یادگیری ماشین | دکتر منصور رزقی

تمرین اول | دانشکده علوم ریاضی ، گروه علوم کامپیوتر، گرایش داده‌کاوی

تمرین ۱

قرار است برای دو متغیر تصادفی X و Y ثابت شود:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\ \mathrm{Cov}(X,Y)$$

تعریف واریانس در صورت سوال:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

حل:

$$\mathrm{Var}(X+Y) = \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2$$

از خطی بودن امید ریاضی میتوان نتیجه گرفت:

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

پس:

$$\mathbb{E}[(X+Y)^2] = \mathbb{E}[X^2 + 2XY + Y^2] = \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2]$$

$$(\mathbb{E}[X+Y])^2 = (\mathbb{E}[X] + \mathbb{E}[Y])^2 = (\mathbb{E}[X])^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + (\mathbb{E}[Y])^2$$

جایگذاری:

$$\begin{aligned}\mathrm{Var}(X+Y) &= \big(\mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2]\big) - \big((\mathbb{E}[X])^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + (\mathbb{E}[Y])^2\big) \\ &= \underbrace{\big(\mathbb{E}[X^2] - (\mathbb{E}[X])^2\big)}_{\mathrm{Var}(X)} + \underbrace{\big(\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2\big)}_{\mathrm{Var}(Y)} + 2\underbrace{\big(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\big)}_{\mathrm{Cov}(X,Y)}\end{aligned}$$

پس:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\ \mathrm{Cov}(X,Y)$$

تمرین ۲

صورت مسأله. تابع هزینه زیر را در نظر بگیرید:

$$J_1(w) = \|y - Xw\|_2^2 + \lambda_2\|w\|_2^2 + \lambda_1\|w\|_1.$$

تعریف:

$$c = (1+\lambda_2)^{-1/2}, \qquad \widetilde{X} = c\begin{pmatrix} X \\ \sqrt{\lambda_2}\, I_d \end{pmatrix}, \qquad \widetilde{y} = \begin{pmatrix} y \\ 0_{d\times 1} \end{pmatrix},$$

و

$$J_2(w) = \|\widetilde{y} - \widetilde{X}w\|_2^2 + c\lambda_1\|w\|_1.$$

حل: از آنجا که

$$\widetilde{X}w = c\begin{pmatrix} Xw \\ \sqrt{\lambda_2}\, w \end{pmatrix},$$

داریم:

$$\begin{aligned}\|\widetilde{y} - \widetilde{X}w\|_2^2 &= \|y - cXw\|_2^2 + \|0 - c\sqrt{\lambda_2}\, w\|_2^2 \\ &= \|y - cXw\|_2^2 + c^2\lambda_2\|w\|_2^2.\end{aligned}$$

در نتیجه:

$$J_2(w) = \|y - cXw\|_2^2 + c^2\lambda_2\|w\|_2^2 + c\lambda_1\|w\|_1.$$

اکنون تغییر متغیر زیر را اعمال می‌کنیم:

$$v = cw \quad \Longleftrightarrow \quad w = \frac{v}{c}.$$

با جایگذاری در تابع هزینه:

$$\begin{aligned}J_2\left(\frac{v}{c}\right) &= \left\|y - cX\frac{v}{c}\right\|_2^2 + c^2\lambda_2\left\|\frac{v}{c}\right\|_2^2 + c\lambda_1\left\|\frac{v}{c}\right\|_1 \\ &= \|y - Xv\|_2^2 + \lambda_2\|v\|_2^2 + \lambda_1\|v\|_1 \\ &= J_1(v).\end{aligned}$$

بنابراین برای هر v رابطه زیر برقرار است:

$$J_1(v) = J_2(v/c).$$

از این تطابق یک‌به‌یک نتیجه می‌گیریم که کمینه‌سازهای دو مسأله با رابطهٔ مقیاسی ساده به هم مرتبط هستند:

$$v^\star = \arg\min_v J_1(v) \quad \Longleftrightarrow \quad \frac{v^\star}{c} = \arg\min_w J_2(w).$$

معادل آن:

$$\boxed{\arg\min_w J_1(w) = c\ \arg\min_w J_2(w).}$$

این یعنی مسألهٔ اصلی J_1 را می‌توان با مسألهٔ J_2 حل کرد که نسخهٔ نرمال‌سازی‌شدهٔ آن است و فقط یک ضریب مقیاس بین پاسخ‌های دو مسأله وجود دارد.

