Сравнение жадных эвристик для решения задачи выбора оптимального подмножества

А.Р. Валеев, А.И. Архипов, Т.В. Кузнецова

Московский физико-технический институт, Долгопрудный

Abstract

В данной работе исследуется *задача выбора оптимального k-подмножества* признаков, с целью получить наилучшее линейное приближение предсказываемой случайной величины. Также будет рассмотрена более общая *задача выбора оптимального словаря*. Такие задачи часто рассматриваются в контексте проблем отбора признаков и разреженной линейной регрессии. Опираясь на предыдущие работы [1, 2], в которых исследовались методы оптимизации субмодулярных функций, мы изучаем работу *жадных алгоритмов* на описанных выше несубмодулярных задачах, для которых приводятся новые оценки сходимости и точности, подтвержденные экспериментами на синтетических и реальных данных.

Keywords: sparse feature selection, subset selection, dictionary selection, submodularity ratio, greedy algorightm, sparse linear regression

1. Введение

В этой работе рассматривается $NP-$трудная [3] задача выбора оптимального подмножества случайных величин из множества признаков для наилучшей линейной аппроксимации целевой случайной величины. В качестве целевой максимизируемой функции рассматривается *коэффициент детерминации* R^2. Формальные определения и постановка задачи приведены в секциях *теоретического введения* и *постановки задачи*.

Задачу выбора оптимального словаря можно описать следующим образом. Даны нормированные ковариации между случайными величинами $X_1, \ldots, X_n$, из которых выбирается оптимальное k-подмножество, и величиной Z, которую требуется аппроксимировать. Задача состоит в выборе подмножества из $k \ll n$ величин X_i, такого что для линейной

комбинации выбранных $X_{i_1}, \ldots, X_{i_k}$ достигается максимальное значение R^2.

Данная формулировка эквивалентна [4] *задаче разреженной аппроксимации по словарным векторам*, которая формулируется следующим образом. Дано множество из n векторов $\mathbf{x}_i \in \mathbf{R}^m$, также целевой вектор $\mathbf{z} \in \mathbf{R}^m$. Требуется выбрать не более k векторов из словаря так, чтобы их линейная комбинация наилучшим образом аппроксимировала $\mathbf{z}$.

Комбинаторное решение такой задачи требует больших вычислительных мощностей [3]. Однако, как показывает опыт [5, 2], существуют эффективные неточные методы решения таких задач. Именно такие методы являются предметом изучения в данной работе. Будут даны теоретические оценки их сходимости и оценка точности по сравнению с результатом, полученным полным перебором.

1.1. Почему это важно?

Данная задача широко применяется задач машинного обучения, таких как отбор признаков, выбор оптимального словаря (Dictionary Selection) [6] и обработка сигналов (Compressed Sensing) [7]. В контексте машинного обучения под оптимальным подмножеством может подразумеваться некоторая часть признаков, по которой можно достаточно точно предсказывать целевое значение. Это позволяет не использовать признаки, не входящие в оптимальное подмножество, при предсказании целевой переменной, что может значительно уменьшить количество вычислений.

1.2. На чем мы фокусируемся

Более конкретно мы работаем с задачей выбора оптимального подмножества случайных величин, заданной мощности, из множества признаков при имеющейся матрице ковариаций. Проведены эксперименты на сгенерированных синтетических данных и датасете *Boston Housing* [8].

1.3. Ожидаемые результаты

В следующих разделах будет приведено сравнение между собой следующих *жадных* алгоритмов: прямая регрессия, метод ортогонального согласованного преследования, забывчивый жадный алгоритм. Ожидается, что результаты согласуются с теоретическими оценками на точность аппроксимации (по коэффициенту детерминации) и покажут выигрыш по времени *жадных* эвристик по сравнению с NP-перебором.

Обзор литературы. Как было сказано ранее, в общем случае комбинаторная задача выбора оптимального k-подмножества является NP-сложной [9]. Таким образом, не существует алгоритма, способного решить такую задачу без приближений за полиномиальное время для всех входных данных. По этой причине для решения чаще всего используются два подхода: жадные алгоритмы [10, 11, 12] и выпуклые релаксационные схемы [13, 14, 15]. Применительно к L_1-релаксационным схемам Tropp [16] показал условия, основанные на когерентности, то есть максимальной корреляции между любой парой величин, при которых гарантируется оптимальное восстановление разреженного сигнала, то есть достигается максимум R^2. Другие результаты [17] также подтверждают условия, при которых L_1-регуляризация достаточно точно восстанавливает разреженный входной сигнал. Однако вышеприведенные результаты не являются непосредственно применимыми к нашей формулировке выбора оптимального подмножества. Цель разреженного восстановления состоит в том, чтобы восстановить истинные коэффициенты разреженности входного сигнала, что отличается от нашей задачи минимизации погрешности предсказания произвольного сигнала при заданном условии уровня разреженности.

