درس : یادگیری ماشین
دانشجو : امیرمحمد خرازی
شماره دانشجویی : ۴۰۱۵۲۵۲۱۰۰۲
استاد درس : دکتر منصور رزقی آهق

پرسش‌های کلاسی سری سوم — دانشکده علوم ریاضی ، گروه علوم کامپیوتر، گرایش داده‌کاوی

گیت‌هاب درس (لینک) — گیت‌هاب این پرسش (لینک)

مقدمه

در این پرسش کلاسی، قصد بر این است که رگرسیون‌های شناخته شده را بر روی داده‌های عملی‌تر امتحان و ارزیابی کنیم. برای این کار ۵ نوع دیتاست مختلف انتخاب شده است. هر دیتاست را می‌توانید از لینک‌هایی که برای آن‌ها در هر بخش مشخص می‌شود دانلود کنید. اطلاعات کامل تر هر دیتاست نیز در هر بخش آورده خواهد شد.

نحوه کد و شیوه استفاده از رگرسیون‌های چند جمله (و یا خطی)، لاجستیک و غیره در کدهای جوپیتر هر دیتاست آورده شده‌اند. همه این مسائل توسط ابزارهایی که Scikit-Learn در اختیار ما قرار داده است حل می‌شوند. البته همه این مسائل بصورت رگرسیون خطی یا چندجمله قابل حل نیستند. لذا علاوه بر این روش‌ها، روش‌های دیگری نیز برای حل ارائه می‌شود.

بطور خلاصه دیتاست‌ها می‌توانند شامل چندین ویژگی و چندین هدف باشند. یعنی اگر هر سطر یا نمونه از دیتاست را با (X, Y) نشان دهیم و هدف در رگرسیون بازسازی تابعی باشد که این X را به Y مرتبط می‌کند، X و Y هر کدام می‌توانند عضو فضای چند بعدی باشند. حالت‌های زیر را بررسی کنید:

یک ویژگی یک هدف :

فرض کنید $X \in \mathbb{R}$ و $Y \in \mathbb{R}$ باشد، آنگاه مانند قبل یک چندجمله رگرسیون با درجه M برای نمونه iام داریم :

$$\hat{Y_i} = W_0 + W_1 X_i + W_2 X_i^M + \dots W_M X_i^M$$

. اگر N نمونه مجموعا داشته باشیم، خواهیم داشت :

$$\Phi(X) = \begin{bmatrix} 1, X_1, X_1^2, \dots, X_1^M \\ 1, X_2, X_2^2, \dots, X_2^M \\ \vdots \\ 1, X_N, X_N^2, \dots, X_N^M \end{bmatrix}_{N \times (1^M+1)} \quad W = \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_M \end{bmatrix}_{(1^M+1)\times 1} \quad \hat{Y} = \begin{bmatrix} \hat{Y_1} \\ \hat{Y_2} \\ \vdots \\ \hat{Y_N} \end{bmatrix}_{N\times 1} \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}_{N\times 1}$$

مشخص است که $\hat{Y} = \Phi(X)W$ و هدف کمینه کردن فاصله Y و $\hat{Y}$ است که در اینجا نرم۲ مد نظر است (اگر چه مشکلاتی نیز دارد) یعنی : $\min_W\{||Y - \hat{Y}||_2^2\} = \min_W\{\sum_{i=1}^{N}(Y_i - \hat{Y_i})^2\}$.

چندین ویژگی یک هدف :

فرض کنید $X \in \mathbb{R}^D$ باشد و $Y \in \mathbb{R}$. در چنین حالتی یک چندجمله رگرسیون با درجه M بصورت زیر است :

$$\hat{Y_i} = W_0 + W_1 X_{i1} + W_2 X_{i2} + W_3 X_{i3} + \cdots + W_{D+1} X_{i1}^2 + W_{D+2} X_{i1} X_{i2} + \dots$$

مثلا اگر $X \in \mathbb{R}^2$ باشد، چند جمله رگرسیون درجه M می‌شود:

$$\hat{Y_i} = W_0 + W_1 X_{i1} + W_2 X_{i2} + W_3 X_{i1}^2 + W_4 (X_{i1} X_{i2} = X_{i2} X_{i1}) + W_5 X_{i2}^2 + \dots$$

اگر N نمونه داشته باشیم، خواهیم داشت:

$$\Phi(X) = \begin{bmatrix} 1, X_{11}, X_{12}, \dots, X_{11}^2 \dots \\ 1, X_{21}, X_{22}, \dots, X_{21}^2 \dots \\ \vdots \\ 1, X_{N1}, X_{N2}, \dots, X_{N1}^2 \dots \end{bmatrix}_{N \times (D^M+1)} \quad W = \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_{D^M} \end{bmatrix}_{(D^M+1)\times 1} \quad \hat{Y} = \begin{bmatrix} \hat{Y_1} \\ \hat{Y_2} \\ \vdots \\ \hat{Y_N} \end{bmatrix}_{N \times 1} \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}_{N \times 1}$$

