概率论与随机过程 (2) ，project2 © 清华大学电子工程系

第二次 project

一. 主题介绍

本次大作业的主题是**基于高斯过程的回归分析—Gaussian Process Regression (GPR)**

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。

有关回归分析的经典知识，在茆书第 8.4 节 [1]，陈书第九章 [2] 都有很好的入门介绍。“回归分析”起源于十九世纪英国生物学家兼统计学家高尔顿研究父与子身高的遗传问题。他发现子代的平均身高有向中心回归的现象，“回归”一词由此而得名。近年来，基于高斯过程的回归分析（RW 书 [3]），得到了广泛的重视，其历史发展在 RW 书第 2.8 节有介绍。基于高斯过程的回归分析的一个突出优点是，对于非线性回归函数的选择，形成了一套行之有效的方法（见 RW 书第 5 章）。RW 书中主要介绍了以下几点内容：

- 方法 A：标准线性回归模型的（非概率）最小二乘求解；
- 方法 B：标准线性回归模型的概率求解（即 RW 书第 2.1.1 节的贝叶斯线性回归模型）；

- 方法 C：引入基函数构成广义线性回归模型（实乃非线性回归模型）；
- 方法 D（即 GPR 方法）：选择某种核函数，完成基于高斯过程的非线性回归分析。

在 RW 书第 5 章以及 [4] 中介绍了一系列核函数以及模型选择方法，本次大作业的核心内容将围绕着这些核函数以及模型选择方法展开。

二. 作业说明

本次大作业的基本要求包括：

(a) 对方法 A、方法 B、方法 C、方法 D 做出阐述。

(b) 实现方法 A、方法 B、使用多项式基函数的方法 C。

(c) 参照 RW 书第 5 章以及文献 [4]，选择使用一种核函数以及一种模型选择方法对给定数据实现方法 D（GPR），详细阐述实现过程并对结果进行分析。

同学们可以进一步选择（但不限于）如下研究内容，并根据自己选择的内容进行详细分析：

- 尽可能好地预测出 30% 同学的“概率论与随机过程 2”的成绩。
- 使用不同形式、不同超参数取值的核函数的 GPR 方法的理论分析和实验比较。
- 研究不同的模型选择方法，可以与第一次大作业中的 RBF 网络模型选择方法（如 RJ-MCMC）相结合。
- 洞察数据、发现规律的自由探索。

本次作业的附件中提供了电子系五个年级的成绩数据。注意各年级所选课程不完全一致。对各个年级，均是约 70% 同学的全部课程成绩提供给大家作为训练集，约 30% 同学隐去了“概率论与随机过程 2”的成绩作为测试集，具体说明以 txt 文档的形式放在附件中。

三. 具体要求

(a) 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。

(b) 最终提交包括:

i. 报告
报告的书写要求参见《Project 报告撰写建议》。

ii. 源程序 务必包含: **自己所有的原始程序、所有方法的输出数据**，输出数据全部放在数据文件**GPR.mat**中，格式如下：

- 将 A 方法在测试集上给出的输出结果（回归值）按年份从小到大一起记于列矢量 a，以此类推将方法 B、C、D 的结果记于列矢量 b、c、d。
- 如果还实现了其他回归方法，也将该方法在测试集上的回归值存成列矢量，命名为 $e1, e2, e3, \ldots$。
- 将你认为最好的回归结果记于 *optimal* 列矢量。

- 所有列矢量（至少包含 a、b、c、d、$optimal$）统一存在取名为 **GPR.mat** 的文件中，进行提交。

将以上两项一起压缩打包，命名为“学号 _ 姓名.rar”进行提交。

(c) 评分标准：报告书写清晰和规范，工作新意及深入程度，工作量及完整程度，模型的输出结果与原始数据的贴近程度。

(d) 一旦发现抄袭，计零分。

(e) 请大家在规定截止时间前提交。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。

参考文献

[1] 茆诗松，程依明，濮晓龙, 概率论与数理统计教程. 高等教育出版社, 2004.

[2] 陈家鼎，郑忠国, 概率与统计. 北京大学出版社, 2007.

[3] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[4] D. Duvenaud, J. R. Lloyd, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani, "Structure Discovery in Nonparametric Regression through Compositional Kernel Search," *Creative Commons Attribution-Noncommercial-Share Alike*, 2013.