Домашнее задание №4 по курсу
«Математическая Статистика в Машинном Обучении»

Федор Ерин

Задача 1 [3 балла]

Пусть данные порождаются моделью

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \varepsilon,$$

где $\varepsilon \sim \mathcal{N}(\varepsilon|0, \beta^{-1})$, $y(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}^T\varphi(\boldsymbol{x})$, $\varphi(\boldsymbol{x}) = (\varphi_1(\boldsymbol{x}), \ldots, \varphi_M(\boldsymbol{x}))$ — некоторый набор функций.

Также обозначим через $\boldsymbol{t} = (t_1, \ldots, t_N)$ – известный набор целевых значений, через $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ – соответствующие значения векторов признаков, $\Phi = (\varphi_j(\boldsymbol{x}_i)), i = 1, \ldots, N, j = 1, \ldots, M$.

Предположим, что априорное распределение $p(\boldsymbol{w}|\alpha) = N(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$. В таком случае легко подсчитать, что апостериорное распределение $p(\boldsymbol{w}|\boldsymbol{t}, \alpha) = N(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N)$, где

$$\boldsymbol{m}_N = \beta\boldsymbol{S}_N\Phi^T\boldsymbol{t},$$

$$\boldsymbol{S}_N^{-1} = \alpha\boldsymbol{I} + \beta\Phi^T\Phi.$$

Апостериорное распределение выходного значения t в новой точке $\boldsymbol{x}$ равно

$$p(t|\boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) = \int p(t|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\boldsymbol{t}, \alpha, \beta)d\boldsymbol{w} = N(t|\boldsymbol{m}_N^T\varphi(\boldsymbol{x}), \sigma_N^2(\boldsymbol{x})),$$

где

$$\sigma_N^2(\boldsymbol{x}) = \frac{1}{\beta} + \varphi(\boldsymbol{x})^T\boldsymbol{S}_N\varphi(\boldsymbol{x}).$$

Используя равенство

$$(\boldsymbol{M} + \boldsymbol{v}\boldsymbol{v}^T)^{-1} = \boldsymbol{M}^{-1} - \frac{(\boldsymbol{M}^{-1}\boldsymbol{v})(\boldsymbol{v}^T\boldsymbol{M}^{-1})}{1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}}, \qquad (1)$$

показать, что

$$\sigma_{N+1}^2(\boldsymbol{x}) \leq \sigma_N^2(\boldsymbol{x}).$$

Решение

- Найдем S_{N+1}^{-1}, для этого рассмотрим новое наблюдение $(\boldsymbol{x}_{N+1}, t_{N+1})$ и выпишем функцию правдоподобия:

  $$p(t_{N+1}|\boldsymbol{x}_{N+1}, \boldsymbol{w}) = N(t_{N+1}|y(\boldsymbol{x}_{N+1}, \boldsymbol{w}), \beta^{-1})$$

  Распишем показатель экспоненты в апостериорном распределении:

  $$(\boldsymbol{w} - \boldsymbol{m}_N)^T\boldsymbol{S}_N^{-1}(\boldsymbol{w} - \boldsymbol{m}_N) + \beta(t_{N+1} - \boldsymbol{w}^T\varphi(\boldsymbol{x}_{N+1}))^2 =$$

  $$= \boldsymbol{w}^T(\boldsymbol{S}_N^{-1} + \beta\varphi(\boldsymbol{x}_{N+1})\varphi(\boldsymbol{x}_{N+1})^T)\boldsymbol{w} - 2\boldsymbol{w}^T(\boldsymbol{S}_N^{-1}\boldsymbol{m}_N + \beta\varphi(\boldsymbol{x}_{N+1})t_{N+1}) + const$$

  Отсюда получили:

  $$\boldsymbol{S}_{N+1}^{-1} = \boldsymbol{S}_N^{-1} + \beta\varphi(\boldsymbol{x}_{N+1})\varphi(\boldsymbol{x}_{N+1})^T$$