Для жадных алгоритмов, решающих задачу разреженного восстановление подбора признака, Lozano [18] (OMP) и Zhang [19] (FR) показали в своих работах, что прямая регрессия и обратная регрессия могут применяться для восстановления разреженного сигнала.

Das и Kempe [9], а также Gilbert [12] исследовали жадные алгоритмы с такой же постановкой задачи, как в нашем случае. Ими были получены результаты оценки на точность аппроксимации: $1 + \Theta\left(\mu^2 k\right)$ для среднеквадратической ошибки (MSE), $1 - \Theta(\mu k)$ для коэффициента детерминации R^2.

2. Теоретическое введение

2.1. О субмодулярных функциях

Определение 1. *Пусть Ω – конечное множество. Функция на множестве $f:\ 2^{\Omega} \to \mathbb{R}$ называется*

- *субмодулярной, если*

$$\forall\ A, B \in \Omega \hookrightarrow f(A \cup B) + f(A \cap B) \leqslant f(A) + f(B)\ ;$$

- *супермодулярной, если* $(-f)$ *– субмодулярная функция;*
- *модулярной, если она субмодулярная и супермодулярная.*

Приведем равносильное определение субмодулярной функции.

Определение 2. *Функция* $f: 2^{\Omega} \to \mathbb{R}$ *называется субмодулярной, если*

$$\forall\ X, Y \in \Omega : X \subseteq Y,\ \forall\ x \in (\Omega \setminus Y) \hookrightarrow$$

$$f(X \cup \{x\}) - f(X) \geqslant f(Y \cup \{x\}) - f(Y).$$

Проиллюстрируем определение субмодулярной функции на наглядном примере оптимизации производства [20]. Пусть имеется производство, на котором может изготавливаться некоторое конечное множество продуктов E. Для того чтобы начать производство нового продукта $e \in E$, необходимо установить новое оборудование. Затраты на его покупку c зависят от множества продуктов $S \subseteq E$, которые уже производятся. Заметим, что при увеличении множества S до множества T, затраты на производство нового продукта e потенциально могут уменьшится, что выражается следующей формулой:

$$\forall S \subset T \subset T \cup \{e\}, c(T \cup \{e\}) - c(T) \leq c(S \cup \{e\}) - c(S),$$

Пусть $p \in \mathbb{R}^E$, в предположении аддитивности p: $p(S)$ – фактическая выручка от производства множества товаров S. Тогда прибыль равна $p(S) - c(S)$, таким образом итоговая задача:

$$\max_{S \subseteq E}(p(S) - c(S)) \Leftrightarrow \min_{S \subseteq E}(c(S) - p(S)),$$

из определения 2, $c(S) - p(S)$ – субмодулярная функция, что приводит нас к задаче минимизации субмодулярной функции.

2.2. Коэффициент субмодулярности

Следующим шагом является введения понятия *коэффициента субмодулярности*. Это величина показывает насколько рассматриваемая функция близка к субмодулярной.

Определение 3 (Коэффициент субмодулярности)**.** *Пусть функция* $f : 2^{\Omega} \to \mathbb{R}$ *неотрицательна. Коэффициентом субмодулярности функции* f *по отношению к множеству* U *при параметре* $k \geqslant 1$ *называется соотношение*

$$\gamma_{U,k}(f) = \min_{\substack{L \subseteq U \\ S:\ |S| \leqslant k \\ S \cap L = \varnothing}} \frac{\sum\limits_{x \in S} f(L \cup \{x\}) - f(L)}{f(L \cup S) - f(L)} \ .$$

Таким образом, коэффициент субмодулярности показывает, как сильно может увеличиться функция f при добавлении к её аргументу любого подмножества $S : |S| \leqslant k$ по отношению суммарному приращению при добавлении отдельно каждого элемента $x \in S$.

