УДК 519.7

О.А. Галкін, аспірант

Проблема перенавчання у процедурі вибору моделі на основі структурної мінімізації ризику

Об'єктом дослідження є методологія вибору моделі, як одна із проблем інтелектуального аналізу даних. Метою дослідження є вивчення проблеми перенавчання, коли доступною є множина гіперпараметрів або моделей. В статті розглядається методологія оптимізації гіперпараметрів з використанням методу градієнтного спуску, а також проблема перенавчання у виборі моделі з використанням критерію помилки перевірки.

Ключові слова: вибір моделі, проблема перенавчання, оптимізація гіперпараметрів.

Київський національний університет імені Тараса Шевченка, 03680, м. Київ, пр. Глушкова 4д., e-mail: oleksandr.galkin@mail.ru

Oleksandr A. Galkin, postgraduate (PhD)

The problem of over-fitting in the procedure of model selection based on structural risk minimization

The object of study is the methodology of model selection as one of the problems of data mining. The aim of the investigation is to study the problem of over-fitting when a set of hyperparameters or models is available. The paper contains the methodology of optimizing of hyperparameters using the gradient descent and also the problem of over-fitting in model selection using the criterion of validation error.

Key words: model selection, over-fitting problem, optimizing hyperparameters.

Taras Shevchenko National University of Kyiv, 03680, Kyiv, Glushkova str., 4d, e-mail: oleksandr.galkin@mail.ru

Статтю представив: чл.-кор. НАНУ, д.ф.-м.н., проф. Анісімов А.В.

Вступ

В рамках статистичної теорії навчання, навчальні дані генеруються незалежно з ідентичним розподілом з деякого невідомого розподілу $P(x,y)$, в якому закодована залежність між входом x та виходом y. Тобто, якщо вхідні дані відповідають ймовірнісному розподілу $P(x)$, а вихідним даними є функція $f(x)$, що знаходиться під впливом дисперсії гауссівського шуму σ^2, тоді:

$$P(x,y) = P(x)\mathrm{N}_\sigma(f(x)-y),$$

де N_σ є розподілом Гаусса із щільністю $\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{t^2}{2\sigma^2}\right)$.

Якість функції, яка моделює дане співвідношення, вимірюється шляхом використання очікування функції втрат відносно $P(x,y)$. Розглядаючи задачу класифікації даних, функція втрат буде мати наступний вигляд:

$$\ell(f(x),y) = \mathrm{I}_{f(x)\neq y},$$

де I є функцією-показником, тобто $\mathrm{I}_A = 1 \Leftrightarrow A$ є істинним. Оскільки, одне з припущень в рамках даного дослідження є те, що дані, вихідні значення яких повинні бути заздалегідь передбачені, також генерується з того ж самого розподілу, метою статистичного навчання є знаходження функції мінімізації втрат

$$R(f) = \int \ell(x, f(x))dP(x,y). \qquad (1)$$

Проблема навчання зводиться до пошуку функції $f \in F$, що мінімізує очікування (1) функції-показника втрат $\ell(f(x),y) = \mathrm{I}_{f(x)\neq y}$. Дане очікування не може бути обчислено, оскільки розподіл $P(x,y)$ є невідомим. Однак, враховуючи навчальну множину $\{(x_i,y_i)\}_{1\le i\le n}$, ми можемо мінімізувати *емпіричний ризик*:

$$R_{емп}(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(f(x_i),y_i). \qquad (2)$$

Використовуючи принцип індукції емпіричної мінімізації ризику, вибір множини функцій F має вирішальне значення: якщо дана множина буде надто великою, ми можемо зіткнутися з проблемою *перенавчання*. Це означає, що емпіричний ризик буде дуже малим, а відповідний реальний ризик дуже великим. З іншої сторони, якщо F буде дуже малим, ми зіткнемося з проблемою *незавершеного навчання*:

$\min_{f\in F} R(f)$ буде знаходитись далеко від $\min R(f)$.

Структурна мінімізація ризику у проблемі вибору моделі

У даній статті ми розглядаємо методологію вибору моделі, як одну із проблем інтелектуального аналізу даних. Розглядаючи задачу класифікації даних, F повинно бути обмежено для того, щоб мати відповідну складність. Вибір класу функцій F називається *вибором моделі*.