  Воспользуемся равенством (1) из условия:

  $$\boldsymbol{S}_{N+1} = (\boldsymbol{S}_N^{-1} + \beta\varphi(\boldsymbol{x}_{N+1})\varphi(\boldsymbol{x}_{N+1})^T)^{-1} =$$

  $$= \boldsymbol{S}_N - \frac{(\boldsymbol{S}_N\sqrt{\beta}\varphi(\boldsymbol{x}_{N+1})(\sqrt{\beta}\varphi(\boldsymbol{x}_{N+1})^TS_N)}{1 + \sqrt{\beta}\varphi(\boldsymbol{x}_{N+1})^T\boldsymbol{S}_N\sqrt{\beta}\varphi(\boldsymbol{x}_{N+1})} = \boldsymbol{S}_N - \frac{\beta\boldsymbol{S}_N\varphi(\boldsymbol{x}_{N+1})\varphi(\boldsymbol{x}_{N+1})^T\boldsymbol{S}_N}{1 + \beta\varphi(\boldsymbol{x}_{N+1})^T\boldsymbol{S}_N\varphi(\boldsymbol{x}_{N+1})}$$

  Теперь можем рассмотреть разность:

  $$\sigma_N^2(\boldsymbol{x}) - \sigma_{N+1}^2(\boldsymbol{x}) = \varphi(\boldsymbol{x})(\boldsymbol{S}_N - \boldsymbol{S}_{N+1})\varphi(\boldsymbol{x}) =$$

  $$= \varphi(\boldsymbol{x})\frac{\beta\boldsymbol{S}_N\varphi(\boldsymbol{x}_{N+1})\varphi(\boldsymbol{x}_{N+1})^T\boldsymbol{S}_N}{1 + \beta\varphi(\boldsymbol{x}_{N+1})^T\boldsymbol{S}_N\varphi(\boldsymbol{x}_{N+1})}\varphi(\boldsymbol{x}) = \frac{\beta(\varphi(\boldsymbol{x})^T\boldsymbol{S}_N\varphi(\boldsymbol{x}_{N+1}))^2}{1 + \beta\varphi(\boldsymbol{x}_{N+1})^T\boldsymbol{S}_N\varphi(\boldsymbol{x}_{N+1})} \geq 0,$$

  так как $\boldsymbol{S}_N$ положительно определена.

Задача 2 [5 баллов]

Рассмотрим модель регрессии

$$t = \boldsymbol{x}^T \boldsymbol{w} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \beta^{-1}),$$

где $p(\boldsymbol{w}|\beta) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_0, \beta^{-1}\boldsymbol{S}_0)$, $p(\beta) = \text{Gamma}(\beta|a_0, b_0)$, т.е. совместное априорное распределение $(\boldsymbol{w}, \beta)$ имеет вид:

$$p(\boldsymbol{w}, \beta) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_0, \beta^{-1}\boldsymbol{S}_0)\text{Gamma}(\beta|a_0, b_0).$$

Покажите, что апостериорное распределение $(\boldsymbol{w}, \beta)$ после наблюдения выборки $(\boldsymbol{X}, \boldsymbol{t}) = \{(\boldsymbol{x}_i, t_i)\}_{i=1}^n$ имеет вид

$$p(\boldsymbol{w}, \beta|\boldsymbol{X}, \boldsymbol{t}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_n, \beta^{-1}\boldsymbol{S}_n)\text{Gamma}(\beta|a_n, b_n).$$

Найдите параметры $\boldsymbol{w}_n$, $\boldsymbol{S}_n$, a_n, b_n.

