Machine learning Project

Abhishek Chandra

February 18, 2021

0.1 Gaussian Process

In this part of the project we have to approximate a gaussian fit for booster function. Booster function is computationally expensive due to which GP fit is a very good to approximate the function. In this report combination of various gaussian configuration with varying parameters(lengthscale and variance). We have presented tables for each kernel functions with 400 and 800 total number of points with lengthscale values are 0.5, 1 and 5. In each table we have highlighted the best fit acording to the R2 score. In total four kinds of errors are estimated RMSE, R2score, MNLL and Std RMSE. ALong with the code we have also presented the box whisker plot for RMSE error for 400 and 800 sample points.

kernel	Data	LS	Var	R2	RMSE	Std RMSE	MNLL
RBF	500	0.1	0.1	0.8428	0.6879	0.1072	0.0293
RBF	400	0.5	0.5	0.8271	0.7701	0.0547	0.1093
RBF	800	0.5	0.5	0.8793	0.6195	0.0396	0.0975
RBF	400	0.5	1.0	0.8065	0.7661	0.0794	0.1131
RBF	800	0.5	1.0	0.8678	0.6508	0.0660	0.0961
RBF	400	0.5	5.0	0.8469	0.6726	0.0640	0.1255
RBF	800	0.5	5.0	0.8622	0.6726	0.0215	0.1124
RBF	400	1.0	0.5	0.8389	0.7000	0.0625	0.1040
RBF	800	1.0	0.5	0.8744	0.6470	0.0506	0.0948
RBF	400	1.0	1.0	0.8403	0.7122	0.0605	0.1076
RBF	800	1.0	1.0	0.8633	0.6689	0.0445	0.0867
RBF	400	1.0	5.0	0.8243	0.7263	0.0390	0.1141
RBF	800	1.0	5.0	0.8723	0.6430	0.0555	0.1010
RBF	400	5.0	0.5	0.8394	0.7268	0.0653	0.1328
RBF	800	5.0	0.5	0.8603	0.6777	0.0589	0.0886
RBF	400	5.0	1.0	0.8542	0.6580	0.0699	0.1069
RBF	800	5.0	1.0	0.8703	0.6524	0.0692	0.0889
RBF	400	5.0	5.0	0.8410	0.7347	0.0616	0.1129
RBF	800	5.0	5.0	0.8710	0.6424	0.0496	0.0957

The table depicts the error estimates while fitting the booster function with RBF kernel with varying parameter choices. From error estimates: R2, RMSE, sqr RMSE, MNLL are used to check the accuracy of fit.
Analysis of table: From the table it is clearly depicts that with the increase in number of samples from 400 to 800 the error is decreasing. For RBF kernel the best configuration which gives most nearer value of R2 score to 1 is 800 samples along with 0.5 lengthscale and variance respectively alongwith standard RMSE error is the least

kernel name	pts	ls	var	R2 score	RMSE	MNLL	Std RMSE
Matern32	400	0.5	0.5	0.8188	0.7471	0.0726	0.2108
Matern32	800	0.5	0.5	0.8732	0.6417	0.0362	0.2079
Matern32	400	0.5	1.0	0.8482	0.7112	0.0538	0.2117
Matern32	800	0.5	1.0	0.8747	0.6421	0.0521	0.2118
Matern32	400	0.5	5.0	0.8656	0.6672	0.0741	0.1986
Matern32	800	0.5	5.0	0.8930	0.5942	0.0288	0.2104
Matern32	400	1.0	0.5	0.8658	0.6600	0.0587	0.2167
Matern32	800	1.0	0.5	0.8817	0.6010	0.0340	0.2103
Matern32	400	1.0	1.0	0.8782	0.6294	0.0705	0.2236
Matern32	800	1.0	1.0	0.8851	0.6159	0.0456	0.2058
Matern32	400	1.0	5.0	0.8833	0.6054	0.0749	0.2215
Matern32	800	1.0	5.0	0.8939	0.5879	0.0341	0.2139
Matern32	400	5.0	0.5	0.8787	0.6591	0.0636	0.2165
Matern32	800	5.0	0.5	0.8862	0.5989	0.0689	0.2212
Matern32	400	5.0	1.0	0.8615	0.6503	0.0650	0.2159
Matern32	800	5.0	1.0	0.9025	0.5507	0.0334	0.2183
Matern32	400	5.0	5.0	0.8626	0.6472	0.0616	0.2175
Matern32	800	5.0	5.0	0.8935	0.5967	0.0537	0.2070