Вибір моделі є досить складною проблемою, однак в рамках теорії Вапника-Червоненкіса (ВЧ) можна спробувати знайти клас функцій, який мінімізує границю наступного твердження:

Твердження 1. Нехай F є класом функцій розмірності ВЧ h, тоді для будь-якого розподілу P та деякої вибірки $\{(x_i, y_i)\}_{1\le i\le n}$, що отримана з цього розподілу, справедливою є наступна нерівність з ймовірністю $1-\eta$:

$$\forall f\in F,\ R(f)\le R_{емп}(f)+\sqrt{\frac{h\left(\log\frac{2n}{h}+1\right)-\log\left(\frac{\eta}{4}\right)}{n}}+\frac{1}{n}.$$

Це є ідеєю принципу індукції структурної мінімізації ризику, що представлена на рисунку 1.

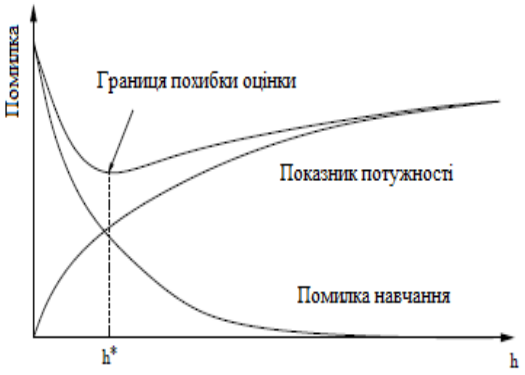

Рисунок 1: *Гіперплощина відповідає вирівнюванню точок у верхній частині рисунку.*

Розглянемо сімейство класу функцій F_i, кожна з яких має розмірність ВЧ h_i. У традиційному формулюванні структурної мінімізації ризику [1], класи функцій є вкладеними ($F_i \subset F_{i+1}$), що призводить до підвищення складності ($h_i \ge h_{i+1}$). Однак, дане припущення не є необхідним.

Нехай для даного i, f_n^i є функцією мінімізації емпіричного ризику по F_i. З твердження 1 ми маємо, що з ймовірністю $1-\eta_i$,

$$R(f_n^i)\le R_{емп}(f_n^i)+\sqrt{\frac{\varphi(h_i)-\log(\eta_i/4)}{n}}+\frac{1}{n}, \quad (3)$$

де $\varphi(h)$ є показником складності,

$$\varphi(h)=h\left(\log\frac{2n}{h}+1\right).$$

Для одного класу функцій ми зафіксували одне значення η_i. Зафіксуємо рівняння (3) рівномірним по i: з ймовірністю $1-\sum\eta_i$,

$$\forall i,\ R(f_n^i)\le R_{емп}(f_n^i)+\sqrt{\frac{\varphi(h_i)-\log(\eta_i/4)}{n}}+\frac{1}{n}. \quad (4)$$

Припустимо, що існує p класів функцій $F_1,\dots,F_p$, а також виберемо $\eta_i=\eta/p$. Якщо $\hat{i}$ є моделлю, що вибрана за принципом системної мінімізації ризику, тобто $\hat{i}$ мінімізує праву частину рівняння (3), тоді з рівняння (4) ми маємо, що з ймовірністю $1-\eta$,

$$R(f_n^{\hat{i}})\le \min_{1\le i\le p} R_{емп}(f_n^i)+ +\sqrt{\frac{\varphi(h_i)+\log(p)-\log(\eta_i/4)}{n}}+\frac{1}{n}. \quad (5)$$

Додатковим фактором є той факт, що протестованим моделям p відповідає лише $\log(p)$. Якщо значення p є малим, використання цього фактору не має сенсу. Однак, коли кількість моделей є експоненціально великою або навіть нескінченною, необхідно встановити для моделі заздалегідь визначену вагу шляхом вибору постійних величин η_i. Для нескінченного числа моделей, двома можливими варіантами є $\eta_i=\eta 2^{-i}$ або $\eta_i=\eta 2^{-i}/(\pi^2/6)$. Перший варіант призводить до більшого відхилення в сторону першої моделі, оскільки показник $\log(\eta_i)$ в рівнянні (3) є лінійним у першому випадку та логарифмічним в другому випадку.

Проблема перенавчання у процедурі вибору моделі

Як уже було зазначено, у випадку коли доступними є багато моделей, границі ризику стають більшими. Ігнорування даного факту

може призвести до феномену перенавчання на стадії вибору моделі. Для вивчення цієї проблеми спочатку узагальнимо процедуру вибору моделі.

У загальному випадку, параметри α та гіперпараметри θ алгоритму навчання повинні бути певним чином оцінені. У випадку емпіричної мінімізації ризику, гіперпараметри θ відповідають за вибір класу функцій F, тоді як параметри α відповідають за опис самої функції в класі.