Примечание. Плотность гамма-распределения $\text{Gamma}(a, b)$ *имеет вид*

$$\text{Gamma}(x|a, b) = \frac{x^{a-1}}{\Gamma(a)b^a} e^{-\frac{x}{b}}.$$

Решение

- Распишем априорную плотность и функцию правдоподобия:

$$p(\boldsymbol{w}, \beta) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_0, \beta^{-1}\boldsymbol{S}_0)\text{Gamma}(\beta|a_0, b_0) \sim \frac{\beta^2}{\boldsymbol{S}_0^2} exp\big(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_0)^T \beta \boldsymbol{S}_0^{-1}(\boldsymbol{w} - \boldsymbol{w}_0)\big) b_0^{a_0} \beta^{a_0 - 1} exp(-b_0\beta)$$

$$p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^n \mathcal{N}(t_i|\boldsymbol{w}^T\varphi(\boldsymbol{x}_i), \beta^{-1}) \sim \prod_{i=1}^n \sqrt{\beta} \cdot exp\big(-\frac{\beta}{2}(t_i - \boldsymbol{w}^T\varphi(\boldsymbol{x}_i))^2\big)$$

Известно, что

$$p(\boldsymbol{w}, \beta|\boldsymbol{t}) \sim p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}, \beta) \cdot p(\boldsymbol{w}, \beta)$$

Для нахождения $\boldsymbol{S}_n$ распишем только квадратичные слагаемые под экспонентой:

$$-\frac{\beta}{2}\boldsymbol{w}^T\boldsymbol{S}_0^{-1}\boldsymbol{w} + \sum_{i=1}^n -\frac{\beta}{2}\boldsymbol{w}^T\varphi(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_i)^T\boldsymbol{w} = -\frac{\beta}{2}\boldsymbol{w}^T(\boldsymbol{S}_0^{-1} + \sum_{i=1}^n \varphi(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_i)^T)\boldsymbol{w}$$

Отсюда получаем:

$$\boldsymbol{S}_n^{-1} = \boldsymbol{S}_0^{-1} + \sum_{n=1}^N \varphi(\boldsymbol{x}_n)\varphi(\boldsymbol{x}_n)^T \Rightarrow \boldsymbol{S}_n = (\boldsymbol{S}_0^{-1} + \sum_{n=1}^N \varphi(\boldsymbol{x}_n)\varphi(\boldsymbol{x}_n)^T)^{-1}$$

Для нахождения $\boldsymbol{w}_n$ распишем линейные слагаемые:

$$\beta\boldsymbol{w}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{w} + \sum_{i=1}^n \beta t_i\varphi(\boldsymbol{x}_i)^T\boldsymbol{w} = \beta\big(\boldsymbol{w}_0^T\boldsymbol{S}_0^{-1} + \sum_{i=1}^n t_i\varphi(\boldsymbol{x}_i)^T\big)\boldsymbol{w}$$

С другой стороны:

$$\boldsymbol{w}_n^T\boldsymbol{S}_n^{-1} = \boldsymbol{w}_0^T\boldsymbol{S}_0^{-1} + \sum_{i=1}^n t_i\varphi(\boldsymbol{x}_i)^T$$

Откуда получаем:

$$\boldsymbol{w}_n = \boldsymbol{S}_n\big(\boldsymbol{S}_0^{-1}\boldsymbol{w}_0 + \sum_{i=1}^n t_i\varphi(\boldsymbol{x}_i)\big)$$

Для нахождения b_n распишем константные слагаемые:

$$(-\frac{\beta}{2}\boldsymbol{w}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{w}_0 - b_0\beta) - \frac{\beta}{2}\sum_{i=1}^n t_i^2 = -\beta(\frac{1}{2}\boldsymbol{w}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{w}_0 + b_0 + \frac{1}{2}\sum_{i=1}^n t_i^2)$$

При этом:

$$\frac{1}{2}\boldsymbol{w}_n^T\boldsymbol{S}_n^{-1}\boldsymbol{w}_n + b_n = \frac{1}{2}\boldsymbol{w}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{w}_0 + b_0 + \frac{1}{2}\sum_{i=1}^n t_i^2$$

Откуда получаем:

$$b_n = b_0 + \frac{1}{2}(\boldsymbol{w}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{w}_0 - \boldsymbol{w}_n^T\boldsymbol{S}_n^{-1}\boldsymbol{w}_n + \sum_{i=1}^n t_i^2)$$

