Определение сложности выборки с помощью универсальной аппроксимирующей модели

Малиновский Григорий Станиславович

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

ММРО - 2019, г. Москва

Постановка задачи

Цель: предложить способ построения набора моделей локальной аппроксимации для решения задач классификации и регрессии.

Задачи

1. Проверка адекватности выборки для обобщенно-линейной модели
2. Оптимизировать структурные параметры выбираемых моделей по порождающей выборке с целью получения выборки с оптимальными свойствами.

Исследуемая проблема

Исследование свойств промежуточного параметрического пространства, строящегося моделями локальной аппроксимации.

Методы решения

Применение параметров универсальной аппроксимирующей модели для определения сложности выборки

Список литературы

- *A.M. Katrutsa, V.V. Strijov* Stress test procedure for feature selection algorithms // Chemometrics and Intelligent Laboratory Systems 142, 172-183.
- *A.A. Aduenko, A.P. Motrenko, V.V. Strijov* Object selection in credit scoring using covariance matrix of parameters estimations // Annals of Operations Research, 2018, 260(1-2) : 3-21.
- *G.V. Cybenko* Approximation by Superpositions of a Sigmoidal function // Mathematics of Control Signals and Systems. – 1989. – Т. 2, № 4. – С. 303–314.
- *L. Ling, Y.S. Abu-Mostafa* Data complexity in machine learning // Computer Science Technical Reports, 2006.003.
- *A.C. Lorena, A.I. Maciel, P.B.C. de Miranda, I.G. Costa, R.B.C. Prudencio* Data complexity meta-features for regression problems // Machine Learning 107 (1), 209-246

Постановка задачи

Задан временной ряд

$$S : T \to \mathbb{R}, \text{где } T = \{t_0, t_0 + d, t_0 + 2d, \ldots\}$$

. Определен сегмент временного ряда

$$\mathbf{x}_i = \left[S\left(t_i\right), S\left(t_i - d\right), S\left(t_i - 2d\right), \ldots, S\left(t_i - (n-1)d\right)\right]^\top, \quad \mathbf{x}_i \in X \equiv \mathbb{R}^n$$

Задана выборка
$\mathfrak{D} = \left\{(\mathbf{x}_i, y_i)\right\}_{i=1}^{l}, \quad y_i \in \{1, 2, \ldots K\}.$
$\mathbf{X} \in \mathbb{X}$ — множество наборов сегментов временного ряда.
$\mathbf{Y} \in \mathbb{Y}$ — множество меток классов движения (бег, ходьба, подъем) в случае задачи классификации; исследуемая величина (температура тела, давление) в случае задачи регрессии.
Требуется найти отображение

$$f : \mathbb{X} \to \mathbb{Y}, \text{ где } \mathbb{Y} = \{1, 2, \ldots K\} \text{ или } \mathbb{R}$$

Локальная аппроксимация

Модель локальной аппроксимации

$$g_i(\mathbf{w}, \mathbf{x}) \in \mathbb{X}, \text{ где } w \in \mathbb{W} = \mathbb{R}^{n_g}$$

Оптимальные параметры находятся из решения задачи

$$\mathbf{w}_i^{opt} = \arg \min_{\mathbf{w} \in \mathbb{R}^{n_\mathbf{g}}} \rho(g_i(\mathbf{w}, \mathbf{x}), \mathbf{x}), \text{ где } \rho \text{ — функция расстояния}$$