Процес навчання є процедурою, що складається з двох етапів:

1. При фіксованому значенні θ, необхідно знайти найкращі параметри α^0,
$$\alpha^0(\theta) = \arg\min_{\alpha} T(\alpha, \theta).$$
2. Знайти найкраще значення θ,
$$\theta^0 = \arg\min_{\theta} V(\alpha^0(\theta), \theta).$$

Перший етап полягає в класичній мінімізації ризику: модель, що описана θ є фіксованою, а емпірична мінімізація ризику є випадком, коли T є емпіричною помилкою. Вибір моделі виконується на другому етапі (вибір θ). Для структурної мінімізації ризику, V є верхньою границею, що представлена в твердженні 1. Зауважимо, що виконання класичної моделі вибору може зайняти багато часу, оскільки для кожного протестованого значення θ вимагається мінімізація по α.

Як правило, критерієм V є оцінка або верхня границя помилки узагальнення, що призводить до прямої залежності від виконання алгоритму. Тим не менш, досить ефективним критерієм є помилка перевірки. Цей критерій не залежить від алгоритму навчання, а також досить легко обчислюється. Враховуючи цей факт, даний критерій буде використовуватися нами в подальших дослідженнях.

Припустимо, що моделі $\theta_1, ..., \theta_p$ знаходяться в режимі тестування. Для кожної моделі θ_i застосовується алгоритм навчання та виводиться функція f_i. Нехай F^* є множиною функцій $\{f_1, ..., f_p\}$. Крок вибору моделі полягає у виборі найкращої функції в F^* за допомогою критерію вибору моделі V.

Припустимо, що V є помилкою перевірки:
$$R_{перев}(f) = \frac{1}{n'} \sum_{i=1}^{n'} \ell(f(x_i'), y_i'),$$
де $\{(x_i', y_i)\}_{1 \le i \le n'}$ є незалежною вибіркою, що взята з того ж розподілу, що і навчальна множина.

Оскільки для всіх $f \in F^*$, $\mathrm{R}_{\text{перев}}(f)$ є незміщеною оцінкою істинного ризику $R(f)$, стандартним способом виконання вибору моделі є вибір функції f_i, що мінімізує $\mathrm{R}_{\text{перев}}(f)$.

Емпіричний ризик $\mathrm{R}_{\text{емп}}$ не є незміщеною оцінкою істинного ризику, оскільки функції в F^* вибираються з використанням навчальних прикладів. У цьому і полягає причина, чому "невидимі" приклади необхідні для того, щоб мати незміщену оцінку ризику.

Нехай f^* є мінімізатором помилки перевірки,
$$\mathrm{f}^* = \arg\min_{f \in F^*} R_{перев}(f).$$
Як і для емпіричної мінімізації ризику, $\mathrm{R}_{\text{навч}}(f^*)$ не є незміщеною оцінкою $\mathrm{R}(\mathrm{f}^*)$. Для визначення верхньої границі $\mathrm{R}(\mathrm{f}^*)$, ми повинні мати єдиний аргумент збіжності. Для цього виведемо верхню границю:
$$P\left\{ \sup_{f \in F^*} \left| R(f) - R_{навч}(f) \right| > \varepsilon \right\}.$$
Використовуючи нерівність Хефдінга [2], ми стверджуємо, що
$$\forall \mathrm{f} \in \mathrm{F}^*,\ \mathrm{P}\left\{ \left| \mathrm{R(f)} - \mathrm{R}_{\text{навч}}(f) \right| > \varepsilon \right\} < 2\exp(-2n'\varepsilon^2).$$
Оскільки потужність множини F^* дорівнює p, об'єднання границь призводить до
$$P\left\{ \sup_{f \in F^*} \left| R(f) - R_{навч}(f) \right| > \varepsilon \right\} \le 2p\exp(-2n'\varepsilon^2).$$
З цього ми можемо зробити висновок, що з упевненістю $1-\eta$,
$$R(f^*) \le R_{перев}(f) + \sqrt{\frac{\log(p) - \log(\eta/2)}{2n'}}. \qquad (6)$$

Поки значення p є не надто великим, $\mathrm{R}_{\text{навч}}(f)$ буде ефективною оцінкою $\mathrm{R(f)}$, а зведення до мінімуму помилки перевірки буде мати сенс. Однак, якщо значення p є великим (тобто $\log(p)$ має порядок n'), може виникнути проблема перенавчання. Дану проблему можна порівняти з класом функцій, що є досить великим при проведенні емпіричної мінімізації ризику. Два етапи є фактично еквівалентними: етап вибору моделі полягає в проведенні емпіричної мінімізації з використанням множини перевірки на множині F^*. Таким чином, якщо значення F^* є надто великим (тобто існує надто багато

моделей), перенавчання буде відбуватися під час цього етапу [3].