Наконец, найдем a_n:

$$2 + a_n - 1 = (2 + a_0 - 1) + \frac{N}{2}$$

$$a_n = a_0 + \frac{N}{2}$$

Ответ:

- $w_n = \boldsymbol{S}_n\big(\boldsymbol{S}_0^{-1}\boldsymbol{w}_0 + \sum_{i=1}^n t_i\varphi(\boldsymbol{x}_i)\big)$,
- $S_n = (\boldsymbol{S}_0^{-1} + \sum_{i=1}^n \varphi(\boldsymbol{x}_n)\varphi(\boldsymbol{x}_i)^T)^{-1}$,
- $a_n = a_0 + N/2$,
- $b_n = b_0 + \frac{1}{2}(\boldsymbol{w}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{w}_0 - \boldsymbol{w}_n^T\boldsymbol{S}_n^{-1}\boldsymbol{w}_n + \sum_{i=1}^n t_i^2)$.

Задача 3 [4 балла]

Пусть дана выборка $(\boldsymbol{X}, \boldsymbol{t})$, $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Предположим, что наблюдаемые данные представляют собой зашумленные значения гауссовского случайного процесса $f(\boldsymbol{x})$, т.е. имеет место следующая модель:

$$t(\boldsymbol{x}) = f(\boldsymbol{x}) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

где $f(\boldsymbol{x})$ — стационарный гауссовский процесс с нулевым средним и функцией ковариации $K(\boldsymbol{x}', \boldsymbol{x}'')$. Через $\hat{K}(\boldsymbol{x}', \boldsymbol{x}'')$ обозначим функцию ковариации зашумленного гауссовского процесса $t(\boldsymbol{x})$:

$$\hat{K}(\boldsymbol{x}', \boldsymbol{x}'') = K(\boldsymbol{x}', \boldsymbol{x}'') + \sigma^2\delta_{\boldsymbol{x}', \boldsymbol{x}''}.$$

Оказалось, что в выборке $\boldsymbol{X}$ все индексирующие параметры идентичны: $\boldsymbol{x}_1 = \cdots = \boldsymbol{x}_n = \boldsymbol{x}$. Найдите мат. ожидание и дисперсию прогноза в произвольной точке $\boldsymbol{y}$. Чему равен прогноз в той же самой точке $\boldsymbol{x}$, при условии, что $f(\boldsymbol{x}) = t$ — истинное значение рассматриваемой реализации в точке $\boldsymbol{x}$? Чему равна дисперсия прогноза? Является ли оценка на $f(\boldsymbol{x})$ смещенной?

Замечание. *К объяснению понятия "переизмерения значения в конкретной точке $\boldsymbol{x}$" можно подходить несколькими способами. Формальный способ предполагает обобщение функции ковариации: рассмотрим i-ое измерение в точке $\boldsymbol{x}'$ и j-ое измерение в точке $\boldsymbol{x}''$. Тогда*

$$Cov(t(\boldsymbol{x}', i), t(\boldsymbol{x}'', j))) = \hat{K}(\boldsymbol{x}', \boldsymbol{x}'', i, j) = K(\boldsymbol{x}', \boldsymbol{x}'') + \sigma^2\delta_{\boldsymbol{x}', \boldsymbol{x}''}\delta_{i,j}.$$

Неформальный способ состоит в том, что измерения происходят в точках $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, очень близких к $\boldsymbol{x}$, но все же не совпадающих с ней. В результате значения шумов в этих точках независимы, а истинные значение целевого гауссовского процесса почти совпадают (в предположении непрерывности функции ковариации $K(\boldsymbol{x}', \boldsymbol{x}'')$).

