دانشگاه اصفهان - رشته علوم کامپیوتر

یادگیری ماشین

پاسخ تمرین ۱: مدل‌های خطی و روش‌های ارزیابی آن‌ها

شماره دانشجویی:

نام و نام خانوادگی:

پرسش ۱.

می‌دانیم پس از آن‌که یک مدل یادگیرنده مرحله آموزش خود را سپری کرد، باید توسط چندین معیار، ارزیابی و سنجیده شود. همچنین اگر مسئله موردنظر از نوع رگرسیون باشد، خروجی مدل یک بردار مانند $\hat{Y}$ و مقادیر برچسب نیز درون بردار دیگری مانند Y قرار دارند که همگی از نوع اعداد پیوسته هستند.

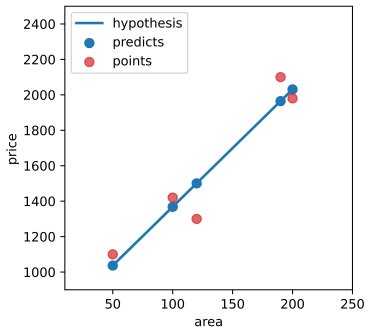
$$Y = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \hat{Y} = \begin{bmatrix} \hat{y}_0 \\ \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

به عنوان مثال یکی از معیارهای ارزیابی مدل‌های رگرسیونی، میانگین مربع خطا [1] می‌باشد که به صورت زیر بیان می‌شود:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{۱}$$

الف. مجموعه داده‌ای از مساحت خانه و قیمت آن به همراه پیش‌بینی مدل به شما داده شده است (جدول ۱). در شکل ۱ نقاط این مجموعه داده با درنظر گرفتن مساحت خانه در محور x و قیمت آن در محور y به همراه فرضیه h که یک خط تخمین زده شده توسط الگوریتم رگرسیون خطی است، نشان داده شده است. باتوجه به خط فرضیه تخمین زده شده، هر نقطه از این مجموعه داده که به خط فرضیه داده شود، دارای خطای e_i است. مقدار MSE را برای این مجموعه داده محاسبه کرده و مقدار حاصل از آن را توصیف کنید و توضیح دهید که این معیار ارزیابی چه معایب و مزایایی دارد. (می‌توانید با ارائه یک مثال، توضیحات کامل‌تری ارائه دهید)

[1] Mean Square Error



شکل ۱: نمایش دو بُعدی داده‌ها به همراه فرضیه تخمین زده شده. برای اطلاعات بیشتر و کد زدن با این مجموعه داده، می‌توانید از کدی که در [این] آدرس نوشته شده است استفاده کنید.

جدول ۱: مجموعه داده قیمت خانه‌ها

مساحت خانه	قیمت واقعی	قیمت پیش‌بینی شده
۱۹۰	۲۱۰۰	۱۹۶۴
۲۰۰	۱۹۸۰	۲۰۳۱
۱۰۰	۱۴۲۰	۱۳۶۸
۵۰	۱۱۰۰	۱۰۳۷
۱۲۰	۱۳۰۰	۱۵۰۰

ب. دو معیار ارزیابی دیگر برای مسائل رگرسیون انتخاب کرده و مقدار آن‌ها را برای مجموعه داده جدول ۱ محاسبه کنید. همچنین توضیح دهید که هرکدام از چه رابطه‌ای بدست می‌آیند و معایب و مزایای آن‌ها نسبت به یکدیگر چگونه است.

پاسخ.

الف. با توجه به جدول ۱، بردارهای Y و $\hat{Y}$ به ترتیب برابرند با:

$$Y = \begin{bmatrix} 2100 \\ 1980 \\ 1420 \\ 1100 \\ 1300 \end{bmatrix}, \quad \hat{Y} = \begin{bmatrix} 1964 \\ 2031 \\ 1368 \\ 1037 \\ 1500 \end{bmatrix}$$

حال با توجه به رابطه ۱ مقدار میانگین مربع خطا برابرست با:

$$\begin{aligned} MSE(Y, \hat{Y}) &= \frac{1}{5}[(1964 - 2100)^2 + (2031 - 1980)^2 + (1368 - 1420)^2 + (1037 - 1100)^2 + (1500 - 1300)^2] \\ &= \frac{1}{5}[18496 + 2601 + 2704 + 3969 + 40000] \\ &= \frac{1}{5}[67770] \\ &= 13554 \end{aligned}$$

