Машинное обучение в экономике

Машинное обучение в эконометрике

Потанин Богдан Станиславович

доцент, научный сотрудник, кандидат экономических наук

2023–2024

- Методы оценивания параметров:
  - Ридж и Лассо регрессии.
  - Пост-Лассо.
  - Двойное машинное обучение.
- Базовые понятия:
  - Регуляризация.
  - Метод моментов.
  - Структурный параметр.
  - Функция шума.
  - Ортогональность по Нейману.
  - Кросс-фиттинг.
  - Эндогенность и неслучайный отбор.

- Машинное обучение, как правило, применяется для прогнозирования различных характеристик распределения, таких как условные математические ожидания и вероятности.
- Обычно методы машинного обучения дают оценки, обладающие малым смещением и большой дисперсией, поскольку не накладывают структурных предпосылок (например, о линейности) на форму связи между переменными модели.
- В задаче прогнозирования эконометрические методы обычно демонстрируют преимущество на выборках малого и среднего объема, поскольку обладают структурой, позволяющей компенсировать недостаток данных реалистичными предположениями, снижающими дисперсию оценок.
- Однако, основной упор в эконометрическом анализе делается на оценивание параметров моделей, имеющих содержательную экономическую интерпретацию.
- Иногда исследователя интересуют не все, а лишь часть параметров модели, характеризующих связь между основными переменными модели. В таком случае можно объединить сильные стороны эконометрики (интерпретабельность) и машинного обучения (высокая точность прогнозирования).
- **Основная идея** – часть модели, не представляющая содержательный интерес для исследователя, оценивается методами машинного обучения, а для оценивания структурных параметров применяются эконометрические методы анализа.

Регуляризация

Основная идея

- **Проблема** – машинное обучение позволяет избегать допущения о линейной связи Y_i с X_i и T_i, тем самым снижая смещение оценок, но часто серьезно повышает дисперсию на малых выборках.
- **Идея** – для того, чтобы снизить дисперсию оценок и избежать переобучения, пусть и ценой повышения смещения, можно воспользоваться **регуляризацией**.
- Одним из наиболее популярных подходов к регуляризации заключается в накладывании штрафов на параметры модели:

$$\underbrace{\mathrm{L}\left(Y, F(X;\theta)\right)}_{\text{функция потерь}} + \underbrace{\text{penalty}\left(\theta\right)}_{\text{штраф}} \qquad \textbf{мнимизируемый функционал}$$

- Функция $\text{penalty}\left(\theta\right)$ накладывает **штраф** (penalty) за определенные, как правило **большие по модулю** значения элементов n_θ-мерного вектора параметров θ модели $F(X;\theta)$.
- **Интуиция** – ограничение $\theta_i = 0$ обычно соответствует исключению параметра θ_i из модели, что приводит к ее упрощению. Регуляризация предлагает в качестве альтернативы накладывать штрафы, приводящие, образно говоря, к естественному отбору среди параметров, когда значительно отличными от 0 оказываются лишь те из них, что оказывают существенное влияние на качество модели.
- В роли параметров θ, например, могут выступать коэффициенты β в обычной линейной или логистической регрессии.

- В большинстве случае функция штрафа задается с помощью Lp-нормы:

$$\text{penalty}(\theta) = ||\theta||_p^p = \sum_{i=1}^{n_\theta} \lambda_i |\theta_i|^p, \text{ где } \lambda_i > 0 \text{ и } p \in \{0, 1, 2, ...\}$$

- Случаи $p = 1$ и $p = 2$ являются наиболее популярными:

$$\text{penalty}(\theta) = \sum_{i=1}^{n_\theta} \lambda_i |\theta_i| \qquad \textbf{Лассо регуляризация}$$

$$\text{penalty}(\theta) = \sum_{i=1}^{n_\theta} \lambda_i \theta_i^2 \qquad \textbf{Ридж регуляризация}$$

- Чем больше значения константа λ_i, тем сильнее накладываемый штраф за большие по абсолютной величине значения параметров θ_i
- Подбор λ_i обычно осуществляется по аналогии с гиперпараметрами, например, с помощью кросс-валидации. Для простоты часто полагают $\lambda_i = \lambda$ для всех i.