تمرین ۳

فرض کنید برای هر مؤلفهٔ وزن w_i یک مدل (برنولی—گوسی) قرار دهیم. معادلِ صریح آن از طریق متغیر $z_i \in \{0, 1\}$ بیان می‌شود:

$$p(z_i) = \text{Bernoulli}(\pi), \qquad p(w_i \mid z_i) = \begin{cases}\delta(w_i), & z_i = 0, \\ \mathcal{N}(0, \sigma^2), & z_i = 1.\end{cases}$$

بنابراین توزیع w_i برابر است با

$$p(w_i) = (1 - \pi)\,\delta(w_i) + \pi\,\mathcal{N}(0, \sigma^2).$$

برای برازش (MAP) معمولاً کمینه‌سازی منفی لگاریتم احتمال انجام می‌دهیم. اگر $L(w)$ منفی لگِ احتمال داده (یا تابع هزینهٔ داده) باشد، MAP متناظر با کمینه‌سازی

$$\min_{w,z} \; L(w) \; - \; \log p(w,z)$$

است. چون

$$p(w,z) = \prod_i p(z_i)\, p(w_i \mid z_i),$$

منفی لگاریتمِ مبحثیِ (w,z) به‌صورت جمع روی مؤلفه‌ها نوشته می‌شود:

$$-\log p(w,z) = \sum_i \Big[-\log p(z_i) - \log p(w_i \mid z_i) \Big].$$

برای هر i داریم (تا ضرب در ثابت‌ها):

- اگر $z_i = 0$ آنگاه $w_i = 0$ و بخش مربوطه $\approx -\log(1-\pi)$.
- اگر $z_i = 1$ آنگاه w_i از گوسی است و بخش مربوطه $\approx -\log \pi + \dfrac{w_i^2}{2\sigma^2} + \text{const.}$

پس منفی لگاریتمِ احتمالِ مشترک (جزء مربوط به prior) را می‌توان به شکل زیر نوشت:

$$-\log p(w,z) = \sum_i \Big[(1-z_i)(-\log(1-\pi)) + z_i\big(-\log \pi + \tfrac{w_i^2}{2\sigma^2}\big) \Big] + \text{const.}$$

اگر حالا به‌صورت مشترک روی w و z مینیمم بگیریم، برای هر مؤلفه i انتخابِ z_i به‌صورت زیر خواهد بود: $z_i = 1$ هنگامی انتخاب می‌شود که سودِ کاهشِ خطا (یا مقدار w_i غیرصفر) بیشتر از هزینهٔ فعال‌سازی $-\log(\pi/(1-\pi))$ باشد. به عبارت دیگر، فعال‌سازی هر مؤلفه هزینه‌ای ثابتِ $\lambda_0 := -\log(\pi/(1-\pi))$ به‌علاوه یک جریمهٔ گوسیِ $w_i^2/(2\sigma^2)$ وارد تابع هزینه می‌کند.
در حالتِ تقریب:

- اگر σ^2 را بسیار بزرگ فرض کنیم (slab خیلی پهن باشد) یا جریمهٔ مربعی را کوچک در نظر بگیریم، آنگاه جملهٔ مربعی $\frac{w_i^2}{2\sigma^2}$ ناچیز می‌شود و تنها هزینهٔ ثابت برای هر وزنِ $(z_i = 1)$ باقی می‌ماند.
- در این صورت مجموع هزینه‌های ثابت برای همهٔ مؤلفه‌های فعال برابر است با $\lambda_0 \cdot \#\{i : w_i \neq 0\}$، یعنی $\lambda_0 \|w\|_0$.