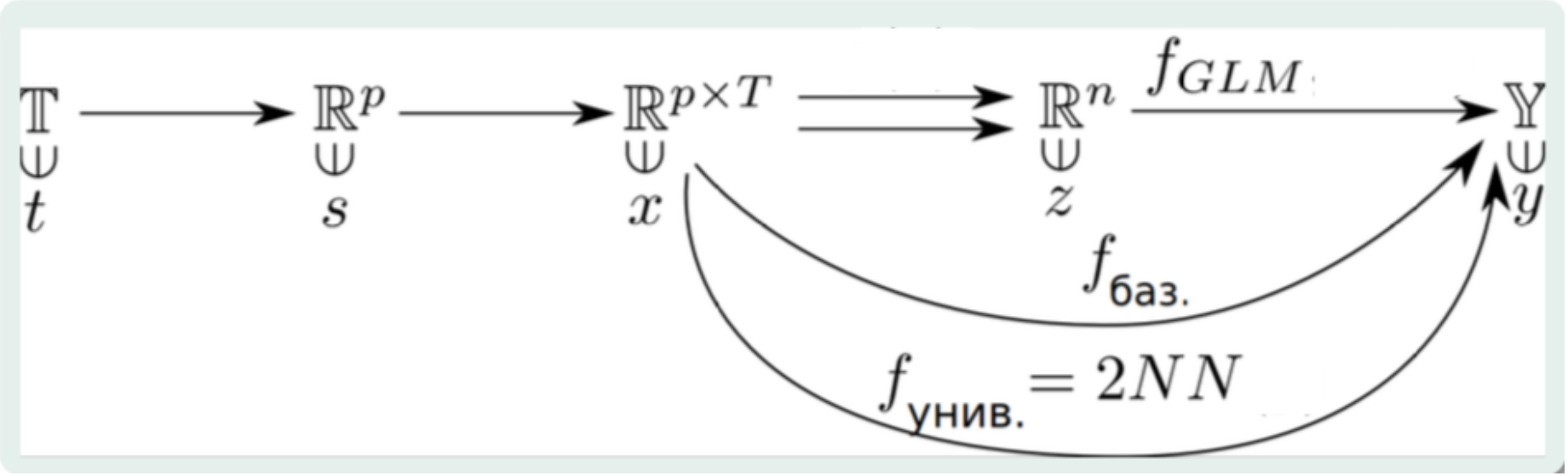
Промежуточное пространство

Данная процедура задает отображение из пространства сегментов временных рядов в пространство параметров.

$$h : \mathbb{X} \to \mathbb{Z} \subseteq \mathbb{W}$$

Пространство $\mathbb{Z}$ будем называть промежуточным пространством признаковых описаний.

Общая схема



$$\mathbb{X} \xrightarrow{f_{\text{баз}}(\mathbf{x})} \mathbb{Y} = \mathbb{X} \xrightarrow{h(\mathbf{x})} \mathbb{Z} \xrightarrow{f_{GLM}(\mathbf{z})} \mathbb{Y}$$

Требуется проанализировать, может ли выборка промежуточного признакового описания быть адекватно описана обобщенно-линейной моделью.

Коэффициенты детерминации

Определение 1

Коэффициент детерминации R^2

В случае задачи регрессии будем определять коэффициент по формуле:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

где y_i — истинные ответы, $\hat{y}_i$ — предсказания модели, $\bar{y}$ — среднее значение.

Определение 2

Коэффициент детерминации pseudo-R^2

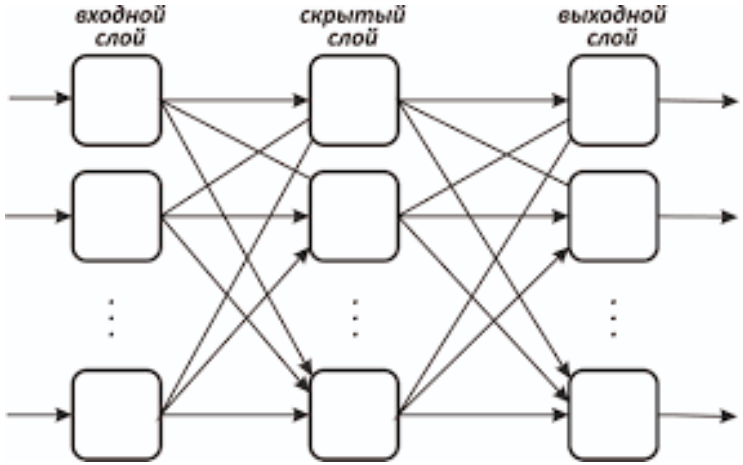
В случае классификации будем определять коэффициент по формуле:

$$R^2 = 1 - \frac{\ln \hat{L}(M)}{\ln \hat{L}(M_{\text{Null}})}$$

где $\hat{L}(M)$ — правдоподобие исследуемой модели, $\hat{L}(M_{\text{Null}})$ — правдоподобие константной модели.

Универсальная модель

Рассмотрим двухслойную полносвязную нейронную сеть.