- Как правило величины коэффициентов θ тесно связаны с единицами измерения признаков X.
- **Проблема** – единицы измерения признаков X влияют на величину штрафа λ, а значит и на оценки параметров θ.
- **Решение** – привести признаки к сопоставимым единицам измерения, например, за счет стандартизации в форме деления на выборочное стандартное отклонение.
- Кроме того, часто стандартизация снижает сложность оптимизационной задачи, тем самым повышая скорость нахождение минимума методами численной оптимизации.

- Даже сохраняя линейную форму связи $E(Y_i|X_i) = X_i\beta$, линейная регрессия может аппроксимировать очень сложные зависимости, за счет того, что X_i могут быть разнообразными функциями (например, полиномы и сплайны) от исходных данных.
- Чем больше функций от исходных данных включает исследователь, тем, как правило, ниже смещение, но выше дисперсия оценок параметров и прогнозов.
- **Проблема** – при включении большого числа функций от исходных данных число оцениваемых коэффициентов β_i может оказаться чрезвычайно велико, что приведет к крайне большой дисперсии оценок.
- **Решение** – воспользоваться, например, Лассо регуляризацией, минимизируя:

$$\sum_{i=1}^{n}(Y_i - X_i\beta)^2 + \sum_{i=1}^{n_\beta}\lambda_i|\beta_i|$$

- **Полезное свойство Лассо регуляризации** – часто оценки коэффициентов при наименее значимых (с точки зрения вклада в прогностическое качество модели) регрессорах обнуляются $\hat{\beta}_i = 0$, что эквивалентно их исключению из модели.

- Преимущество Ридж регуляризации в линейной регрессии заключается в возможности получения аналитических оценок коэффициентов и их характеристик:

$$\hat{\beta} = \lambda \left(X^T X + \Lambda \right)^{-1} X^T Y, \text{ где } \Lambda = \text{diag}\left(\lambda, ..., \lambda\right)$$

$$\mathsf{E}\left(\hat{\beta}|X\right) = \beta - \underbrace{\lambda \left(X^T X + \Lambda \right)^{-1} \beta}_{\text{смещение}}$$

$$\text{Cov}\left(\hat{\beta}|X\right) = \left(X^T X + \Lambda \right)^{-1} X^T \text{Cov}\left(\varepsilon|X\right) X \left(X^T X + \Lambda \right)^{-1}$$

- Можно показать, что смещение увеличивается по мере роста штрафа λ.
- Производная $\text{Cov}\left(\hat{\beta}|X\right)$ по λ является отрицательно определенной матрицей, поэтому увеличение штрафа приводит к уменьшению дисперсии оценок.
- Если случайные ошибки ε_i гетероскедастичны, то существует такая константа c, что при $\lambda \in (0, c)$ оценки Ридж регрессии более эффективны, чем МНК.

- Если в модели используется лишь один регрессор (без константы), то легко показать, что смещение возрастает вместе со штрафом λ:

$$\partial \text{bias}\left(\hat{\beta}|X\right)/\partial\lambda = \partial\left|\lambda\beta/\left(\sum_{i=1}^{n} X_i^2 + \lambda\right)\right|/\partial\lambda = \left|\beta\sum_{i=1}^{n} X_i^2/\left(\sum_{i=1}^{n} X_i^2 + \lambda\right)^2\right| > 0$$

- Поскольку $\text{Cov}\left(\varepsilon|X\right)$ положительно определена, то дисперсия падает с ростом λ:

$$\partial \text{Var}\left(\hat{\beta}|X\right)/\partial\lambda = \partial\left(X^T \text{Cov}\left(\varepsilon|X\right) X/\left(\sum_{i=1}^{n} X_i^2 + \lambda\right)^2\right)\partial\lambda =$$

$$= \underbrace{\left(-2/\left(\sum_{i=1}^{n} X_i^2 + \lambda\right)^3\right)}_{<0} \underbrace{X^T \text{Cov}\left(\varepsilon|X\right) X}_{>0} < 0$$