بنابراین MAP با این prior تقریباً معادل حداقل‌سازیِ

$$\min_{w} \; L(w) \; + \; \lambda_0 \|w\|_0 \; + \; \text{(یک جریمهٔ کوچکِ مربعی برای مقادیر غیرصفر)},$$

که نشان می‌دهد مدل برنولی—گوسی می‌تواند به‌عنوان منبع منطقیِ ظهورِ منظم‌ساز ℓ_0 در تابع هزینه تفسیر شود.

تمرین ۴

چرا منظم‌ساز ℓ_1 در لاسو باعث صفر شدن وزن‌ها و ایجاد پراکندگی (sparsity) می‌شود؟

منظم‌ساز ℓ_1 به صورت زیر تعریف می‌شود:

$$\lambda \|w\|_1 = \lambda \sum_i |w_i|.$$

تابع قدرمطلق دارای یک گوشه (non-differentiable point) در مقدار صفر است. این خاصیت باعث می‌شود حل بهینه در بسیاری موارد دقیقا روی محور مختصات بیفتد—یعنی برخی ضرایب w_i دقیقاً صفر شوند. بنابراین لاسو مدل‌هایی با تعداد پارامتر کمتر و ضرایب صفرشونده تولید می‌کند.

تفسیر هندسی و احتمالاتی: چرا ℓ_1 نسبت به ℓ_2 تمایل بیشتری به صفر کردن ضرایب دارد؟

دیدگاه هندسی

مجموعه‌ی قیود $\|w\|_1 \leq c$ یک چندضلعی لوزی‌شکل است:

$$|w_1| + |w_2| \leq c.$$

گوشه‌های این چندضلعی دقیقاً روی محورهای مختصات قرار دارند. وقتی بیضی ناشی از خطای مربعی با این چندضلعی برخورد می‌کند، احتمال این‌که نقطه‌ی تماس روی گوشه‌ها قرار گیرد زیاد است، و گوشه‌ها همان نقاطی هستند که یک یا چند مؤلفه w_i برابر صفر می‌شوند.

دیدگاه احتمالاتی

در تفسیر بیزی، لاسو متناظر با یک پیش‌فرض لاپلاس (Laplace prior) روی وزن‌ها است:

$$p(w_i) \propto e^{-\lambda|w_i|}.$$

این پیش‌فرض شدیداً وزن‌های کوچک را ترجیح می‌دهد و احتمال زیادی برای مقدار دقیقاً صفر فراهم می‌کند. در مقابل، ریج از پیش‌فرض گاوسی استفاده می‌کند:

$$p(w_i) \propto e^{-\frac{\lambda}{2}w_i^2},$$

که وزن‌ها را به صفر *نزدیک* می‌کند اما به ندرت مقدار آن‌ها را دقیقاً صفر می‌سازد.

چه کاربردهایی برای منظم‌سازی ℓ_1 وجود دارد؟

- انتخاب ویژگی (Feature Selection) به صورت خودکار
- ایجاد مدل‌های ساده‌تر و قابل‌تفسیرتر
- کاربرد در داده‌های با ابعاد بالا (High-dimensional data) مانند ژنومیکس
- جلوگیری از بیش‌برازش در مدل‌هایی با تعداد ویژگی زیاد

ت) تحلیل رفتاری منظم‌ساز ℓ_2 (روش ریج)

منظم‌ساز ℓ_2 به صورت زیر است:

$$\lambda\|w\|_2^2 = \lambda \sum_i w_i^2.$$

این منظم‌ساز وزن‌ها را کوچک می‌کند اما هرگز آن‌ها را به صفر نمی‌رساند. هندسی آن یک کره (ball) است و هیچ گوشه‌ای ندارد؛ بنابراین نقطه بهینه معمولاً روی محورها نمی‌افتد.

مزایای ℓ_2

- مناسب برای داده‌های چندخطی (multicollinearity)
- توزیع هموارتر وزن‌ها
- بهبود پایداری مدل
- هیچ ویژگی حذف نمی‌شود، بنابراین برای وظایف پیش‌بینی (نه تفسیر) بهتر است