روش حداقل مربعات با تنظیم خودکار

مهدی اکرمی و مرضیه عبدالحمدی

تابستان ۱۴۰۱

فهرست موضوعی

مسئله حداقل مربعات با تنظیم خودکار

- **تفاوت هایپرپارامتر و پارامتر:** در یادگیری ماشین، هایپرپارامتر به پارامتری گفته می‌شود که مقدار آن، قبل از آغازِ فرآیند یادگیری، تعیین می‌شود. برخلاف دیگر پارامترهای مدل، که حین فرآیند یادگیری تعیین می‌شوند. در یادگیری عمیق، هایپرپارامترها شامل متغیرهایی هستند که برای تنظیم شبکه عصبی استفاده می‌شوند، مانند رگولاریزاسیون و نرخ یادگیری.

- **هایپرپارامتر پیوسته:** هایپرپارامتری که مقادیر آن بطور پیوسته (مثلا در یک بازه) میتواند تغییر کند.

- **هایپرپارامتر در مسئله حداقل مربعات:** مسئله حداقل مربعات میتواند به هایپرپارامتر از طریق رگولاریزاسیون یا درجه چندجمله ای و ... مجهز شود.
مثال: رگرسیون ریج

- **مزیت‌ها:**
  - ✓ انتخاب مدل بهتر
  - ✓ گذر به فراتر از تابع زیان مربعی

مسئله حداقل مربعات

- **صورت کلی مسئله حداقل مربعات:**

$$\underset{x}{\text{minimize}} \|Ax - b\|_2^2$$

که در آن A ماتریسی $m \times n$ و لاغر $(m > n)$ است.

- **پاسخ مسئله حداقل مربعات:**

با فرض مستقل بودن ستون‌های ماتریس A، جواب یکتا برای مسئله فوق بصورت زیر است:

$$\hat{x} = (A^T A)^{-1} A^T b = A^\dagger b$$

تابع هدف اصلی و عملگر proximal

- **تابع هدف:** یافتن ابرپارامتری که مسئله زیر را بهینه کند

$$minimize\ F(\omega) = \psi\left(\theta^{ls}(\omega)\right) + r(\omega)$$

- **ایده حل:** استفاده از روشی ابتکاری مانند proximal gradient که روشی تکراری بصورت زیر است

$$\omega^{k+1} = prox_{t^k r}(\omega^k - t^k \nabla_\omega \psi\left(\theta^{ls}(\omega^k)\right)$$

- **تعریف عملگر proximal:**

$$prox_{tr}(v) = \underset{\omega}{\operatorname{argmin}}(r(\omega) + \frac{1}{2t}\|\omega - v\|_2^2)$$

(Proximal Gradient) الگوریتم‌های بر پایه گرادیان

Given initial hyper-parameter vector $\omega^1 \in \Omega$**,** initial step size t^1**,** number of iterations n_{iter}**,** tolerance ε**.**

for $k = 1, \dots, n_{iter}$

1. Solve the least squares problem. $\theta^{ls}\left(\omega^k\right) = \left(A^T\left(\omega^k\right)A\left(\omega^k\right)\right)^{-1} A^T\left(\omega^k\right)B\left(\omega^k\right)$
2. Compute the gradient. $g^k = \nabla_\omega \psi\left(\theta^{ls}\left(\omega^k\right)\right)$
3. Compute the gradient step. $\omega^{k+\frac{1}{2}} = \omega^k - t^k g^k$
4. Compute the proximal operator. $\omega^{tent} = prox_{t^k r}\left(\omega^{k+\frac{1}{2}}\right)$
5. **if** $F\left(\omega^{tent}\right) \le F\left(\omega^k\right)$

   increase step size and accept update. $\omega^{k+1} = \omega^{tent}, \quad t^{k+1} = 1.2t^k$

   stopping criterion. **quit** if $\left\| \dfrac{\omega^k - \omega^{k+1}}{t^k} + \left(g^{k+1} - g^k\right) \right\|_2 \le \varepsilon$