Регуляризация в линейном регрессионном анализе

Пост-Лассо

- Напомним, что при Лассо регуляризации в линейных регрессионных моделях некоторые из коэффициентов β могут обращаться в 0.
- **Проблема** – включение большого числа регрессоров с нулевыми коэффициентами может привести к снижению эффективности оценок вследствие серьезного смещения.
- **Решение** – применить двухшаговую процедуру, на первом шаге которой оценивается Лассо регрессия, а на втором – обычная МНК регрессия, в которой в качестве объясняющих переменных используются лишь те, при которых коэффициенты оказались отличными от нуля в Лассо регрессии.
- Поскольку МНК регрессия используется после Лассо, описанный метод именуется **пост-Лассо**.
- **Примечание** – эффективность оценок пост-Лассо может быть ниже, чем у обычной Лассо регрессии, поскольку они обладают большей дисперсией.

Метод моментов

Повторение основ

- Рассмотрим i.i.d. выборку $X_1, ..., X_n$ из распределения с параметром $\theta_0 \in R$ (по аналогии для векторов).
- Рассмотрим непрерывную по θ функцию $\psi(X_i, \theta)$, часто именуемую вкладом (**score**), такую, что θ_0 является единственным решением равенства:

$$E(\psi(X_i, \theta)) = 0$$

- Оценка $\hat{\theta}$, полученная из решения следующего равенства, является, при определенных условиях регулярности, состоятельной для параметра θ_0 и асимптотически нормальной:

$$\frac{1}{n}\sum_{i=1}^{n}\psi(X_i; \hat{\theta}) = 0$$

- Иногда из равенства $E(\psi(X_i, \theta_0)) = 0$ или некоторым иным образом можно выразить θ_0 как функцию от моментов, не зависящих от θ_0:

$$\theta_0 = g(E(\psi_1(X_i)), ..., E(\psi_m(X_i))).$$

- В таком случае, при некоторых условиях регулярности, состоятельная и асимптотически нормальная оценка $\hat{\theta}$ параметра θ_0 может быть получена как:

$$\hat{\theta} = g(\hat{E}(\psi_1(X_i)), ..., \hat{E}(\psi_m(X_i))), \text{ где } \hat{E}(\psi_k(X_i)) = \frac{1}{n}\sum_{i=1}^{n}\psi_k(X_i; \hat{\theta}) \xrightarrow{p} E(\psi_k(X_i))$$

- Если некоторые параметры функции ψ_k неизвестны, то иногда ее заменяют состоятельной оценкой $\hat{\psi}_k$, однако в таком случае изучение свойств оценки $\hat{\theta}$ оказывается нетривиальной задачей.

Двойное машинное обучение (DML)

Пример с частично линейной регрессии

- Рассмотрим частично линейную модель:

$$Y_i = \alpha T_i + g(X_i) + \varepsilon_i, \text{ где } (T_i, X_i, \varepsilon_i) \text{ i.i.d.}$$

- В качестве основного параметра интереса для исследователя выступает $\alpha \in R$.
- Введем стандартное допущение $\mathrm{E}(\varepsilon_i | T_i, X_i) = 0$, которое можно ослабить до $\mathrm{E}(\mathrm{Cov}(\varepsilon_i, T_i | X_i)) = 0$.
- Используя введенные допущения и закон чередующихся математических ожиданий можно показать, что:

$$\mathrm{E}[\psi] = 0, \text{ где } \psi(\alpha, g_Y, g_T) = (Y_i - g_Y(X_i) - \alpha[T_i - g_T(X_i)])(T_i - g_T(X_i))$$

$$g_Y(X_1) = \mathrm{E}(Y_i | X_i), \; g_T(X_i) = \mathrm{E}(T_i | X_i)$$

Где аргумент (T_i, X_i, Y_i) функции ψ опущен для краткости и i может быть произвольным в силу i.i.d.