3. Постановка задачи

Опишем следующую задачу выбора k случайных величин из множества *(Subset Selection problem)*. Требуется аппроксимировать некоторую интересующую нас величину Z, используя линейную регрессию на небольшом подмножестве исследуемых случайных величин $V = \{X_1, \ldots, X_n\}$.

Для величин $X_i \in V$ известны дисперсии $\mathbb{D}X_i = \mathrm{Cov}(X_i, X_i)$ и ковариации $\mathrm{Cov}(X_i, X_j)$. При соответствующей нормализации можно предположить, что все случайные величины имеют математическое ожидание 0 и дисперсию 1. Таким образом, может быть составлена матрица ковариаций C:

$$c_{ij} = \mathrm{Cov}(X_i, X_j).$$

Также обозначим вектор ковариаций Z и X_i как

$$\mathbf{b} = \{\mathrm{Cov}(Z, X_i)\}_i \ .$$

Запишем формальную постановку задачи Subset Selection следующим образом.

Задача 1 (Subset Selection problem)**.** *Пусть даны попарные ковариации между всеми величинами множества* V*, а также параметр* k*. Найти подмножество* $S \subseteq V$*, состоящее из не более, чем* k *элементов, а также линейную аппроксимацию* $Z' = \sum\limits_{i \in S} \alpha_i X_i$ *величины* Z*, максимизируя*

коэффициент детерминации

$$R^2_{Z,S} \doteq \frac{\mathbb{D}(Z) - \mathbb{E}\left[(Z - Z')^2\right]}{\mathbb{D}(Z)} \ .$$

Коэффициент детерминации R^2 широко используется в статистике. Его рассматривают как универсальную меру зависимости одной случайной величины от множества других. При предположении, что величина Z нормализована, так что дисперсия $\mathbb{D}Z = 1$, то запись можно упростить до следующего вида

$$R^2_{Z,S} = 1 - \mathbb{E}\left[(Z - Z')^2\right] \ .$$

Для подмножества S обозначим за C_S подматрицу C с множеством индексов строк и столбцов S, а также вектор $\mathbf{b}_s = \{b_i\}_{i \in S}$.

Полагаем, что матрица C_S не вырождена. Оптимальные коэффициенты регрессии:

$$\alpha_S = (\alpha_i)_{i \in S} = C_S^{-1} \mathbf{b}_S \ .$$

Таким образом, если даны C, $\mathbf{b}$, k, задача принимает следующий вид

$$\max_S R^2_{Z,S} = \max_S \mathbf{b}_S^T \left(C_S^{-1}\right) \mathbf{b}_S$$

$$\text{s.t.} \quad S \subseteq V$$

$$|S| \leqslant k$$

Задачу выбора k случайных величин можно обобщить на задачу *Dictionary Selection*, если вместо одной величины Z рассмотреть множество из s прогнозируемых величин $Z_1, \ldots, Z_s$. Другими словами, требуется выбрать набор D из d исследуемых величин $X_1, ..., X_d$, чтобы максимизировать общий коэффициент детерминации R^2 для Z_i, используя не более k векторов из D для каждого случая. Формально это записать можно следующим образом

Задача 2 (Dictionary Selection problem)**.** *Пусть заданы параметры d и k и известны попарные ковариации между $X_i \in V$ и Z_j. Найти такое множество $D \subseteq V : \ |D| \leqslant d$, которое будет максимизировать функцию*

$$F(D) = \sum_{j=1}^{s} \max_{S \subset D, |S| = k} R^2_{Z_j,S} \ .$$

Введем также понятие *вычета*

$$\mathrm{Res}(Z, S) = Z - \sum_{i \in S} \alpha_i X_i,$$

то есть той части вектора Z, которая не коррелирует с $X_i \; \forall i \in S$.

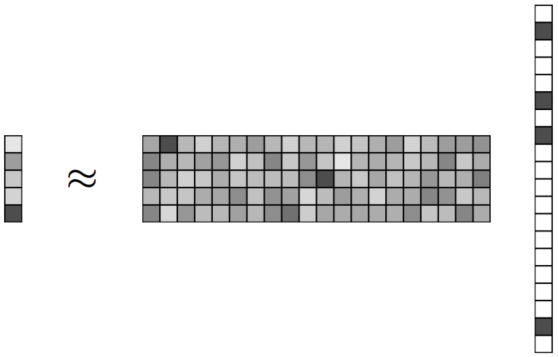


Рис. 1: Разреженная линейная регрессия $Z \approx X\alpha$

Определение 4. *L_1 – регуляризацией задачи линейной регрессии [21] называется добавление дополнительного слагаемого $\sum_{j=1}^{n} |w_j|$ к целевой функции, и исходная задача принимает вид:*