Подсказка. *Для обращения матрицы вида $\alpha\boldsymbol{I} + \boldsymbol{E}$, где $\boldsymbol{I} \in \mathbb{R}^{n\times n}$ — единичная матрица, и $\boldsymbol{E} \in \mathbb{R}^{n\times n}$ — матрица из единиц, можно воспользоваться тождеством Шермана-Моррисона-Вудберри, положив $\boldsymbol{E} = \boldsymbol{1}\cdot\boldsymbol{1}^T$, $\boldsymbol{1} = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$.*

Задача 4 [5 баллов]

мотрим регрессию на основе гауссовских процессов. Пусть нам дана выборка $(\boldsymbol{x}, \boldsymbol{t}) = \{(x_i, t_i)\colon x_i, t_i \in \mathbb{R}\}_{i=1}^n$, где $x_i = x_1 + (i-1)h$, $h > 0$, т.е. n точек расположены равномерно на вещественной оси. Предположим, что выборка $(\boldsymbol{x}, \boldsymbol{t})$ представляет собой реализацию некоторого гауссовского случайного процесса $f(\cdot)$, т.е.

$$t_i = f(x_i).$$

Будем считать, что шума в наблюдениях нет. Пусть ковариационная функция $K(\cdot, \cdot)$ имеет вид:

$$K(x', x'') = \alpha \exp\left(-\frac{|x' - x''|}{\gamma}\right),$$

где α и γ — известные параметры. Найдите апостериорное среднее и дисперсию гауссовского случайного процесса в точке $x_* \geq x_n$.

Задача 5 [3 балла]

Рассмотрим задачу построения адаптивного дизайна на основе гауссовских процессов. Пусть $(\boldsymbol{X}, \boldsymbol{t}) = \{(\boldsymbol{x}_i, t_i) \colon \boldsymbol{x}_i \in \mathbb{X}, t_i \in \mathbb{R}\}$, где $\{\boldsymbol{x}_i\}_{i=1}^n$ — известные точки дизайна из рассматриваемого множества $\mathbb{X} \subset \mathbb{R}^d$, и $\{t_i\}_{i=1}^n$ — измеренные значения аппроксимируемой зависимости. Одним из критериев выбора следующей точки дизайна является *критерий максимальной дисперсии*:

$$\mathcal{I}_{\mathrm{MV}}(\boldsymbol{x}) \triangleq \hat{\sigma}^2(\boldsymbol{x}|\boldsymbol{X}),$$
$$\boldsymbol{x}_* = \arg\max_{\boldsymbol{x}\in\mathbb{X}} \mathcal{I}_{\mathrm{MV}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{x}\in\mathbb{X}} \hat{\sigma}^2(\boldsymbol{x}|\boldsymbol{X}),$$

где $\hat{\sigma}^2(\boldsymbol{x}|\boldsymbol{X})$ — апостериорная дисперсия в точке $\boldsymbol{x} \in \mathbb{X}$ при заданном дизайне $\boldsymbol{X}$ (заметим, что в случае гауссовских процессов $\hat{\sigma}^2(\boldsymbol{x}|\boldsymbol{X})$ не зависит от $\boldsymbol{t}$). Такой критерий легко вычислим и включает в себя информацию о поведении истинной зависимости, однако учитывает только локальное поведение и склонен выдавать точки близкие к границе множества $\mathbb{X}$. Поэтому более разумным является *критерий для минимизации ожидаемой среднеквадратичной ошибки аппроксимации на следующей итерации*:

$$\mathcal{I}_{\rho_2}(\boldsymbol{x}) \triangleq \frac{1}{|\mathbb{X}|} \int\limits_{\mathbb{X}} \left(\hat{\sigma}^2(\boldsymbol{v}|\boldsymbol{X}) - \hat{\sigma}^2(\boldsymbol{v}|\boldsymbol{X}\cup\boldsymbol{x})\right) d\boldsymbol{v},$$
$$\boldsymbol{x}_* = \arg\min_{\boldsymbol{x}\in\mathbb{X}} \mathcal{I}_{\rho_2}(\boldsymbol{x}),$$

