یادگیری ماشین

نیم‌سال دوم ۱۴۰۲-۱۴۰۱

مدرس: دکتر سید ابوالفضل مطهری

دانشکده‌ی مهندسی کامپیوتر

درس‌نامه سوم

همانطور که بیان شد، در مسئله رگرسیون خروجی‌ها از جنس عددی هستند. برای سادگی فرض می‌نماییم که ورودی‌ها نیز عددی هستند. در حالت ساده ورودی را یک بعدی و خروجی را نیز یک بعدی در نظر می‌گیریم. همچنین واقعیت را به صورت زیر فرض می‌کنیم.

x → واقعیت → y

$$y = \beta_0^* + \beta_1^* x + \epsilon$$

که در آن ϵ یک مقدار تصادفی است و مستقل از ورودی می‌باشد. همچنین: $\mathbb{E}[\epsilon] = 0$

با توجه به داده‌های موجود $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ می‌خواهیم برای یک ورودی جدید x خروجی مناسب $\hat{y}$ را داشته باشیم. بنابراین هدف را قرار می‌دهیم که $\mathbb{E}[(y - \hat{y})^2]$ کوچک باشد و همچنین رابطه بین $\hat{y}$ و x را خطی در نظر می‌گیریم: $\hat{y} = \beta_0 + \beta_1 x$. در نتیجه به دنبال حل مسئله زیر هستیم:

$$\mathbb{E}[(y - \beta_0 - \beta_1 x)^2] \tag{1}$$

اگر توزیع مشترک x و y یعنی $\mathbb{P}(x, y)$ را در اختیار داشتیم، می‌توانستیم مسئله فوق را حل نماییم. ولی باید از روی داده این کار را انجام دهیم. طبق قانون اعداد بزرگ، اگر $\hat{\beta}_0$ و $\hat{\beta}_1$ را به دست آوریم، از روی m نمونه می‌توانیم تخمین خوبی از رابطه ۱ داشته باشیم. یعنی:

$$\mathbb{E}[(y - \hat{\beta}_0 - \hat{\beta}_1 x)^2] \approx \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

بنابراین m نمونه را برای ارزیابی و آزمون نگاه می‌داریم. برای رسیدن به $\hat{\beta}_0$ و $\hat{\beta}_1$ از بقیه نمونه‌ها که با n نشان می‌دهیم استفاده می‌کنیم:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1)}{\operatorname{argmin}} \operatorname{RSS}(\beta_0, \beta_1)$$

که در آن باقیمانده مجموع مربعات به صورت زیر تعریف می‌گردد:

$$\operatorname{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

برای یافتن $\hat{\beta}_0$ و $\hat{\beta}_1$ از رابطه فوق مشتق گرفته و برابر صفر قرار می‌دهیم:

$$\frac{\partial \operatorname{RSS}}{\partial \beta_0}\Big|_{\hat{\beta}_0, \hat{\beta}_1} = 0$$

$$\frac{\partial \mathrm{RSS}}{\partial \beta_1}|_{\hat{\beta}_0,\hat{\beta}_1} = 0$$

بنابراین خواهیم داشت:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

که به صورت یک دستگاه معادلات قابل بیان است:

$$n\hat{\beta}_0 + (\sum_{i=1}^{n} x_i)\hat{\beta}_1 = \sum_{i=1}^{n} y_i$$

$$(\sum_{i=1}^{n} x_i)\hat{\beta}_0 + (\sum_{i=1}^{n} x_i^2)\hat{\beta}_1 = \sum_{i=1}^{n} x_i y_i$$

از رابطه اول داریم:

$$\hat{\beta}_0 = \underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}_{\bar{y}} - \hat{\beta}_1 \underbrace{(\frac{1}{n}\sum_{i=1}^{n} x_i)}_{\bar{x}} = \bar{y} - \hat{\beta}_1 \bar{x}$$

اگر رابطه فوق را در معادله دوم قرار دهیم، خواهیم داشت:

$$\bar{x}\hat{\beta}_0 + (\frac{1}{n}\sum x_i^2)\hat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \implies \bar{x}\bar{y} - \hat{\beta}_1 \bar{x}^2 + \hat{\beta}_1 \overline{x^2} = \overline{xy}$$

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

بنابراین مقادیر $\hat{\beta}_0$ و $\hat{\beta}_1$ به دست می‌آیند. حال به مبحث ارزیابی می‌رسیم.