6. **else** decrease step size and reject update. $\omega^{k+1} = \omega^k, \quad t^{k+1} = \dfrac{t^k}{2}$

end for

محاسبه گرادیان

- **هدف:** محاسبه گرادیان $g=\nabla_{\omega}\psi\left(\theta^{ls}(\omega)\right)$
- **فرض:** مقدار $\theta=\theta^{ls}(\omega)$ محاسبه شده

حال باید عبارت‌های زیر محاسبه شوند:

$$\nabla_{\omega}\psi(\theta)$$

$$C=\left(A^{T}A\right)^{-1}\nabla_{\omega}\psi(\theta)\in R^{n\times m}$$

پس از انجام محاسبات و استفاده از قاعده مشتق زنجیره‌ای داریم:

$$\nabla_{A}\psi=(B-A\theta)C^{T}-AC\theta^{T} \qquad \nabla_{B}\psi=AC$$

و نهایتا گرادیان بصورت زیر بدست می‌آید:

$$g=\sum_{i,j}\left(\nabla_{A}\psi\right)_{i,j}\left(\nabla_{\omega}A\right)_{i,j}+\sum_{i,j}\left(\nabla_{B}\psi\right)_{i,j}\left(\nabla_{\omega}B\right)_{i,j}$$

مسئله حداقل مربعات با قید خطی

- تعمیم مسئله حداقل مربعات به حالت مقید:

$$\begin{array}{ll} \min & \dfrac{\| A(\omega)\theta - B(\omega) \|_F^2}{2} \\ subject.to & C(\omega)\theta = D(\omega) \end{array}$$

در مسئله فوق داریم: $\theta \in R^{n\times m}, C: \quad \Omega \quad \rightarrow \quad R^{d\times n}, D: \quad \Omega \quad \rightarrow \quad R^{d\times m}$

دوتایی متغیرهای اولیه و دوگان $(\theta,\upsilon) \in R^{n\times m} \times R^{d\times m}$ بهینه هستند اگروتنها اگر در شرایط KKT بصورت زیر صدق کنند:

$$\underbrace{\begin{bmatrix} 0 & A(\omega)^T & C(\omega)^T \\ A(\omega) & -I & 0 \\ C(\omega) & 0 & 0 \end{bmatrix}}_{M(\omega)} \begin{bmatrix} \theta \\ q \\ \upsilon \end{bmatrix} = \begin{bmatrix} 0 \\ B(\omega) \\ D(\omega) \end{bmatrix}$$

مسئله حداقل مربعات با قید خطی

تابع هدف اصلی، تابعی از متغیرهای θ,υ خواهد شد. فرض میکنیم $\nabla_\theta\psi,\nabla_\upsilon\psi$ قابل محاسبه باشند. ابتدا محاسبه میکنیم:

$$\begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix} = M(\omega)^{-1} \begin{bmatrix} \nabla_\theta\psi \\ 0 \\ \nabla_\upsilon\psi \end{bmatrix}$$

حال گرادیان ψ نسبت به A, B بصورت زیر قابل محاسبه است:

$$\nabla_A\psi = -(rg_1^T + g_2\theta^T), \nabla_B\psi = g_2$$

و مشابها برای C, D داریم:

$$\nabla_C\psi = -\left(\upsilon g_1^T + g_3\theta^T\right), \nabla_D\psi = g_3$$

