Small Exercises 9

Gaussian Processes

These exercises are meant to prepare the inverted classroom lecture. Keep your answers short: two or three sentences, sometimes even less, should suffice.

Problem 1: Why do we often chose the mean function m of a Gaussian Process to be zero, i.e. $m(\boldsymbol{x}) = 0$?

The mean function can be used to incorporate domain-knowledge. However, we often do not have any and more importantly it is always possible to mean normalize our data.

Problem 2: Basically all calculations involving Gaussian Processes require K^{-1}, but we specify the covariance function K instead. Why?

Specifying K for Gaussian processes or multivariate normals simplifies many computations e.g. we can marginalize by just taking a sub-matrix. This is not possible when we specify the inverse K^{-1} as can be seen when looking at the Schur complement for a block matrix:

$$\begin{aligned}\begin{pmatrix}\boldsymbol{A} & \boldsymbol{C}^T\\ \boldsymbol{C} & \boldsymbol{B}\end{pmatrix}^{-1} &= \begin{pmatrix}\boldsymbol{I} & \boldsymbol{0}\\ -\boldsymbol{B}^{-1}\boldsymbol{C} & \boldsymbol{I}\end{pmatrix}\begin{pmatrix}(\boldsymbol{A}-\boldsymbol{C}^T\boldsymbol{B}^{-1}\boldsymbol{C})^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{B}^{-1}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{I} & -\boldsymbol{C}^T\boldsymbol{B}^{-1}\\ \boldsymbol{0} & \boldsymbol{I}\end{pmatrix}\\ &= \begin{pmatrix}(\boldsymbol{A}-\boldsymbol{C}^T\boldsymbol{B}^{-1}\boldsymbol{C})^{-1} & -(\boldsymbol{A}-\boldsymbol{C}^T\boldsymbol{B}^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{B}^{-1}\\ -\boldsymbol{B}^{-1}\boldsymbol{C}(\boldsymbol{A}-\boldsymbol{C}^T\boldsymbol{B}^{-1}\boldsymbol{C})^{-1} & \boldsymbol{B}^{-1}+\boldsymbol{B}^{-1}\boldsymbol{C}(\boldsymbol{A}-\boldsymbol{C}^T\boldsymbol{B}^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{B}^{-1}\end{pmatrix}\end{aligned}$$

If we wanted the marginal that has covariance $\mathbf{A}$, we would need $\boldsymbol{A}^{-1}$ which cannot be easily extracted from the expression above.

Problem 3: Why do large datasets cause problems for traditional Gaussian Processes?

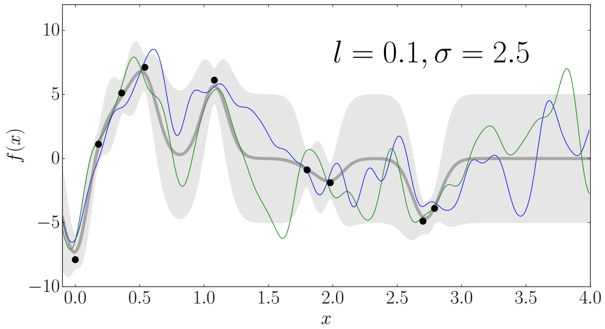
Each additional datapoint adds an additional row and column to the covariance matrix. This is problematic as we have to invert that matrix which has complexity $O(n^3)$ (or a little better for some optimized algorithms) for an $n \times n$ matrix.

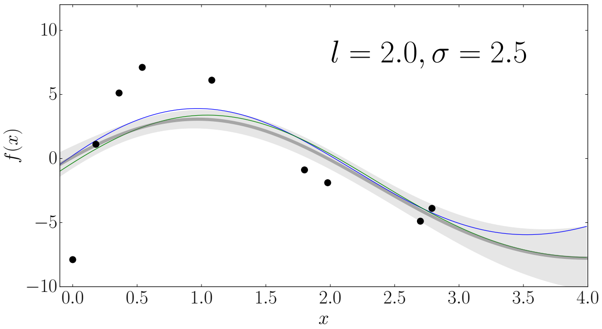
Problem 4: What is more likely to cause overfitting when using the squared exponential kernel for the covariance function in a Gaussian Process: a small or a large length scale l? Why?

In the noise-free case, all training data will be fitted exactly by the GP regardless of the length-scale. The draws becomes much smoother for a larger length scale, with little variation between the draws which is a disadvantage but not necessarily overfitting.

Submit to `http://ml1.brml.org/hw8` *by 2017/01/09, 11:59AM CET*

Overfitting is more likely to occur in the noisy case. Data point are modeled more closely for a small length scale. A larger length scale tends to cause underfitting. See the figures below.


$l = 0.1, \sigma = 2.5$
$f(x)$
x


$l = 2.0, \sigma = 2.5$
$f(x)$
x