1 作业题目：高斯过程回归

回归分析是统计学、信号处理、机器学习等多领域中的基础研究问题之一。回归分析研究的是变量与变量间的关系。记其中一个变量称为自变量 $\mathbf{x} \in \mathbf{S} \subset \mathbb{R}^d$，另一个变量称为因变量 $y \in \mathbb{R}$，假设两者存在如下的关系

$$y = f(\mathbf{x}) + e$$

其中，e 为表示误差的随机变量，$f(\mathbf{x})$ 称为回归函数（或预测函数、拟合函数）。给定一组观测 $\mathcal{D} = \{(\mathbf{x}_i, y_i)|i = 1, 2, , n\}$，$\mathcal{D}$ 又被称为训练数据，回归分析希望能就此找出在某个准则下最好的回归函数 $f(*)$。传统的回归分析包括了两个层次的问题，一是确定合适的回归函数形式；二是在给定回归函数形式下，依据训练数据求出具体的回归函数。

一类广泛应用的回归模型是广义线性模型，它假定回归函数为基函数的线性组合，即：

$$f(\mathbf{x}) = \sum_{i=1}^{M} w_i \phi_i(\mathbf{x})$$

其中 $\phi_i(\mathbf{x})$ 称为基函数，利用简单的基函数可以构成较复杂的回归函数形式，比如: $\phi_i(\mathbf{x}) = x_i$ 可以构成线性回归函数;$\phi_i(\mathbf{x}) = x_i^k \quad (k = 1, \cdots, N)$ 可以构成多项式回归函数；$\phi_i(\mathbf{x}) = \cos(w_i \mathbf{a}_i^T \mathbf{x})$ 可以形成三角函数等。在基于广义线性模型的回归分析中，首先要利用训练数据集确定基函数中的参数以及基函数系数，从而求出具体的回归函数 (求解过程通常要处理基于观测数据与预测数据间差别最小导出的优化问题)。

获取具体回归函数以后，可以以此预测未知自变量 x_* 处的因变量值 y_*。因此，可以在一组有别于训练数据的测试数据 $\mathcal{T} = \{(x_i^*, y_i^*)|i = 1, 2, \cdots, m\}$ 上，通过计算预测值与观测值间的差异来评价回归函数的好坏。本次作业中采用均方误差 (Mean Squared Error, MSE) 来衡量，即：

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} (y_{*,i} - y_i^*)^2$$

上式中 $y_{*,i}$ 表示所测试的回归函数在 $\mathbf{x}_i^*$ 的预测值。在测试数据 $\mathcal{T}$ 上的 MSE 越小，则表示所求的回归函数的推广能力越强，对两变量间的关系拟合得越好。

高斯过程回归 (GPR) 是近年来广受关注的一种回归分析方法，这种方法的基本想法是将 $\{f(\mathbf{x})|\mathbf{x} \in \mathbf{S}\}$ 建模为高斯过程 $\mathcal{N}\{m(\mathbf{x}), K(\mathbf{x}, \mathbf{x})'\}$，其

中 $\mathbb{E}\{f(\mathbf{x})\} = m(\mathbf{x}), \text{cov}(f(\mathbf{x}), f(\mathbf{x}')) = K(\mathbf{x}, \mathbf{x}')$，若已知此过程中在某些点处的带噪声观测值 (由训练数据集给出)，利用 $f(\mathbf{x})$ 为高斯过程及观测噪声 e 的 pdf(由此确定了似然函数)，可以估计其他任何点 $\mathbf{x}_*$ 处 $f(\mathbf{x}_*)$ 的后验分布，即 $p(f(\mathbf{x}_*)|x_*, \mathcal{D})$。GPR 方法中不必直接给定回归函数形式，但需要确定核函数 $K(\mathbf{x}, \mathbf{x}')$ 的函数形式，实质上核函数 $K(\mathbf{x}, \mathbf{x}')$ 给出了两个随机变量 $f(\mathbf{x})$ 和 $f(\mathbf{x}')$ 间的相关性，因此所指定的 $m(\mathbf{x})$ 与 $K(\mathbf{x}, \mathbf{x}')$ 直接决定了高斯过程模型的推广能力。

常用的简单核函数 $K(\mathbf{x}, \mathbf{x}')$ 有 (见文献 1 第 5 章和文献 2)：(1). 各向同性的平方指数核函数 $\exp(-||\mathbf{x}-\mathbf{x}'||^2/\sigma^2)$；(2). 各向异性的平方指数核函数 $\exp[-\sum_{i=1}^{d}\frac{1}{\alpha_i^2}(x_i - x_i')^2]$；(3). 线性核函数 $\sigma_0^2 + \sum_{i=1}^{d}\sigma_i^2 x_i x_i'$；(4). 各向同性有理核函数 $(1+\frac{||\mathbf{x}-\mathbf{x}'||^2}{\alpha l^2}/2)^{-\alpha}$；(5). 各向异性有理核函数 $(1+\sum_{i=1}^{d}\frac{(x_i - x_i')^2}{2\alpha l_i^2})^{-\alpha}$；等。通过对核函数进行加、乘或者函数复合等操作，可以构造非常复杂的核函数，使 GPR 具有非常强大的建模能力。选定恰当的核函数形式后，需要通过训练数据从中估计出核函数的参数，通常可以通过求解下述优化问题：

$$\Theta = \arg\max_{\Theta} \ln p(y_1, y_2, \cdots, y_n | \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n, \Theta)$$

