Выбор моделей в машинном обучении

В. В. Стрижов, А. А. Адуенко,
О. Ю. Бахтеев, Р. В. Исаченко, О. В. Грабовой

Московский физико-технический институт
Кафедра интеллектуальных систем

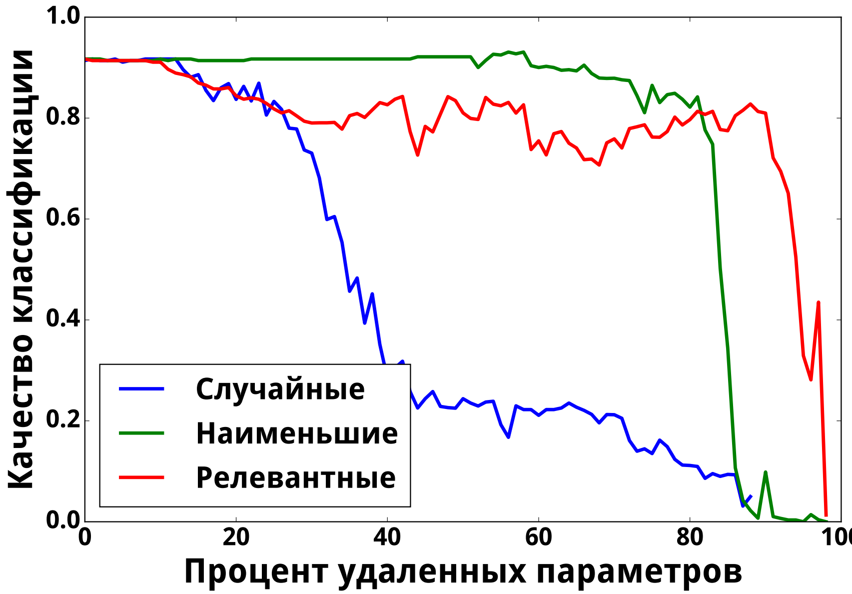
2019

Значительное повышение сложности и скромный прирост точности

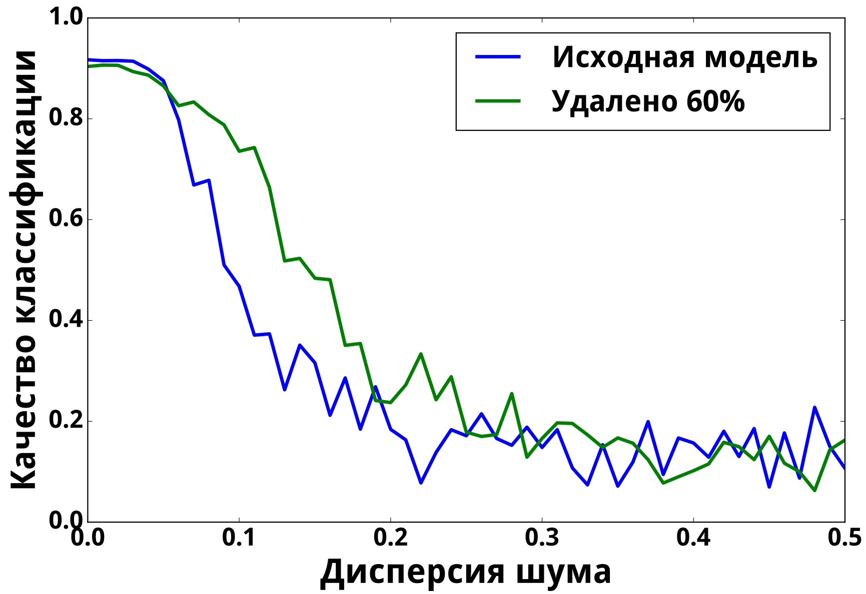
	train	test	Number of parameters
Логистическая регрессия	53,08%	55,18%	= 12
Нейронная сеть	59,85%	57,04%	~ 240
Случайный лес	61,85%	57,01%	> 1000
Градиентный бустинг	**63,58%**	**58,31%**	> 10 000

... это был скоринг

Правдоподобие моделей с избыточным числом параметров не изменяется значимо при их удалении



Избыточность параметров модели

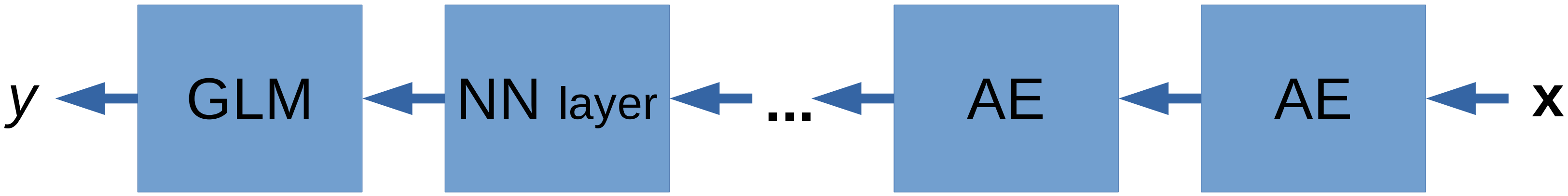


Устойчивость модели

Глубокое обучение предполагает оптимизацию моделей с заведомо избыточной сложностью.

Bakhteev, Strijov. 2019. Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Annals of Operations Research

Линейная модель, (глубокая) нейросеть и автоэнкодер



$$f = \sigma_k \circ \underset{1\times 1}{\mathbf{w}_k^{\mathsf{T}}} \boldsymbol{\sigma}_{k-1} \circ \mathbf{W}_{k-1} \boldsymbol{\sigma}_{k-2} \circ \cdots \circ \underset{n_2 \times 1}{\mathbf{W}_2} \boldsymbol{\sigma}_1 \circ \underset{n_1 \times n}{\mathbf{W}_1} \underset{n\times 1}{\mathbf{x}}$$

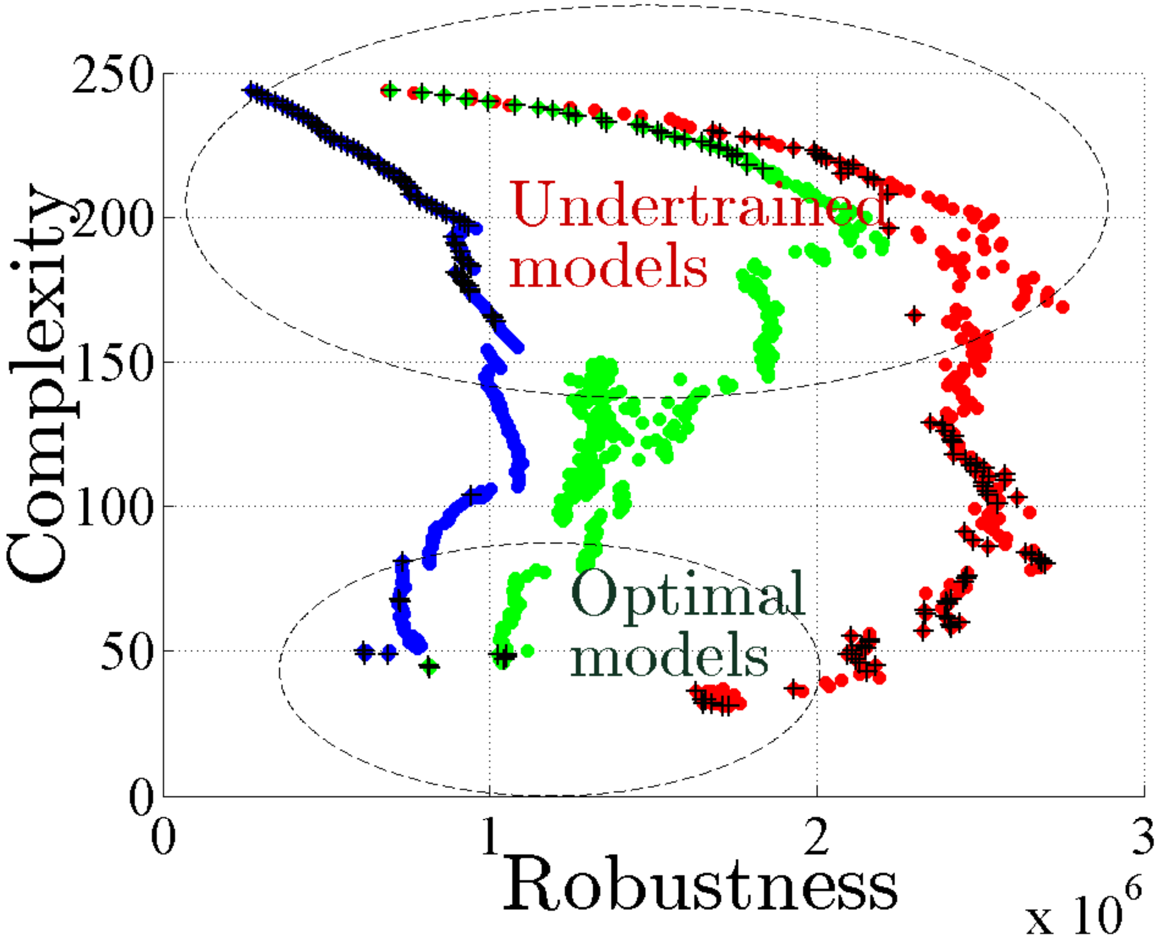
$$S = \sum_{(\mathbf{x}_i, y_i) \in \mathfrak{D}} \bigl(y_i - f(\mathbf{x}_i)\bigr)^2 \qquad E_{\mathbf{x}} = \sum_{\mathbf{x}_i \in \mathfrak{D}} \|\mathbf{x}_i - \mathbf{r}(\mathbf{x}_i)\|_2^2$$