Среднеквадратичная ошибка (MSE) [21] может быть переписана в матричном представлении

$$Q(\mathbf{w}; X) = \frac{1}{m} \|X\mathbf{w} - \boldsymbol{y}\|^2$$

С учетом L_1-регуляризации

$$Q(\mathbf{w}; X) + \lambda \|\mathbf{w}\| \to \min_{\mathbf{w}}$$

Аналитическим решением будет являться

$$\mathbf{w}_* = \arg\min_{\mathbf{w}} \left(\frac{1}{m} \sum_{i=1}^{m} \left(\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i \right)^2 + \lambda \sum_{j=1}^{n} |w_j| \right)$$

4. Причем тут субмодулярность?

Стоит отметить, что в общем случае коэффициент детерминации $R^2_{Z,S}$ не является субмодулярной функцией. В то же время исследуемые нами алгоритмы были хорошо изучены применительно к оптимизации субмодулярных функций.

Именно по этой причине в рассмотрение вводится понятие коэффициента субмодулярности, с помощью которого можно получить новые оценки на точность аппроксимации несубмодулярной функции $R^2_{Z,S}$ алгоритмами прямой регрессии и ортогонального согласованного преследования.

Рассмотрим, каким будет коэффициент субмодулярности для функции R^2.

$$\gamma_{U,k} = \min_{\substack{L\subseteq U \\ S:|S|\leq k \\ S\cap L=\varnothing}} \frac{\sum_{i\in S}\left(R^2_{Z,L\cup\{X_i\}} - R^2_{Z,L}\right)}{R^2_{Z,S\cup L} - R^2_{Z,L}} = \min_{\substack{L\subseteq U \\ S:|S|\leq k \\ S\cap L=\varnothing}} \frac{\left(\mathbf{b}^L_S\right)^T \mathbf{b}^L_S}{\left(\mathbf{b}^L_S\right)^T \left(C^L_S\right)^{-1} \mathbf{b}^L_S},$$

где C^L и $\mathbf{b}^L$ – нормированные матрица ковариаций и вектор ковариаций для множества $\{\mathrm{Res}\left(X_1, L\right), \mathrm{Res}\left(X_2, L\right), \ldots, \mathrm{Res}\left(X_n, L\right)\}$.

Приведем следующую лемму для иллюстрации связи коэффициента субмодулярности с субмодулярностью функции.(Доказательство приведено в статье [9])

Лемма 1. *Функция f субмодулярна тогда и только тогда, когда*

$$\forall\ U, k \hookrightarrow \gamma_{U,k} \geqslant 1$$

Введем обозначение

$$\lambda_{\min}(C, k) = \min_{S:|S|=k} \lambda_{\min}\left(C_S\right)\ .$$

Сформулируем следующее утверждение, содержащее спектральную оценку снизу для величины $\gamma_{U,k}$ (доказательство приведено в статье [9]).

Лемма 2.

$$\gamma_{U,k} \geq \lambda_{\min}(C, k + |U|) \geq \lambda_{\min}(C)$$

Для того, чтобы понять, какие оценки на аппроксимацию R^2 существуют, нам необходимо разобраться в устройстве изучаемых алгоритмов.

Следующим шагом опишем принцип работы рассматриваемых в работе алгоритмов, после чего приведем теоретические оценки точности их работы.

5. Методы

В общем случае комбинаторная задача выбора оптимального k-подмножества является NP-сложной [9]. Таким образом, не существует алгоритма, способного решить такую задачу за полиномиальное время для всех входных данных. По этой причине для решения чаще всего используются два подхода: жадные алгоритмы [10, 11, 12] и выпуклые релаксационные схемы [13, 14, 15]. Для нашей формулировки недостаток методов выпуклой релаксации заключаются в том, что они не обеспечивают явного контроля над целевым уровнем разреженности k.

Более простой и интуитивно понятный подход, широко используемый на практике для решения задач выбора оптимального подмножества (например, он интегрирован во все пакеты коммерческой статистики), заключается в использовании жадных алгоритмов, которые итеративно добавляют (или удаляют) величины X_i в текущее оптимальное подмножество на основе меры соответствия их комбинации с Z.