где $\hat{\sigma}^2(\boldsymbol{v}|\boldsymbol{X})$ и $\hat{\sigma}^2(\boldsymbol{v}|\boldsymbol{X}\cup\boldsymbol{x})$ — апостериорные дисперсии в точке $\boldsymbol{v}$ на дизайнах $\boldsymbol{X}$ и $\boldsymbol{X}\cup\boldsymbol{x}$ соответственно. Покажите, что критерий $\mathcal{I}_{\rho_2}(\boldsymbol{x})$ может быть записан в виде:

$$\boxed{\mathcal{I}_{\rho_2}(\boldsymbol{x}) = \frac{1}{|\mathbb{X}|} \int\limits_{\mathbb{X}} \frac{\hat{K}^2(\boldsymbol{v},\boldsymbol{x})}{\hat{\sigma}^2(\boldsymbol{x}|\boldsymbol{X})} d\boldsymbol{v}},$$

где $\hat{K}(\boldsymbol{v},\boldsymbol{x}) = K(\boldsymbol{v},\boldsymbol{x}) - \boldsymbol{k}^T(\boldsymbol{v})\Sigma_{\boldsymbol{X}}^{-1}\boldsymbol{k}(\boldsymbol{x})$, и

$$\Sigma_{\boldsymbol{X}} = \begin{pmatrix} K(\boldsymbol{x}_1,\boldsymbol{x}_1) & \dots & K(\boldsymbol{x}_1,\boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ K(\boldsymbol{x}_n,\boldsymbol{x}_1) & \dots & K(\boldsymbol{x}_n,\boldsymbol{x}_n) \end{pmatrix}.$$

Задача 6 [5 баллов]

Допустим, что в нашем распоряжении имеется выборка $(\boldsymbol{X}, \boldsymbol{Y})$, где $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times d}$ и $\boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_n]^T \in \mathbb{R}^{n\times k}$, т.е. каждому объекту $\boldsymbol{x} \in \mathbb{R}^d$ соответствует вектор $\boldsymbol{y} \in \mathbb{R}^k$. Требуется по новому вектору $\boldsymbol{x}^*$ предсказать вектор $\boldsymbol{y}^*$. Ниже рассматривается алгоритм, основанный на использовании PCA.

Этап обучения. Для объединенной матрицы $\boldsymbol{Z} = [\boldsymbol{X}, \boldsymbol{Y}] \in \mathbb{R}^{n\times(d+k)}$ находим первые r (*параметр алгоритма*) главных компонент $\boldsymbol{u}_1, \dots, \boldsymbol{u}_r$, где $\boldsymbol{u}_i \in \mathbb{R}^{d+k}$. Расположим эти векторы в качестве строк матрицы $\boldsymbol{U}$:

$$\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_r] \in \mathbb{R}^{(d+k)\times r}$$

Каждый из векторов $\{\boldsymbol{u}_i\}_{i=1}^r$ представляется как объединение двух подвекторов: $\boldsymbol{u}_i^T = (\boldsymbol{u}_{X,i}, \boldsymbol{u}_{Y,i})$, где $\boldsymbol{u}_{X,i} \in \mathbb{R}^d$ соответствует пространству признаков, а $\boldsymbol{u}_{Y,i} \in \mathbb{R}^k$ — пространству целевых переменных.

Введем обозначения

$$\langle\boldsymbol{X}\rangle \triangleq \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \in \mathbb{R}^d, \qquad \langle\boldsymbol{Y}\rangle \triangleq \frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i \in \mathbb{R}^k, \qquad \langle\boldsymbol{Z}\rangle \triangleq \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i \in \mathbb{R}^{d+k}.$$

В таких обозначениях

$$\langle\boldsymbol{Z}\rangle = \begin{pmatrix} \langle\boldsymbol{X}\rangle \\ \langle\boldsymbol{Y}\rangle \end{pmatrix} \in \mathbb{R}^{d+k}.$$