Варианты

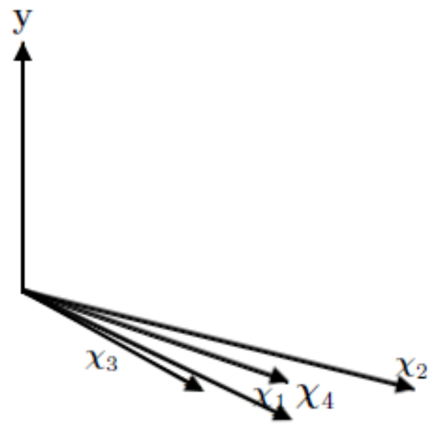
- $E_{\mathbf{x}}$ — ошибка реконструкции автоэнкодера
- метод главных компонент: $\mathbf{W}^{\mathsf{T}}\mathbf{W} = \mathbf{I}_n$
- нет блока: $\mathbf{W} = \mathbf{I}_n$, $\sigma = \mathrm{id}$
- классификация: $\sigma = \bigl(1 + \exp(-\cdot)\bigr)^{-1}$

$+\mathbf{b}$

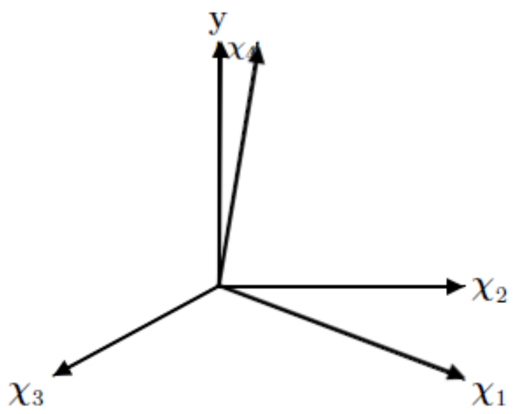
Последовательный выбор моделей: точность, сложность, устойчивость



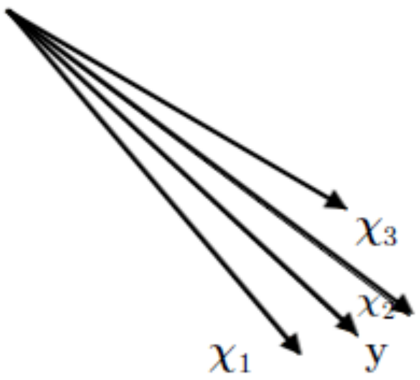
Конфигурации признакового пространства



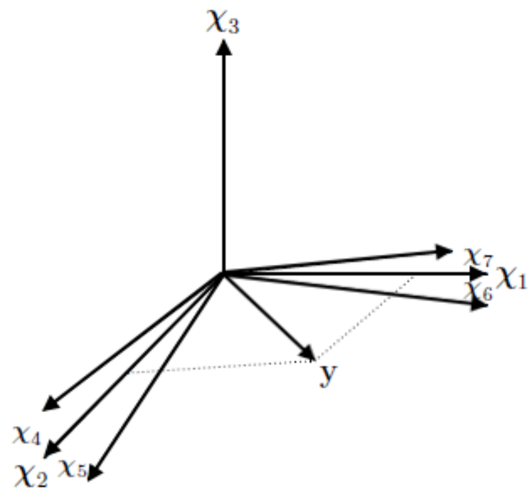
Неадекватный коррелированный



Адекватный случайный



Адекватный избыточный

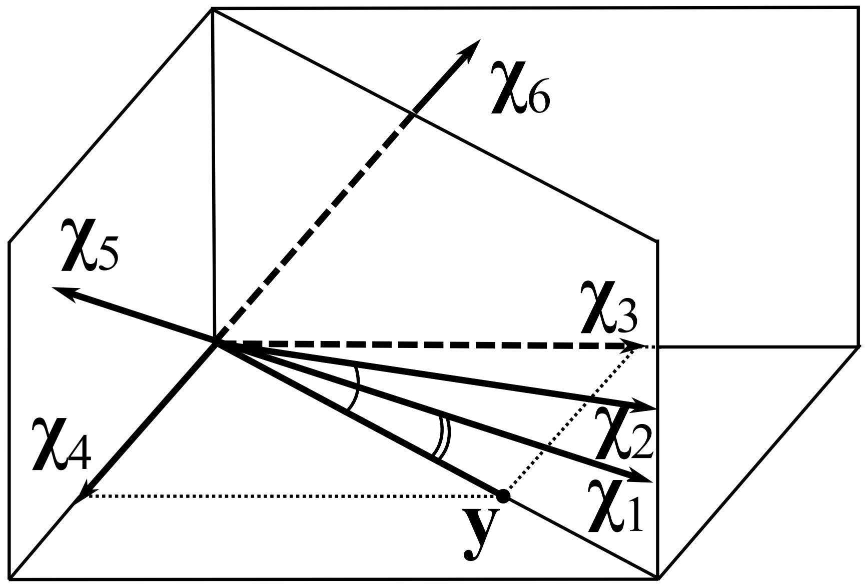


Адекватный коррелированный

Katrutsa, Strijov. 2017. Comprehensive study of feature selection methods to solve multicollinearity problem // Expert Systems with Applications

Выбор устойчивого и точного набора признаков

Признаки $\chi_1, \ldots, \chi_6$ — столбцы матрицы плана $\underset{3\times 6}{\mathbf{X}}$.

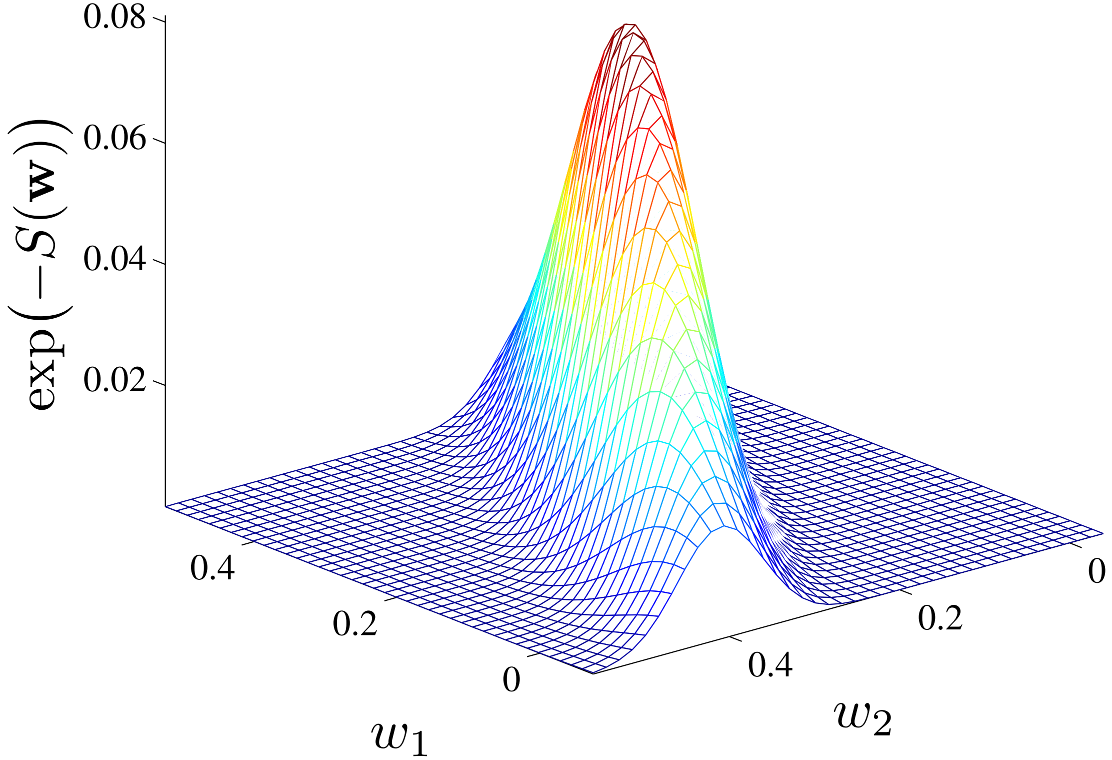


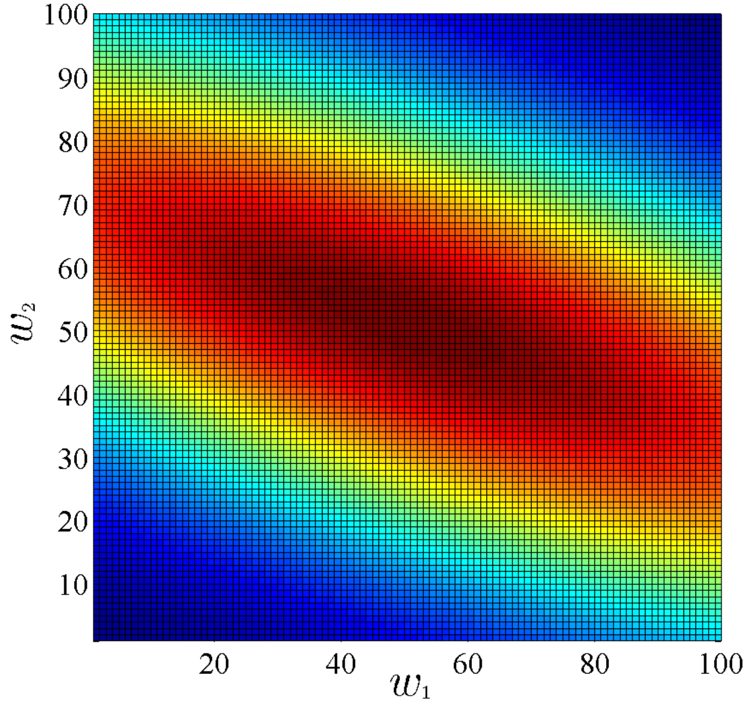
Решение: χ_3, χ_4 ортогональны, их комбинация приближает $\mathbf{y}$, минимизируя ошибку.