اولین مزیت این تابع هزینه، فهم آسان آن است. این تابع هزینه از مجذور اختلاف مقدار پیش‌بینی شده و برچسب بدست می‌آید که درواقع برای هر خطای بدست آمده، مساحت ناحیه اختلاف این دو مقدار را بیان می‌کند. دومین و یکی از مهم‌ترین مزیت‌های این تابع، **مشتق‌پذیر** بودن آن است. می‌دانیم که در بسیاری از الگوریتم‌های یادگیری، از الگوریتم گرادیان کاهشی برای بهینه‌سازی مسئله استفاده می‌کنیم در این

روش برای بروزرسانی پارامترهای قابل یادگیری مدل، نیاز است که در هر مرحله مشتق تابع هزینه نسبت به تمامی پارامترهای قابل یادگیری[۲] مدل محاسبه شود. اما از معایب این تابع، می‌توان به حساس بودن به داده‌های پرت اشاره کرد. در بسیاری از مسائل یادگیری، داده‌ها به همراه یک نویز جمع‌آوری می‌شوند که این پدیده ممکن است روی مقدار بدست آمده از این تابع تأثیر چشم‌گیری بگذارد. از طرفی این تابع، به شدت به مقیاس داده‌ها نیز حساس است. بنابراین بهتر است از روش‌های نرمال‌سازی داده متناسب با مسئله نیز استفاده کرد.

ب.

معیار R^2.

رابطه این معیار ارزیابی برابرست با:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{۲}$$

که در این‌جا SS_{res} مجموع مربعات باقیمانده‌ها و SS_{tot} مجموع مجذورات متناسب با واریانس داده‌ها می‌باشد که هرکدام نیز برابرند با:

$$SS_{res} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{۳}$$

$$SS_{tot} = \sum_{i=1}^{N} (y_i - \bar{y})^2, \qquad \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{۴}$$

که در این رابطه، پراکندگی داده‌ها نسبت به میانگین آن‌ها محاسبه شده و سپس نسبت گرفته می‌شود. خروجی این معیار طبیعتاً عددی بین ۰ و ۱ است و هرچه این عدد به سمت ۱ میل کند، به معنای حداکثر واریانس[۳] و حداقل بایاس[۴] و هرچقدر به سمت ۰ میل کند، برعکس. به عبارت دیگر، هرچقدر این معیار به سمت ۱ برود، احتمال آن‌که مدل تخمین زده شده دچار بیش‌برازش[۵] شده است بیشتر می‌شود. مقدار این معیار برای جدول ۱ برابرست با:

$$\begin{aligned}
SS_{res} &= [(1964-2100)^2 + (2031-1980)^2 + (1368-1420)^2 + (1037-1100)^2 + (1500-1300)^2] \\
&= [18496 + 2601 + 2704 + 3969 + 40000] \\
&= 67770 \\
\bar{y} &= \frac{1}{5}[2100 + 1980 + 1420 + 1100 + 1300] = \frac{1}{5}[7900] = 1580 \\
SS_{tot} &= [(2100-1580)^2 + (1980-1580)^2 + (1420-1580)^2 + (1100-1580)^2 + (1300-1580)^2] \\
&= [270400 + 160000 + 25600 + 230400 + 78400] \\
&= 764800
\end{aligned}$$

بنابراین داریم:

$$\begin{aligned}
R^2 &= 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{67770}{764800} \\
&= 1 - 0.08861140 \\
&= 0.91138859
\end{aligned}$$

[2] Trainable Parameters
[3] High Variance
[4] Low Bias
[5] Overfitting

معیار *NMSE*.

