362	4.5916329	2.6110586
O3504.mean	29.9223688	8.7826482	38.1124111	7.6215776	36.1765368	9.4885172	38.9206969	7.4458559
O3504.std	3.3379269	2.1808745	3.6198586	2.0168362	2.9581700	2.1895254	3.6856651	2.1572885
O3672.mean	30.2570159	8.8153168	38.4638542	7.5423407	36.4179137	9.4671344	39.1708386	7.4083422
O3672.std	3.2974286	2.1320519	3.4793625	1.8999729	2.8855798	2.2111996	3.5270694	2.0859228
O384.mean	28.2950117	8.6449395	36.4847757	8.0918533	35.2079586	9.5844538	37.6702692	7.8834566
O384.std	3.6818176	2.4325495	4.2321932	2.3653424	3.1852934	2.1710969	4.2537071	2.4787015
Pamb0.mean	989.5739182	10.4595733	992.4344221	7.5447690	993.0831075	12.7955733	993.8398014	9.2653530
Pamb0.std	0.9036629	0.7414164	1.1832659	0.8141691	1.0277771	0.6889135	1.1659888	0.8041929
PAR.mean	252.1034293	219.8362910	521.8252647	189.1382262	415.7491902	262.7742736	533.1584717	184.6378523
PAR.std	201.6516309	180.6231724	387.9777895	139.0148940	286.6517870	191.6229248	386.9933144	139.2281505
PTG.mean	0.0013185	0.0072903	- 0.0005988	0.0047814	- 0.0007118	0.0053309	- 0.0016611	0.0031918
PTG.std	0.0067485	0.0064188	0.0120327	0.0058381	0.0102281	0.0069848	0.0115483	0.0054228
RGlob.mean	18.2423426	14.4447614	36.2972617	12.9405799	32.9937881	16.9661898	38.3190568	11.9587230
RGlob.std	13.7137875	10.0052002	23.9939113	6.5384811	20.5618458	9.8098910	24.4001760	6.4951044
RHIRGA168.mean	79.9616469	15.9356044	57.2879848	15.6409709	60.9485167	19.8261856	55.8371629	13.4217013
RHIRGA168.std	5.5691925	4.7117565	11.6976471	4.5173211	9.6863927	6.1905295	12.1942132	4.0480959
RHIRGA336.mean	80.4897203	16.2525874	57.4444120	15.8227254	61.3027904	20.0627248	56.0733062	13.6805601
RHIRGA336.std	5.5792926	4.6981200	11.4247132	4.4972056	9.5752945	6.1246428	11.9648446	4.0725800
RHIRGA42.mean	80.3492081	15.1864570	58.4772003	15.3820059	61.1773670	18.9236909	56.8580399	13.2335082
RHIRGA42.std	5.7424427	5.0783234	12.4831364	4.7587657	9.9832382	6.4979431	13.0917220	4.2715860
RHIRGA504.mean	80.5896461	16.3846137	57.3573156	15.8184975	61.3350061	20.1122624	56.1103848	13.8601198
RHIRGA504.std	5.5425971	4.6848747	11.1521853	4.4914608	9.2819531	6.0207258	11.5776651	4.1202661
RHIRGA672.mean	81.4285506	16.7576380	57.8459091	16.1270668	62.0233782	20.3404874	56.6574265	14.2507381
RHIRGA672.std	5.5541586	4.6333869	10.9344953	4.5554081	9.1514718	5.9629777	11.2599693	4.2834540
RHIRGA84.mean	80.0840781	15.5444730	57.6708747	15.5608801	61.0015883	19.5211092	56.0668416	13.3454606
RHIRGA84.std	5.7412174	4.9777942	12.1984752	4.6677814	9.9443157	6.3240611	12.7443155	4.1544103
RPAR.mean	14.1556559	12.5023017	22.5765995	13.2180729	23.3232753	11.0389199	24.4094019	9.8784467