Katrutsa, Strijov. 2015. Stress-test procedure for feature selection // Chemometrics

Эмпирическое распределение параметров модели

Значение функции ошибки $S(\mathbf{w}|\mathfrak{D}, f)$ зависит от параметров.





Kuznetsov, Tokmakova, Strijov. 2016. Analytic methods of structure parameter // Informatica

Байесовский вывод, первый уровень

$$p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}) = \frac{p(\mathfrak{D}|\mathbf{w}, \mathbf{B})p(\mathbf{w}|\mathbf{A})}{p(\mathfrak{D}|\mathbf{A}, \mathbf{B})}.$$

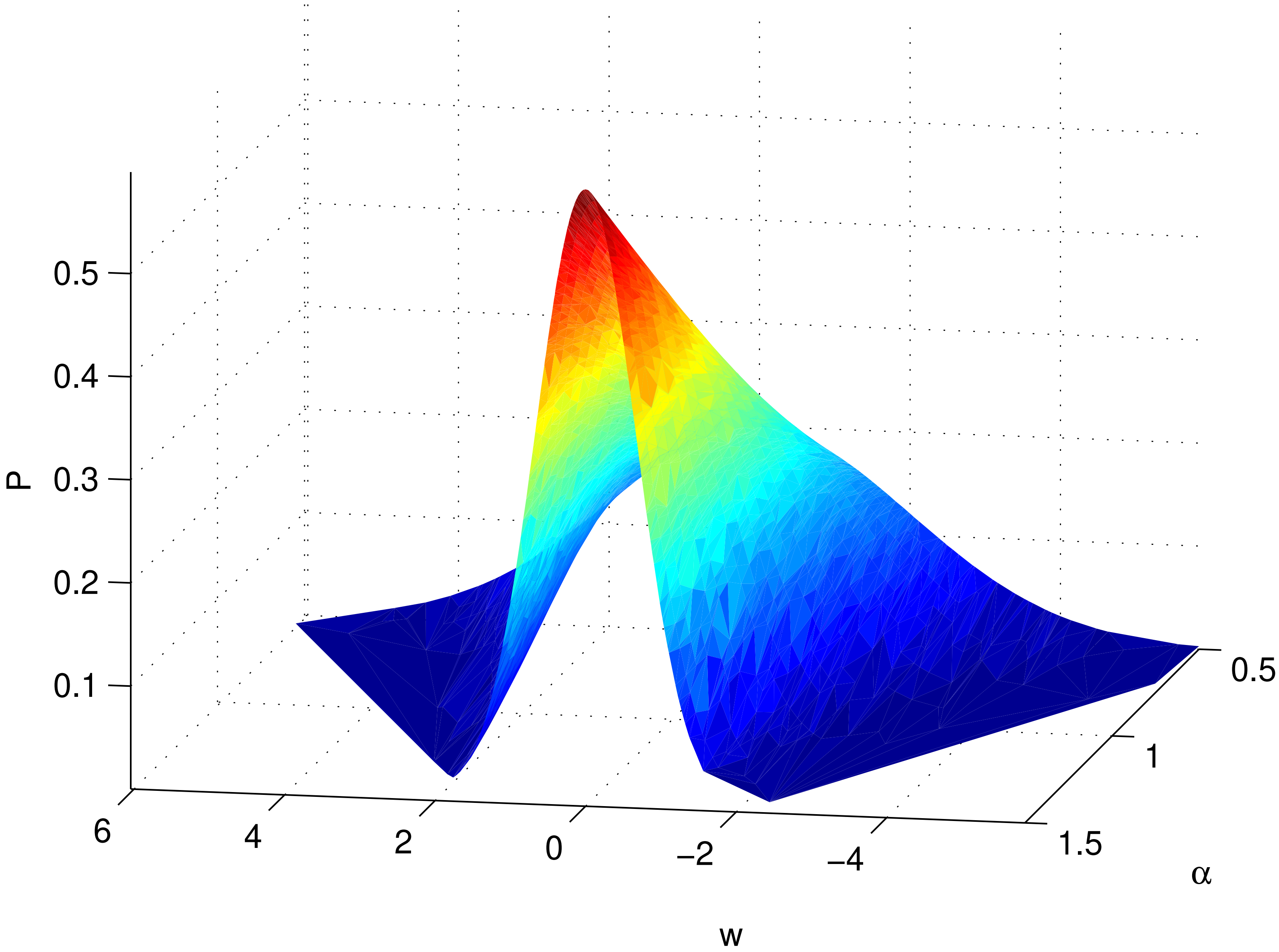
Элементы этого выражения и соответствующие им параметры:

- $p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B})$ — апостериорное распределение параметров,
- $p(\mathfrak{D}|\mathbf{w}, \mathbf{B})$ — функция правдоподобия данных,
- $p(\mathbf{w}|\mathbf{A})$ — априорное распределение параметров,
- $p(\mathfrak{D}|\mathbf{A}, \mathbf{B})$ — функция правдоподобия модели (обоснованность).

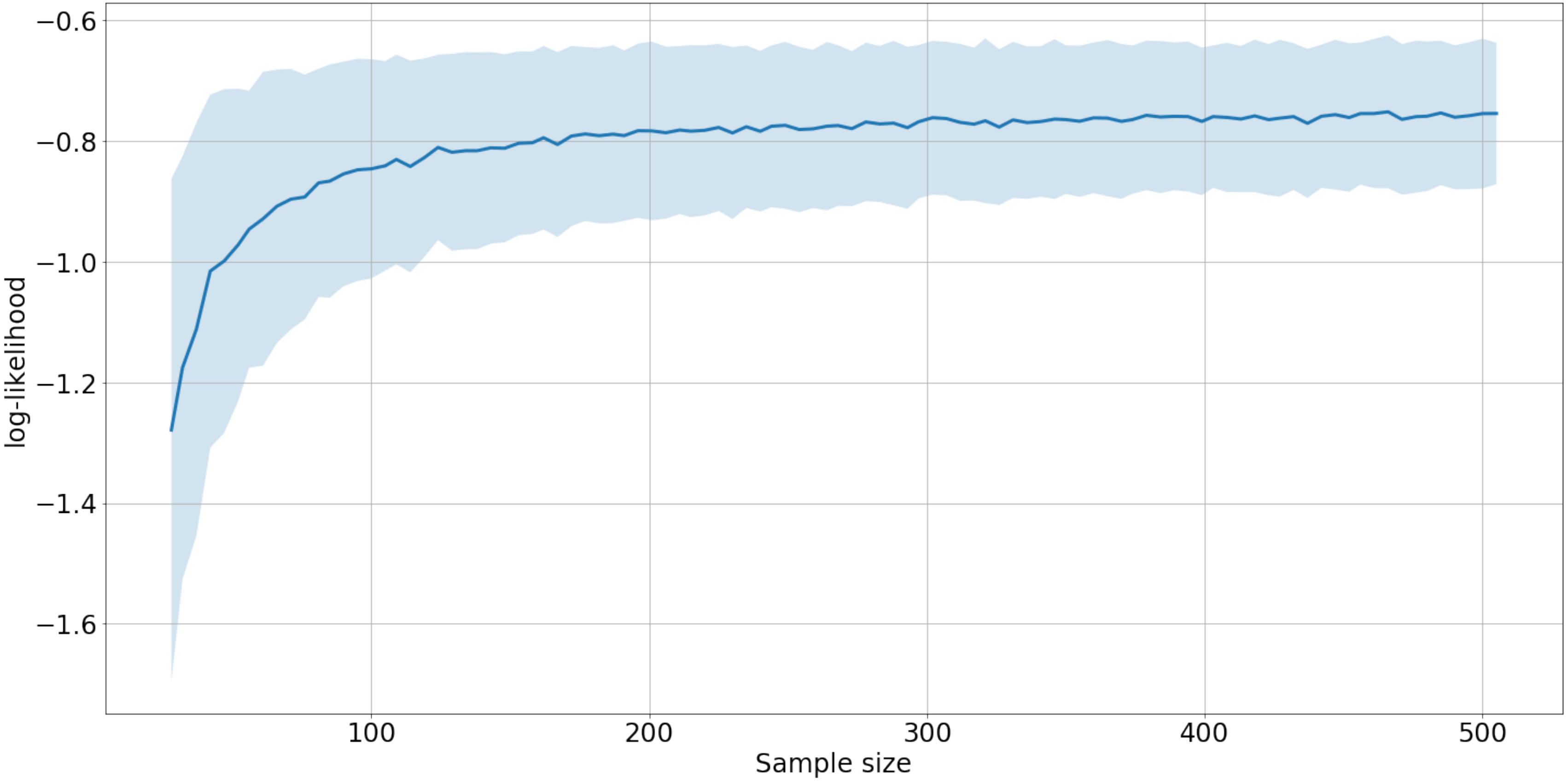
Функция ошибки $S = E_{\mathbf{w}} + E_{\mathfrak{D}}$, пример для регрессии

$$S(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^{\mathsf{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_0) + \frac{1}{2}(\mathbf{y} - \mathbf{f})^{\mathsf{T}}\mathbf{B}(\mathbf{y} - \mathbf{f}),$$