The table depicts the error estimates while fitting the booster function with Matern32 kernel with varying parameter choices. From error estimates: R2, RMSE, sqr RMSE, MNLL are used to check the accuracy of fit.

Analysis of table: From the table it is clearly depicts that with the increase in number of samples from 400 to 800 the error is decreasing. For matern32 the best configurations is 800 samples along with 5 as a lengthscale and 1 as a variance which clearly shows the most closer value to 1 of R2 score.

kernel name	pts	ls	var	R2 score	RMSE	MNLL	Std RMSE
Exponential	400	0.5	0.5	0.8686	0.6542	0.0359	0.2993
Exponential	800	0.5	0.5	0.8991	0.5696	0.0245	0.2968
Exponential	400	0.5	1.0	0.8103	0.7715	0.0903	0.2943
Exponential	800	0.5	1.0	0.8864	0.6148	0.0631	0.2905
Exponential	400	0.5	5.0	0.8491	0.6973	0.0448	0.2905
Exponential	800	0.5	5.0	0.8904	0.5832	0.0246	0.2964
Exponential	400	1.0	0.5	0.8604	0.6646	0.0486	0.3076
Exponential	800	1.0	0.5	0.8739	0.6270	0.0676	0.2994
Exponential	400	1.0	1.0	0.8398	0.6719	0.0495	0.3009
Exponential	800	1.0	1.0	0.8732	0.6407	0.0513	0.2879
Exponential	400	1.0	5.0	0.8236	0.7528	0.0850	0.2928
Exponential	800	1.0	5.0	0.8987	0.5627	0.0147	0.3073
Exponential	400	5.0	0.5	0.8675	0.6691	0.0586	0.2986
Exponential	800	5.0	0.5	0.8887	0.5983	0.0378	0.3027
Exponential	400	5.0	1.0	0.8964	0.5454	0.0600	0.3057
Exponential	800	5.0	1.0	0.8835	0.6231	0.0202	0.2929
Exponential	400	5.0	5.0	0.8458	0.6889	0.0812	0.2979
Exponential	800	5.0	5.0	0.8892	0.5826	0.0214	0.2995

The table depicts the error estimates while fitting the booster function with exponential kernel with varying parameter choices. From error estimates: R2, RMSE, sqr RMSE, MNLL are used to check the accuracy of fit.
Analysis of table: From the table it is clearly depicts that with the increase in number of samples from 400 to 800 the error is decreasing. For exponential kernel the best configuration which gives most nearer value of R2 score to 1 is 800 samples along with 0.5 lengthscale and variance respectively alongwith standard RMSE error is the least .