- *Преобразование произвольного объединенного вектора* $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ *в сжатое описание* $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_d)$ происходит согласно следующей формуле:
$$\lambda_i = \boldsymbol{u}_i^T(\boldsymbol{z} - \langle\boldsymbol{Z}\rangle), \quad i = 1, \dots, d,$$
или в векторном виде
$$\boldsymbol{\lambda} = \boldsymbol{U}^T(\boldsymbol{z} - \langle\boldsymbol{Z}\rangle).$$

- *Восстановление* $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ *по его сжатому описанию* $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)$ происходит следующим образом:

$$\tilde{\boldsymbol{x}}(\boldsymbol{\lambda}) = \langle \boldsymbol{X} \rangle + \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_{X,i} = \langle \boldsymbol{X} \rangle + \boldsymbol{U}_X \boldsymbol{\lambda},$$

$$\tilde{\boldsymbol{y}}(\boldsymbol{\lambda}) = \langle \boldsymbol{Y} \rangle + \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_{Y,i} = \langle \boldsymbol{Y} \rangle + \boldsymbol{U}_Y \boldsymbol{\lambda}.$$

Этап предсказания. Теперь опишем алгоритм восстановления $\boldsymbol{y}$ по признакам $\boldsymbol{x}$:

1. По заданному $\boldsymbol{x}$ определяется вектор коэффициентов $\boldsymbol{\lambda}^*$:

$$\boldsymbol{\lambda}^*(\boldsymbol{x}) = \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}^r} \|\tilde{\boldsymbol{x}}(\boldsymbol{\lambda}) - \boldsymbol{x}\|_2^2.$$

2. По вычисленному вектору $\boldsymbol{\lambda}^*$ оценивается вектор целевых переменных по формуле

$$\boldsymbol{y}^*(\boldsymbol{x}) = \langle \boldsymbol{Y} \rangle + \boldsymbol{U}_Y \boldsymbol{\lambda}^*(\boldsymbol{x}).$$

Задание. Предполагается, что известны обучающая выборка $\boldsymbol{X}$, $\boldsymbol{Y}$ и матрица главных компонент $\boldsymbol{U}$.

- Найдите формулы для коэффициентов $\boldsymbol{\lambda}^*$, подсчитанных по вектору признаков $\boldsymbol{x}$.
- Формулы для $\boldsymbol{y}$, восстановленному по $\boldsymbol{x}$ согласно описанной выше процедуре.

Внимание. Ответы будут отличаться для случаев когда $r \leq d$ *и* $r > d$. *В решении следует предусмотреть оба случая. Для простоты при решении следует полагать, что* $\mathrm{rg}\boldsymbol{Z} = d + k$.

Задача 7 [10 баллов]

В данной задаче требуется реализовать подбор параметров алгоритма классификации с помощью байесовской оптимизации на основе гауссовских процессов.

