یادگیری ماشین

دکتر منصور رزقی

تمرین اول - پاسخ تمارین نظری

دانشکده علوم ریاضی، گروه علوم کامپیوتر، گرایش داده‌کاوی

۱ تمرین ۱: واریانس مجموع متغیرهای تصادفی

۱.۱ صورت مسئله

نشان دهید که برای دو متغیر تصادفی X و Y داریم:

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y) \tag{۱}$$

۲.۱ حل

از تعریف واریانس شروع می‌کنیم:

$$\text{var}(X+Y) = \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2 \tag{۲}$$

ابتدا جمله اول را بازنویسی می‌کنیم:

$$\begin{aligned}\mathbb{E}[(X+Y)^2] &= \mathbb{E}[X^2 + 2XY + Y^2] \\ &= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2]\end{aligned} \tag{۳}$$

و جمله دوم:

$$\begin{aligned}(\mathbb{E}[X+Y])^2 &= (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\ &= (\mathbb{E}[X])^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + (\mathbb{E}[Y])^2\end{aligned} \tag{۴}$$

حال با جایگذاری:

$$\begin{aligned}\text{var}(X+Y) &= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] \\ &\quad - (\mathbb{E}[X])^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - (\mathbb{E}[Y])^2 \\ &= \left(\mathbb{E}[X^2] - (\mathbb{E}[X])^2\right) + \left(\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2\right) \\ &\quad + 2\left(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\right) \\ &= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y)\end{aligned} \tag{۵}$$

۲ تمرین ۲: کاهش مسئلهٔ الاستیک نت به لاسو

۱.۲ صورت مسئله

نشان دهید که با تعریف:

$$J_1(\mathbf{w}) = \|\mathbf{y} - X\mathbf{w}\|_2^2 + \lambda_2\|\mathbf{w}\|_2^2 + \lambda_1\|\mathbf{w}\|_1 \tag{۶}$$

$$J_2(\mathbf{w}) = \|\tilde{\mathbf{y}} - \tilde{X}\mathbf{w}\|_2^2 + c\lambda_1\|\mathbf{w}\|_1 \tag{۷}$$

که در آن $c = (1+\lambda_2)^{-1/2}$ و:

$$\tilde{X} = c\begin{pmatrix} X \\ \sqrt{\lambda_2} I_d \end{pmatrix}, \quad \tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_{d\times 1} \end{pmatrix} \tag{۸}$$

آنگاه داریم:

$$\arg\min J_1(\mathbf{w}) = c\,(\arg\min J_2(\mathbf{w})) \tag{۹}$$

۲.۲ حل

ابتدا J_1 را بازنویسی می‌کنیم. حال با جایگذاری $\mathbf{w} = c\mathbf{u}$ (که در آن $c = (1+\lambda_2)^{-1/2}$):

$$\begin{aligned} J_1(c\mathbf{u}) &= \|\mathbf{y} - Xc\mathbf{u}\|_2^2 + \lambda_2\|c\mathbf{u}\|_2^2 + \lambda_1\|c\mathbf{u}\|_1 \\ &= \|\mathbf{y} - cX\mathbf{u}\|_2^2 + \lambda_2 c^2\|\mathbf{u}\|_2^2 + c\lambda_1\|\mathbf{u}\|_1 \end{aligned} \tag{۱۰}$$

اکنون جمله اول را بسط می‌دهیم. می‌دانیم که:

$$\begin{aligned} \|\mathbf{y} - cX\mathbf{u}\|_2^2 + \lambda_2 c^2\|\mathbf{u}\|_2^2 &= \left\| \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} cX \\ c\sqrt{\lambda_2} I_d \end{pmatrix} \mathbf{u} \right\|_2^2 \\ &= \|\tilde{\mathbf{y}} - \tilde{X}\mathbf{u}\|_2^2 \end{aligned} \tag{۱۱}$$

بنابراین:

$$J_1(c\mathbf{u}) = \|\tilde{\mathbf{y}} - \tilde{X}\mathbf{u}\|_2^2 + c\lambda_1\|\mathbf{u}\|_1 = J_2(\mathbf{u}) \tag{۱۲}$$

در نتیجه می‌توان مسئلهٔ الاستیک نت را با حل‌کنندهٔ لاسو برای داده‌های تغییر یافته حل کرد.