kernel name	pts	ls	var	R2 score	RMSE	MNLL	Std RMSE
RBF*Matern32	400	0.5	0.5	0.8503	0.6937	0.0728	0.1906
RBF*Matern32	800	0.5	0.5	0.8931	0.5837	0.0492	0.2243
RBF*Matern32	400	0.5	1.0	0.8479	0.7064	0.0716	0.2170
RBF*Matern32	800	0.5	1.0	0.8894	0.5987	0.0278	0.2103
RBF*Matern32	400	0.5	5.0	0.8634	0.6575	0.0582	0.2112
RBF*Matern32	800	0.5	5.0	0.8959	0.5851	0.0534	0.1982
RBF*Matern32	400	1.0	0.5	0.8648	0.6650	0.0640	0.1944
RBF*Matern32	800	1.0	0.5	0.9066	0.5476	0.0547	0.2080
RBF*Matern32	400	1.0	1.0	0.8802	0.6455	0.0533	0.2197
RBF*Matern32	800	1.0	1.0	0.8919	0.5924	0.0422	0.2270
RBF*Matern32	400	1.0	5.0	0.8612	0.6492	0.0491	0.2000
RBF*Matern32	800	1.0	5.0	0.8869	0.6157	0.0224	0.2007
RBF*Matern32	400	5.0	0.5	0.8164	0.7174	0.1398	0.2099
RBF*Matern32	800	5.0	0.5	0.8625	0.6298	0.0382	0.2145
RBF*Matern32	400	5.0	1.0	0.8450	0.7056	0.0818	0.2141
RBF*Matern32	800	5.0	1.0	0.8750	0.6099	0.0640	0.1929
RBF*Matern32	400	5.0	5.0	0.8813	0.6162	0.0506	0.2194
RBF*Matern32	800	5.0	5.0	0.8751	0.6293	0.0483	0.2066

The table depicts the error estimates while fitting the booster function with RBF*Matern32 kernel with varying parameter choices. From error estimates: R2, RMSE, sqr RMSE, MNLL are used to check the accuracy of fit.
Analysis of table: From the table it is clearly depicts that with the increase in number of samples from 400 to 800 the error is decreasing. For RBF*Matern32 kernel the best configuration which gives most nearer value of R2 score to 1 is 800 samples along with 1 lengthscale and 0.5 variance respectively alongwith RMSE error is the least .

kernel name	pts	ls	var	R2 score	RMSE	MNLL	Std RMSE
Exponential+Matern52+RBF	400	0.5	0.5	0.8525	0.6723	0.0391	0.2394
Exponential+Matern52+RBF	800	0.5	0.5	0.8844	0.6041	0.0536	0.2745
Exponential+Matern52+RBF	400	0.5	1.0	0.8572	0.7020	0.0430	0.2793
Exponential+Matern52+RBF	800	0.5	1.0	0.9061	0.5509	0.0266	0.2935
Exponential+Matern52+RBF	400	0.5	5.0	0.8798	0.6024	0.0680	0.3039
Exponential+Matern52+RBF	800	0.5	5.0	0.8659	0.6530	0.0589	0.2626
Exponential+Matern52+RBF	400	1.0	0.5	0.8588	0.6725	0.0836	0.2963
Exponential+Matern52+RBF	800	1.0	0.5	0.8846	0.5998	0.0637	0.2937
Exponential+Matern52+RBF	400	1.0	1.0	0.8493	0.7044	0.0981	0.2543
Exponential+Matern52+RBF	800	1.0	1.0	0.8626	0.6787	0.0273	0.2772
Exponential+Matern52+RBF	400	1.0	5.0	0.8866	0.5850	0.0807	0.2294
Exponential+Matern52+RBF	800	1.0	5.0	0.8632	0.6509	0.0578	0.1800
Exponential+Matern52+RBF	400	5.0	0.5	0.8643	0.6304	0.1061	0.1833
Exponential+Matern52+RBF	800	5.0	0.5	0.8897	0.6103	0.0324	0.1847
Exponential+Matern52+RBF	400	5.0	1.0	0.8408	0.6995	0.0487	0.1821
Exponential+Matern52+RBF	800	5.0	1.0	0.8812	0.6010	0.0484	0.2025
Exponential+Matern52+RBF	400	5.0	5.0	0.7907	0.7765	0.0923	0.1853
Exponential+Matern52+RBF	800	5.0	5.0	0.8623	0.6591	0.0240	0.1858

The table depicts the error estimates while fitting the booster function with exponential+Matern52+RBF kernel with varying parameter choices. From error estimates: R2, RMSE, sqr RMSE, MNLL are used to check the accuracy of fit.

