Анализ метода отбора признаков QPFS

Александр Адуенко

5 апреля 2021 г.

Рассматриваемый метод предложен в [1] и дается следующей оптимизационной задачей.

$$\begin{aligned} \frac{1}{2}(1-\alpha)\mathbf{a}^\intercal\mathbf{Q}\mathbf{a} - \alpha\mathbf{b}^\intercal\mathbf{a} &\to \min_{\mathbf{a}} \\ \text{s.t. } \mathbf{a} \geq 0,\ \sum_i a_i &= 1. \end{aligned} \tag{1}$$

Решение задачи (1) $\mathbf{a}^*$ определяет, какие признаки используются при построении модели. Признак j активен $\iff a_j > 0$.

Эта задача эквивалетна

$$\begin{aligned} \frac{1}{2}\mathbf{a}^\intercal\mathbf{Q}\mathbf{a} - \frac{\alpha}{1-\alpha}\mathbf{b}^\intercal\mathbf{a} &\to \min_{\mathbf{a}} \\ \text{s.t. } \mathbf{a} \geq 0,\ \sum_i a_i &= 1. \end{aligned} \tag{2}$$

Обозначим эту задачу $S\left(\underbrace{\frac{\alpha}{1-\alpha}}_{\beta^{-1}},\ 1\right)$, где 1 указывает на норму решения. Рассмотрим задачу $S(\beta^{-1},\ \gamma)$ и сделаем замену переменной $\mathbf{a} = \gamma\tilde{\mathbf{a}}$, получим

$$\begin{aligned} \gamma^2\left(\frac{1}{2}\tilde{\mathbf{a}}^\intercal\mathbf{Q}\tilde{\mathbf{a}} - \frac{1}{\beta\gamma}\mathbf{b}^\intercal\tilde{\mathbf{a}}\right) &\to \min_{\tilde{\mathbf{a}}} \\ \text{s.t. } \tilde{\mathbf{a}} \geq 0,\ \sum_i \tilde{a}_i &= 1, \end{aligned}$$

откуда задача $S(\beta^{-1},\ \gamma)$ эквивалентна задаче $S((\beta\gamma)^{-1},\ 1)$ в терминах активных признаков (сами же решения отличаются в γ раз), а потому мы имеем дело именно с однопараметрическим семейством и задание нормы решения, равной одному, не ограничивает общности.

Рассмотрим теперь замену переменной $\mathbf{a} = \beta^{-1}\tilde{\mathbf{a}}$ в (2). Получим, что (2) эквивалентна при $\alpha \in (0,\ 1)$ задаче

$$\begin{aligned} \frac{1}{2}\tilde{\mathbf{a}}^\intercal\mathbf{Q}\tilde{\mathbf{a}} - \mathbf{b}^\intercal\tilde{\mathbf{a}} &\to \min_{\tilde{\mathbf{a}}} \\ \text{s.t. } \tilde{\mathbf{a}} \geq 0,\ \sum_i \tilde{a}_i &= \beta. \end{aligned} \tag{3}$$

Наряду с задачей (3) можно рассмотреть задачу

$$\begin{aligned} \frac{1}{2}\tilde{\mathbf{a}}^\intercal\mathbf{Q}\tilde{\mathbf{a}} - \mathbf{b}^\intercal\tilde{\mathbf{a}} &\to \min_{\tilde{\mathbf{a}}} \\ \text{s.t. } \tilde{\mathbf{a}} \geq 0,\ \sum_i \tilde{a}_i &\leq \beta \end{aligned} \tag{4}$$

и соответствующую задачу без ограчения на норму вектора $\tilde{\mathbf{a}}$

$$\frac{1}{2}\tilde{\mathbf{a}}^\intercal \mathbf{Q}\tilde{\mathbf{a}} - \mathbf{b}^\intercal\tilde{\mathbf{a}} \to \min_{\tilde{\mathbf{a}}} \quad \text{s.t. } \tilde{\mathbf{a}} \geq 0. \tag{5}$$

Задача (1) не позволяет выбросить все признаки, поскольку $\|\mathbf{a}\| = 1 > 0$. Задача (4) эквивалентна ограничению неравенства в исходной задаче и, например, при $\alpha = 0$ будет иметь решением исключение всех признаков. Далее приведем анализ свойств решения (1) с ограничением равенства и неравенства для разных значений α.