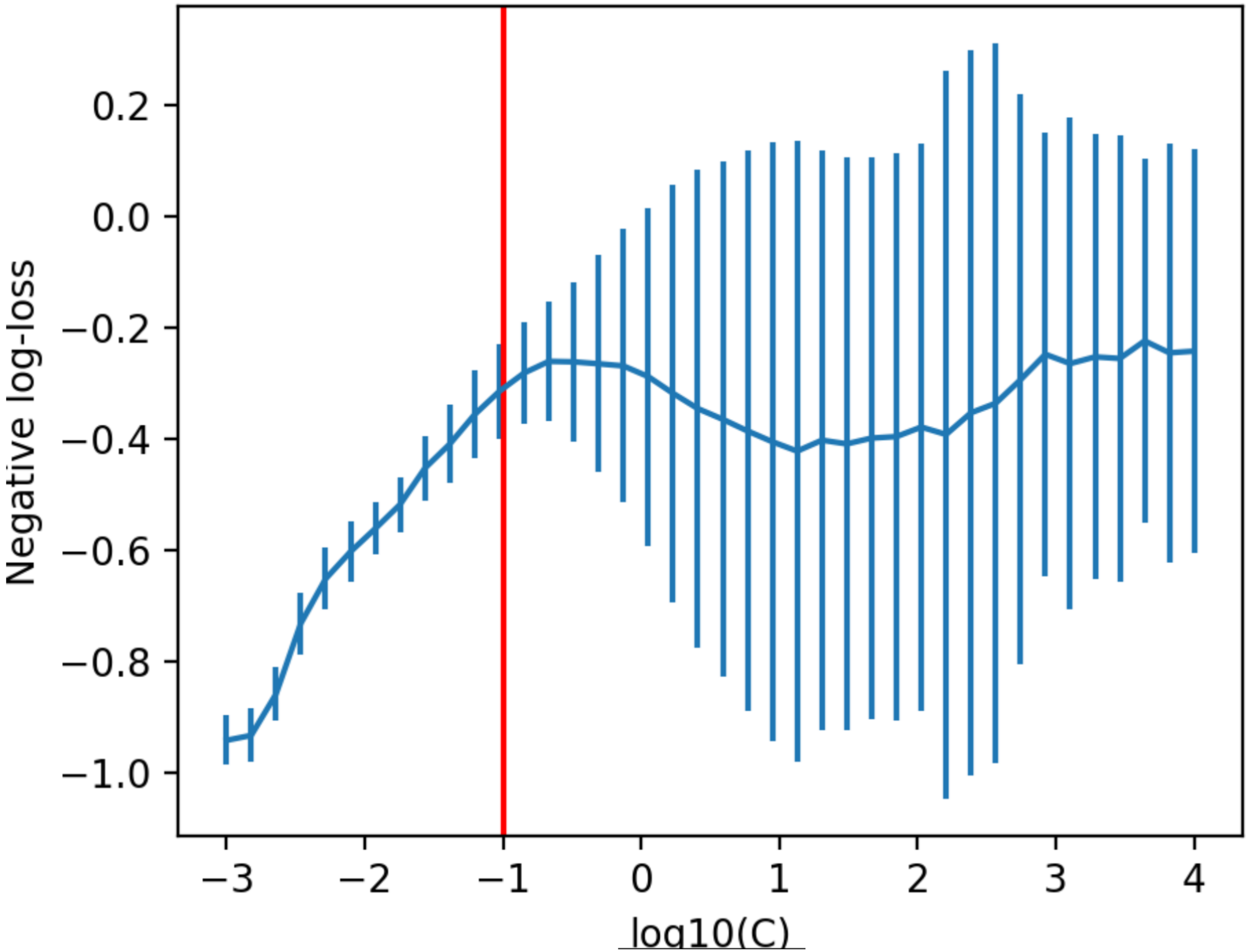
Точность или устойчивость



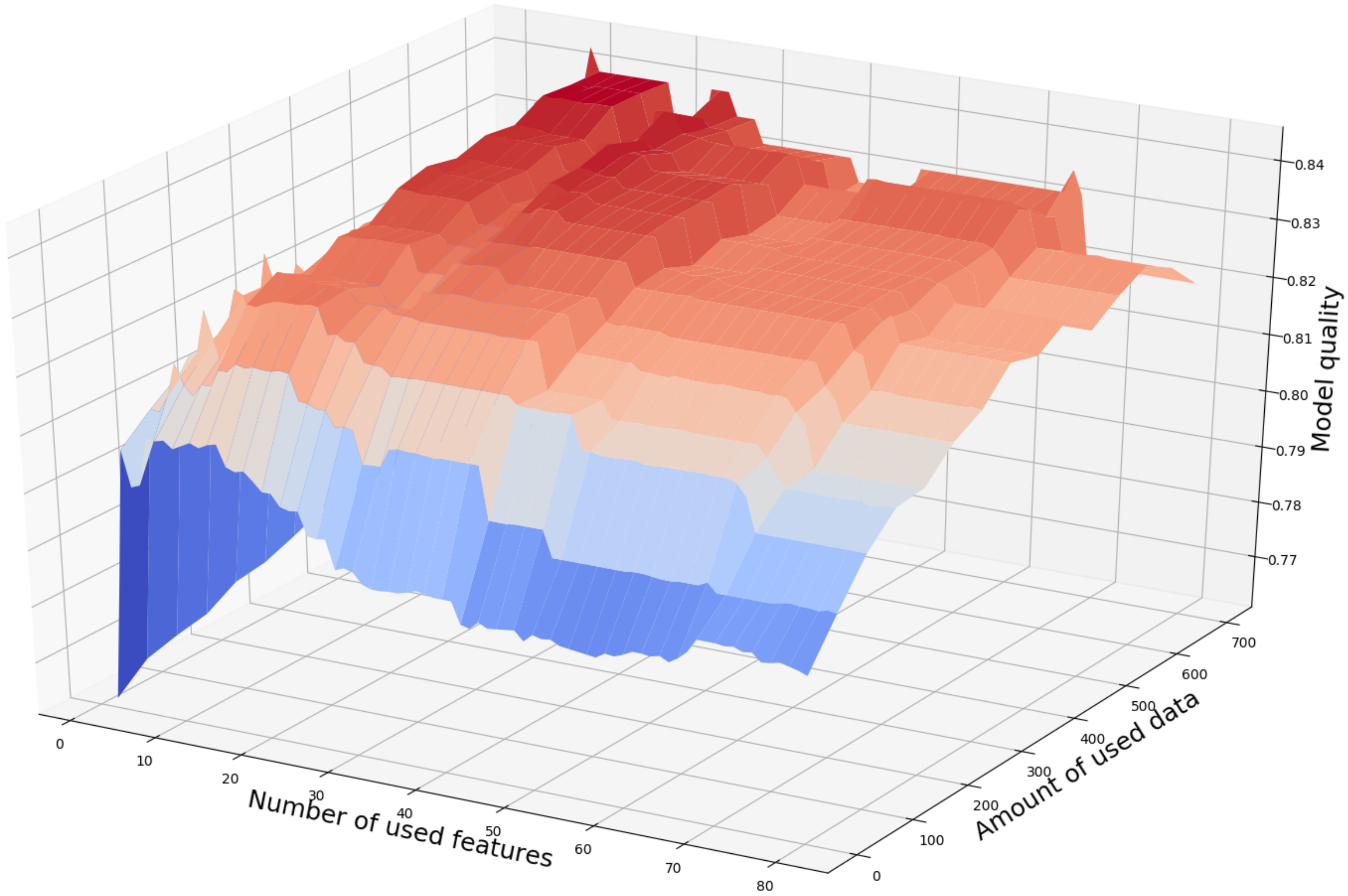
– Ошибка и её дисперсия при пополнении выборки



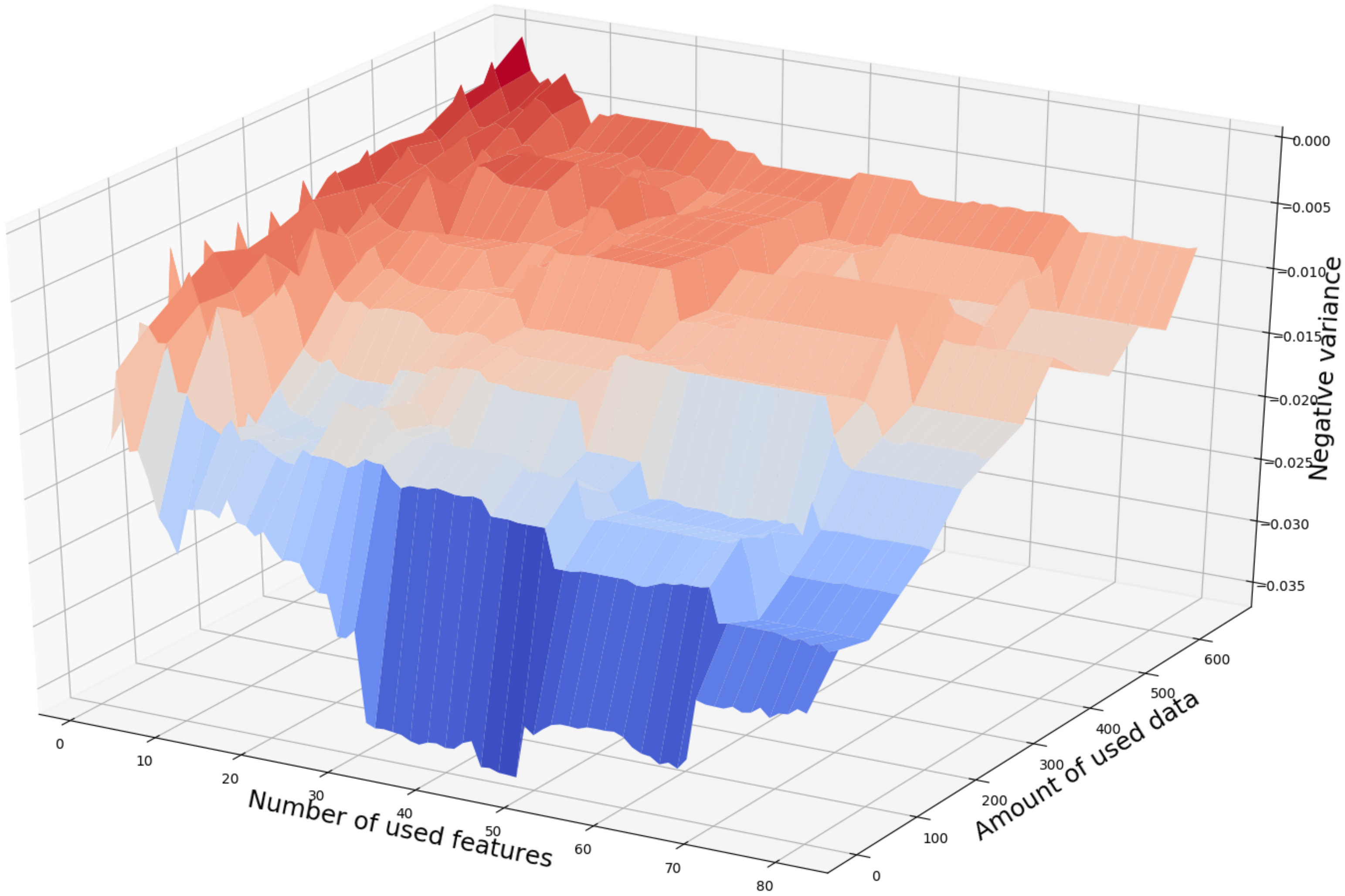
Дисперсия ошибки при повышении сложности модели