В этой работе будут рассмотрены следующие жадные алгоритмы: Прямая регрессия (Forward Regression) [10], метод Ортогонального согласованного Преследования (Orthogonal Matching Pursuit)[11], Забывчивый жадный алгоритм (Oblivious greedy algorithm), также будет рассмотрена L_1 – регуляризация.

5.1. Forward Regression (Прямая регрессия)

Данный алгоритм для задачи *SubsetSelection* выбирает множество S размера k, искомое в задаче.

Algorithm 1 Forward Regression

Require: $k \in \mathbb{N}$

1: $i := 0$;
2: $S := \emptyset$;
3: **while** $i < k$ **do**
4: Выбирается случайная величина $X_m \in V \backslash S$ максимизирующая $R^2_{Z,S+\{X_m\}}$;
5: $S \leftarrow S \cup X_m$;
6: $i \leftarrow i + 1$;
7: **end while**

5.2. Orthogonal Matching Pursuit (Метод ортогонального согласованного преследования)

Данный алгоритм для задачи *SubsetSelection* выбирает множество S размера k, искомое в задаче.

Algorithm 2 Orthogonal Matching Pursuit

Require: $k \in \mathbb{N}$

1: $i := 0$;
2: $S := \emptyset$;
3: **while** $i < k$ **do**
4: Выбирается случайная величина $X_m \in V \backslash S$ максимизирующая $|Cov(Res(Z, S_i), X_m)|$;
5: $S \leftarrow S \cup X_m$;
6: $i \leftarrow i + 1$;
7: **end while**

5.3. Oblivious greedy algorithm

Выбирается k случайных величин X_i с наибольшими значениями $\mathbf{b_i}$.

Данный алгоритм является самым тривиальным из исследуемых, так как в его работе вообще не используется матрица ковариаций C.

Algorithm 3 Oblivious greedy algorithm

Require: $k \in \mathbb{N}$

$i := 0$;
$S := \emptyset$;
while do$i < k$
    Выбирается случайная величина $X_m \in V \backslash S$ максимизирующая $|b_i|$;
    $S \leftarrow S \cup X_m$;
    $i \leftarrow i + 1$;
end while

5.4. L_1 – регрессия

В описанном ранее *определении 4* подбирается такой минимальный коэффициент λ, чтобы не более чем k координат вектора α оказались по модулю больше некоторого малого заданного ε, соответствующие им признаки образуют искомое множество. Это делается для контроля количества значимых признаков, то есть размера искомого множества.

Algorithm 4 L_1 – регрессия

Require: $k \in \mathbb{N}$, $\alpha > 1$, $\varepsilon > 0$

$\lambda := \lambda_0$;
$S := V$;
while $|S| > k$ **do**
    Выбирается подмножество случайных величин $\{X_{n_m}\}_{m=1}^{k'} \subset V$ таких, что $|w_{n_m}| > \varepsilon$, где w - вектор весов в решении задачи минимизации MSE с L_1-регуляризацией с параметром λ;
    $S \leftarrow \{X_{n_m}\}_{m=1}^{k'}$;
    $\lambda \leftarrow \lambda \cdot \alpha$
end while

6. Анализ работы алгоритмов

В этой секции мы оценим сходимость описанных выше алгоритмов. Введем величину $OPT = \max_{S:|S|=k} R^2_{Z,S}$ для обозначения оптимального значения R^2, достигаемого для любого подмножества размера k.

6.1. Forward Regression (Метод прямой регрессии)

Для оптимального подмножества, выбранного в результате работы алгоритма прямой регрессии, введем обозначение S^{FR}. Основные результаты оценки работы метода приведены в следующей теореме.

Теорема 1 (Оценки на множество S^{FR}[9])**.** *Для подмножества S^{FR}, выбранного методом прямой регрессии, справедливы следующие оценки:*

$$\begin{aligned} R^2_{Z,S^{FR}} &\geq \left(1 - e^{-\gamma_{S^{FR},k}}\right) \cdot OPT \\ &\geq \left(1 - e^{-\lambda_{\min}(C,2k)}\right) \cdot OPT \\ &\geq \left(1 - e^{-\lambda_{\min}(C,k)}\right) \cdot \Theta\left(\left(\frac{1}{2}\right)^{1/\lambda_{\min}(C,k)}\right) \cdot OPT \end{aligned} \tag{1}$$

6.2. Orthogonal Matching Pursuit (Метод ортогонального согласованного преследования)

Следующий в рассмотрении алгоритм – метод ортогонального согласованного преследования, часто используемый в обработке сигналов. Аналогично прошлому пункту, введем обозначение S^{OMP}.