۱ ارزیابی

در ابتدا m نمونه را برای آزمون نگاه داشتیم. اگر $\frac{1}{n}\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_1)$ که از درون نمونه می‌آید، با مقدار $\frac{1}{m}\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_1)$ که از نمونه‌های آزمون می‌آید نزدیک بود، آنگاه تعمیم صورت پذیرفته و احتمالا رابطه قابل تعمیمی را به دست آورده‌ایم.

با توجه به آن که مدل واقعی را می‌دانیم، می‌توانیم رابطه اصلی را نیز بررسی کنیم. یعنی

$$L = \mathbb{E}[(y - \hat{y})^2] = \mathbb{E}[l(y, \hat{y})]$$

$$\mathbb{E}[l(y, \hat{y})] = \mathbb{E}[(\beta_0^* - \beta_0 + (\beta_1^* - \beta_1)x + \epsilon)^2]$$

توجه کنید که β_0 و β_1 به داده مرتبط می‌باشند و خود متغیر تصادفی هستند. در نتیجه اگر متوسط آن‌ها را به ترتیب با $\mathbb{E}[\beta_0]$ و $\mathbb{E}[\beta_1]$ نشان دهیم، خواهیم داشت:

$$L = \mathbb{E}[\beta_0^* - \mathbb{E}[\beta_0] + (\beta_1^* - \mathbb{E}[\beta_1])x + \mathbb{E}[\beta_0] - \beta_0 + (\mathbb{E}[\beta_1] - \beta_1)x + \epsilon]^2$$

$$= \mathbb{E}[[\beta_0^* - \mathbb{E}[\beta_0] + (\beta_1^* - \mathbb{E}[\beta_1])x]^2] + \mathbb{E}[[(\mathbb{E}[\beta_0] - \beta_0) + (\mathbb{E}[\beta_1] - \beta_1)x]^2] + \sigma_\epsilon^2 = [\mathrm{bias}(\beta_0, \beta_1)]^2 + \mathrm{var}[\beta_0, \beta_1] + \sigma_\epsilon^2$$

به رابطه فوق مصالحه بین بایاس و واریانس[1] اطلاق می‌گردد. دقت نمایید که در رابطه فوق β_0 و β_1 هر تخمین‌گری می‌تواند باشد و بنابراین بهترین تخمین باید سعی نماید که جمع هر دو را کمینه کند. برای مثال:

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

از طرفی

$$\overline{xy} = \frac{1}{n}\sum x_i y_i = \frac{1}{n}\sum x_i(\beta_0^* + \beta_1^* x_i + \epsilon_i) = \beta_0^*\bar{x} + \beta_1^*\bar{x^2} + \frac{1}{n}\sum x_i\epsilon_i$$

همچنین

$$\bar{x}\bar{y} = \bar{x}(\frac{1}{n}\sum_{i=1}^{n} y_i) = \bar{x}(\beta_0^* + \beta_1^*\bar{x} + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i) = \beta_0^*\bar{x} + \beta_1^*\bar{x}^2 + (\frac{1}{n}\sum_{i=1}^{n}\epsilon_i)\bar{x}$$

بنابراین خواهیم داشت:

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \beta_1^* + \frac{\frac{1}{n}\sum x_i\epsilon_i - (\frac{1}{n}\sum\epsilon_i)\bar{x}}{\bar{x^2} - \bar{x}^2}$$

بنابراین:

$$\mathbb{E}[\hat{\beta}_1] = \beta_1^*$$

از طرف دیگر:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

بنابراین:

$$\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\hat{y}] - \mathbb{E}[\hat{\beta}_1\bar{x}] = \beta_0^*$$

روابط فوق نشان می‌دهند که $(\hat{\beta}_1, \hat{\beta}_0)$ تخمین‌های ناااریب از β_0^* و β_1^* می‌باشند و در نتیجه داریم:

$$L = \mathrm{var}[\hat{\beta}_1, \hat{\beta}_0] + \sigma_\epsilon^2$$

حال سوال این است که آیا کمینه‌کردن بایاس همواره بهتر است؟ در آینده خواهیم دید که کم‌کردن بایاس همواره مطلوب نیست.