روش حداقل مربعات برای برازش داده‌ها

- ✓ **در مسئله برازش داده‌ها** ⟵ در دسترس بودن داده‌های ورودی به همراه مقدار مربوط به هر ورودی.
- ✓ **هدف** ⟵ برازش پارامترهای تابع پیش بینی کننده زیر:

$$\hat{y} = \phi\left(u, \omega^{feat}\right)^T \theta$$

- ✓ **انتخاب مدل** ⟵ باید مسئله حداقل مربعات زیر حل شود:

$$A(\omega) = \begin{bmatrix} e^{\omega_1^{data}} \phi\left(u, \omega^{feat}\right)^T \\ . \\ . \\ . \\ e^{\omega_N^{data}} \phi\left(u, \omega^{feat}\right)^T \\ e^{\omega_1^{reg}} R_1 \\ . \\ . \\ . \\ e^{\omega_d^{reg}} R_1 \end{bmatrix}, B(\omega) = \begin{bmatrix} e^{\omega_1^{data}} y_1 \\ . \\ . \\ . \\ e^{\omega_N^{data}} y_1 \\ 0 \\ . \\ . \\ . \\ 0 \end{bmatrix}$$

$\omega^{data} \in \Omega^{data} \subseteq R^N \longrightarrow$ هایپرپارامترهای وزنی

$R_1, \ldots, R_d \longrightarrow$ ماتریس‌های منظم‌ساز با سایز مناسب

$\omega^{reg} \in \Omega^{reg} \subseteq R^d \longrightarrow$ هایپرپارامترهای منظم‌ساز

$\omega = \left(\omega^{feat}, \omega^{data}, \omega^{reg}\right) \in \Omega^{feat} \times \Omega^{data} \times \Omega^{reg} \longrightarrow$ مجموعه کل هایپرپارامترها

تابع هدف حقیقی

فرض: داده‌های اعتبارسنجی از ورودی‌های $u_1^{val},...,u_{N_{val}}^{val} \in U$ و $y_1^{val},...,y_{N_{val}}^{val} \in R^m$ تشکیل شده‌اند.

پیش‌بینی: ابتدا پیش‌بینی زیر را انجام می‌دهیم:

$$\hat{y}_i^{val} = \phi(u_i^{val})\theta^{ls}(\omega)$$

در اینجا تابع $\phi\left(u,\omega^{feat}\right)$ مشخص و ثابت درنظر گرفته می‌شود.

تابع هدف حقیقی: تابع هدف حقیقی ψ در مسئله برازش داده‌ها با حداقل مربعات را میانگین زیان برای پیش‌بینی داده‌های اعتبارسنجی بصورت زیر در نظر می‌گیریم:

$$\psi(\theta) = \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} l\left(\hat{y}_i^{val}, y_i^{val}\right)$$

در عبارت فوق، $l: R^m \times R^m \to R$ تابع جریمه است که فرض می‌شود نسبت به متغیر اول خود مشتق پذیر است.

تابع جریمه رگرسیون

صورت کلی:

$$l\left(\hat{y}, y\right) = \pi\left(r\right)$$

که $r = \hat{y} - y$ مانده و $\pi : R^m \rightarrow R$ تابع جریمه اعمال شده روی مانده است.

رایج‌ترین توابع جریمه

$$\pi\left(r\right) = \|r\|_2^2$$

تابع مربعی

$$\pi\left(r\right) = \begin{cases} \|r\|_2^2 & \|r\|_2 \le M \\ M\left(2\|r\|_2^2 - M\right) & \|r\|_2 > M \end{cases}$$

تابع هوبر

$$\pi\left(r\right) = \begin{cases} \dfrac{M^2}{6}\left(1 - \left(1 - \dfrac{\|r\|_2^2}{M^2}\right)^3\right) & \|r\|_2 \le M \\ \dfrac{M^2}{6} & \|r\|_2 > M \end{cases}$$

تابع دو مربعی

تابع جریمه طبقه‌بندی

برای طبقه‌بندی، پیش‌بینی $\hat{y} \in R^m$ متناسب با توزیع احتمال روی m برچسب بصورت زیر داده می‌شود:

$$\Pr\left(y = e_i\right) = \frac{e^{\hat{y}_i}}{\sum_{j=1}^{m} e^{\hat{y}_j}}, \quad i = 1, 2, \ldots, m$$

پیش‌بینی $\hat{y}$ به عنوان یک توزیع احتمال شرطی برروی برچسب‌های y برای x داده شده، تعبیر می‌شود.