Перед введением основной теоремы для текущего метода рассмотрим следующую лемму.

Лемма 3. *Пусть $A \in \mathbb{R}^{(n+1)\times(n+1)}$ – матрица ковариаций между случайными величинами $Z, X_1, \ldots, X_n$. Тогда*

$$\mathbb{D}\left(\mathrm{Res}\left(Z, \{X_1, \ldots, X_n\}\right)\right) \geq \lambda_{\min}(A)$$

Доказательство. Напомним, что $\mathrm{Res}(Z, S) = Z - \sum\limits_{i \in S} \alpha_i X_i$. Обозначим $\mathbf{Y} = (Z, X_1, \ldots, X_n)^\top$Рассмотрим данную матрицу

$$A = \mathrm{cov}(\mathbf{Y}) = \mathbb{E}\left[\mathbf{Y}\mathbf{Y}^\top\right] - \mathbb{E}[\mathbf{Y}] \cdot \mathbb{E}\left[\mathbf{Y}^\top\right] = \begin{pmatrix} 1 & \mathbf{b}^T \\ \mathbf{b} & C \end{pmatrix}.$$

Для всякой матрицы M будем обозначать как $M[i, j]$ подматрицу, полученной из нее удалением i-ой строки и j-го столбца. Пользуясь этим обозначением, распишем детерминант матрицы A по первой строке:

$$
\begin{aligned}
\det(A) &= \sum_{j=1}^{n+1}(-1)^{1+j}a_{1,j}\det(A[1,j]) \\
&= \det(C) + \sum_{j=1}^{n}(-1)^{j}b_j\det(A[1,j+1]) \\
&= \det(C) + \sum_{j=1}^{n}(-1)^{j}b_j\sum_{i=1}^{n}(-1)^{i+1}b_i\det(C[i,j]) \\
&= \det(C) - \sum_{j=1}^{n}\sum_{i=1}^{n}(-1)^{i+j}b_ib_j\det(C[i,j]) \\
&= \det(C)\left(1-\mathbf{b}^TC^{-1}\mathbf{b}\right)
\end{aligned}
\tag{2}
$$

Таким образом, имеем $\dfrac{\det(A)}{\det(C)} = 1 - \mathbf{b}^TC^{-1}\mathbf{b}$.

Рассмотрим теперь дисперсию

$$
\mathbb{D}\operatorname{Res}(Z,\mathbf{X}) = \mathbb{D}(Z - \sum_{i\in S}\alpha_iX_i) = 1 - \mathbb{D}(\sum_{i\in S}\alpha_iX_i) = 1 - \mathbf{b}^TC^{-1}\mathbf{b} = \frac{\det(A)}{\det(C)}.
$$

Так как $\det(A) = \prod_{i=1}^{n+1}\lambda_i^A$ и $\det(C) = \prod_{i=1}^{n}\lambda_i^C$, а также, что по теореме Коши о переплетении [22] $\lambda_1^A \le \lambda_1^C \le \lambda_2^A \le \lambda_2^C \le \ldots \le \lambda_{n+1}^A$, получаем

$$
\mathbb{D}\operatorname{Res}(Z,\mathbf{X}) = \frac{\det(A)}{\det(C)} \ge \lambda_1^4,
$$

лемма доказана. □

Теперь сформулируем известные оценки точности метода OMP.

Теорема 2 (Оценки на множество S^{OMP} [9])**.** *Подмножество S^{OMP}, выбранное алгоритмом OMP, имеет следующие оценки аппроксимации коэффициента детерминации:*