- Выражая α из равенства $E(\psi) = 0$ получаем:

$$\alpha = \frac{\mathrm{E}((Y_i - g_Y(X_i))(T_i - g_T(X_i)))}{\mathrm{E}\left((T_i - g_T(X_i))^2\right)} = \frac{E(\psi_1(X_i, Y_i, T_i; g_Y, g_T))}{E(\psi_2(X_i, T_i; g_T))}$$

- Следовательно, параметр α можно оценить двухшаговой процедурой, на первом шаге которой **методами машинного обучения** оцениваются **функции шума** $\hat{g}_Y(x) = \hat{\mathrm{E}}(Y_i | X_i = x)$ и $\hat{g}_T(x) = \hat{\mathrm{E}}(T_i | X_i = x)$, а на втором шаге – параметр α:

$$\hat{\alpha} = \frac{\sum_{i=1}^n (Y_i - \hat{g}_Y(X_i))(T_i - \hat{g}_T(X_i))}{\sum_{i=1}^n (T_i - \hat{g}_T(X_i))^2} = \frac{\sum_{i=1}^n \hat{\psi}_1(X_i, Y_i, T_i; g_Y, g_T)}{\sum_{i=1}^n \hat{\psi}_2(X_i, T_i; g_T)} = \frac{\sum_{i=1}^n \psi_1(X_i, Y_i, T_i; \hat{g}_Y, \hat{g}_T)}{\sum_{i=1}^n \psi_2(X_i, T_i; \hat{g}_T)}$$

Двойное машинное обучение (DML)

Ортогональность по Нейману

- **Проблема** – поскольку оценивание $\hat{\alpha}$ происходит в два шага, то смещение оценок $\hat{\mathrm{E}}(Y_i|X_i)$ и $\hat{\mathrm{E}}(T_i|X_i)$, часто возникающие вследствие использования регуляризации (**regularization bias**), может приводить к серьезному смещению $\hat{\alpha}$, так как $E(\psi(\alpha, \hat{g}_T, \hat{g}_Y)) \neq 0$.
- **Решение** – частично данная проблема смягчается за счет формы функции ψ, удовлетворяющей условию **ортогональности по Нейману**:

$$\partial E \left(\psi \left(\alpha, g_Y(X_1) + q \underbrace{(\hat{g}_Y(X_1) - g_Y(X_1))}_{\text{смещение}}, g_T(X_1) + q \underbrace{(\hat{g}_T(X_1) - g_T(X_1))}_{\text{смещение}} \right) \right) / \partial q|_{q=0} = 0$$

- **Интуиция** – при небольших отклонениях оцененных условных математических ожиданий от истинных моментное тождество $E(\psi) = 0$ продолжает соблюдаться с малой погрешностью:

$$\alpha \approx \frac{\mathrm{E}\left((Y_i - \hat{g}_Y(X_i)) (T_i - \hat{g}_T(X_i)) \right)}{\mathrm{E}\left((T_i - \hat{g}_T(X_i))^2 \right)}$$

- **Примечание** - в работе (ссылка) приводятся полезные алгоритмы построения функций ψ, обладающих свойством ортогональности по Нейману. Например, ортогональной по Нейману будет также следующая функция:

$$\psi(\alpha, g, g_T) = (Y_1 - \alpha T_1 - g(X_1)) (T_1 - g_T(X_1))$$

Двойное машинное обучение (DML)