مشخص است که مانند قبل $\hat{Y} = \Phi(X) W$ و هدف کمینه کردن $\min_W \{||Y - \hat{Y}||_2^2\} = \min_W \{\sum_{i=1}^{N} (Y_i - \hat{Y_i})^2\}$ است. در اینجا چون Y_i ، $\hat{Y_i}$ ها عدد هستند، اختلاف آن‌ها را گرفته و توان ۲ رساندیم، مجموع این توان ۲ اختلاف‌ها برابر است با نرم ۲. اما اگر از حالت عددی خارج شوند، برای اختلاف، مانند بردار عمل می‌کنند. در بخش بعدی این مورد بررسی می‌شود.

چندین ویژگی چندین هدف

در این حالت فرض کنید $X \in \mathbb{R}^D$ و $Y \in \mathbb{R}^d$ باشد. چند جمله رگرسیون برابر خواهد بود با :

$$\hat{Y_{ij}} = W_{0j} + W_{1j} X_{i1} + W_{2j} X_{i2} + W_{3j} X_{i3} + \cdots + W_{D+1,j} X_{i1}^2 + W_{D+2,j} X_{i1} X_{i2} + \dots$$

اگر N نمونه داشته باشیم، خواهیم داشت:

$$\Phi(X) = \begin{bmatrix} 1, X_{11}, X_{12}, \dots, X_{11}^2 \dots \\ 1, X_{21}, X_{22}, \dots, X_{21}^2 \dots \\ \vdots \\ 1, X_{N1}, X_{N2}, \dots, X_{N1}^2 \dots \end{bmatrix}_{N \times (D^M+1)} \quad W = \begin{bmatrix} W_{01}, W_{02}, \dots, W_{0d} \\ W_{11}, W_{12}, \dots, W_{1d} \\ \vdots \\ W_{D^M 1}, W_{D^M 2}, \dots, W_{D^M d} \end{bmatrix}_{(D^M+1)\times d}$$

$$\hat{Y} = \begin{bmatrix} \hat{Y_{11}}, \hat{Y_{12}}, \dots, \hat{Y_{1d}} \\ \hat{Y_{21}}, \hat{Y_{22}}, \dots, \hat{Y_{2d}} \\ \vdots \\ \hat{Y_{N1}}, \hat{Y_{N2}}, \dots, \hat{Y_{Nd}} \end{bmatrix}_{N \times d} \quad Y = \begin{bmatrix} Y_{11}, Y_{12}, \dots, Y_{1d} \\ Y_{21}, Y_{22}, \dots, Y_{2d} \\ \vdots \\ Y_{N1}, Y_{N2}, \dots, Y_{Nd} \end{bmatrix}_{N \times d}$$

در این حالت نیز مانند قبل داریم : $\hat{Y} = \Phi(X)W$ و هدف کمینه کردن فاصله Y با $\hat{Y}$ است . اما می‌دانیم که هر دوی آن‌ها دیگر بردار نیستند و ماتریس‌اند. لذا برای اینکار از نرم دیگری استفاده می‌کنیم. این نرم $||.||_F^2$ است.

$$\min_W\{\sum_{i=1}^{N}||Y_i - \hat{Y_i}||_2^2\} = \min_W\{\sum_{i=1}^{N}\sum_{j=1}^{d}(Y_{ij} - \hat{Y_{ij}})^2\} = \min_W\{||Y - \hat{Y}||_F^2\}$$

با توجه به این موارد می‌توانیم دید بهتری نسبت به رگرسیون در زمان‌های که با چنیدن بعد سروکار داریم داشته باشیم.
در ادامه اطلاعات هر دیتاست آورده خواهد شد و موارد مربوط به کد آن ذکر می‌شود.