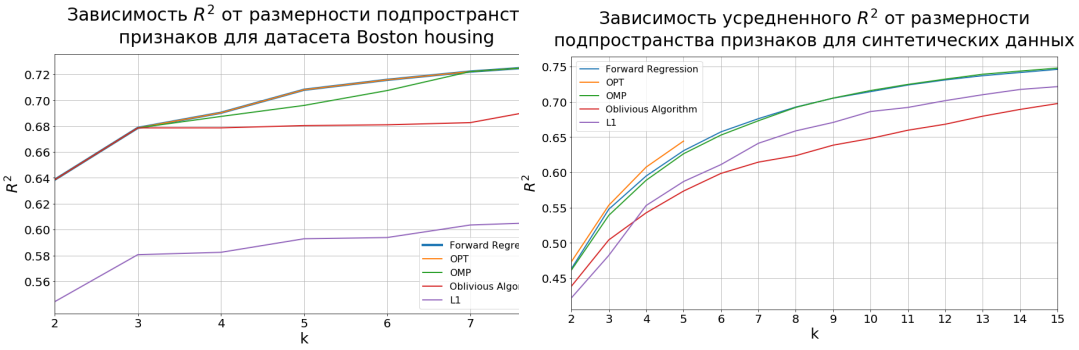
$$
\begin{aligned}
R^2_{Z,S^{OMP}} &\ge \left(1 - e^{-\left(\gamma_S Ou,k\cdot\lambda_{\min}(C,2k)\right)}\right)\cdot OPT \\
&\ge \left(1 - e^{-\lambda_{\min}(C,2k)^2}\right)\cdot OPT \\
&\ge \left(1 - e^{-\lambda_{\min}(C,k)^2}\right)\cdot \Theta\left(\left(\frac{1}{2}\right)^{1/\lambda_{\min}(C,k)}\right)\cdot OPT
\end{aligned}
\tag{3}
$$

6.3. Oblivious Greedy Algorithm

Теорема 3 (Оценка на множество $S^{Oblivious}$ [9]). *Пусть подмножество $S^{Oblivious}$ было выбрано в результате работы забывчивого жадного алгоритма. Тогда справедливы следующие оценки аппроксимации коэффициента детерминации:*

$$R^2_{Z,S^{OBL}} \geq \frac{\gamma_{0,k}}{\lambda_{\max}(C,k)} \cdot OPT \geq \frac{\lambda_{\min}(C,k)}{\lambda_{\max}(C,k)} \cdot OPT \tag{4}$$

7. Эксперименты



Исследуемые методы были проверены на двух наборах данных: *Boston Housing Dataset* и синтетических данных (X_i сгенерированы из многомерного нормального распределения с матрицей ковариации C, близкой к вырожденной, и $Z = \sum_{i=1}^{n} \alpha_i X_i + \varepsilon$, где ε - нормально распределенная случайная величина).

Из графиков видно, что наилучший результат аппроксимации показывает метод прямой регрессии. Зависимость $R^2_{S_{FR},Z}(k)$ наиболее близка к единице среди результатов, полученных остальными методами.

На реальных данных хуже всего себя показывает L_1 – регуляризация:

$$R^2_{S_{L_1},Z}(k) < 0.61 \ , \ \forall \ k \in \{1, \ldots, 8\}.$$

Самый простой из методов – забывчивый жадный алгоритм – логично восстанавливает самую грубую зависимость среди прочих жадных

методов. Особенно это характерно на живых данных, так как в данном случае заранее ничего не известно о связи между признаками.

Приведем результаты работы для некоторых k на синтетических и реальных данных в следующих таблицах:

Таблица значений R^2 на синтетических данных для разных алгоритмов при разных k					
k	Forward Regression	OMP	Oblivious Algorithm	L_1	OPT
2	0.45	0.44	0.42	0.40	0.47
5	0.63	0.62	0.56	0.58	0.64
10	0.71	0.71	0.65	0.68	-
15	0.74	0.74	0.69	0.72	-

Таблица значений R^2 на реальных данных для разных алгоритмов при разных k					
k	Forward Regression	OMP	Oblivious Algorithm	L_1	OPT
2	0.64	0.64	0.64	0.54	0.64
5	0.71	0.70	0.68	0.59	0.71
8	0.73	0.73	0.69	0.61	0.73

8. Выводы

- Методу прямой регрессии соответствует наибольшее приближение R^2 (причем на обоих видах данных), то есть этот метод восстанавливает наиболее сильную зависимость $Z' = Z'(X_{i_1}, \ldots, X_{i_k}),\ i_k \in S$.

- На реальных данных при $k \geqslant 7$ результат $R^2_{FR}(k) > 0.7$, что позволяет считать, что результат аппроксимации достаточно качественный.

- При работе на данных *Boston Housing* L_1-регуляризация (Lasso) демонстрирует наименьший результат $R^2(k)$ среди всех методов.

  Возможной причиной может быть то, что была реализована модификация алгоритма L_1-регуляризации, которая позволяет задать

конкретное количество k искомых признаков, хотя исходный алгоритм для этого не предназначен. Данный алгоритм увеличивает λ, пока нужное количество компонент вектора весов w не станут достаточно малы, что может привести к слишком большому значению λ и понижению точности минимизации целевой функции.