- ✓ **تابع جریمه پیشنهادی در طبقه‌بندی:**

$$l\left(\hat{y}, y\right) = -\hat{y}_i + \log\left(\sum_{j=1}^{m} e^{\hat{y}_j}\right) \quad i = 1, 2, \ldots, m$$

تابع آنتروپی متقابل

مجموعه تست و توقف زودهنگام

افزایش تعداد هایپرپارامترها ⟸ افزایش ریسک برازش بیش ازحد

- **چگونگی تشخیص برازش بیش از حد:** در نظر گرفتن یک مجموعه سوم از داده‌ها و محاسبه روی آن ← **مجموعه تست**

> **توجه!**
> مقدار زیان مجموعه اعتبارسنجی لزومی ندارد که تخمین خوبی از اندازه دقیق عملکرد مدل بر روی داده های دیده نشده باشد.

- **روشی متفاوت برای رویایی با مسئله برازش بیش از حد:** محاسبه مقدار زیان بر روی داده تست در هر تکرار و توقف الگوریتم در صورت افزایش مقدار زیان ⟵ **توقف زودهنگام**
  - پیش‌نیاز این روش، مجموعه داده چهارم است که پس از توقف الگوریتم، روی این داده محاسبات صورت گرفته و بعنوان عملکرد نهایی الگوریتم، گزارش می‌شود ← **مجموعه داده تست نهایی**

داده‌های MNIST

داده آموزش

۶۰۰۰۰

داده آزمایش

۱۰۰۰۰

رگرسیون ریج (Ridge Regression)

- مسئله رگرسیون ریج با یک هایپرپارامتر:

$$\| A\theta - B \|_F^2 + \exp(2\lambda) \| \theta \|_F^2$$

- حل مسئله با استفاده از:
  - ✓ تابع آنتروپی متقابل
  - ✓ تابع هدف مسئله حداقل مربعات

- جواب مسئله :

$$X = \begin{bmatrix} A \\ \exp(\lambda) I \end{bmatrix} \quad C = \begin{bmatrix} B \\ 0 \end{bmatrix} \quad X^T X \theta^{ls} = X^T C$$

- محاسبه مشتق θ^{ls} برحسب λ

$$\frac{d\theta^{ls}}{d\lambda} = D^{-1}\left(\frac{dX^T}{d\lambda} C - \frac{dD}{d\lambda}\left(D^{-1} X^T C\right)\right)$$

رگرسیون ریج (Ridge Regression)

- در نظر گرفتن **تابع حقیقی آنتروپی**:

$$l\left(\hat{y}, y=e_i\right) = -\hat{y}_i + \log\left(\sum_{j=1}^{m} e^{\hat{y}_j}\right)$$

- اعمال تابع آنتروپی متقابل روی داده‌های اعتبارسنجی و میانگین گیری:

$$\hat{y}_{val} = A_{val}\theta^{ls}, \quad \psi(\theta) = \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} l\left(\hat{y_i^{val}}, y_i^{val}\right)$$

- به منظورمحاسبه $g = \nabla_\lambda \psi$ کافی است عبارت زیر محاسبه شود:

$$\frac{dl}{d\lambda}\left(\hat{y}_j, y=e_i\right) = -\left(A_{val}\frac{d\theta^{ls}}{d\lambda}\right)_{ji} + \frac{\sum_{l=1}^{m} e^{\hat{y}_{li}}\left(A_{val}\frac{d\theta^{ls}}{d\lambda}\right)_{li}}{\sum_{l=1}^{m} e^{\hat{y}_{li}}}$$

رگرسیون ریج (Ridge Regression)