来获得，Θ 表示核函数、似然函数、均值函数中的全体参数体。

GPR 回归中另一个重要问题是在给定均值 $m(\mathbf{x})$ 和核函数 $K(\mathbf{x}, \mathbf{x}')$ 及数据集 $\mathcal{D}$ 后，对某点 $\mathbf{x}_*$ 处的 $f(\mathbf{x}_*)$ 推断其后验概率，在似然函数取高斯函数时，后验概率也是高斯形式（因为高斯分布是高斯分布的共轭先验），当似然函数取其他函数形式或者数据维度很高时，不能计算其后验分布的解析式或者计算量很大，此时准确获取后验分布不太可能，因而需要采用其他的贝叶斯推断近似方法，比如 Lapalace 近似、变分贝叶斯、MCMC 等。

本次作业需要对于不同的训练数据集，选择合适的核函数、均值函数、似然函数和贝叶斯推断方法，建立相应的模型，并在测试数据上进行比较。下面分别说明本作业中涉及到的两个数据集。

(1) (50%) question1.mat 中包含了 xtrain, ytrain, xtest，其中训练数据共有 543 个，测试数据有 87 个。利用训练数据 xtrain, ytrain 训练合适的高斯过程回归器，预测 xtest 上各点处的值，并与实际观测值进行比较，计算 MSE。本问题关键是通过核函数的加、乘等操作，得到

足以描述数据内在规律的核函数。

(2) (50%) planecontrol.mat 中的数据源于 F16 飞机的副翼控制问题。其中 xtrain(对应于 $\{\mathbf{x}_i|i=1,\cdots,10000\}$) 和 xtest ($\{\mathbf{x}_i^*|i=1,\cdots,3750\}$) 中的数据都有 40 个分量 (每一行有 40 个数，表示一组数据有 40 个分量)，这些分量描述了飞机的 40 种个状态值；ytrain 中第 i 行表示飞机处于 xtrain 中第 i 行 $\mathbf{x}_i$ 的状态时应该采用的飞机副翼操作量。利用 GPR 实现解决飞机副翼控制问题并在 xtest 上进行预测，预测结果与实际数据进行比较，计算其 MSE。本问的研究内容：核函数、均值函数、似然函数的选择，并采用多种贝叶斯推断近似方法对测试数据进行预测。

2 作业要求与说明

1. 希望同学充分调研和阅读相关文献，积极动脑 + 动手，取得有自己见解的结果，整理成最终报告。

2. 本次作业提供了 data_read_and_MSE.m 文件，用于读取数据和说明如何调用已经封装好的函数计算问题 (1) 和问题 (2) 中在测试数据下的 MSE。注：MSE_question2.p 和 MSE_plane_control.p 定义了计算 MSE 的函数。

3. 最终提交包括：

   (a) 报告：报告的书写要求参见《Project 报告撰写建议》。

   (b) 源程序：必须将**最终版本的程序** (针对问题已经确定好核函数、似然函数的形式和贝叶斯推断近似方法的源程序) **填补到** **data_read_and_MSE.m** 中，使得一键运行此脚本文件能够获得你所认为最好的预测结果。注意，如果用到了非 Matlab 自带库函数，需用将所有涉及到的函数一起打包，保证 data_read_and_MSE.m 能够**一键运行**。提交的结果中不用放置任何数据文件。

4. 本次作业中问题 (1) 和问题 (2) 各占本次作业总成绩的 50%，最终成绩将依据工作新意及深入程度、工作量及完整程度、在测试数据得到的 MSE 这几个方面。

5. 一旦发现抄袭，计零分。

6. 请大家在规定截止时间 (2017 年 1 月 27 日 23:59) 前提交。晚交的处理方法如下：按晚交天数，以 90% 的几何级数进行折扣。晚交时间在 (0, 24 小时]，按 90% 折扣。晚交时间在 (24 小时, 48 小时]，按 90% × 90% 折扣。以此类推。

3 参考文献

1. C. E. Rasmussen, C. K. I. Williams, "Gaussian Processes for Machine Learning", MIT press, 2006.

2. Duvenaud D, Lloyd J R, Grosse R, et al. Structure Discovery in Nonparametric Regression through Compositional Kernel Search[J]. Creative Commons Attribution-Noncommercial-Share Alike, 2013.

3. Lloyd J R, Duvenaud D, Grosse R, et al. Automatic construction and natural-language description of nonparametric regression models[C]// Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI Press, 2014.

4. The Kernel Cookbook:Advice on Covariance functions. `http://www.cs.toronto.edu/~duvenaud/cookbook/index.html`

5. David Kristjanson Duvenaud.Automatic Model Construction with Gaussian Processes.PhD Thesis. University of Cambridge. 2014