- Забывчивый жадный алгоритм (Oblivious greedy Algorithm) и L_1-регуляризация показывают результаты, существенно уступающие Прямой регрессии (Forward Regression) и Ортогональному согласованному преследованию (Orthogonal Matching Pursuit).

- Были экспериментально подтверждены теоретические гарантии аппроксимации (см. теоремы *1*, *2*, *3*).

9. Заключение

В данной статье были рассмотрены такие задачи, как *Выбор оптимального подмножества* и *Выбор оптимального словаря*. Для их решения были рассмотрены следующие жадные алгоритмы: Прямая регрессия (Forward Regression), Ортогональное согласованное преследование (Orthogonal Matching Pursuit), Забывчивый жадный алгоритм (Oblivious greedy Algorithm). Так же была рассмотрена L_1-регуляризация (Lasso Regression). С использованием субмодулярного анализа и спектральных свойств матрицы ковариаций были получены гарантии аппроксимации коэффициента детерминации для выбранного оптимального подмножества. Полученные оценки помогают понять, почему жадные алгоритмы решают задачу выбора подмножества эффективно даже с наличием коррелированных данных и подтверждаются экспериментально на реальных и синтетических данных.

10. Направление исследования в будущем

В дальнейших исследованиях планируется исследовать возможность решения схожей задачи с помощью эвристик, например Прямо-обратная регрессия (Forward-backward Regression), Поиск в ширину (Breadth-first Search), Поиск в глубину (Depth-first Search), и Генетические алгоритмы (Genetic Algorithm) [23]. Также необходимо протестировать работу методов на других метриках: в первую очередь с помощью логарифмической функции правдоподобия, AUC ROC [21].

Список литературы

[1] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

[2] László Lovász. Submodular functions and convexity. In *Mathematical Programming The State of the Art*, pages 235–257. Springer, 1983.

[3] Bin Chen and Guangri Quan. Np-hard problems of learning from examples. In *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, volume 2, pages 182–186. IEEE, 2008.

[4] Abhimanyu Das and David Kempe. Algorithms for subset selection in linear regression. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 45–54, 2008.

[5] Andrew An Bian, Joachim M Buhmann, Andreas Krause, and Sebastian Tschiatschek. Guarantees for greedy maximization of non-submodular functions with applications. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 498–507. JMLR. org, 2017.

[6] Andreas Krause and Volkan Cevher. Submodular dictionary selection for sparse representation. In *International Conference on Machine Learning (ICML)*, 2010.

[7] Mario Coutino, Sundeep Prabhakar Chepuri, and Geert Leus. Submodular sparse sensing for gaussian detection with correlated observations. *IEEE Transactions on Signal Processing*, 66(15):4025–4039, 2018.

[8] Data set from Kaggle competition boston housing. `https://www.kaggle.com/c/boston-housing`.

[9] Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.

[10] AJ Miller. Subset selection in regression, chapman and hall, crc. *Florida.*, 2002.

[11] Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.

[12] Anna C Gilbert, Shanmugavelayutham Muthukrishnan, and Martin J Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 243–252. Society for Industrial and Applied Mathematics, 2003.

[13] Guillaume Obozinski and Francis Bach. Convex relaxation for combinatorial penalties. *arXiv preprint arXiv:1205.1240*, 2012.

[14] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[15] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

[16] Joel A Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, 52(3):1030–1051, 2006.

[17] Shuheng Zhou. Thresholding procedures for high dimensional variable selection and statistical estimation. In *Advances in Neural Information Processing Systems*, pages 2304–2312, 2009.

[18] Grzegorz Swirszcz, Naoki Abe, and Aurelie C Lozano. Grouped orthogonal matching pursuit for variable selection and prediction. In *Advances in Neural Information Processing Systems*, pages 1150–1158, 2009.

[19] Tong Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10(Mar):555–568, 2009.

[20] S Thomas McCormick. Submodular function minimization. *Discrete Optimization*, 12:321–391, 2005.

[21] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons, 2012.

[22] M Seetharama Gowda and J Tao. The cauchy interlacing theorem in simple euclidean jordan algebras and some consequences. *Linear and Multilinear Algebra*, 59(1):65–86, 2011.

[23] КВ Воронцов. Лекции по методам оценивания и выбора моделей. *Режим доступа: http://www. Machi neLearning. ru*, 2007.