- Создайте искусственный датасет для классификации. Возьмите 15 информативных признаков и 5 шумовых. Количество сэмплов возьмите от 1000 и выше: 500 сэмплов отведите на обучающую выборку, а оставшиеся на тестовую.
- Используйте Kernel SVM с RBF-ядром для того, чтобы обучиться на данных. Для параметров C и γ в некоторой сетке (например, в сетке `np.logspace(-4, 3, 71)` $\times$ `np.logspace(-4, 3, 71)`) подсчитайте значения ROC-AUC на тестовой выборке (не забудьте использовать для подсчета ROC AUC метод `decision_function`, а не `predict`). Отрисуйте в виде heat map-ы значения ROC AUC (см. для примера код с семинара или рис. 1 ниже). В качестве альтернативного подхода можно провести grid search по сетке с помощью класса `GridSearchCV`, который хранит в себе значения качества во всех подсчитанных узлах. Это сэкономит время, так как для ускорения можно передать параметр `jobs = -1`. Но при указанных выше параметрах подсчет в один поток не занимает много времени.
- Реализуйте следующий алгоритм выбора параметров (C, γ) с помощью байесовской оптимизации:
  1. На вход алгоритма подается множество пар значений C и γ, среди которых будет проводится поиск наилучших параметров. Для простоты пусть данные параметры образуют сетку `np.logspace(-4, 3, 701)` $\times$ `np.logspace(-4, 3, 701)`. Стоит заметить, что эта сетка гораздо плотнее построенной в предыдущем пункте.
  2. На первом шаге просэмплируйте из сетки 5-10 пар значений (C, γ), на которых явно подсчитайте ROC AUC.
  3. Обучите гауссовскую регрессию на полученных данных. Подсчитайте дисперсии предсказаний на всем множестве значений пар (C, γ), поданных на вход алгоритма. Выберите параметры C_* и γ_*, дисперсия прогноза ROC AUC для которых максимальна.
  4. Подсчитайте ROC AUC для C_* и γ_* и добавьте ее в выборку известных соответствий (C, γ) и ROC AUC.
  5. Далее вновь обучите регрессию на обновленных данных и т.д. Повторяем процесс итеративно до тех пор, пока не будет достигнуто заданное количество итераций `max_iter`, либо дисперсия прогноза ROC AUC в каждой точке не станет меньше некоторого порога `var_threshold` (параметры `max_iter` и `var_threshold` стоит подобрать так, чтобы получить более-менее хорошее качество).
  6. В тот момент, когда подсчеты остановились, в качестве (C, γ) выдаем точку с максимальным ROC AUC.
- **[10 баллов]** Отрисуйте все полученные в процессе оптимизации точки на heatmap-е с ROC AUC. Выделите цветом точку с лучшим качеством. Получилось ли с помощью байесовской оптимизации добиться того же качества, что и с помощью поиска по сетке? Помните, что сетка параметров для байесовской оптимизации плотнее, поэтому хотя чисто теоретически результат и может получаться один в один, но на практике это далеко не факт. К тому же в данной задаче не требуется получить результат, лучший, чем поиск по сетке. Требуется лишь показать, что этот результат приближается к результату, полученному по сетке.

Используйте ядро следующего вида

$$K(\boldsymbol{x}, \boldsymbol{x}') = \theta_1 \exp\left\{-\frac{1}{2}\left(\frac{(x_1 - x_1')^2}{2\eta_1} + \frac{(x_2 - x_2')^2}{2\eta_2}\right)\right\} + \theta_2 \delta_{\boldsymbol{x}, \boldsymbol{x}'}$$

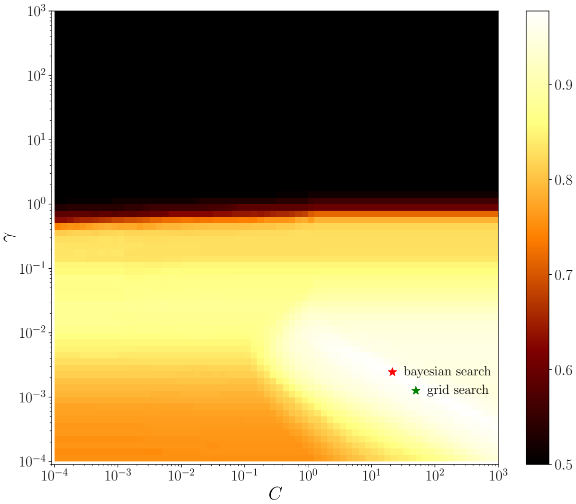

Рис. 1:

Так как на каждой итерации алгоритма выборка увеличивается, то вполне разумно переобучать параметры θ_1, θ_2, η_1, η_2 перед тем, как выбирать очередную точку $\boldsymbol{x} = (C, \gamma)$ для подсчета ROC AUC. В данной задаче это допустимо, так как количество точек и размерность невелики. В практических задачах можно просто перестать оптимизировать данные параметры как только их изменение от итерации к итерации становится незначительным.

Замечание. *Конкретные числа (число сэмплов, размер сетки, порог* `var_threshold`*) остаются на ваше усмотрение.*

Замечание. *Код должен запускаться от и до, чтобы была возможность полностью воспроизвести эксперимент. Ну и чем лучше логическое разделение кода, тем проще его проверять.*