	class_non.mean	class_non.sd	classII.mean	classII.sd	classIa.mean	classIa.sd	classIb.mean	classIb.sd
RPAR.std	10.5616485	8.7993333	15.8318446	7.3983849	15.2332906	6.8970832	16.9022109	6.0921322
SO2168.mean	0.2969173	0.4872757	0.1926844	0.1883541	0.1695263	0.1507744	0.2362119	0.3058720
SO2168.std	0.1529806	0.1386084	0.1577387	0.1276472	0.1289616	0.0805043	0.1684688	0.1246219
SWS.mean	901.2928793	39.4040349	915.2222334	18.7440739	923.1286425	9.5964837	919.5428844	13.0544876
SWS.std	28.9825589	43.5319573	16.4777475	35.0055353	5.8773344	17.8563535	12.6442312	27.7441377
T168.mean	6.0779466	10.9514407	8.5962921	8.1833693	2.8687104	8.3095384	7.6928053	8.0038911
T168.std	1.3599286	0.9772331	2.3834288	0.8861056	2.0180805	1.1075181	2.4789441	0.9679934
T42.mean	6.1653770	10.9224699	8.6499311	8.2085590	2.9865494	8.2421630	7.7541718	8.0064785
T42.std	1.4646568	1.1112941	2.6472802	1.0024390	2.1859501	1.2825396	2.7393903	1.0755007
T504.mean	5.8160029	10.8954991	8.2703335	8.2021538	2.5531536	8.2892509	7.3368535	8.0393041
T504.std	1.2650820	0.9053966	2.1760562	0.8340721	1.8472449	1.0164464	2.2836336	0.9179909
T672.mean	5.6329681	10.8540661	8.0609689	8.1864121	2.3391703	8.2753919	7.1069544	8.0400739
T672.std	1.2171312	0.8754616	2.0874301	0.8033751	1.7835714	0.9649712	2.1963858	0.8942363
T84.mean	6.1500669	10.9556256	8.6976018	8.2085735	2.9538699	8.2793216	7.8076448	8.0135334
T84.std	1.4406672	1.0637270	2.5498165	0.9428052	2.1481955	1.1791491	2.6451795	1.0221479
UV_A.mean	7.6842886	6.1089685	14.5738046	5.0739380	11.3026124	6.9411607	14.7907621	5.0123836
UV_A.std	5.6597847	4.8468768	10.3911167	3.9771654	7.5590838	5.1088458	10.3899427	3.9910193
UV_B.mean	0.3258228	0.3044106	0.6028462	0.2732618	0.4214711	0.2929315	0.5940769	0.2781953
UV_B.std	0.2887586	0.2824756	0.5214276	0.2522659	0.3465650	0.2574466	0.5066760	0.2556018
CS.mean	0.0037101	0.0026296	0.0024940	0.0015237	0.0017976	0.0013724	0.0024524	0.0016285
CS.std	0.0006865	0.0005781	0.0006405	0.0006305	0.0004566	0.0004533	0.0006782	0.0004962

Laplace smoothing of 1 was used for the data. The estimated class probabilities for training data:

```
nb_class
```

```
##
##   nonevent         II         Ia         Ib
## 0.49783550 0.25541126 0.06493506 0.18181818
```

Then I applied NB classifier to compute the class probabilities for all rows of testing data. The variables were considered as conditionally independent. The classifier predicted the probabilities of each class for each row. The row was identified as `nonevent`, if the probability was higher than the `nb_class` probability for that class.

The formula of NB Gaussian density was $\frac{e^{(-(x-\mu)^2/(2*\sigma^2))}}{\sqrt{2*\pi*\sigma^2}}$

I used some small coefficient adjustments and modifications for the multivariate classification problem. I quessed that the accuracy of the binary classification could be 0.73. The head of estimated classes and probabilities:

```
head(df, 10)
```

```
##
## 1       0.73
## 2     class4                    p
## 3         Ia    0.727504450479219
## 4   nonevent    0.080463843559317
## 5         Ib    0.990897111743513
## 6         II    0.991238939899939
## 7   nonevent    0.120253990842419
## 8         II    0.978951849227886
## 9   nonevent 0.000677548460571997
## 10  nonevent  0.00245307701477748
```

This whole data frame was exported as a csv file. The predicted class distributions for testing data for classes `nonevent`, `II`, `Ia`, `Ib`:

```
## [1] 0.4611399
```

```
## [1] 0.2683938
```

```
## [1] 0.09948187
```

```
## [1] 0.1709845
```

Regarding the methods, I considered using cross-validation or SVM. The pros of NB are that it is a highly scalable and simple generative classifier. NB can usually be trained efficiently in supervised learning, and it often requires only a small number of training data.

After making some tweaks, I got the NB classifier to predict reasonable results, but there were initially some problems with the class distributions. I learned the effectiviness and usability of NB classifier.