INF367: Spring 2020
Exercise 7

Instructions:

You can either return the solutions electronically via MittUiB by Monday 10.00 or show them on paper on Monday's meeting. Grades are awarded for effort so scanned notes are fine if you solve exercises by hand (no need to make fancy latex files).

Students are encouraged to write computer programs to derive solutions whenever appropriate.

Tasks

We want to use linear models. But the data are not always linear and thus we may need to use basis functions to handle non-linearity. However, it is not clear which basis functions are good for any particular data. Furthermore, selection of the noise precision β is crucial for to get good performance.

In this task, we model (again) bicycling in New York. You should download the data set `new_york_bicycles3.csv` from MittUiB. The goal is to predict the highest daily temperature given the other variables.

In this exercise, we select a set of basis functions and the value of β. To this end, you should try different basis functions and different values of β. In other words, a model is specified by the basis functions and *beta*, i.e., $M_i = (\phi_i, \beta_i)$. Note: If the selected β is either smallest or the largest value that you tried, then you should probably try more values.

Use the following model selection strategies:

1. Use the posterior ratio to select a model. To do this, you need to compute marginal likelihoods. (Note: you want to compare $P(D \mid M_i)P(M_i)$ values of different models.)

2. Use BIC to select a model. (Note: you want to maximize BIC.)

3. Use AIC to select a model. (Note: you want to minimize AIC.)

4. Use cross-validation to select a model. (Note: you want to minimize the validation loss.)

Hints:

Typically, one projects data to higher dimensions. However, it may be useful to consider projecting data also to a lower dimensional space (i.e. not to use all features).

Remember that when you compute marginal likelihood, you should use the predictive distribution $P(y \mid \hat{f}_\theta(\mathbf{x})) = N(y \mid \mathbf{m}^T\phi(\mathbf{x}), 1/\beta + \phi(\mathbf{x})^T\mathbf{S}\phi(\mathbf{x}))$. In case of AIC, BIC, and the cross-validation loss, you should use the the usual likelihood $P(y \,|\, \mathbf{w}, \mathbf{x}, \beta) = N(y \,|\, \mathbf{w}^T\phi(\mathbf{x}), 1/\beta)$.

Remember that the MAP estimate of a Gaussian distribution is its mean.

Scipy's Gaussian distribution, `norm` has a `scale` parameter that takes in a standard deviation, i.e., $\sqrt{1/precision}$.

If you get totally unsensible results, first sanity check is to check whether you are minimizing when you should be maximizing or vice versa.