Таблица 1. Свойства решения в методе QPFS в зависимости от параметра α

α	Равенство	Неравенстсво
$\alpha = 0$	Используются все признаки, $\mathbf{Qa}^* = \eta\mathbf{e}$, $\eta > 0$	Выброшены все признаки
$\alpha \to 0$	Используются все признаки, $\mathbf{Qa}^* \to \eta\mathbf{e}$, $\eta > 0$	Решение (5)
$\alpha \to 1$	Сходимся к отбору одного признака с максимальным b_j	То же, что и в «равенство»
$\alpha = 1$	Отбор одного признака с максимальным b_j	То же, что и в «равенство»

В случае $\alpha = 0$ для неравенства ($\sum_i a_i \leq 1$) решением является $\mathbf{a}^* = \mathbf{0}$. В случае равенства ($\sum_i a_i = 1$), чтобы минимизировать потери от необходимости иметь ненулевой $\mathbf{a}$, в оптимальном $\mathbf{a}^*$ оптимизируемая функция $\frac{1}{2}\mathbf{a}^\intercal\mathbf{Qa}$ должна иметь одинаковый градиент по всем направлениям (так как иначе можно уменьшить одну координату, увеличить другую, оставив норму $\mathbf{a}$ неизменной, уменьшив значение функции). Тот же результат можно получить и другим способом - из рассмотрения Лагранжиана задачи.

Анализ решения QPFS в зависимости от параметра α

Задачи (3) и (4) эквивалентны для некоторых η задаче

$$\frac{1}{2}\tilde{\mathbf{a}}^\intercal \mathbf{Q}\tilde{\mathbf{a}} - \mathbf{b}^\intercal\tilde{\mathbf{a}} + \eta\sum_j \mathbf{a}_j \to \min_{\tilde{\mathbf{a}}} \text{ s.t. } \tilde{\mathbf{a}} \geq 0,$$

что эквивалентно

$$\frac{1}{2}\tilde{\mathbf{a}}^\intercal \mathbf{Q}\tilde{\mathbf{a}} - \tilde{\mathbf{b}}^\intercal\tilde{\mathbf{a}} \to \min_{\tilde{\mathbf{a}}} \text{ s.t. } \tilde{\mathbf{a}} \geq 0,$$

где $\tilde{b}_j = b_j - \eta$ и η монотонно убывает по $\|\tilde{\mathbf{a}}\|_1 = \beta$. При этом для случая неравенства $\eta \geq 0$, в для равенства $\eta < 0$, если $\beta > \|\mathbf{a}^*\|_1$, где $\mathbf{a}^*$ есть решение задачи без ограничения на норму (5).

Таким образом, добавление ограничения на норму, фактически штрафует релевантность и происходит исключение тех признаков, у которых $b_j < \eta$, поскольку у них поправленная релевантность $\tilde{b}_j$ становится отрицательной.

Связь с лассо для линейной регрессии

Стандартная задача линейной регрессии с l_1 регуляризацией (лассо) имеет вид

$$\frac{1}{2}\|\mathbf{y} - \mathbf{Xw}\|_2^2 + \tau\|\mathbf{w}\|_1 \to \min_{\mathbf{w}}. \tag{6}$$

Предположим, что $\mathbf{x}_j^\intercal\mathbf{x}_j = 1$, $\mathbf{y}^\intercal\mathbf{y} = 1$, то есть признаки и целевая переменная нормированы. В качестве функций сходства (similarity) и релевантности (relevance) в QPFS рассмотрим

корреляцию Пирсона (не модуль как обычно в QPFS!). Имеем

$$\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2+\tau\|\mathbf{w}\|_1=\frac{1}{2}\underbrace{\mathbf{y}^\intercal\mathbf{y}}_{=1}+\frac{1}{2}\mathbf{w}^\intercal\underbrace{\mathbf{X}^\intercal\mathbf{X}}_{\tilde{\mathbf{Q}}}\mathbf{w}-(\underbrace{\mathbf{X}^\intercal\mathbf{y}}_{\tilde{\mathbf{b}}})^\intercal\mathbf{w},$$

где учтена нормированность всех признаков и целевой переменной, откуда корреляция и ковариация совпадают; $\tilde{\mathbf{Q}}$, $\tilde{\mathbf{b}}$ корреляции со знаком между признаками и признаками и целевой переменной соответственно. Отсюда задачу (6) можно переписать в виде

$$\frac{1}{2}\mathbf{w}^\intercal\tilde{\mathbf{Q}}\mathbf{w}-\tilde{\mathbf{b}}^\intercal\mathbf{w}+\tau\|\mathbf{w}\|_1\to\min_{\mathbf{w}},$$

что может быть переписано в эквивалентную задачу с ограничением равенства (так же для неравенства) для некоторого η

$$\begin{gathered}\frac{1}{2}\mathbf{w}^\intercal\tilde{\mathbf{Q}}\mathbf{w}-\tilde{\mathbf{b}}^\intercal\mathbf{w}\to\min_{\mathbf{w}}\\ \text{s.t. }\|\mathbf{w}\|_1=\eta.\end{gathered}\tag{7}$$