[1] Bias-Variance Trade-off

۲ استنتاج

ارزیابی‌های صورت‌گرفته در قسمت قبل همگی بر پایه تعمیم و پیش‌بینی بوده است. در این قسمت دیدگاه را عوض نموده و هدف را تخمین پارامترها می‌گذاریم. این بدان معنی است که کیفیت تخمین $\hat{\beta}_0$ و $\hat{\beta}_1$ برای ما اهمیت دارد. همچنین از روی این تخمین‌ها می‌خواهیم به رابطه درون سیستم پی ببریم. مثلا متوجه شویم که آیا x و y با یکدیگر رابطه دارند یا خیر. در قسمت قبل دیدیم که $\mathbb{E}[\hat{\beta}_1] = \beta_1^*$ و $\mathbb{E}[\hat{\beta}_0] = \beta_0^*$ می‌باشند. حال می‌خواهیم واریانس آن ها را نیز محاسبه کنیم (x را مفروض در نظر می‌گیریم).

$$(\mathrm{SE}(\hat{\beta}_1))^2 = \mathbb{E}[(\hat{\beta}_1 - \beta_1^*)^2]$$

$$= \mathbb{E}[(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \beta_1^*)^2] = \mathbb{E}[(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^{n}(x_i - \bar{x})^2})^2] = \frac{\sigma_\epsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

از طرف دیگر برای $\hat{\beta}_0$ داریم:

$$(\mathrm{SE}(\hat{\beta}_0))^2 = \mathbb{E}[(\hat{\beta}_0 - \beta_0^*)^2] = \mathbb{E}[(\bar{y} - \hat{\beta}_1\bar{x} - \beta_0^*)^2] = \mathbb{E}[((\beta_1^* - \hat{\beta}_1)\bar{x} + \bar{\epsilon})^2] = \frac{\sigma_\epsilon^2 \bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\sigma_\epsilon^2}{n}$$

رابطه‌های فوق نشان می‌دهند که اگر n به سمت بی‌نهایت میل کند، آنگاه واریانس‌ها به سمت صفر میل می‌کنند و تخمین دقیق خواهد بود. در روابط فوق $\mathrm{SE}(\hat{\beta}_0)$ و $\mathrm{SE}(\hat{\beta}_1)$ را نمی‌توان محاسبه کرد و دلیل آن این است که σ_ϵ^2 را معمولا در اختیار نداریم. بدین منظور σ_ϵ^2 را تخمین می‌زنیم.

$$\hat{\sigma}_\epsilon^2 = \frac{\mathrm{RSS}}{n-2} \tag{۲}$$

دلیل وجود $n - 2$ در مخرج آن است که تخمین σ_ϵ^2 نا اریب گردد. یعنی:

$$\mathbb{E}[\hat{\sigma}_\epsilon^2 = \sigma_\epsilon^2 = \mathrm{RSE} \quad \text{(sum of errors)}$$