Analysis of table: From the table it is clearly depicts that with the increase in number of samples from 400 to 800 the error is decreasing. For exponential+Matern32+RBF kernel the best configuration which gives most nearer value of R2 score to 1 is 800 samples along with 0.5 lengthscale and 1 variance respectively .

kernel name	pts	ls	var	R2 score	RMSE	MNLL	Std RMSE
M52*RBF	400	0.1	0.1	0.84811	0.67704	0.18697	0.06817
M52*RBF	800	0.1	0.1	0.88086	0.61822	0.17519	0.05866
M52*RBF	400	0.1	1.0	0.84911	0.68811	0.17345	0.05219
M52*RBF	800	0.1	1.0	0.86920	0.66004	0.15261	0.02731
M52*RBF	400	0.1	10.0	0.88082	0.60672	0.17307	0.06903
M52*RBF	800	0.1	10.0	0.89026	0.60922	0.16295	0.03251
M52*RBF	400	1.0	0.1	0.87882	0.63020	0.18200	0.04477
M52*RBF	800	1.0	0.1	0.90353	0.56771	0.17681	0.03259
M52*RBF	400	1.0	1.0	0.86653	0.65585	0.17483	0.02630
M52*RBF	800	1.0	1.0	0.90669	0.54171	0.17949	0.05095
M52*RBF	400	1.0	10.0	0.89088	0.59123	0.17907	0.04817
M52*RBF	800	1.0	10.0	0.90452	0.54962	0.16507	0.04630
M52*RBF	400	10.0	0.1	0.86159	0.66368	0.18149	0.05352
M52*RBF	800	10.0	0.1	0.89130	0.59609	0.16376	0.01426
M52*RBF	400	10.0	1.0	0.86956	0.64317	0.17958	0.08067
M52*RBF	800	10.0	1.0	0.87805	0.62143	0.17710	0.06333
M52*RBF	400	10.0	10.0	0.86213	0.66745	0.15652	0.07453
M52*RBF	800	10.0	10.0	0.89591	0.58045	0.17955	0.02549

The table depicts the error estimates while fitting the booster function with M52*RBF kernel with varying parameter choices. From error estimates: R2, RMSE, sqr RMSE, MNLL are used to check the accuracy of fit.
Analysis of table: From the table it is clearly depicts that with the increase in number of samples from 400 to 800 the error is decreasing. For M52*RBF kernel the best configuration which gives most nearer value of R2 score to 1 is 800 samples along with 1 lengthscale and variance respectively with RMSE error is the least.

kernel name	pts	ls	var	R2 score	RMSE	MNLL	Std RMSE
Exp+RBF	400	0.1	0.1	0.85159	0.67709	0.27678	0.04326
Exp+RBF	800	0.1	0.1	0.90152	0.55204	0.29381	0.03203
Exp+RBF	400	0.1	1.0	0.86841	0.64169	0.29044	0.04605
Exp+RBF	800	0.1	1.0	0.88193	0.59507	0.29184	0.04891
Exp+RBF	400	0.1	10.0	0.86999	0.61748	0.29482	0.03867
Exp+RBF	800	0.1	10.0	0.89485	0.59442	0.29006	0.04349
Exp+RBF	400	1.0	0.1	0.89710	0.57461	0.28661	0.04112
Exp+RBF	800	1.0	0.1	0.88224	0.61559	0.29722	0.01925
Exp+RBF	400	1.0	1.0	0.87219	0.65022	0.26917	0.05402
Exp+RBF	800	1.0	1.0	0.88494	0.60293	0.29361	0.01369
Exp+RBF	400	1.0	10.0	0.88073	0.59334	0.28472	0.05993
Exp+RBF	800	1.0	10.0	0.89834	0.57735	0.28273	0.04122
Exp+RBF	400	10.0	0.1	0.87260	0.63284	0.27844	0.07450
Exp+RBF	800	10.0	0.1	0.89014	0.59406	0.28919	0.05952
Exp+RBF	400	10.0	1.0	0.84972	0.66180	0.27939	0.05610
Exp+RBF	800	10.0	1.0	0.88655	0.60151	0.28508	0.04149
Exp+RBF	400	10.0	10.0	0.83283	0.69272	0.28705	0.03904
Exp+RBF	800	10.0	10.0	0.88838	0.60101	0.28705	0.03641