۱ Weather in Szeged :

این دیتاست توسط لینک زیر قابل دسترسی است:

Weather in Szeged 2006-2016

اطلاعات کامل تر شامل ۱۲ ستون این دیتاست و نوع هر ستون و غیره در کدهای این گزارش قابل مشاهده است. هدف استفاده از این دیتاست این است که میان رطوبت و دمای هما رابطه ای درست کنیم. مثلا اگر رطوبت را به ما دادن بتوانیم دما را تخمین بزنیم . بدین منظور یک دما به عنوان متغییر وابسه داریم و یک میزان روطوبت به عنوان متغییر مستقل خواهیم داشت. دو نوع دما در این دیتاست آورده شده است لذا دو نوع مدل (کار) انجام می‌دهیم یکی برای دمای اولی و دیگری برای دمای نوع دوم. می‌دانیم که به راحتی می‌توانیم این مسئله را به حالت چند ویژگی چند هدفه در آوریم. یعنی فرض کنید با میزان رطوبت بخواهیم مقدار دو نوع دما را تخمین بزنیم. با این روش (تخمین هر کدام جداگانه)، ستون‌های W ساخته می‌شوند. و در نهایت می‌توانیم W کامل را از روی آن‌ها بسازیم. همچنین این دیتاست ها بر روی چندین چند جمله‌ای آزمون و آموزش داده شده‌اند و نتیجه ی آن‌ها در کدهای این بخش موجود است. لذا برای بررسی $RMSE$ های نتیجه می‌توانید به کدهای این بخش مراجعه فرمائید.
لازم به ذکر است که در این بخش علاوه بر ۵ نوع رگرسیون خطی و چند جمله با درجات مختلف از Spline Regression و XGBoost نیز برای مدل‌سازی استفاده شده است. در بخش Spline مشکلاتی وجود دارد که موفقا آن‌را بیخیال می‌شویم.

۲ Weather Conditions in World War Two :

این دیتاست توسط لینک زیر قابل دسترسی است:

Weather Conditions in World War Two

در این لینک دو فایل موجود است . هدف استفاده از این دیتاست این است که رابطه‌ای میان حداقل و حداکثر دما پیدا کنیم. همانطور که گفته شد، دو فایل در این دیتاست موجود است . یکی مربوط به مکان‌هایی است که این اطلاعات آب و هوا (شامل دما و غیره) ثبت شده است . دیگری مربوط به اطلاعاتی است که در هر مکان ثبت شده. یعنی یکی از فایل‌ها، اطلاعات مربوط به محل ثبت را داراست و فایل دیگر اطلاعات مربوط به آب و هوا را داراست. می‌توانیم این دو فایل را از روی ستون‌های WBAN و STA با هم مرتبط کنیم. البته در این مسئله کار ما ساده است و فقط با یکی از این فایل‌ها کار داریم. البته درست است که مواردی چون ارتفاع محل ثبت این دما و غیره روی دما تاثیر دارد ولی در این چالش از ما خواسته شده است تا ارتباط میان دمای حداقل و دمای حداکثر را بیابیم. برای اطلاعات بیشتر از نحوه این کار و همچنین نتیجه $RMSE$ بدست آمده از مدل‌ها می‌توانید به کدهای این بخش مراجعه فرمائید. لازم به ذکر است در این بخش علاوه بر ۱۰ نوع رگرسیون خطی و چند جمله‌ای با درجه‌های مختلف، از XGBoost نیز برای مدل‌سازی استفاده شده است.

۳ The Ultimate Halloween Candy Power Ranking :

این دیتاست توسط لینک زیر قابل دسترسی است :

The Ultimate Halloween Candy Power Ranking

در این فایل حدود ۸۵ نمونه با ۱۳ ویژگی وجود دارد. هدف ما در این مسئله بررسی شکلاتی بودن یا نبودن آبنبات بر اساس سایر ویژگی‌های آن است. تفاوت این دیتاست با سایر دیتاست‌هایی که تا بدین جا از آن‌ها استفاده کردیم در این است که این دیتاست ماهیتی Categorical دارد. بدین معنی که شکلات بودن یا نبودن می‌تواند ۰ و ۱ باشد و عددهای حقیقی بین یا بیشتر از این‌ها معنی ندارند. مثلا اگر با رگرسیون‌های معمولی این موارد را حل کنیم، مقدارهایی مثل 0.86 و غیره خواهیم داشت که بی معنی هستند. برای حل این مشکل باید به مسئله به شکل کلاس‌بندی نگاه کنیم. یعنی در این مرحله فرض می‌کنیم دو کلاس ۰ و ۱ داریم و می‌خواهیم تصمیم بگیریم با توجه به ویژگی‌ها، نمونه در کدام کلاس قرار می‌گیرد. یکی از روش‌های حل این مسئله رگرسیون لاجستیک است. لذا در حل این سوال تنها از رگرسیون لاجستیک استفاده شده است. خروجی‌های این رگرسیون به صورت واضح و کامل (۰ و ۱ ای) بوده و از آنجایی که دیگر اختلاف‌های y و $\hat{y}$ برای هر نمونه مقداری ۰ یا ۱ است، دیگر نمی‌توان از معیارهای $RMSE$ یا غیره استفاده است و باید از معیارهای کلاس‌بندی برای ارزیابی این مدل استفاده شود. روش کامل حل و اطلاعات مربوط به کدهای این بخش را می‌توانید در کد مربوطه مشاهده فرمائید.