– Ошибка при различных объемах выборки



– Дисперсия ошибки при различных объемах выборки



Модель глубокого обучения

Определение

Моделью $\mathbf{f}(\mathbf{w}, \mathbf{x})$ назовем дифференцируемую по параметрам $\mathbf{w}$ функцию из множества признаковых описаний объекта во множество меток:

$$\mathbf{f} : \mathbb{X} \times \mathbb{W} \to \mathbb{Y},$$

где $\mathbb{W}$ — пространство параметров функции $\mathbf{f}$.

Особенность задачи выбора модели *глубокого обучения* — значительное число параметров моделей приводит к неприменимости ряда методов оптимизации и выбора структуры модели (AIC, BIC, кросс-валидация).

Модель определяется параметрами $\mathbf{W}$ и структурой $\mathbf{\Gamma}$.
Структура задает набор суперпозиций, входящих в модель и выбирается согласно статистическим критериям сложности модели.
Эмпирические оценки статистической сложности модели:

1. число параметров;
2. число суперпозиций, из которых состоит модель.

Выбор структуры: двуслойная нейросеть

Модель $\mathbf{f}$ задана **структурой** $\boldsymbol{\Gamma} = [\gamma^{0,1}, \gamma^{1,2}]$.

$$\text{Модель: } \mathbf{f}(\mathbf{x}) = \mathbf{softmax}\left((\mathbf{w}_0^{1,2})^\mathsf{T}\mathbf{f}_1(\mathbf{x})\right), \quad \mathbf{f}(\mathbf{x}) : \mathbb{R}^n \to [0,1]^{|\mathbb{Y}|}, \quad \mathbf{x} \in \mathbb{R}^n.$$

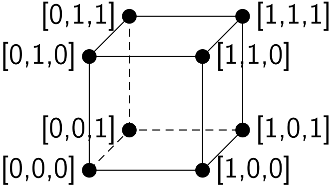
$$\mathbf{f}_1(\mathbf{x}) = \gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x}) + \gamma_1^{0,1}\mathbf{g}_1^{0,1}(\mathbf{x}),$$

где $\mathbf{w} = [\mathbf{w}_0^{0,1}, \mathbf{w}_1^{0,1}, \mathbf{w}_0^{1,2}]^\mathsf{T}$ — матрицы параметров, $\{\mathbf{g}_{0,1}^0, \mathbf{g}_{0,1}^1, \mathbf{g}_{1,2}^0\}$ — обобщенно-линейные функции скрытых слоев нейросети.

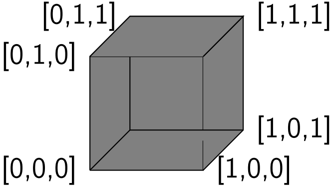
$$\begin{tikzcd}[column sep=huge]
\mathbf{f}_0(\mathbf{x}) = \mathbf{x} \arrow[r, bend left, "\gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x}) = \gamma_0^{0,1}\boldsymbol{\sigma}\left((\mathbf{w}_0^{0,1})^\mathsf{T}\mathbf{x}\right)"] \arrow[r, bend right, "\gamma_1^{0,1}\mathbf{g}_1^{0,1}(\mathbf{x}) = \gamma_1^{0,1}\boldsymbol{\sigma}\left((\mathbf{w}_1^{0,1})^\mathsf{T}\mathbf{x}\right)"'] & \mathbf{f}_1(\mathbf{x}) \arrow[r, "\gamma_0^{1,2}\mathbf{g}_0^{1,2}(\mathbf{x}) = \gamma_0^{1,2}\mathbf{softmax}\left((\mathbf{w}_0^{1,2})^\mathsf{T}\mathbf{x}\right)"] & \mathbf{f}_2(\mathbf{x})
\end{tikzcd}$$

Ограничения на структурные параметры

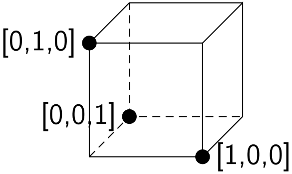
Примеры ограничений для одного структурного параметра $\gamma, |\gamma| = 3$.



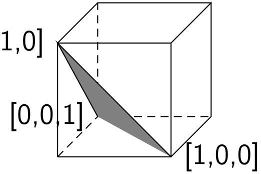
На вершинах куба



Внутри куба

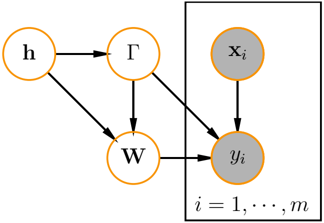


На вершинах симплекса



Внутри симплекса

Априорное распределение параметров



Определение

Априорным распределением параметров $\mathbf{w}$ и структуры $\mathbf{\Gamma}$ модели $\mathbf{f}$ назовем вероятностное распределение $p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{h}, \mathbf{f}) : \mathbb{W} \times \mathbb{\Gamma} \times \mathbb{H} \to \mathbb{R}^+$, где $\mathbb{W}$ — множество значений параметров модели, $\mathbb{\Gamma}$ — множество значений структуры модели.

Определение

Гиперпараметрами $\mathbf{h} \in \mathbb{H}$ модели назовем параметры распределения $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \mathbf{f})$ (параметры распределения параметров модели $\mathbf{f}$).

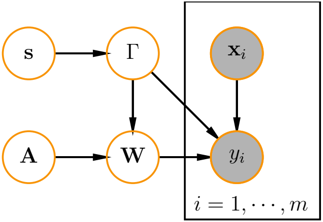
Модель $\mathbf{f}$ задается следующими величинами:

- **Параметры** $\mathbf{w} \in \mathbb{W}$ задают суперпозиции $\mathbf{f}_v$, из которых состоит модель $\mathbf{f}$.
- **Структурные параметры** $\mathbf{\Gamma} = \{\gamma^{j,k}\}_{(j,k)\in E} \in \mathbb{\Gamma}$ задают вклад суперпозиций $\mathbf{f}_v$ в модель $\mathbf{f}$.
- **Гиперпараметры** $\mathbf{h} \in \mathbb{H}$ задают распределение параметров и структурных параметров модели.
- **Метапараметры** $\boldsymbol{\lambda} \in \mathbb{\Lambda}$ задают вид оптимизации модели.

Байесовский выбор модели

Базовая модель:

- **параметры** модели $\mathbf{w} \sim \mathcal{N}(0, \alpha^{-1})$,
- **гиперпараметры** модели $\mathbf{h} = [\alpha]$.



Предлагаемая модель:

- **параметры** модели $\mathbf{w}_r^{j,k} \sim \mathcal{N}(0, (\gamma_r^{j,k})^2(\mathbf{A}_r^{j,k})^{-1})$, $\mathbf{A}_r^{j,k}$ — диагональная матрица параметров, соответствующих базовых функций $\mathbf{g}_r^{j,k}$,
  $(\mathbf{A}_r^{j,k})^{-1} \sim \text{inv-gamma}(\lambda_1, \lambda_2)$,
- **структурные параметры** модели $\mathbf{\Gamma} = \{\gamma^{j,k}, (j, k) \in E\}$,
  $\gamma^{j,k} \sim \text{GS}(\mathbf{s}^{j,k}, \lambda_{\text{temp}})$,
- **гиперпараметры** модели $\mathbf{h} = [\text{diag}(\mathbf{A}), \mathbf{s}]$,
- **метапараметры** $\lambda_1, \lambda_2, \lambda_{\text{temp}}$.

Вариационная нижняя оценка обоснованности

Интеграл обоснованности невычислим аналитически.

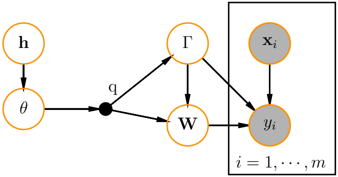
Обоснованность модели:

$$p(\mathbf{y}|\mathbf{X}, \lambda_{\text{temp}}, \mathbf{f}) = \iint\limits_{\mathbf{w},\mathbf{\Gamma}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{f}) p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) d\mathbf{w} d\mathbf{\Gamma}.$$

Определение

Вариационными параметрами модели $\boldsymbol{\theta} \in \mathbb{R}^u$ назовем параметры распределения q, приближающие апостериорное распределение параметров и структуры $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{X}, \mathbf{y}, \mathbf{h}, \mathbf{f}, \lambda_{\text{temp}})$:

$$q \approx \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{f}) p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})}{\iint\limits_{\mathbf{w}',\mathbf{\Gamma}'} p(\mathbf{y}|\mathbf{X}, \mathbf{w}', \mathbf{\Gamma}', \mathbf{f}) p(\mathbf{w}', \mathbf{\Gamma}'|\mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) d\mathbf{w}' d\mathbf{\Gamma}'}$$



Получим нижнюю оценку $\log \hat{p}(\mathbf{y}|\mathbf{X}, \lambda_{\text{temp}}, \mathbf{f})$ интеграла

$$\log p(\mathbf{y}|\mathbf{X}, \lambda_{\text{temp}}, \mathbf{f}) \geq \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{f}) - \mathsf{D}_{KL}\big(q(\mathbf{w}, \mathbf{\Gamma})||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})\big).$$

Она совпадает с интегралом обоснованности при

$$D_{\text{KL}}\big(q(\mathbf{w}, \mathbf{\Gamma})|p\left(\mathbf{w}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \lambda_{\text{temp}}, \mathbf{f}\right)\big) = 0.$$

Задача выбора модели

Зададим вариационное распределение $q = q_{\mathbf{w}} q_{\mathbf{\Gamma}}$ с параметрами $\boldsymbol{\theta}$, приближающие апостериорное распределение $p(\mathbf{w}, \mathbf{\Gamma} | \mathbf{X}, \mathbf{y}, \mathbf{h}, \mathbf{f})$ параметров и структуры.