Разбиение выборки

- **Проблема** – даже несмотря на регуляризацию, многие методы машинного обучения склонны к переобучению (**overfitting bias**), из-за чего по крайней мере внутривыборочные оценки $Y_i - \hat{g}_Y(X_i)$ и $T_i - \hat{g}_T(X_i)$ могут существенно отклоняться от $Y_i - g_Y(X_i)$ и $T_i - g_T(X_i)$, тем самым снижая точность оценок второго шага.
- **Решение** – применить разбиение выборки (**sample splitting**) на две части – первая часть выборки используется на первом шаге, то есть для оценивания g_Y и g_T, а вторая – на втором шаге для оценивания α с использованием полученных на первом шаге оценок $\hat{g}_Y$ и $\hat{g}_T$.
- **Проблема** – мы используем лишь по половине выборки для каждого из шагов, что может снижать эффективность наших оценок.
- **Решение** – воспользоваться **кросс-фиттингом**.
- Обозначим через $\hat{g}_Y^{(1)}$, $\hat{g}_T^{(1)}$ и $\hat{g}_Y^{(2)}$, $\hat{g}_T^{(2)}$ оценки функций g_Y и g_T, полученные на первой и второй половинах выборки соответственно. То есть обе половины выборки поочередно используются на первом шаге.
- Введем вспомогательную переменную q_i, такую, что $q_i = 1$ если наблюдение i **не** вошло в первую половину выборки, и $q_i = 2$ в противном случае.
- Оценим $\hat{\alpha}$ таким образом, чтобы для каждого наблюдения i на втором шаге использовались оценки функций g_Y и g_T, которые были получены без использования i-го наблюдения:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{g}_Y^{(q_i)}(X_i) \right) \left(T_i - \hat{g}_T^{(q_i)}(X_i) \right)}{\sum_{i=1}^{n} \left(T_i - \hat{g}_T^{(q_i)}(X_i) \right)^2}$$

Двойное машинное обучение (DML)

Кросс-фиттинг

- **Проблема** – использование лишь половины выборки может существенно снизить эффективность оценок функций g_Y и g_T.
- **Решение** – реализовать кросс-фиттинг по аналогии с кросс-валидацией, разбив выборку на K (примерно) равных частей, где $\hat{g}_Y^{(k)}$ и $\hat{g}_T^{(k)}$ оцениваются на данных, **не** вошедших в k-ю из этих выборок (обычно $K = 5$):

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{g}_Y^{(q_i)}(X_i) \right) \left(T_i - \hat{g}_T^{(q_i)}(X_i) \right)}{\sum_{i=1}^{n} \left(T_i - \hat{g}_T^{(q_i)}(X_i) \right)^2}$$

  Где $q_i = k$, если наблюдение i **не** вошло в k-ю выборку.
- **Проблема** – результаты оценивания могут быть чувствительны к конкретному разбиению на K частей.
- **Решение** – повторить кросс-фиттинг m раз и либо усреднить все полученные оценки, либо взять ту из них, что является выборочной медианой.

Двойное машинное обучение (DML)

Резюме

- Описанный метод именуется **двойным машинным обучением** (DML), поскольку предполагает применение методов машинного обучения при оценивании функций $\hat{g}_Y$ и $\hat{g}_T$, а также кросс-фиттинга.
- При достаточно слабых допущениях DML метод дает состоятельную и асимптотически нормальную оценку $\hat{\alpha}$.
- Идейно DML опирается на метод моментов, поскольку выражение, используемое для оценивания α, выводится из равенства $E(\psi) = 0$.
- **Проблема** – использование оценок $\hat{g}_Y$ и $\hat{g}_T$ вместо истинных значений g_Y и g_T может приводить к неточностям в оценивании $\hat{\alpha}$.
- **Решение** – кросс-фиттинг и подбор функции ψ, удовлетворяющей ортогональности по Нейману.
  - Ортогональность по Нейману позволяет сгладить смещение вследствие регуляризации.
  - Кросс-фиттинг помогает снизить смещение, обусловленное переобучением.
- Иногда кросс-фиттинг реализуется упрощенным образом – параметр α оценивается на каждой из K подвыборок и полученный результат усредняется. Такой подход называется DML, а рассмотренный ранее – DML2.
- Авторы метода рекомендуют применять DML2, особенно на малых выборках.