این معیار، حالت نرمال شده معیار MSE (رابطه ۱) می‌باشد [۱]. به نظر می رسد NMSE از میزان بایاس نسبت به مدل‌هایی که بیش از حد بیش‌برازش یا کم‌برازش شده‌اند اجتناب می کند و به پراکندگی مجموعه داده‌ها تاکید می‌کند [۲]. کمترین مقدار این معیار ارزیابی صفر و بیشترین آن یک است، همچنین رابطه آن به صورت زیر بیان می‌شود:

$$NMSE = \frac{\overline{(\hat{Y} - Y)^2}}{\overline{\hat{Y}}\,\overline{Y}} \tag{۵}$$

که در این‌جا، Y و $\hat{Y}$ به ترتیب برابر با بردار برچسب و پیش‌بینی و $\overline{Y}$ و $\overline{\hat{Y}}$ نیز به ترتیب میانگین این دو بردار هستند. همچنین مقدار $\overline{(\hat{Y} - Y)^2}$ میانگین تفاضل این دو بردار است که در صورت کسر قرار می‌گیرد. عدد حاصل از این معیار عددی نرمال شده و بین صفر و یک است که می‌تواند در تحلیل و میزان حساسیت به داده‌های پرت از اهمیت بالایی برخوردار باشد. مقدار این معیار ارزیابی برای جدول ۱ برابرست با:

$$\overline{(\hat{Y} - Y)^2} = \frac{1}{5}\sum_{i=1}^{5}(\hat{y}_i - y)^2 = 13554$$

$$\overline{Y} = 1580$$

$$\overline{\hat{Y}} = 1580$$

$$NMSE = \frac{13554}{1580 \times 1580} = 0.00542941$$

که این مقدار هر چه به صفر نزدیک باشد، نشان از آن است که عملکرد مدل خوب بوده است.

پرسش ۲.

فرض کنید یک مجموعه داده دارای دو یا چند برچسب باشد و قرار است یک مدل یادگیرنده، مقدار آن‌ها را پیش‌بینی کند. چه راهکاری برای حل چنین مسئله‌ای ارائه می‌دهید؟ (به عنوان مثال، فرض کنید که قرار است در یک مسئله تشخیص اشیا در پردازش تصویر، تمامی حیوانات در تصویر را تشخیص داده و دور تصویر آن حیوانات، یک باکس مستطیل مانند رسم کنید. برای این‌کار نیاز است که مختصات دو نقطه از این مستطیل را در صفحه پیش‌بینی کنید که هرکدام دارای مؤلفه x و y هستند)

پاسخ.

در این حالت، فرض می‌کنیم که $\mathcal{S}$ مجموعه نمونه‌ها باشد، بطوری که:

$$\mathcal{S} = ((x_1, y_1), \ldots, (x_m, y_m)), \qquad x_i \in \mathbb{R}^d, \quad y_i \in \mathbb{R}^p$$

که در این‌جا، d و p به ترتیب ابعاد فضای ویژگی و فضای برچسب هستند. همان‌طور که مشاهده می‌کنید، در این حالت فضای برچسب (خروجی) p بُعدی است، یعنی به ازای هر نمونه، به تعداد p برچسب در اختیار داریم. حال برای حل چنین مسئله‌ای، می‌توان از دو رویکرد کلی استفاده کرد که در ادامه آن‌ها را شرح خواهیم داد.

روش اول: رگرسیون چند خروجی مستقیم. یکی از ساده‌ترین رویکردهای قابل ارائه برای حل چنین مسئله‌ای، استفاده از رگرسیون چند خروجی مستقیم[۶] است. در این سناریو، به ازای هر بُعد از فضای برچسب، یک مدل رگرسیون ساخته می‌شود که روی مجموعه داده $\mathcal{X}$ آموزش می‌بیند (شکل ۲). بنابراین باتوجه به فضای ویژگی $\mathcal{X}$ که برابرست با:

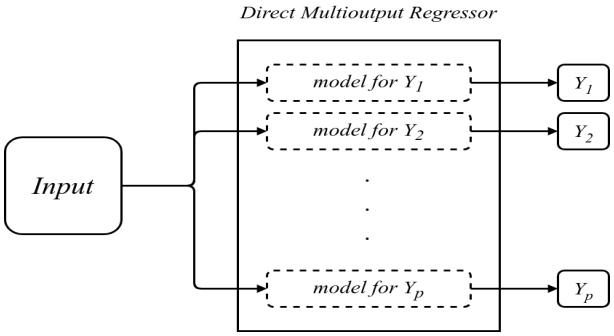
$$\mathcal{X} = (x_1, \ldots, x_m), \qquad x_i \in \mathbb{R}^d$$

یک مجموعه از فرضیه‌ها با نام $\hat{\mathcal{H}}$ در اختیار داریم، به‌طوری که:

$$\hat{\mathcal{H}} = \{\phi : \mathcal{X} \times \Theta \to \mathbb{R}\}, \qquad |\hat{\mathcal{H}}| = p$$

که در این مجموعه، ϕ_i مدل iـام برای تخمین بُعد iـام فضای برچسب‌ها است.