تا بدین جا مسئله‌های اول و دوم بصورت تشکیل یک ارتباط میان دو متغیر بود، مسئله سوم به صورت کلاس‌بندی حل می‌شد ولی همچنان رابطه را میان یک مغییر و چندین متغییر برقرار می‌کرد. از این به بعد علاوه بر امکان وجود چندین متغییر مستقل، امکان وجود چندین متغییر وابسته نیز وجود دارد.

۴ ATP1D

این دیتاست توسط لینک زیر قابل دسترسی است:

Airline Ticket Price dataset - ATP1D

این دیتاست شامل حدود ۳۰۰ نمونه و بیش از ۴۰۰ ویژگی است. از بین این ویژگی‌ها ۴۱۱ ویژگی به عنوان متغییرهای مستقل و ۶ ویژگی به عنوان متغییر هدف انتخاب شده‌اند. نحوه بررسی و انجام این نوع رگرسیون‌ها را قبل در بخش مقدمه شرح داده‌ام و لذا در اینجا به راحتی می‌توانیم مانند قبل عمل کنیم و رگرسیون خود را ساخته و جواب خود را از مدل‌ها، بدست آوریم. تنها تفاوت اصلی آن با حالت‌های قبل این است که دیگر با بردار سروکار نداریم، لذا ماتریسی خواهیم داشت که شامل آن ۶ ستون هدف است. برای اینکار نرم ۲ سطرها را گرفته و با هم جمع می‌کنیم. این کار مانند نرم F می‌ماند. نرم F که در اینجا بکار می‌رود،، مجموع توان ۲ درایه‌های ماتریس است که در این بخش بصورت تفاضل هر یک از مولفه‌های بردار هدف با برداد تخمین زده شده هدف از روی مدل، در آمده است. در این بخش، تنها از رگرسیون درجه ۱ و درجه ۲ استفاده شده است چرا که با توجه به زیاد بودن ویژگی‌ها به مشکلاتی چون کرش کردن برنامه و خطاهای حافظه بر خواهیم خورد که علاوه بر آن، می‌توانند Overfitting هم رخ بده، همانطور که در کد مشاهده می‌کنید، بیش برازش رخ می‌دهد. برای اطلاعات بیشتر در خصوص این بخش، می‌توانید به کدهای مربوطه مراجعه فرمائید. لازم به ذکر است که در این بخش علاوه بر روش‌های گفته شده، از XGBoost نیز استفاده شده است و این روش ثابت کرد که نتیجه‌ای بهتر از رگرسیون‌ها به ما ارائه می‌دهد.

۵ RF1 :

این دیتاست توسط لینک زیر قابل دسترسی است:

River Flow-RF1

این دیتاست شامل ۶۴ ستون ویژگی و ۸ ستون هدف است. در این دیتاست همچنین مقادیر گمشده وجود دارد. تصمیم بر این شد که از آنجایی که تعداد این نمونه ها نسبت به کل مجموعه داده کم است، آن‌ها را نادیده بگیریم. در این صورت حدود ۱۲۰ رکورد از مجموعه داده‌های ما حذف می‌شوند ولی مجموعه نهایی بدون داده گمشده خواهد بود. سپس می‌توانیم مانند قبل روی این نمونه‌ها رگرسیون انجام دهیم. رگرسیون را تنها با درجات ۱ و ۲ انجام دادیم. برای کسب اطلاعات کامل‌تر در خصوص این بخش می‌توانید از کدهای مربوطه استفاده فرمائید. لازم به ذکر است که علاوه بر رگرسیون از XGBoost نیز برای یادگیری استفاده شده است. نتایج نیز به طور مشخص، دیداری‌سازی شده‌اند. البته در این روش‌های دیگر نیز بیش برازش داشتیم که با XGBoost مدل بهتری ارائه دادیم.

بطور کلی فرقی نمی‌کند داده چند بعدی باشد و یا هدف بصورت برداری باشد یا عدد، نتیجه نهایی همان فرمول ساده رگرسیون است که مسئله را با بهینه کردن نرم ۲ یا نرم F یا نرم‌های دیگر حل می‌کند و وزن‌های لازم را بدست می‌آورد.

از آنجایی که کدها بطور نسبتا کامل با توضیح در حوپیتر پایتون نوشته شده‌اند، کدها در این گزارش آورده نشده‌اند و برای بررسی بیشتر پیشنهاد می‌شود تا به خود کدها مراجعه شود