Определение

Функцией потерь $L(\boldsymbol{\theta}|\mathbf{h}, \mathbf{X}, \mathbf{y}, \mathbf{f})$ назовем дифференцируемую функцию, качество модели на обучающей выборки при параметрах $\boldsymbol{\theta}$ распределения q.

Функцией валидации $Q(\mathbf{h}|\boldsymbol{\theta}, \mathbf{X}, \mathbf{y}, \mathbf{f})$ назовем дифференцируемую функцию, качество модели при векторе $\boldsymbol{\theta}$, заданном неявно.

Задачей выбора модели $\mathbf{f}$ назовем двухуровневую задачу оптимизации:

$$\mathbf{h}^* = \arg\max_{\mathbf{h} \in \mathbb{H}} Q(\mathbf{h}|\boldsymbol{\theta}^*, \mathbf{X}, \mathbf{y}, \mathbf{f}),$$

где $\boldsymbol{\theta}^*$ — решение задачи оптимизации

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \mathbb{U}} L(\boldsymbol{\theta}|\mathbf{h}, \mathbf{X}, \mathbf{y}, \mathbf{f}).$$

Обобщающая задача

Задачу выбора модели $\mathbf{h}^*, \boldsymbol{\theta}^*$ назовем обобщаюшей на множестве $U_\theta \times U_h \times U_\lambda \subset \mathbb{R}^u \times \mathbb{H} \times \mathbb{\Lambda}$, если выполнены условия:

1. Область параметров, гиперпараметров и метапараметров не является пустым или точкой.
2. Для каждого $\mathbf{h} \in U_h$ и каждого $\boldsymbol{\lambda} \in U_\lambda$ решение $\boldsymbol{\theta}^*$ определено однозначно.
3. **Критерий непрерывности:** $\mathbf{h}^*, \boldsymbol{\theta}^*$ непрерывны по метапараметрам.
4. **Критерий перехода между структурами:** существует константа $K_3 > 0$, такая, что существует хотя бы одна пара гиперпараметров $\mathbf{h}_1, \mathbf{h}_2$, и набор метапараметров $\boldsymbol{\lambda}$, такие, что для произвольных локальных оптимумов $\mathbf{h}_1, \mathbf{h}_2$ задачи оптимизации Q, полученных при метапараметрах $\boldsymbol{\lambda}$ и удовлетворяющих неравенствам

$$D_{\mathrm{KL}}\left(p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})|p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})\right) > K_3, D_{\mathrm{KL}}\left(p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})|p(\boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})\right) > K_3,$$

$$Q(\mathbf{h}_1|\boldsymbol{\lambda}) > Q(\mathbf{h}_2|\boldsymbol{\lambda}),$$

существует значение метапараметров $\boldsymbol{\lambda}' \neq \boldsymbol{\lambda}$, такое, что

   1. соответствие между вариационными параметрами $\boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\theta}^*(\mathbf{h}_2)$ сохраняется при $\boldsymbol{\lambda}'$,
   2. выполняется неравенство $Q(\mathbf{h}_1|\boldsymbol{\lambda}') < Q(\mathbf{h}_2|\boldsymbol{\lambda}')$.

Обобщающая задача

Задачу выбора модели $\mathbf{h}^*, \boldsymbol{\theta}^*$ назовем обобщающей на множестве $U_\theta \times U_h \times U_\lambda \subset \mathbb{R}^u \times \mathbb{H} \times \mathbb{\Lambda}$, если выполнены условия:

5. **Критерий максимизации правдоподобия выборки:** существует $\boldsymbol{\lambda} \in U_\lambda$ и $K_1 \in \mathbb{R}_+$, такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_1$: выполнено: $\mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_1), \lambda_{\text{temp}}, \mathbf{f}) > \log \mathsf{E}_q\ p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_2), \lambda_{\text{temp}}, \mathbf{f})$.
6. **Критерий минимизации параметрической сложности модели:** существует $\boldsymbol{\lambda} \in U_\lambda$ и $K_2 \in \mathbb{R}_+$, такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_2$, $\mathsf{E}_q \log p(\mathbf{y}|\boldsymbol{\theta}_1, \lambda_{\text{temp}}, \mathbf{f}) = \log \mathsf{E}_q\ p(\mathbf{y}|\boldsymbol{\theta}_2, \lambda_{\text{temp}}, \mathbf{f})$, сложность первой модели меньше, чем второй.
7. **Критерий максимизации обоснованности модели:** существует значение гиперпараметров $\boldsymbol{\lambda}$, такое что оптимизация задачи эквивалента оптимизации вариационной оценки обоснованности модели:
$$\mathbf{h}^* \propto \arg\max p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) p(\mathbf{h}|\boldsymbol{\lambda}), \quad \boldsymbol{\theta}^* = \arg\min D_{\text{KL}}(q|p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \lambda_{\text{temp}}, \mathbf{f})).$$

Анализ задач выбора моделей

Теорема [Бахтеев, 2019]

Следующие задачи выбора модели не являются обобщающими:

1. критерий максимума правдоподобия: $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f})$;
2. критерий максимума апостериорной вероятности $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathbf{f}) p(\boldsymbol{\theta}|\mathbf{h}, \lambda_{\text{temp}})$;
3. метод максимума вариационной оценки обоснованности модели $\max_{\mathbf{h}} \max_{\boldsymbol{\theta}} \mathsf{E}_q \log\ p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}, \mathbf{f}) - \mathsf{D}_{KL}\big(q(\mathbf{w}, \boldsymbol{\Gamma})||p(\mathbf{w}, \boldsymbol{\Gamma}, \lambda_{\text{temp}})\big) + \log\ p(\mathbf{h}|\mathbf{f})$;
4. кросс-валидация $\max_{\mathbf{h}} \mathsf{E}_q \log p(\mathbf{y}_{\text{valid}}|\mathbf{X}_{\text{valid}}, \boldsymbol{\theta}^*, \lambda_{\text{temp}}, \mathbf{f})$, $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f}) p(\boldsymbol{\theta}|\mathbf{h})$.
5. AIC: $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f}) - |\theta_i : \mathsf{D}_{\text{KL}}\left(q(w_i)|p(w_i|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}\right) < \lambda|$;
6. BIC: $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f}) - \frac{1}{2}\log(|\mathbb{W}||\theta_i : \mathsf{D}_{\text{KL}}\left(q(w_i)|p(w_i|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}\right) < \lambda|$;
7. перебор структуры модели: $\max \boldsymbol{\Gamma}' \max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f}) \mathbb{I}(q(\boldsymbol{\Gamma}\Gamma = p')$, где p' — распределение на структуре.

Предлагаемая задача оптимизации

Теорема [Бахтеев, 2018]

Пусть функции потерь и валидации L, Q являются непрерывно-дифференцируемыми на компакте U. Тогда следующая задача является обобщающей на U.

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q = \tag{Q^*}$$

$$= \lambda^{\mathbf{Q}}_{\text{likelihood}} \mathsf{E}_{q^*} \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) -$$

$$-{}^{\text{prior}}_{\mathbf{Q}} \mathrm{D}_{KL}\big(q^*(\mathbf{w}, \mathbf{\Gamma}) || p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})\big) -$$

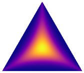
$$- \sum_{p' \in \mathfrak{P}, \lambda \in \boldsymbol{\lambda}^{\text{struct}}_{\mathbf{Q}}} \lambda \mathrm{D}_{KL}(\mathbf{\Gamma}|p') + \log p(\mathbf{h}|\mathbf{f}),$$

где

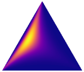
$$q^* = \arg\max_{q} L = \mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) \tag{L^*}$$

$$-{}^{\text{prior}}_{\mathbf{L}} \mathrm{D}_{KL}\big(q^*(\mathbf{w}, \mathbf{\Gamma}) || p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})\big).$$

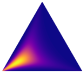
Оптимизационная задача обобщает алгоритмы оптимизации: оптимизация правдоподобия и обоснованности, последовательное увеличение и снижение сложности модели, полный перебор структуры.



$\boldsymbol{\lambda}^{Q}_{\text{struct}} = [0; 0; 0].$



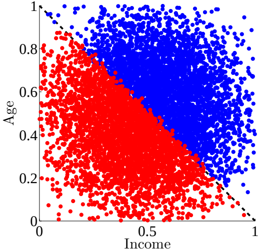
$\boldsymbol{\lambda}^{Q}_{\text{struct}} = [1; 0; 0].$



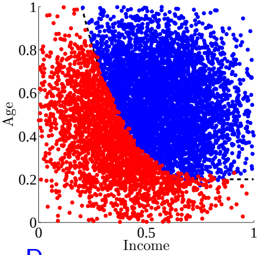
$\boldsymbol{\lambda}^{Q}_{\text{struct}} = [1; 1; 0].$

Мультимоделирование в задачах классификации

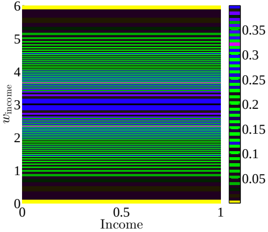
Проблема: выборка $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I} = \overline{1, m}$, $\mathbf{x}_i \in \mathbb{X}$, $y_i \in \mathbb{Y}$ не соответствует гипотезе порождения данных из одиночной модели.