با قرار دادن ۲ در روابط $\mathrm{SE}(\hat{\beta}_0)$ و $\mathrm{SE}(\hat{\beta}_1)$ به تخمین‌های $\hat{\mathrm{SE}}(\hat{\beta}_0)$ و $\hat{\mathrm{SE}}(\hat{\beta}_1)$ می‌رسیم. اگر توزیع ϵ گوسی باشد، آنگاه می‌توانیم در مورد توزیع $\hat{\beta}_0$ و $\hat{\beta}_1$ نیز صحبت کنیم. در این صورت داریم:

$$\hat{\beta}_0 \sim \mathcal{N}(\beta_0^*, (\mathrm{SE}(\hat{\beta}_0))^2)$$

$$\hat{\beta}_1 \sim \mathcal{N}(\beta_1^*, (\mathrm{SE}(\hat{\beta}_1))^2)$$

و می‌توانیم بگوییم که بازه

$$\hat{\beta}_0 \pm 2\mathrm{SE}(\hat{\beta}_0)$$

با احتمال ۹۵ درصد β_0^* را شامل می‌شود و به طور مشابه

$$\hat{\beta}_1 \pm 2\mathrm{SE}(\hat{\beta}_1)$$

با احتمال ۹۵ درصد β_1^* را شامل می‌گردد. البته به دلیل آن که $\mathrm{SE}(\hat{\beta}_0)$ و $\mathrm{SE}(\hat{\beta}_1)$ را در اختیار نداریم، از تخمین آن‌ها استفاده می‌کنیم.

آزمون فرض: می‌خواهیم تصمیم بگیریم که آیا مثلا x و y با هم رابطه‌ای دارند یا خیر. در این شرایط می‌توانیم آزمون زیر را در نظر بگیریم:

$$H_0: \quad \beta_1^* = 0$$

$$H_a: \quad \beta_1^* \neq 0$$

اگر $\hat{\beta}_1$ را آماره موردنظر فرض نماییم، آنگاه $\hat{\beta}_1$ دارای توزیع گوسی است و

$$z = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

تحت فرض صفر دارای توزیع گوسی با متوسط صفر و واریانس یک می‌باشد. در این شرایط می‌توانیم z را با یک مقدار مقایسه نماییم و فرض صفر را قبول و یا رد نماییم.

با توجه به آن که مقدار $\mathrm{SE}(\hat{\beta}_1)$ را در اختیار نداریم، می‌توانیم از تخمین آن استفاده نماییم که در این صورت توزیع $t - \mathrm{student}$ با $n - 2$ درجه آزادی خواهیم داشت:

$$t = \frac{\hat{\beta}_1 - 0}{\hat{\mathrm{SE}}(\hat{\beta}_1)} \tag{3}$$

و می‌توانیم مقدار فوق را با یک آستانه مقایسه نماییم. به طور معادل می‌توانیم مقدار P را بازگردانیم:

$$P = \mathbb{P}[|T| > t]$$

که در آن T توزیع $t - \mathrm{student}$ با $n - 2$ درجه آزادی می‌باشد و t مقدار مشاهده‌شده در ۳.

۳ آماره R^2

می‌توانیم کل مدل را نیز ارزیابی نماییم. یعنی به این سوال پاسخ دهیم که «آیا $f_{\hat{\beta}}(x)$ مناسب است؟». برای این کار می‌توانیم از مفهوم تغییرات استفاده کنیم. داده خروجی دارای تغییرات مشخصی است:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathrm{TSS}$$

اگر تخمین $\hat{y}$ از y عالی باشد، آنگاه پس از کم‌کردن $\hat{y}$ از y دیگر تغییراتی دیده نمی‌شود. به طور کلی مقدار زیر، مقدار تغییراتی است که در خروجی پس از استفاده از $f_{\hat{\beta}}(x)$ هنوز پابرجاست:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \mathrm{RSS}$$

بنابراین توانسته‌ایم به اندازه $\text{TSS} - \text{RSS}$ تغییرات داخل y را توجیه نماییم. به همین خاطر مقدار $R^{۲}$ را به شکل زیر تعریف می‌کنیم:

$$R^{۲} = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = ۱ - \frac{\text{RSS}}{\text{TSS}}$$

توجه کنید که $R^{۲}$ همواره بین صفر و یک است. اگر $R^{۲} = ۰$ آنگاه رگرسیون نتوانسته است تفسیری از رابطه x و y داشته باشد و اگر $R^{۲} = ۱$ آنگاه از روی x می‌توان y را به طور کامل به دست آورد.