

تئوری تخمین (۱-۲۵۱۶۳)



تمرین سری ششم

بهار ۱۴۰۴-۱۴۰۳

دانشکده‌ی مهندسی برق

دانشگاه صنعتی شریف

اساتید: روح الله امیری و محمدمهدی مجاهدیان

موعد تحویل: دوشنبه ۳۱ خرداد

۱ تخمینگر بیشترین درست‌نمایی برای رگرسیون

در یک مسئله رگرسیون خطی مدلی که داده‌ها از روی آن تولید شده‌اند به صورت زیر است:

$$f(y|x, \theta) = \mathcal{N}(y|x, \theta).$$

تخمین بیشترین درست‌نمایی $\hat{\theta}_{ML}$ را به دست آورید؟

۲ تجزیه‌ی بایاس-واریانس

فرض کنید مدل تولید داده به صورت زیر باشد:

$$y_i = f(\mathbf{x}_i) + \epsilon_i,$$

که در آن ϵ_i به صورت مستقل و با توزیع یکسان از $\mathcal{N}(0, \sigma^2)$ می‌آیند. فرض کنید از روی مجموعه داده‌های آموزشی

$$\mathcal{D} = \{(\mathbf{x}_i, y_i), \quad i = 1, \dots, n\}$$

تابع $\hat{f}(\mathbf{x})$ تخمین زده شده است. خطای متوسط پیش‌بینی برای داده‌ی جدید را به صورت زیر تعریف می‌کنیم.

$$\text{Err}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}, \epsilon} \left[\left(y - \hat{f}(\mathbf{x}) \right)^2 \right].$$

۱. این خطا را به ترم‌های بایاس و واریانس تجزیه کنید؟

۲. فرض کنید مدل تولید داده به صورت خطی باشد. یعنی داریم:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

فرض کنید از روی داده‌ی آموزشی زیر

$$\mathcal{D} = \{(x_i, y_i), \quad i = 1, \dots, n\}$$

به طوری که

$$\frac{1}{n} \sum_{i=1}^n x_i = 0$$

$$\frac{1}{n} \sum_{i=1}^n x_i^2 = s_x^2.$$

باشد. دو تخمین‌گر زیر به دست آمده باشند:

$$\hat{f}_1(x) = \frac{1}{n} \sum_{i=1}^n y_i$$

$$\hat{f}_2(x) = \hat{\beta}_0 + \hat{\beta}_1 x,$$

که در رگرسیون خطی پارامترها بر اساس بیشترین درستنمایی به دست آمده‌اند. بایاس و واریانس را برای این دو تخمین‌گر محاسبه کرده و ارتباط پیچیدگی مدل با بایاس و واریانس را تفسیر نمایید.

۳. برای مدل تولید داده‌ی خطی در قسمت قبل و استفاده از رگرسیون سه‌تیغی^۱ با پارامتر λ بایاس و واریانس را حساب کرده و نتیجه را برحسب تغییرات λ تفسیر نمایید.

۴. به مقدار متوسط خطای پیش‌بینی روی توزیع ورودی خطای تعمیم گفته می‌شود. با فرض مدل داده‌ی زیر

$$y_i = \sin(\pi x_i) + \epsilon_i.$$

و تخمین‌گر

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n y_i.$$

خطای تعمیم را با فرض توزیع یکنواخت ورودی x روی بازه‌ی $[0, 1]$ حساب کنید.

۳ پس انتشار

شبکه‌ی عصبی با یک لایه‌ی پنهان و ورودی $\mathbf{x} \in \mathbb{R}^d$ را در نظر گرفته که رابطه‌ی مقادیر گره‌ها به صورت زیر است:

$$\mathbf{h} = \sigma(\mathbf{U}\mathbf{x} + \mathbf{b})$$

$$\hat{y} = \mathbf{v}^T \mathbf{h} + \mathbf{c},$$

که در آن تابع فعال‌ساز^۲ به صورت زیر است:

$$\sigma(z) = \max(z, 0),$$

و تابع هزینه را به صورت زیر در نظر می‌گیریم:

$$\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2.$$

۱. مشتقات $\frac{\partial \ell}{\partial \mathbf{v}}, \frac{\partial \ell}{\partial \mathbf{c}}, \frac{\partial \ell}{\partial \mathbf{w}}, \frac{\partial \ell}{\partial \mathbf{b}}$ را به دست آورید؟

۲. به تابع فعال‌ساز بالا یکسوساز خطی^۳ گفته می‌شود. توضیح دهید چرا عدم مشتق‌پذیری این تابع در $z = 0$ در روش‌های مبتنی بر گرادیان در دسرساز نخواهد شد؟

۴ منظم‌سازی پنهان در روش کاهش گرادیان

(نیاز به تحویل این سؤال نیست)

فرض کنید β_{GD} راه حل

$$\arg \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2$$

به دست آمده از روش کاهش گرادیان با اندازه‌ی گام η و نقطه‌ی اولیه‌ی $\beta^{(0)} = 0$ باشد. به سؤال‌های زیر جواب دهید:

۱. نشان دهید اگر تعداد گام‌ها به سمت بی‌نهایت میل کند جواب روش کاهش گرادیان به جواب بیشترین درستنمایی میل می‌کند؟

۲. نشان دهید توقف در تعداد گام محدود معادل رگرسیون سه‌تیغی است و پارامتر λ مربوطه را به دست آورید؟

۳. نتیجه‌گیری شما برای آموزش شبکه‌ی عصبی با روش‌های مبتنی بر گرادیان چیست؟

^۱ Ridge regression

^۲ Activation function

^۳ Rectified Linear Unit (ReLU)

فرض کنید d مشاهده مستقل $y_i \in \mathbb{R}$ از پارامترها $\theta_i \in \mathbb{R}$ به صورت زیر داریم:

$$y_i \sim \mathcal{N}(\theta_i, 1), \quad i = 1, \dots, d.$$

تخمینگر زیر از پارامترها را در نظر بگیرید:

$$\hat{\theta} = \left(1 - \frac{c}{\|y\|^2}\right) y.$$

۱. با استفاده از لم اشتاین مقدار خطای تخمین زیر را محاسبه نمایید؟ (راهنمایی: بردار y را به عبارت داخل نرم اضافه و کم کرده و از لم اشتاین استفاده کنید!)

$$\mathbb{E} \left[\|\hat{\theta} - \theta\|^2 \right] = ?$$

۲. مقدار بهینه c برای کمینه کردن خطا را محاسبه نمایید؟

۳. از مقایسه تخمینگر بالا با تخمینگر بیشترین درستنمایی چه نتیجه‌ای می‌گیرید؟