- **Проблема** – если T_i является эндогенной переменной, то $E(\varepsilon_i|T_i, X_i) \neq 0$, откуда $E(\psi) \neq 0$, что не позволяет оценить α описанным ранее способом.
- **Решение** – найти **инструментальную переменную** Z_i (случай с несколькими инструментами рассматривается по аналогии), то есть такую, что $E(\varepsilon_i|X_i, Z_i) = 0$ и $E(Cov(T_i, Z_i|X_i)) \neq 0$. После этого рассмотреть такую ψ, что $E(\psi) = 0$ и соблюдается ортогональность по Нейману, например:

$$\psi = (Y_i - g_X(X_i) - \alpha\,(T_i - g_T(X_i)))\,(Z_i - g_Z(X_i))\,, \text{ где } g_Z(X_i) = E(Z_i|X_i)$$

- По аналогии с предыдущим примером применив кросс-фиттинг получаем:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{g}_Y^{(q_i)}(X_i)\right)\left(Z_i - \hat{g}_Z^{(q_i)}(X_i)\right)}{\sum_{i=1}^{n}\left(T_i - \hat{g}_T^{(q_i)}(X_i)\right)\left(Z_i - \hat{g}_Z^{(q_i)}(X_i)\right)}$$

Двойное машинное обучение (DML)

Неслучайный отбор

- Наблюдаемость Y_i может зависеть от некоторого правила, например, запрлата Y_i наблюдается лишь для работающих $Z_i = 1$ индивидов и ненаблюдается для безработных $Z_i = 0$:

$$\underbrace{Y_i^* = \alpha T_i + g(X_i) + \varepsilon_i}_{\text{целевое уравнение}} \qquad \underbrace{Z_i^* = r_1(W_i) + u_i}_{\text{уравнение отбора}}$$

$$Y_i = \begin{cases} Y_i^*, \text{ если } Z_i = 1 \\ \text{ненаблюдаем, в противном случае} \end{cases} \qquad Z_i = \begin{cases} 1, \text{ если } Z_i^* \geq 0 \\ 0, \text{ в противном случае} \end{cases}$$

- Поскольку в данных мы наблюдаем лишь $(Y_i | Z_i = 1)$, а не Y_i, то нарушается допущение о нулевом условном математическом ожидании случайной ошибки:

$$E(\varepsilon_i | X_i, Z_i = 1) = E(\varepsilon_i | X_i, u_i \geq -r_1(W_i)) = r(W_i) \implies E(Y_i | X_i, Z_i = 1) = \alpha T_i + g(X_i) + r(W_i)$$

- **Проблема** – функция $r(W_i)$ является пропущенной переменной, что приведет к несостоятельности DML оценки $\hat{\alpha}$.
- **Решение** – если T_i не входит в W_i, то можно объединить переменные X_i и W_i, получив регрессионное уравнение, в котором α можно оценить DML методом:

$$Y_i = \alpha T_i + g^*(X_i^*) + v_i, \text{ где } g^*(X_i^*) = g(X_i) + r(W_i) \text{ и } X_i^* = (X_i, W_i)$$

Двойное машинное обучение (DML)

Несколько структурных параметров

- **Проблема** – иногда исследователю необходимо оценить не один, а сразу несколько структурных параметров a_j, где $j \in \{1, ..., n_T\}$.

$$Y_i = \alpha_1 T_{1i} + ... + \alpha_{n_T} T_{n_T i} + g(X_i) + \varepsilon_i$$

- Например, параметры α_j могут отражать отдачу от различных уровней образования: базовый, бакалавриат и магистратура.
- **Решение** – оценить каждый из параметров α_j поочередно, используя DML метод для следующего уравнения:

$$Y_i = \alpha_j T_{ji} + g_j(X_i, T_{1i}, ..., T_{(j-1)i}, T_{(j+1)i}, ..., T_{n_T i}) + \varepsilon_i$$

- Для тестирования гипотез о связи между параметрами α_j можно применить бутстрап.