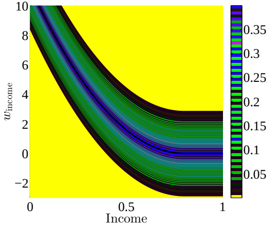
Модель порождения данных



Реальные данные



Модельное апостериорное распределение $p(\mathbf{w}|\mathbf{X},\ \mathbf{y})$

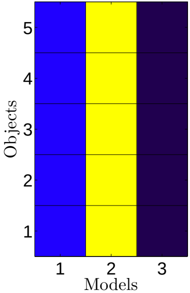


Реальное апостериорное распределение $p(\mathbf{w}|\mathbf{X},\ \mathbf{y})$

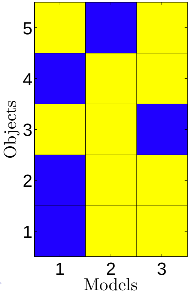
Определение 1. Смесь регрессионных моделей — регрессионная модель вида

$$f = \sum_{k=1}^{K} \pi_k f_k(\mathbf{w}_k), \text{ где}$$

$$\sum_{k=1}^{K} \pi_k = 1, \; \pi_k \geq 0.$$

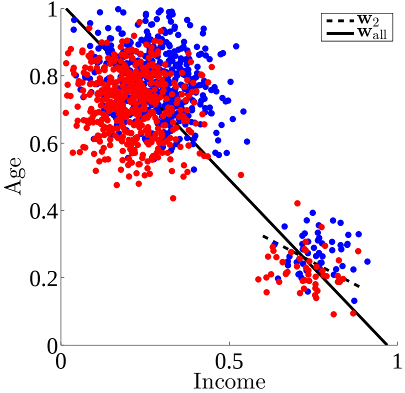


Определение 2. Многоуровневая регрессионная модель — набор регрессионных моделей f_k, $k = 1, \ldots, K$ такой, что при разбиении множества индексов объектов $\mathcal{I} = \sqcup_{k=1}^{K} \mathcal{I}_k$ для всех объектов с индексами из $\mathcal{I}_k$ используется модель f_k.

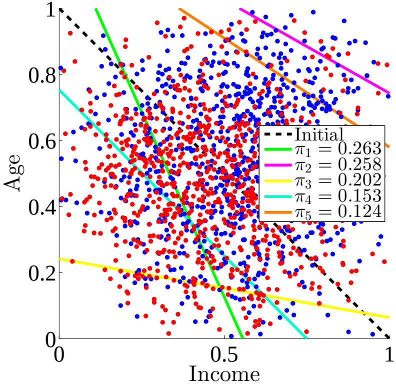


Близость моделей в мультимодели

Проблема: большое число близких или совпадающих моделей ведет к неинтерпретируемости и низкому качеству прогноза.



Неадекватная многоуровневая модель



Неадекватная смесь моделей

Определение 3. Мультимодель с совместным распределением $p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ (\boldsymbol{\pi})|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ (\mu))$ называется $(s,\ \alpha)$-**адекватной**, если модели, ее составляющие, являются попарно статистически различимыми с помощью функции сходства s на уровне значимости α.

Обучение мультимодели

$$[\mathbf{w}_1^*, \ldots, \mathbf{w}_K^*, (\boldsymbol{\pi}^*)] = \operatorname*{arg\,max}_{\mathbf{w}_1, \ldots, \mathbf{w}_K, (\boldsymbol{\pi})} p(\mathbf{w}_1, \ldots, \mathbf{w}_K, (\boldsymbol{\pi})|\mathbf{X},\ \mathbf{y},\ \mathbf{A}_1, \ldots, \mathbf{A}_K, (\mu)).$$

Определение 4. Мультимодель называется **оптимальной**, если она обладает наибольшей обоснованностью

$$[\mathbf{A}_1^*, \ldots,\ \mathbf{A}_K^*,\ (\mu^*)] = \operatorname*{arg\,max}_{\mathbf{A}_1, \ldots,\ \mathbf{A}_K,\ (\mu)} p\big(\mathbf{y}|\mathbf{X},\ \mathbf{A}_1, \ldots,\ \mathbf{A}_K,\ (\mu)\big) =$$

$$\operatorname*{arg\,max}_{\mathbf{A}_1, \ldots, \mathbf{A}_K, (\mu)} \int p\big(\mathbf{y}, \mathbf{w}_1, \ldots, \mathbf{w}_K, (\boldsymbol{\pi})|\mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K, (\mu)\big) d\mathbf{w}_1 \ldots d\mathbf{w}_K (d\boldsymbol{\pi}),$$

где $\mathbf{A}_1 \in Q_{\mathbf{A}_1},\ \ldots,\ \mathbf{A}_K \in Q_{\mathbf{A}_K},\ (\mu \in Q_\mu)$.

Проблема

Несмотря на прореживание мультимодели, она может не являться (s, α) – адекватной, то есть может содержать похожие модели.

Дано

- Две модели f_1 и f_2, векторы параметров моделей $\mathbf{w}_1$, $\mathbf{w}_2$.
- Выборки $(\mathbf{X}_1,\ \mathbf{y}_1)$ и $(\mathbf{X}_2,\ \mathbf{y}_2)$,
  $y_{1,i} = f_1(\mathbf{x}_{1,i},\ \mathbf{w}_1),\ \ y_{2,i} = f_2(\mathbf{x}_{2,i},\ \mathbf{w}_2)$.
- Априорные распределения параметров моделей
  $\mathbf{w}_1 \sim p_1(\mathbf{w}),\ \mathbf{w}_2 \sim p_2(\mathbf{w})$.
- Апостериорные распределения $p(\mathbf{w}_1|\mathbf{X}_1,\ \mathbf{y}_1)$ и $p(\mathbf{w}_2|\mathbf{X}_2,\ \mathbf{y}_2)$, обозначаемые далее $g_1(\mathbf{w})$ и $g_2(\mathbf{w})$.

Требуется: построить функцию сходства, определенную на паре распределений $g_1(\mathbf{w})$ и $g_2(\mathbf{w})$, удовлетворяющую ряду требований.

Корректная функция сходства s должна быть

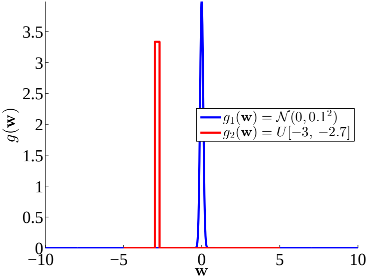
1. определена в случае несовпадения носителей,
2. $s(g_1,\ g_2) \leq s(g_1,\ g_1)$,
3. $s \in [0,\ 1]$,
4. $s(g_1,\ g_1) = 1$,
5. близка к 1, если $g_2(\mathbf{w})$ — малоинформативное распределение,
6. симметрична, $s(g_1,\ g_2) = s(g_2,\ g_1)$.

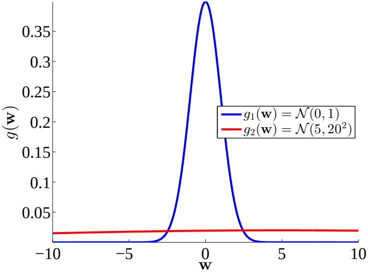
Теорема 3 (Адуенко, 2014)

Функции сходства, порожденные расстояниями Кульбака-Лейблера, Дженсона-Шеннона, Хеллингера, Бхаттачарайа,
не являются корректными.

Важно, чтобы значение функции s

было близко к 1, если $g_2(\mathbf{w})$ — малоинформативное распределение.





Теорема 4 (Адуенко, 2014)

Функции сходства, порожденные дивергенциями Брегмана, симметризованными дивергенциями Брегмана и f-дивергенциями, не являются корректными.

Предлагаемая функция сходства

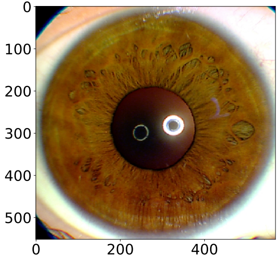
В качестве меры сходства распределения предлагается мера сходства s-score:

$$s(g_1,\ g_2) = \frac{\int_{\mathbf{w}} g_1(\mathbf{w})g_2(\mathbf{w})d\mathbf{w}}{\max_{\mathbf{b}} \int_{\mathbf{w}} g_1(\mathbf{w}-\mathbf{b})g_2(\mathbf{w})d\mathbf{w}}.$$

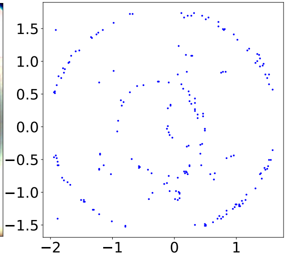
Теорема 5 (Адуенко, 2014). Предлагаемая функция сходства является корректной.

Примеры:

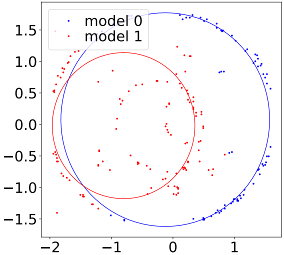
$g_1(\mathbf{w})$	$g_2(\mathbf{w})$	$s(g_1,\ g_2)$
$U[0,\ 1]$	$U[0.5,\ 1.5]$	0.5
$U[0,\ 1]$	$U[0,\ 1]$	1
$\mathcal{N}(0,\ 1)$	$\mathcal{N}(10, 10^{10})$	1



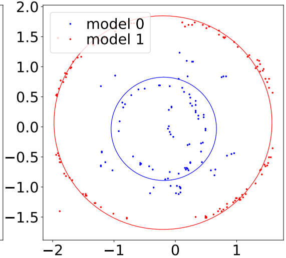
(a) Исходное изображение



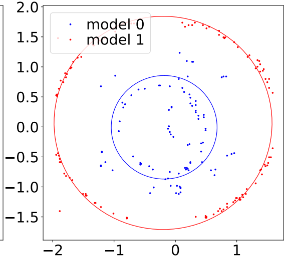
(b) Бинаризованное изображение



(c) Без задания априорного распределения



(d) С заданием априорного распределения на модели



(e) Регуляризация априорных распределений

Задана выборка:

$$\mathbf{X} \in \mathbb{R}^{N \times n},$$

где N — число объектов в выборке, а n — размерность признакового пространства.