Если все корреляции Пирсона между признаками и между признаками и целевой переменной неотрицательны (то есть векторы $\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{y}$ лежат в одном многомерном квадранте), то $\tilde{\mathbf{Q}}=\mathbf{Q}$, $\tilde{\mathbf{b}}=\mathbf{b}$, то есть задача (7) тождественна QPFS, но без ограничения $\mathbf{w}\geq 0$. Таким образом, QPFS можно рассматривать как lasso ограничением на неотрицательность весов, если все корреляции Пирсона между признаками и между признаками и целевой переменной неотрицательны. Далее, если истинный вектор весов в линейной регрессии $\mathbf{w}\geq 0$, то условие не неотрицательность оценки весов тоже избыточно (так как $\mathbf{w}^*$ в задаче (7) и так будет неотрицательным, начиная с некоторого размера выборки), и QPFS будет полностью тождественен lasso. Таким образом, получаем условия тождественности QPFS(α) методу lasso(τ), где между α и τ существует некоторая связь:

- Rel = Sim = |Pearson correlation|;
- Нормированность признаков и целевой переменной: $\mathbf{y}^\intercal\mathbf{y}=\mathbf{x}_j^\intercal\mathbf{x}_j=1,\ j=1,\ldots,n$;
- Неотрицательность попарных корреляций: $\mathbf{y}^\intercal\mathbf{x}_j\geq 0,\ \mathbf{x}_j^\intercal\mathbf{x}_l\geq 0,\ j,\ l=1,\ldots,n$;
- Неотрицательность истинного вектора весов $\mathbf{w}^*\geq 0$.

Отметим, что в lasso реализуется ситуация, когда выбрасываются все признаки, когда $\tau\to\infty$, что соответствует QPFS с ограничением неравенства (а не равенства) при $\alpha=0$. Кроме того, «закритический» режим из QPFS с равенством, когда $\|\mathbf{a}\|_1=\beta>\|\mathbf{a}^*\|$, где $\mathbf{a}^*$ есть решение (5), в lasso не реализуется, поскольку, штраф за норму приводит всегда к ее сокращению, а потому $\tau\to 0$ в lasso соотетствует $\beta\to\|\mathbf{a}^*\|$ в QPFS с ограничением типа равенства.

Проблемы QPFS

Благодаря условию на неотрицательность коэффициентов $\mathbf{a}\geq 0$ в задаче QPFS (3), штраф на $\|\mathbf{a}\|_1$ становится штрафом на сумму коэффициентов, что упрощает оптимизацию. Авторы оригинального метода [1] прямо указывают на скорость оптимизации как на основное преимущество метода QPFS при сопоставимом качестве прогноза (например, с lasso) на тестовых выборках в рассмотренных наборах данных.

Как показано в предыдущей главе, при выполнении некоторых условий (в частности, неотрицательности истинного вектора параметров модели $\mathbf{w}^*$), QPFS будет в точности эквивалентен методу lasso, то есть метод обладает лучшими свойствами с точкиы зрения оптимизации, при этом давая то же решение, что и lasso. Однако, когда условия эквивалентности не выполнены, метод начинает проигрывать lasso, поскольку не учитывает, например, что релевантность пары признаков может быть значительно выше, чем релевантность каждого из них.

Пример. Рассмотрим выборку $(\mathbf{X},\ \mathbf{y})$ в признаковом пространстве размерности $n = 2$.

$$x_{1i} \sim \mathcal{N}(x_{1i}|0,\ 1),\ y_i \sim \mathcal{N}(y_i|0,\ 1),\ x_{2i} = x_{1i} + \varepsilon y_i,\ \varepsilon > 0.$$

Истинная корреляция Пирсона первого признака и целевой переменной y равна 0, а корреляция со вторым – равна $\varepsilon/\sqrt{1+\varepsilon^2}$, схожесть двух признаков $1/\sqrt{1+\varepsilon^2}$. При малом ε выборочная корреляция обоих признаков с целевой переменной будет мала, а схожесть двух признаков - велика. При этом надежное восстановление целевой переменной y возможно только при наличии обоих признаков в выборке, что соответствует ситуации с отсутствием отбора признаков (малое α).

Добавим теперь в выборку N шумовых признаков. Истинное сходство каждого из таких признаков с целевой переменной равно 0 (такое же как и для признака 1) и с учетом того, что QPFS не учитывает взаимодействия между признаками, признаки 1 и 2 не имеют значительного преимущества по отношению к шумовым в терминах релевантности целевой переменной (признак 1 в точности шумовой в изоляции, так как независим от y). При этом признаки 1 и 2 получают штраф за похожесть друг на друга. По этой причине при работе QPFS либо происходит исключение одного или обоих признаков 1 и 2 при исключении некоторых или вех шумовых, или оба признака 1 и 2 остаются, но вместе с ними остаются почти все или все шумы (см. эксперимент). В то же время Lasso учитывает взаимосвязи между признаками и не требует нетрицательности коэффициентов (ссылка на сравнение на этом датасете QPFS и lasso). В рассматриваемом примере не выполнено одно из условий эквивалентности lasso QPFS:

- $\mathbf{w}^* = (-1/\varepsilon,\ 1/\varepsilon)^\intercal$ содержит отрицательные значения.