Застосування методу градієнтного спуску для оптимізації гіперпараметрів

Використовуючи результати попередніх досліджень, оптимізація параметрів моделі займає багато часу, а використання різних моделей має непомірно високу обчислювальну складність. Використовуваний нами підхід полягає в оптимізації гіперпараметрів з використанням методу градієнтного спуску. У цьому випадку, як і для емпіричної мінімізації ризику, значення p в рівнянні (6) не повинно бути числом моделей, що проходять тестування (наприклад, число кроків градієнта). Однак, значення p може бути числом різних можливих значень гіперпараметрів, яке є дуже великим [4].

При намаганні оптимізувати m гіперпараметрів $(\beta_1,...,\beta_m)$ на множині перевірки, кожен з яких може приймати q різних значень, число функцій в F^* буде дорівнювати $p = q^m$, а рівняння (6) буде вказувати на те, що додаткова похибка оцінки з етапу вибору моделі буде мати порядок $\sqrt{m/n'}$,

$$R(f^*) \le R_{перев}(f) + \sqrt{\frac{m\ln(q) - \ln(\eta/2)}{2n'}}. \qquad (7)$$

У випадку, якщо гіперпараметри приймають неперервні значення, побудова верхньої границі є більш складним завданням, однак це можна зробити за допомогою чисел покриття та традиційних границь ВЧ.

Однак розмірність F^* взагалі неможливо обчислити, оскільки функції в цій множині є власне розв'язками задачі оптимізації. З інтуїтивної точки зору, необхідно провести заміну розмірності ВЧ F^* на число гіперпараметрів m , як і в дискретному випадку. Це можна зробити з припущенням того, що функції в F^* "гладко" залежать від параметрів моделі θ . Приведемо лише скорочене доведення.

Нехай Ω є множиною можливих значень параметрів моделі. Для кожного з них, $f_\theta = f_{\alpha^0(\theta)}$ є оптимальною функцією в моделі θ , тобто мінімізацією навчального критерію T . Ми маємо, що

$$F^* = \{f_\theta,\ \theta \in \Omega\}.$$

Будемо вважати, що f_θ суттєво не змінюється з θ , тобто виконується наступна умова Ліпшиця:

$$\forall(\theta,\theta') \in \Omega^2,\ \left\| f_\theta - f_{\theta'} \right\|_\infty \le C \left\| \theta - \theta' \right\|_\infty .$$

Для надання сенсу такому припущенню, ми повинні розглянути клас дійсних функцій. Число покриття F^* може бути обмежено постійною величиною (в залежності від C), помноженою на число покриття Ω . Нарешті, стандартні результати класифікації даних з використанням дійсних функцій забезпечують оцінки помилки узагальнення в контексті чисел покриття та похибок поля.

Висновки

Підводячи підсумки викладеного матеріалу, зазначимо, що нашим завданням було визначити в чому полягає небезпека перенавчання, коли доступними є багато гіперпараметрів (моделей). Встановлено, що додаткова похибка оцінки зростає відповідно до квадратного кореня з числа гіперпараметрів. Зауважимо, що в даній статті ми вивчали проблему перенавчання у виборі моделі з використанням критерію помилки перевірки, однак те ж саме відбувається з будь-яким визначеним критерієм. Використовуючи отримані нами результати експериментальних досліджень для проблеми вибору характеристик з використанням методу структурної мінімізації ризику, можна зробити висновок, що коли число вибраних характеристик є великим, границя очікуваного ризику стає більшою.

Список використаних джерел

1. *Vapnik V., Chervonenkis A.* Theory of Pattern Recognition. Nauka, 1974.
2. *Hoeffding W.* Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association. 58(301): 13-30, 1963.
3. *Ng A.Y.* Preventing over-fitting of cross-validation data. In Proceedings of the 14th International Conference on Machine Learning. Morgan Kaufmann, 1997.
4. *Bengio Y.* Gradient-based optimization of hyper-parameters. Neural Computation, 12(8), 2000.

Надійшла до редколегії 28.01.2013