Пятое задание по курсу «Байесовские методы статистического оценивания»

Бурнаев Е., Зайцев А., Янович Ю.

В задачах будем использовать следующие обозначения. Задана выборка $D = (X, \mathbf{y}) = \{(\mathbf{x}_i, y_i = y(\mathbf{x}_i))\}_{i=1}^n$ — выборка из n значений функции $y(\mathbf{x}_i)$ в точках $\mathbf{x}_i, i = \overline{1,n}$. Задана ковариационная функция гауссовского процесса $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ из параметрического семейства $\{k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}'), \boldsymbol{\theta} \in \Theta\}$.

1. (2 балла) Пусть

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Lambda^{-1}),$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|A\mathbf{x} + \mathbf{b}, L^{-1}).$$

Показать, что тогда маргинальное распределение $\mathbf{y}$ и условное распределение $\mathbf{x}$ для заданного $\mathbf{y}$ имеют вид:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|A\boldsymbol{\mu} + \mathbf{b}, L^{-1} + A\Lambda^{-1}A^T),$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\Sigma A^T L(\mathbf{y} - \mathbf{b}) + \Lambda\boldsymbol{\mu}, \Sigma),$$

где $\Sigma = (\Lambda + A^T LA)^{-1}$.

2. (1 балл) Показать, что для $\theta < 0$ функция $k_\theta(x, x') = \exp\left(-\theta(x - x')^2\right)$ не может быть ковариационной функцией гауссовского процесса для $x \in \mathbb{R}$.

3. (1 балл) Показать, что модель регрессии на основе гауссовских процессов с ковариационной функцией $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}'$ эквивалента линейной регрессии.

4. (2 балла) Выборку, из которой исключили j-ый элемент, обозначим $D_{-j} = \{(\mathbf{x}_i, y_i = y(\mathbf{x}_i))\}_{i=1,2,\ldots,j-1,j+1,\ldots,n}$. Для заданной апостериорной функции назовем $\hat{y}_{-j} = \mathbb{E}p(y(\mathbf{x}_j)|D_{-j}, \boldsymbol{\theta})$ апостериорное среднее гауссовского процесса в точке $\mathbf{x}_j$ с ковариационной функцией $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ и выборкой D_{-j}. Получить выражение для эффективного подсчета ошибки скользящего контроля: $\sum_{j=1}^n (\hat{y}_{-j} - y_j)^2$.

5. (5 баллов) Строится регрессионная модель для $y(\mathbf{x})$, $\mathbf{x} \in [0,1]^2$. Предлагается строить регрессионные модели с помощью адаптивного планирования эксперимента. Алгоритм адаптивного планирования эксперимента со-

стоит в том, что на каждом шаге на основании некоторого критерия к текущей выборке D_n размера n добавляется еще одна точка. Рассматривается четыре различных способа добавления новой точки на n-ом шаге:

- Новая точка выбирается случайно.
- Новая точка максимизирует минимальное расстояние до точек текущей выборки D_n.
- Новая точка максимизирует апостериорную дисперсию гауссовского процесса для выборки D_n.
- Новая точка минимизирует ошибку аппроксимации на заданной тестовой выборке D_{test}, если добавить ее к текущей выборке D_n. Этот критерий не может быть реализован в реальных условиях, так как он требует многократного вычисления целевой функции. Здесь он используется для того, что бы сравнить перечисленные выше критерии с наилучшим жадным критерием.

С помощью численного моделирования нужно сравнить представленные четыре критерия: как для выбранной вами двумерной функции $\mathbf{x} \in \mathbb{R}^2$ убывает среднеквадратичная ошибка на тестовой выборке после добавления новой точки. Реализация каждого критерия дает один балл, реализация всех критериев дает полный балл за задачу.

В качестве алгоритма построения модели регрессии на основе гауссовских процессов можно использовать библиотеку `sklearn` для языка `Python`. Новую точку нужно выбирать из заданного множества точек из $[0, 1]^2$, то есть с самого начала у нас зафиксировано множество точек-кандидатов, из которых на каждом шаге мы выбираем новую точку в соответствии с используемым критерием. Как начальный дизайн можно использовать выборку размером 30 точек из равномерного распределения на $[0, 1]^2$.