Подобный пример (нужно добавить в статью вместе с соответствующим экспериментом с $N = 10$ или $N = 100$ шумами):

1. $x_{2i} = -x_{1i} + \varepsilon y_i \rightarrow$ не выполнено условие неотрицательной корреляции между признаками, а остальные условия эквивалентности выполнены.

Замечание: Подумай, можно ли построить пример, чтобы все условия кроме одного были выполнены для оставшихся условий эквивалентности QPFS и lasso.

Стабильность модели

В статье Катруцы (ссылка) стабильность модели (выраженная, например, в терминах $\lambda_{\max}(\mathbf{X}^\intercal\mathbf{X})/\lambda_{\min}(\mathbf{X}^\intercal\mathbf{X})$ имеет самостоятельную ценность и наряду с качеством прогноза на тестовой выборке определяет решение о превосходстве одного метода отбора признаков над другим. В этой статье мы предлагаем рассматривать стабильность как априорное знание, которое указывает на то, что априори мы считаем, что выборки с меньшим числом обусловленности на множестве активных признаков появляются чаще в рассматриваемой задаче, чем выборки с большим числом обусловленности. При отсутствии такого знания стабильность

стоит рассматривать в контексте повышения качества прогноза: если низкая стабильность модели ведет к снижению качества прогноза, стоит добавить штраф за низку стабильность модели, если качество не снижается, а повышается при уменьшении стабильности, то не стоит предпочитать менее качественную, но более стабильную модель.

Обозначим $\mathbf{X}(\mathbf{w})$ сужение матрицы признаков на множество признаков $j:\ w_j \neq 0$. Примером соответствующей задачи оптимизации, где есть априорное знание о том, что низкое число обусловленности более предпочтительно является

$$\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \tau\lambda_{\max}(\mathbf{X}(\mathbf{w})^\intercal\mathbf{X}(\mathbf{w}))/\lambda_{\min}(\mathbf{X}(\mathbf{w})^\intercal\mathbf{X}(\mathbf{w})) \to \min_{\mathbf{w}},$$

что соответствует показательному распределению на число обусловленности активных признаков выборки с гиперпараметром τ. В таком виде задача является тяжелой для оптимизации и требует перебора наборов активных признаков, например, с помощью генетического алгоритма.

Пример 1. Пусть $y_i \sim \mathcal{N}(y_i|0,\ 1)$, $x_{ij} = y_i + \nu\varepsilon_{ij}$, $\varepsilon_{ij} \sim \mathcal{N}(\varepsilon_{ij}|0,\ 1)$.
Оптимальная модель использует все признаки и осредняет их для получения наилучшего прогноза y_i: $\mathbf{w}^* = (1/n, \ldots, 1/n)^\intercal$. При этом число обусловленности $\eta = \lambda_{\max}(\mathbf{X}^\intercal\mathbf{X})/\lambda_{\min}(\mathbf{X}^\intercal\mathbf{X})$ при малом ν может быть большим.

Пример 2. Пусть в выборке есть дубликат признака $y_i \sim \mathcal{N}(y_i|0,\ 1)$, $x_{i1}x_{i2} = y_i + \nu\varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(\varepsilon_i|0,\ 1)$. В этом случае число обусловленности равно ∞ и имеется неоднозначность решения, минимизирующего $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$. Однако все эти решения имеют $w_1 + w_2 = \text{const}$, а потому если на тестовой выборке признаки 1 и 2 останутся идентичными, качество прогноза будет одинаковым независимо от того, какое разбиение этой константы между w_1, w_2, мы предпочтем. Обычно для того, чтобы сделать решение однозначным, добавляют слабую квадратичную регуляризацию на $\mathbf{w}$, что из всех решения предпочитается то, где $w_1^2 + w_2^2$ минимально, то есть $w_1 = w_2 = const/2$.

Заметим, что если есть основания полагать, что сильная мультиколлинеарность в обучающей выборке не будет продолжена в тестовой [2], то специальная обработка муьтиколлинеарности приобретает важность, а конкретный вид поправок зависит от предположений об эволюции корреляций.

Список литературы

[1] *Rodriguez-Lujan I. et al.* Quadratic programming feature selection // Journal of Machine Learning Research. – 2010.

[2] *Belsley D. A.* Collinearity and forecasting // Journal of Forecasting. – 1984. – Vol. 3. – No. 2. – P. 183-196.