Definition

Смесь экспертов — мультимодель, определяющая правдоподобие веса π_k каждой локальной модели $\mathbf{f}_k$ на признаковом описании объекта $\mathbf{x}$.

$$\hat{\mathbf{f}} = \sum_{k=1}^{K} \pi_k \mathbf{f}_k, \qquad \pi_k(\mathbf{x}, \mathbf{V}) : \mathbb{R}^{n \times |\mathbf{V}|} \to [0, 1], \qquad \sum_{k=1}^{K} \pi_k(\mathbf{x}, \mathbf{V}) = 1,$$

где $\hat{\mathbf{f}}$ — мультимодель, а $\mathbf{f}_k$ является локальной моделью, π_k — шлюзовая функция, $\mathbf{w}_k$ — параметры k-й локальной модели, $\mathbf{V}$ — параметры шлюзовой функции.

В качестве локальных моделей $\mathbf{f}_k$ и шлюзовой функции $\boldsymbol{\pi}$ рассматриваются следующие функции:

$$\mathbf{f}_k(\mathbf{x}) = \mathbf{w}_k^{\mathsf{T}} \mathbf{x}, \quad \boldsymbol{\pi}(\mathbf{x}, \mathbf{V}) = \mathrm{softmax}\big(\mathbf{V}_1^{\mathsf{T}} \boldsymbol{\sigma}\big(\mathbf{V}_2^{\mathsf{T}} \mathbf{x}\big)\big),$$

где $\mathbf{V} = \big\{\mathbf{V}_1, \mathbf{V}_2\big\}$ — параметры шлюзовой функции.

Параметры локальных моделей оптимизируются согласно принципу максимального правдоподобия модели:

$$p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}\bigr) = \prod_{k=1}^{K} p^{k}\left(\mathbf{w}_k\right) \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k p_k\left(y_i|\mathbf{w}_k, \mathbf{x}_i\right)\right),$$

где $\mathbf{W} = \bigl[\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K\bigr]^{\mathsf{T}}$.
Задача оптимизации параметров локальных моделей и параметров смеси:

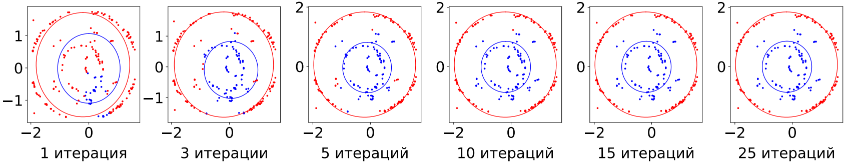
$$\hat{\mathbf{W}}, \hat{\mathbf{V}} = \arg\max_{\mathbf{W}, \mathbf{V}} p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}\bigr).$$

Рассматривается вероятностная постановка задачи:

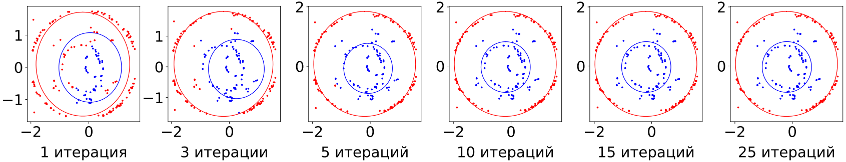
1) правдоподобие выборки $p_k\left(y_i|\mathbf{w}_k, \mathbf{x}_i\right) = \mathcal{N}\left(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\right)$, где β уровень шума,
2) априорное распределение параметров $p^k\left(\mathbf{w}_k\right) = \mathcal{N}\left(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\right)$, где $\mathbf{w}_k^0$ — вектор размера $n \times 1$, $\mathbf{A}_k$ — ковариационная матрица параметров,
3) регуляризация априорного распределения $p\left(\boldsymbol{\varepsilon}_{k,k'}|\boldsymbol{\alpha}\right) = \mathcal{N}\left(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi}\right)$, где $\boldsymbol{\Xi}$ — ковариационная матрица общего вида, $\boldsymbol{\varepsilon}_{k,k'} = \mathbf{w}_k^0 - \mathbf{w}_{k'}^0$.

Обучения на реальных данных

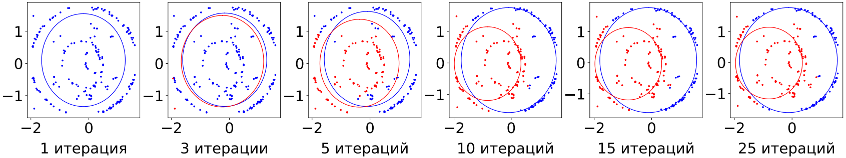
Регуляризация априорных распределений

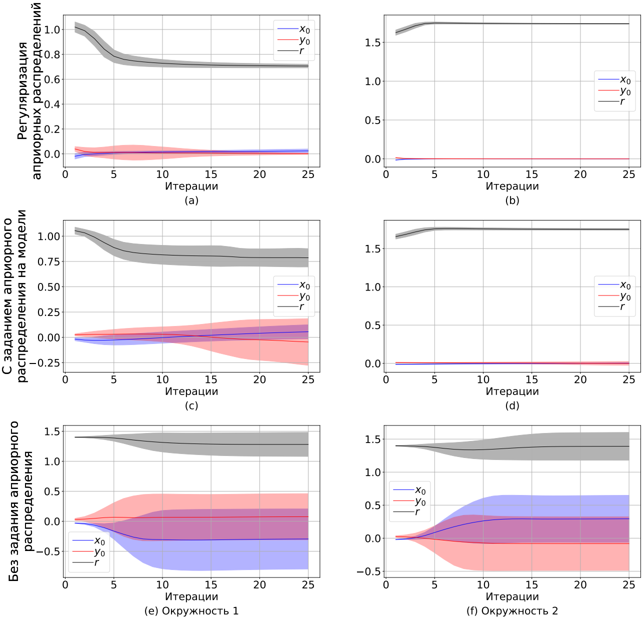


С заданием априорного распределения на модели



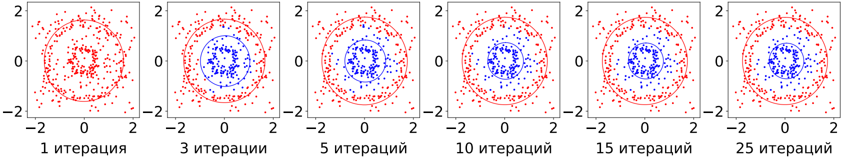
Без задания априорного распределения



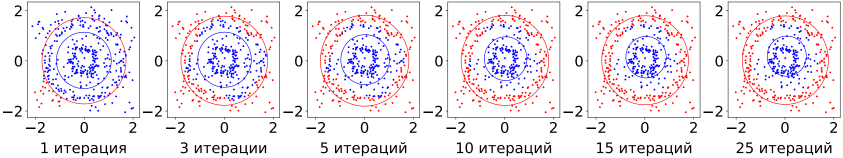

Регуляризация априорных распределений
С заданием априорного распределения на модели
Без задания априорного распределения
x_0
y_0
r
Итерации
(a)
(b)
(c)
(d)
(e) Окружность 1
(f) Окружность 2

Обучения на синтетических данных

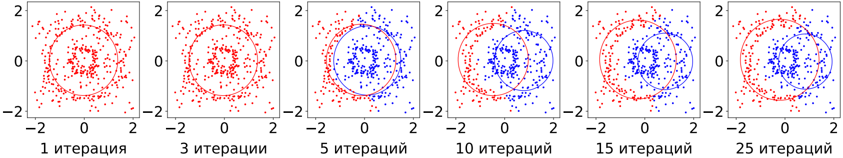
Регуляризация априорных распределений



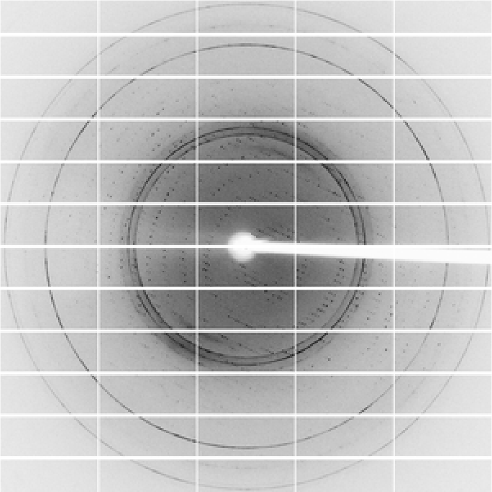
С заданием априорного распределения на модели



Без задания априорного распределения



Планы: смесь локальных моделей в кристаллографии



Crystal structure of a SusD homolog at 2.00 Å resolution

Выбор моделей и мультимоделирование

- Обобщен ряд методов выбора моделей с использованием байесовского подхода.
- Построена смесь моделей с разородными носителями функции распределения параметров.
- Построена смесь экспертов с пространствами параметров малой размерности.

Планируется развивать методы байесовского выбора разнородных моделей в задачах теоретической физики

Спасибо преподавателям Кафедры интеллектуальных систем МФТИ:
А.А. Адуенко, О.Ю. Бахтееву, Р.В. Исаченко, О.В. Грабовому