Теорема Цыбенко

Пусть φ любая непрерывная сигмоидная функция, например, $\varphi(\xi) = 1/(1 + e^{-\xi})$. Тогда, если дана любая непрерывная функция действительных переменных f на $[0,1]^n$ (или любое другое компактное подмножество $\mathbb{R}^n$) и $\varepsilon > 0$, то существуют векторы $\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_N}, \alpha$ и θ и параметризованная функция $G(\cdot, \mathbf{w}, \alpha, \theta) : [0,1]^n \to \mathbb{R}$ такая, что для всех $\mathbf{x} \in [0,1]^n$ выполняется : $\left|G(\mathbf{x}, \mathbf{w}, \alpha, \theta) - f(\mathbf{x})\right| < \varepsilon$, где : $G(\mathbf{x}, \mathbf{w}, \alpha, \theta) = \sum_{i=1}^{N} \alpha_i \varphi(\mathbf{w}_i^T \mathbf{x} + \theta_i)$, и $\mathbf{w}_i \in \mathbb{R}^n$, $\alpha_i, \theta_i \in \mathbb{R}$, $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)$, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$, и $\theta = (\theta_1, \theta_2, \ldots, \theta_N)$.

Определение сложности

Определение 3

Сложностью универсальной модели будем называть количество нейронов n_{uni} на скрытом слое

Определение 4

Сложностью переобучения для выборки будем называть минимальное число нейронов на скрытом слое универсальной модели, при котором коэффициент детерминации равен 1.

$$comp_{fit}(\mathfrak{D}) = \arg\min_{n_{uni} \in \mathbb{N}} n_{uni} | R^2(\mathfrak{D}, n_{uni}) = 1$$

Следствие 1

Для любой возможной выборки сложность переобучения меньше бесконечности.
Доказательство следует из определения и теоремы Цыбенко.

Пусть задана выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m} = \{(\mathbf{X}, \mathbf{y})\}$$

Представим нашу выборку в виде:

$$\mathbf{X} = [X_1, \ldots, X_j, \ldots, X_n], \mathbf{X} \in \mathbb{R}^{m \times n} \text{ and } j \in \mathcal{J} = \{1, \ldots, n\}$$

$$\mathbf{y} = [y_1, \ldots, y_m] \in \mathbb{Y} \subset \mathbb{R}^m$$

Полагаем, что все признаки и вектор ответов нормализованы

$$\|\mathbf{y}\|_2 = 1 \text{ and } \|x_j\|_2 = 1, j \in \mathcal{J}$$

Далее введем определения мультиколлинеарности и скоррелированности

Мультиколлинеарность и скоррелированность

Определение 5

Признаки с индексами из множества A мультиколлениарны, если существует индекс j, коэффициенты a_k и множество индексов $\{k\} \subset A \setminus j$, существует достаточно малое число $\delta > 0$ — степень мультиколлинеарности, такое что

$$\left\| x_j - \sum_{k=A\setminus j} a_k x_k \right\|_2^2 < \delta$$

Определение 6

Признаки с индексами i, j скоррелированны, если существует достаточно малое число $\delta_{i,j} > 0$, такое что

$$\|x_i - x_j\|_2^2 < \delta_{i,j}$$

Определение 7

Признак с индексом j скоррелирован с вектором ответов, если существует существует достаточно малое число $\delta > 0$, такое что

$$\|\mathbf{y} - x_j\|_2^2 < \delta_{yj}$$

Конфигурации данных

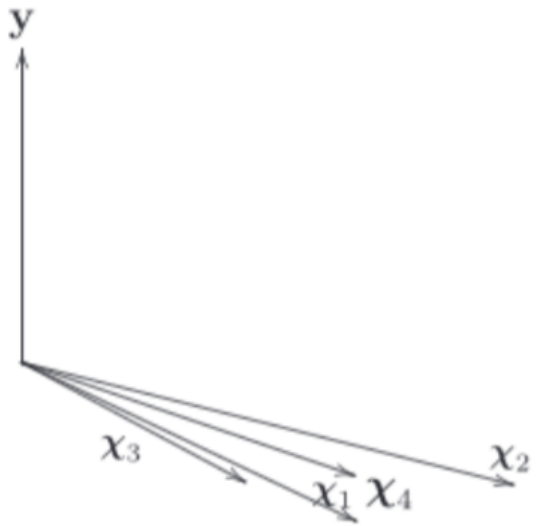


Fig. 1. The inadequate and correlated data set.

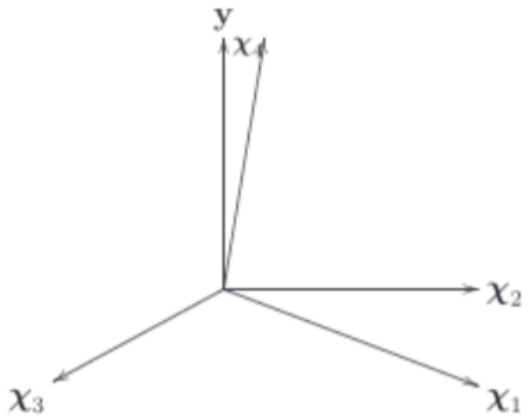


Fig. 2. The adequate and random data set.