[6] Direct Multioutput Regression

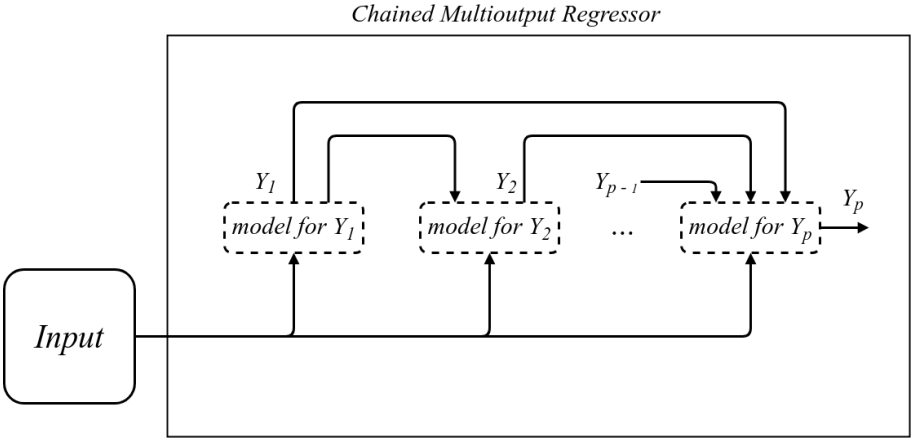


شکل ۲: معماری کلی یک رگرسیون چند خروجی مستقیم.

روش دوم: رگرسیون چندخروجی زنجیره‌ای. در این معماری، مدل‌ها به صورت زنجیره‌ای و پشت سرهم ساخته می‌شوند، به‌طوری که هر مدل با استفاده از فضای ویژگی $\mathcal{X}$ و پیش‌بینی مدل قبل آموزش می‌بیند. به چنین روشی، رگرسیون چندخروجی زنجیره‌ای[۷] گوییم. این روش در مسائلی که ابعاد خروجی از نظر مفهومی با یکدیگر مرتبط هستند می‌تواند کارا باشد. بنابراین در این روش نیز مجموعه‌ای از فرضیه‌ها با نام $\hat{\mathcal{H}}$ در اختیار داریم، به‌طوری که:

$$\hat{\mathcal{H}} = \{(\forall_{0\leq i\leq p})\phi_i : \mathcal{X} \times Y_{i-1} \times \Theta_i \to \mathbb{R}\}, \qquad |\hat{\mathcal{H}}| = p$$

بنابراین آخرین مدل در این معماری، فضای ویژگی‌ای با ابعاد $d + p - 1$ در اختیار دارد که این خاصیت در این معماری، می‌تواند به تدریج تأثیر مدل‌های قبلی را دریافت کند. البته این خاصیت می‌تواند جز معایب این روش نیز باشد. اگر مدل‌های آغازین عملکرد خوبی در پیش‌بینی برچسب متناظرشان نداشته باشند، آن‌وقت به تدریج مدل‌های آخر فضای ویژگی‌شان دارای ویژگی‌های غیرمفید می‌شود و از نظر محاسباتی نیز هزینه بیشتری دربردارد.



شکل ۳: معماری کلی یک رگرسیون چند خروجی زنجیره‌ای.

مراجع

[1] A. A. Poli and M. C. Cirillo, "On the use of the normalized mean square error in evaluating dispersion model performance," *Atmospheric Environment. Part A. General Topics*, vol. 27, no. 15, pp. 2427–2434, 1993.

[2] J. C. Chang and S. R. Hanna, "Air quality model performance evaluation," *Meteorology and Atmospheric Physics*, vol. 87, no. 1, pp. 167–196, 2004.

[7] Chained Multioutput Regression