The table depicts the error estimates while fitting the booster function with exponential+RBF kernel with varying parameter choices. From error estimates: R2, RMSE, sqr RMSE, MNLL are used to check the accuracy of fit.
Analysis of table: From the table it is clearly depicts that with the increase in number of samples from 400 to 800 the error is decreasing. For Exp+RBF kernel the best configuration which gives most nearer value of R2 score to 1 is 800 samples along with 0.1 lengthscale and variance respectively with RMSE error is the least.

kernel name	pts	ls	var	R2 score	RMSE	MNLL	Std RMSE
RBF	800	0.5	0.5	0.8793	0.6195	0.0396	0.0975
Matern32	800	5.0	1.0	0.9025	0.5507	0.0334	0.2183
Exponential	800	0.5	0.5	0.8991	0.5696	0.0245	0.2968
RBF*Matern32	800	1.0	0.5	0.9066	0.5476	0.0547	0.2080
Exponential+Matern52+RBF	800	0.5	1.0	0.9061	0.5509	0.0266	0.2935
M52*RBF	800	1.0	1.0	0.90669	0.54171	0.17949	0.05095
Exp+RBF	800	0.1	0.1	0.90152	0.55204	0.29381	0.03203

This table illustrates the best kernel along with best constraints from which we clearly seen that the best from the best is in highlighted form.
Analysis of table: From the table it is clearly depicts that with the increase in number of samples from 400 to 800 the error is decreasing. The best combination from all the possible configurations M52*RBF which is clearly shows the most nearer value to 1.

0.2 Part-II

As we know Acquisition functions are mathematical techniques that guide how the parameter space should be explored during Bayesian optimization. They use the predicted mean and predicted variance generated by the Gaussian process model. The predicted variance is very close to zero at or very near to an existing result.

In our problem, we have used acquisitions functions such as Expected improvement, probability improvement, upper confidence level, direct approach and POF. we have performed our compuations on different iterations 10, 20, 30, 40, 50 and 60 and from all the four cases we clearly observed that optimal value is near to almost 5. First three tables are without constraints and last two tables shows the optimal along with constraints.

iter	x1	x2	x3	x4	y
10	0.339	0.452	0.167	0.461	4.656
20	0.187	0.478	0.150	0.569	4.837
30	0.672	0.746	0.011	0.931	5.026
40	0.316	0.682	0.193	0.750	5.143
50	0.696	0.765	0.242	0.926	4.984
60	0.318	0.309	0.043	0.770	5.091

EI

10	0.395	0.901	0.180	0.764	5.173
20	0.106	0.173	0.336	0.769	4.939
30	0.054	0.099	0.310	0.946	5.251
40	0.088	0.323	0.320	0.609	4.725
50	0.472	1.000	0.144	0.857	5.329
60	0.124	0.509	0.391	0.803	4.916

PI

iter	x1	x2	x3	x4	y
10	0.500	0.633	0.118	0.943	5.395
20	0.195	0.066	0.126	0.948	5.375
30	0.824	0.944	0.283	0.748	4.637
40	0.390	0.819	0.216	0.641	4.943
50	0.188	0.896	0.310	0.779	4.906
60	0.878	0.876	0.162	0.875	4.952

UCB

iter	x1	x2	x3	x4	y
10	0.248	0.215	0.074	0.858	5.299
20	0.497	0.748	0.296	0.757	4.871
30	0.436	0.705	0.085	0.830	5.279
40	0.051	0.468	0.382	0.582	4.492
50	0.360	0.622	0.104	0.625	4.997
60	0.280	0.236	0.016	0.865	4.955

Direct approach

iter	x1	x2	x3	x4	y
10	0.537	0.771	0.118	0.604	4.946
20	0.907	0.947	0.027	0.967	5.171
30	0.367	0.668	0.092	0.909	5.368
40	0.960	0.838	0.031	0.954	5.029
50	0.351	0.391	0.105	0.741	5.149
60	0.651	0.717	0.146	0.823	5.086

POF