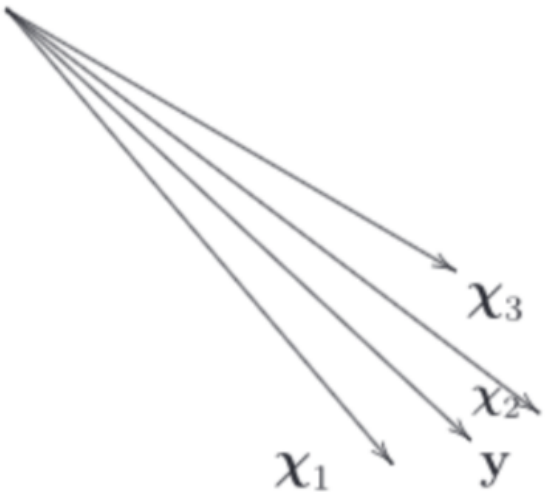


Fig. 3. The adequate and redundant data set.

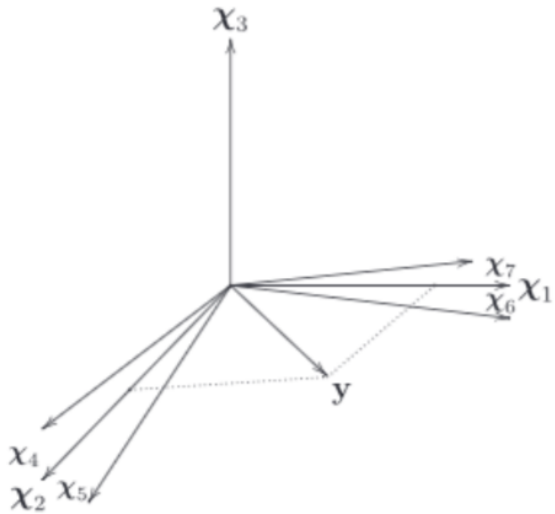
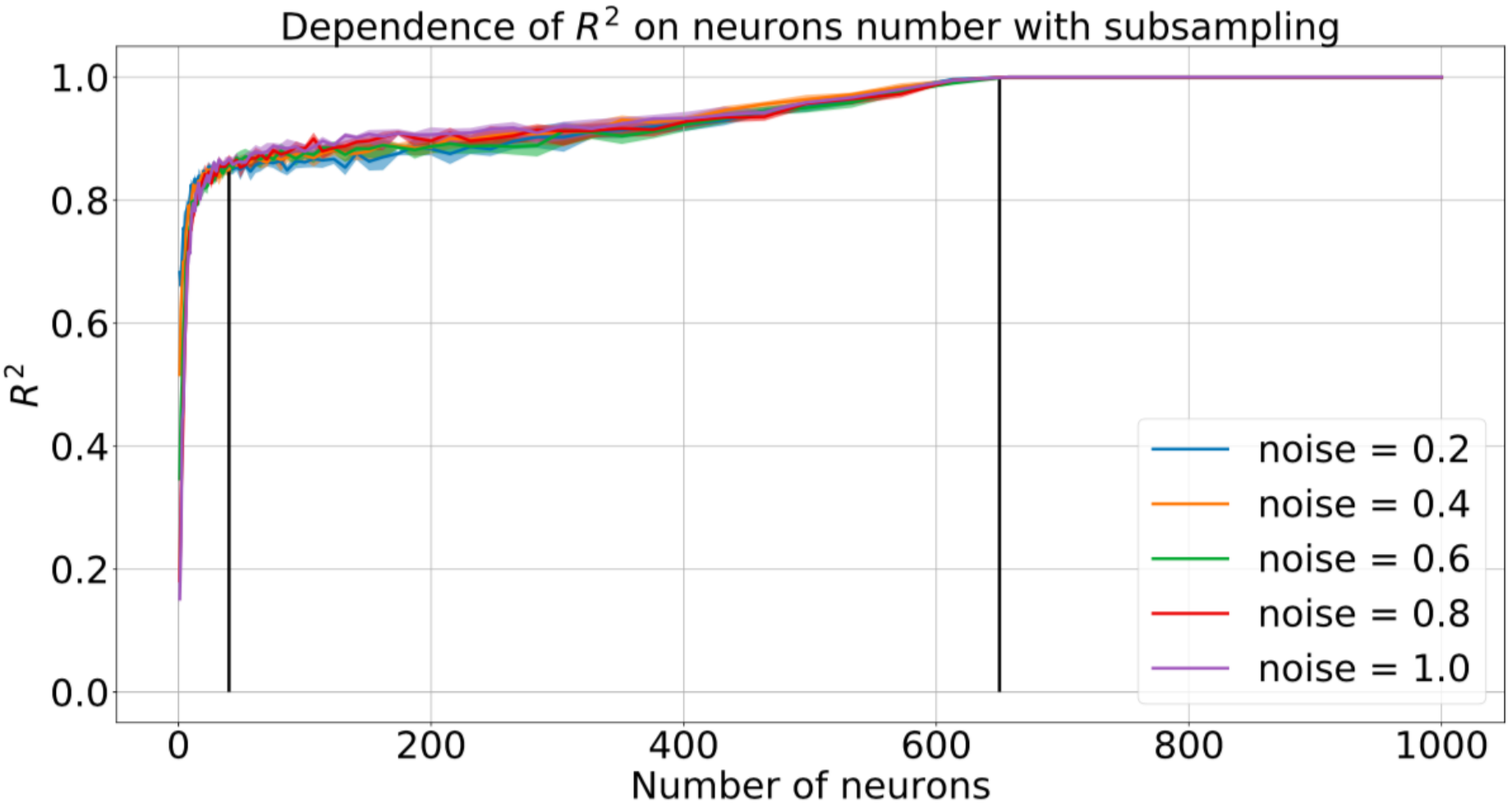
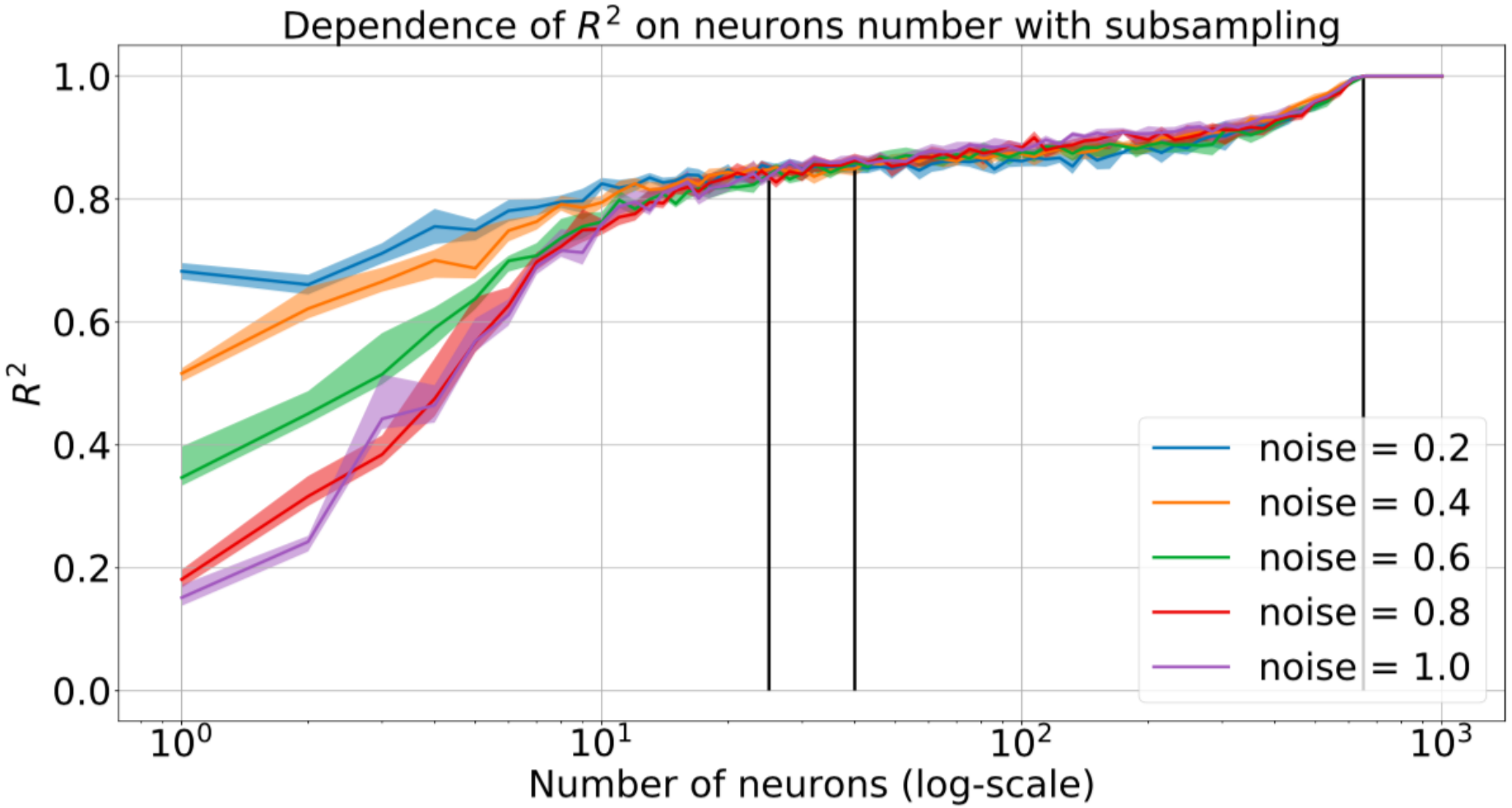