- در نظر گرفتن **تابع هدف مسئله حداقل مربعات** به عنوان تابع حقیقی:

$$\psi\left(\theta^{ls}\right) = \| X\theta^{ls} - b \|_F^2 + \exp(2\lambda) \| \theta^{ls} \|_F^2$$

- مشتق گیری:

$$g = \frac{d\psi}{d\lambda} = 2\left(X\theta^{ls} - b\right)^T \frac{d\theta^{ls}}{d\lambda} + 2\exp\left(2\lambda\right) \| \theta^{ls} \|_F^2 + 2\exp\left(2\lambda\right)\left(\theta^{ls}\right)^T \frac{d\theta^{ls}}{d\lambda}$$

پیاده‌سازی الگوریتم: رگرسیون ریج

رویکرد اول — **تابع هدف حداقل مربعات**

ساخت classifier + الگوریتم Proximal Gradient

رویکرد اول — **آنتروپی متقابل**

ساخت classifier + الگوریتم Proximal Gradient

رویکرد دوم — **آنتروپی متقابل**

ساخت بردار باینری به طول ۱۰ از هر برچسب + الگوریتم Proximal Gradient

نتایج

- مقایسه زمان محاسبه θ با استفاده از دو روش Convex Optimization و Least Squares:

- Convex Optimization: 15.594337 seconds
- Least Squares: 7.582242 seconds

- مهندسی ویژگی‌ها:
  - ✓ حذف ویژگی‌هایی که کمتر از ۶۰۰ درایه فعال دارند
  - ✓ افزودن ویژگی تصادفی

نتایج

- نتایج پیاده‌سازی با رویکرد اول + تابع هدف حداقل مربعات بعنوان تابع هدف حقیقی:

Random features	Iterations	Accuracy
0	10	0.86
1200	5	0.9555
1500	5	0.9587

- نتایج پیاده‌سازی با رویکرد اول + تابع حقیقی آنتروپی متقابل:

Train set	Random features	Validation set	Test set	Iteration	Accuracy
20000	0	1000	10000	2	0.85
40000	1000	4000	10000	2	0.94
40000	1000	4000	10000	6	0.95

نتایج

- نتایج پیاده‌سازی با رویکرد دوم + تابع حقیقی آنتروپی متقابل:

Train set	Random features	Validation set	Test set	Iteration	Accuracy
40000	1200	4000	10000	5	0.9573
50000	1200	5000	10000	10	0.9576
55000	1500	5000	10000	10	0.9581
59000	2000	1000	10000	5	0.9637
59000	5000	1000	10000	5	0.9726

- نتایج پیاده‌سازی با استفاده از شبکه‌های عصبی ساده در ۱۰ تکرار:

test_loss = 0.22957107, test_accuracy = 0.9352

- نتایج پیاده‌سازی با استفاده از شبکه عصبی پیچشی در ۱۰ تکرار:

Test: (loss = 0.0445f0, acc = 98.53)

همکاری در پروژه

نام	تئوری و مرور ادبیات	پیاده‌سازی	گزارش‌نویسی	ساخت اسلاید
مرضیه عبدالحمدی	۵۰	۵۰	۳۰	۷۰
مهدی اکرمی	۵۰	۵۰	۷۰	۳۰

منابع

1) S. Barratt. On the differentiability of the solution to convex optimization problems. arXiv preprint arXiv:1804.05098, 2018.

2) Shane Barratt, Stephen Boyd: Least Squares Auto-Tuning. arXiv:1904.05460v1, 2019.

3) A. Baydin, B. Pearlmutter, A. Radul, and J. Siskind. Automatic differentiation in machine learning: a survey. Journal of Machine Learning Research, 18:1-43, 2018.

4) R. Eigenmann and J. Nossek. Gradient based adaptive regularization. In Proc. Neural Networks for Signal Processing, pages 87-94, 1999.

5) Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

6) Parikh, N., and S. Boyd. 2014. \Proximal algorithms." Foundations and Trends R in Optimization 1 (3): 127-239.

سپاس از توجه شما