13 May

Recap

We have setup a mocked up task based on cheap docking data for a different task (docking to D4 Dopamine receptor). In our mocked up task this cheap docking data counts as the ground truth.

To create results from cheap docking/expensive docking/FEP oracles for our mocked up task we are adding the outputs of randomly initialized NNs (after tuning so that they have suitable variances) to our "ground truth" values.

This allows us to check the BayesOpt pipeline before using it on the more complicated task.

Bayes Model

Implementing a Bayesian Linear Regression from fingerprints to dopamine score — just to ensure that it gets calibrated uncertainties and get an idea of how much data it needs to train.

In particular we want to make sure that the inference scheme which we are using works and whether we need to setup any hyperpriors etc.

The weirdness at 1000 training points is probably related to the fact that the fingerprints are 1000 dimensional?

Bayes Model

```
Name                                                Training set size    MSE (↓)    Avg Loglikelihood (↑)
--------------------------------------------------  -------------------  ---------  -----------------------
Dummy Gaussian (var=1)                              10                   19.85      -10.84
Linear Regression/w Gaussian likelihood (var=1)     10                   19.93      -10.89
Bayesian Regression                                 10                   19.88      -2.98
Sklearn Bayesian Ridge Regression                   10                   19.43      -2.91
Bayesian Regression with the sklearn                10                   19.43      -2.91
 learnt precisions (weights: 1534.742,noise:0.058)
---                                                 ---                  ---        ---
Dummy Gaussian (var=1)                              20                   19.77      -10.80
Linear Regression/w Gaussian likelihood (var=1)     20                   18.08      -9.96
Bayesian Regression                                 20                   18.08      -2.94
Sklearn Bayesian Ridge Regression                   20                   19.40      -2.92
Bayesian Regression with the sklearn                20                   19.40      -2.92
 learnt precisions (weights: 248.117,noise:0.065)
---                                                 ---                  ---        ---
Dummy Gaussian (var=1)                              50                   19.46      -10.65
Linear Regression/w Gaussian likelihood (var=1)     50                   15.46      -8.65
Bayesian Regression                                 50                   15.44      -2.87
Sklearn Bayesian Ridge Regression                   50                   15.48      -4.03
Bayesian Regression with the sklearn                50                   15.48      -2.79
 learnt precisions (weights: 2.445,noise:0.380)
---                                                 ---                  ---        ---
Dummy Gaussian (var=1)                              100                  19.51      -10.67
Linear Regression/w Gaussian likelihood (var=1)     100                  15.21      -8.52
Bayesian Regression                                 100                  15.07      -2.83
Sklearn Bayesian Ridge Regression                   100                  14.66      -3.56
Bayesian Regression with the sklearn                100                  14.66      -2.77
 learnt precisions (weights: 2.555,noise:0.330)
---                                                 ---                  ---        ---
Dummy Gaussian (var=1)                              500                  19.53      -10.68
```

```
Linear Regression/w Gaussian likelihood (var=1)    500          20.88     -11.36
Bayesian Regression                                500          17.84     -2.93
Sklearn Bayesian Ridge Regression                  500          12.21     -2.68
Bayesian Regression with the sklearn               500          12.21     -2.67
 learnt precisions (weights: 3.421,noise:0.135)
---                                                ---          ---       ---
Dummy Gaussian (var=1)                             1000         19.43     -10.64
Linear Regression/w Gaussian likelihood (var=1)    1000         367.00    -184.42
Bayesian Regression                                1000         24.10     -4.26
Sklearn Bayesian Ridge Regression                  1000         11.22     -2.63
Bayesian Regression with the sklearn               1000         11.22     -2.63
 learnt precisions (weights: 4.101,noise:0.119)
---                                                ---          ---       ---
Dummy Gaussian (var=1)                             2500         19.42     -10.63
Linear Regression/w Gaussian likelihood (var=1)    2500         14.44     -8.14
Bayesian Regression                                2500         13.64     -5.29
Sklearn Bayesian Ridge Regression                  2500         10.26     -2.58
Bayesian Regression with the sklearn               2500         10.26     -2.58
 learnt precisions (weights: 5.372,noise:0.110)
---                                                ---          ---       ---
Dummy Gaussian (var=1)                             5000         19.42     -10.63
Linear Regression/w Gaussian likelihood (var=1)    5000         11.23     -6.54
Bayesian Regression                                5000         11.12     -5.46
Sklearn Bayesian Ridge Regression                  5000         9.83      -2.56
Bayesian Regression with the sklearn               5000         9.83      -2.56
 learnt precisions (weights: 5.978,noise:0.111)
---                                                ---          ---       ---
Dummy Gaussian (var=1)                             7500         19.42     -10.63
Linear Regression/w Gaussian likelihood (var=1)    7500         10.26     -6.05
Bayesian Regression                                7500         10.21     -5.39
Sklearn Bayesian Ridge Regression                  7500         9.50      -2.54
Bayesian Regression with the sklearn               7500         9.50      -2.54
 learnt precisions (weights: 6.127,noise:0.112)
---                                                ---          ---       ---
Dummy Gaussian (var=1)                             10000        19.42     -10.63
Linear Regression/w Gaussian likelihood (var=1)    10000        9.72      -5.78
Bayesian Regression                                10000        9.70      -5.32
Sklearn Bayesian Ridge Regression                  10000        9.27      -2.53
Bayesian Regression with the sklearn               10000        9.27      -2.53
 learnt precisions (weights: 6.242,noise:0.114)
---                                                ---          ---       ---
```

⇒ Seems a Bayesian linear model does not that work well unless we put hyperpriors on the weight variance and noise variance parameters. How is it best to do?

BayesOpt vs Traditional VS baseline

Running the Bayesian model in a VS pipeline to compare against not using it and instead just filtering on the previous results.

Note that this model is not currently taking in the cheaper docking scores as input, just the molecular fingerprints

```
Method         Top 1 Found    Top 2 Found    Top 3 found    Top 5 found    Top 10 found
------------  -------------  -------------  -------------  -------------  --------------
Full Dataset       -9.79673       -9.74911       -9.74911       -9.56339        -9.43006
BayesVS            -9.6253        -9.45149       -9.44911       -8.9872         -8.62768
PlainVS            -9.6253        -9.56339       -9.45149       -9.32768        -8.90863
```

⇒ Very preliminary — need to actually implement Bayesian model properly (ie give it good feats and use actually sensible priors)

Misc Notes

- changed it to a minimization problem.
- this subset of data we are operating on (~250k in size) has been preprocessed to uniformly sample over dockscores, this means that it may be an easier task to do well randomly on than the original data....?

4. Here is the plot blue is the original distribution, yellow is the sampled distribution