Fig. 4. The adequate and correlated data set.

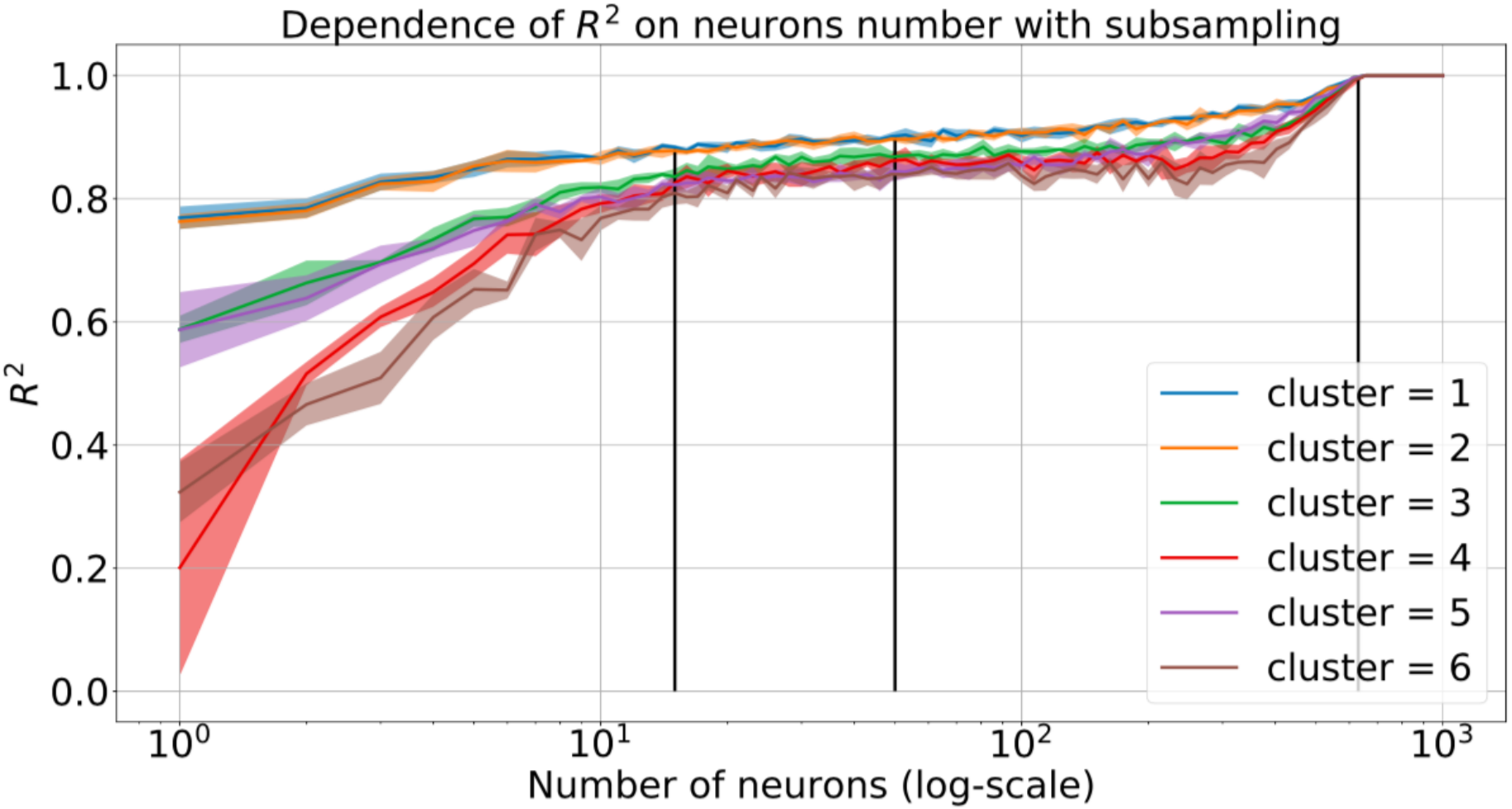
Эксперимент с уровнем шума