۳ تمرین ۳: مدل برنولی–گوسی و منظم‌سازی ℓ_0

۱.۳ صورت مسئله

توضیح دهید چگونه مدل برنولی–گوسی (Bernoulli-Gaussian) (prior برای وزن‌های w_i منجر به استفاده از نُرم ℓ_0 به‌عنوان منظم‌سازی می‌شود.

۲.۳ حل

فرض کنید برای هر وزن w_i داریم:

$$w_i = \begin{cases} \mathcal{N}(0, \sigma^2) & \text{با احتمال } \pi \\ 0 & \text{با احتمال } 1-\pi \end{cases} \tag{۱۳}$$

می‌توان این را به صورت زیر نوشت:

$$p(w_i) = (1-\pi)\delta(w_i) + \pi\mathcal{N}(w_i|0, \sigma^2) \tag{۱۴}$$

که در آن δ تابع دلتای دیراک است.
با استفاده از رویکرد MAP (Maximum Posteriori): A

$$\begin{aligned} \hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{y}) \\ &= \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{w})p(\mathbf{w}) \\ &= \arg\max_{\mathbf{w}} \log p(\mathbf{y}|\mathbf{w}) + \log p(\mathbf{w}) \end{aligned} \tag{۱۵}$$

برای prior داریم:

$$\log p(\mathbf{w}) = \sum_i \log p(w_i) \tag{۱۶}$$

برای π کوچک، وقتی $w_i \neq 0$، جریمه‌ای ثابت به ازای هر وزن غیرصفر اضافه می‌شود. در نتیجه:

$$\log p(\mathbf{w}) \approx -\lambda\|\mathbf{w}\|_0 \tag{۱۷}$$

که $\|\mathbf{w}\|_0$ تعداد عناصر غیرصفر $\mathbf{w}$ است.
نتیجه: مدل برنولی– گوسی منجر به منظم‌سازی ℓ_0 می‌شود که تعداد وزن‌های غیرصفر را جریمه می‌کند.

۴ تمرین ۴: پرسش مفهومی دربارهٔ لاسو و ریج

۱.۴ چرا ℓ_1 منجر به sparse شدن می‌شود؟

۱.۱.۴ دیدگاه هندسی

در بهینه‌سازی مقید:

$$\min \|\mathbf{y} - X\mathbf{w}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{w}\|_1 \leq t \tag{۱۸}$$

ناحیهٔ قابل قبول برای ℓ_1 (الماس در ۲ بعد) دارای گوشه‌های تیز روی محورها است. سطوح هم‌ارز تابع هدف (بیضی‌ها) معمولاً در این گوشه‌ها با ناحیهٔ مجاز برخورد می‌کنند، که این نقاط دارای مختصات صفر هستند.
در مقابل، ℓ_2 (دایره/کره) گوشهٔ تیزی ندارد و کمتر منجر به صفر شدن دقیق ضرایب می‌شود.

۲.۱.۴ دیدگاه احتمالاتی

- **منظم‌ساز** ℓ_1: معادل prior Laplace است:

$$p(w_i) \propto \exp(-\lambda |w_i|)$$

- **منظم‌ساز** ℓ_2: معادل prior Gaussian است:

$$p(w_i) \propto \exp(-\lambda w_i^2)$$

توزیع Laplace در صفر دارای چگالی بیشتر و دم‌های سنگین‌تر است، پس ترجیح می‌دهد بسیاری از وزن‌ها دقیقاً صفر باشند.

۲.۴ کاربردهای sparsity

- **انتخاب ویژگی**: شناسایی مهم‌ترین ویژگی‌ها
- **تفسیرپذیری**: مدل ساده‌تر با تعداد کمتر متغیر
- **کاهش overfitting**: کم کردن پیچیدگی مدل
- **کاهش هزینهٔ محاسباتی**: ذخیره‌سازی و محاسبهٔ سریع‌تر

۳.۴ تحلیل ℓ_2 (Ridge)

منظم‌ساز ℓ_2:

- وزن‌ها را کوچک می‌کند ولی به‌ندرت دقیقاً صفر می‌شوند
- برای مسائل ill-conditioned مناسب است
- در حضور multicollinearity عملکرد بهتری دارد
- دارای جواب closed-form است
- همهٔ ویژگی‌ها را نگه می‌دارد (غیر_sparse)