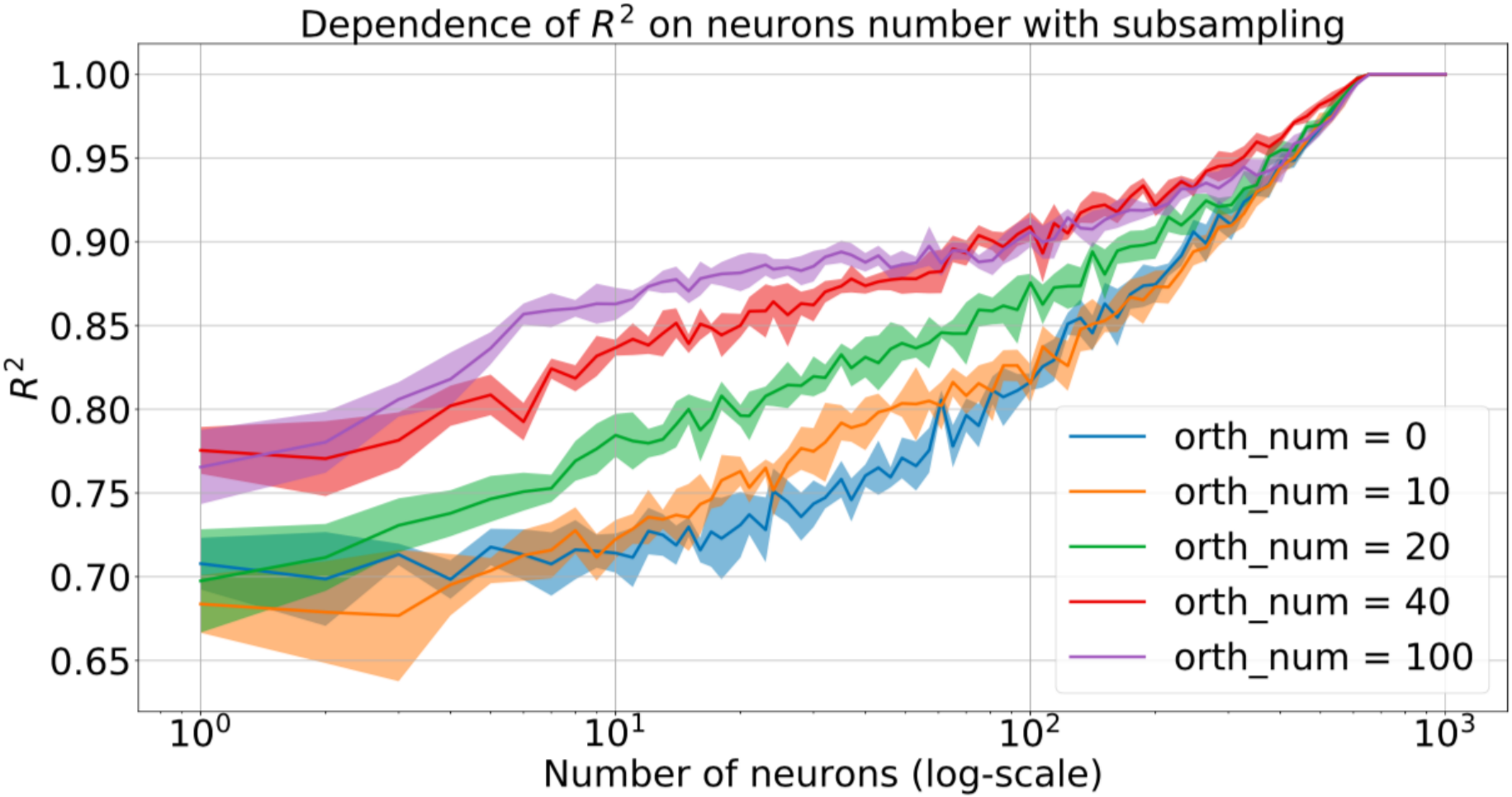
Эксперимент с уровнем шума



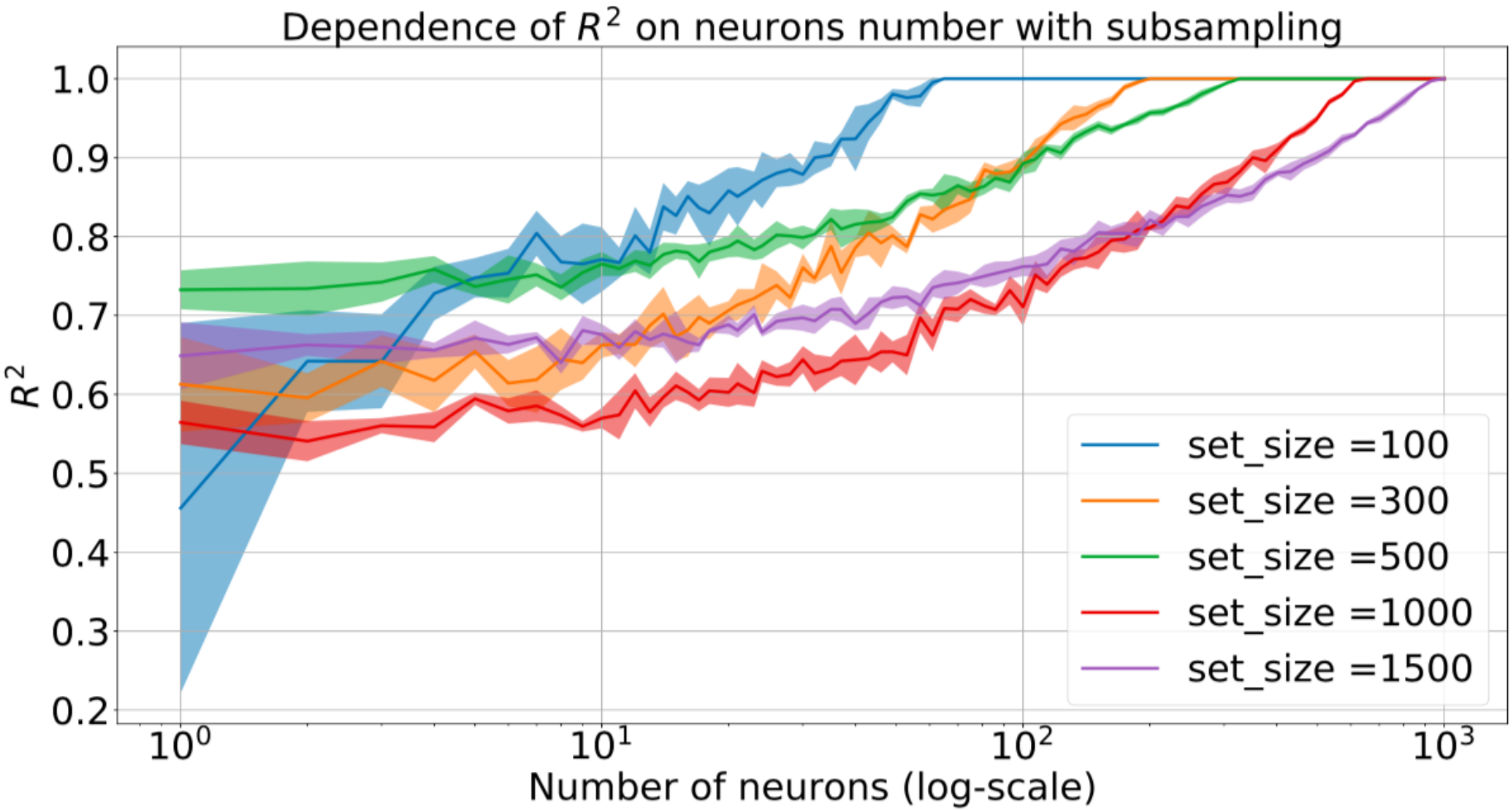
Эксперимент с количеством кластеров



Эксперимент с количеством ортогональных признаков



Эксперимент с количеством объектов



Эксперимент с количеством важных признаков