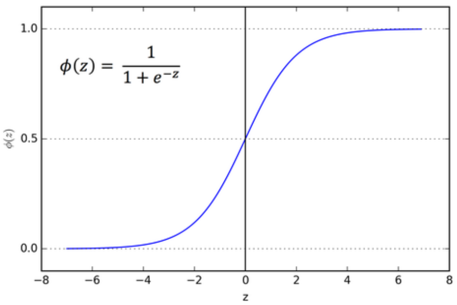
شرح کلی :

تابع سیگموید در نظرگرفته میشود.بیشترین احتمال را دربین احتمال های موجود درنظر بگیرید(likelihood).برای این کار باید مقدار این احتمال را ماکسیمم کنیم که به دلیل سخت بودن مشتق گرفتن از رابطه زیر، با تابع لگاریتمی آن فرایند کار را ادامه میدهم.



$$w^* = argmax_w \prod_{i=1}^{m} P\big(y(i)\big|x(i)\big) = argmax \prod_{i=1}^{m} \frac{1}{1+\exp(-z)}$$

$$=> \quad \sum_{i=1}^{m} \log \Pr\big(y(i)\big|x(i)\big) = -\sum_{i=1}^{m} \log(1+\exp(-z))$$

سمت راست رابطه بالا همان تابع logistic loss میباشد :

$$L_{logistic} = (y , z) = \log (1 + \exp(-z))$$

به این ترتیب به راحتی میتوان گفت که ماکسیمم کردن احتمال لایکلی هود هم ارز با مینیموم کردن loss function میباشد.

حال بررسی حالت چند کلاسه که خواسته سوال میباشد:

میدانیم زمانی که تعداد کلاس ها بیش از دو باشد احتمال و پیش بینی به صورت زیر مدل میشود.

$$\Pr(y = k|x) = \frac{\exp(w_k^T x)}{\sum_{j=1}^{k} \exp(w_j^T x)}$$

فرم تابع از سوی دیگر بیانگر تابع softmax میباشد.

$$softmax(z_1, \ldots, z_K) = \left(\frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)}\right)_{k=1}^{K}$$

تقریباً به پیدا کردن عنصر حداکثر در وکتور یک تا k میپردازد.حال اگر یکی از این عناصر به طرز قابل ملاحظه ای از سایرین بیشتر باشد در اینصورت احتمال آن به نزدیک یک و احتمال سایرین به نزدیک صفر نگاشت میشود.

برای توضیح بیشتر و ارتباط کلاس ها باهم ، دو کلاس k و k’ را درنظر بگیرید.

$$\log \frac{Pr(y = k|\mathbf{x})}{Pr(y = k'|\mathbf{x})} = \log \frac{\exp(\mathbf{w}_k^T \mathbf{x})/Z}{\exp(\mathbf{w}_{k'}^T \mathbf{x})/Z} = \mathbf{w}_k^T \mathbf{x} - \mathbf{w}_{k'}^T \mathbf{x}$$

حال به عنوان مثال در این عبارت احتمال کلاس k بیشتر از کلاس دیگر است.

Mohammadmahdi Mohammadi-964416

حال بافرض عدم وجود احتمال کوچکتر از صفر داریم برای پاسخ نهایی داریم :

$$Pr(y = k^*|\mathbf{x}) = \max_{k=1}^{k}(\exp(\mathbf{w}_k^T\mathbf{x})/Z)$$

بدین ترتیب با مقایسه بدست برای هر کلاس از رابطه صفحه قبل ماکسیمم آنهارا انتخاب میکند. در واقع با تمامی کلاس ها به صورت تک به تک مقایسه میکند.

حال برای مینیموم کردن loss function یا همان ماکسیمم کردن likelihood با درنظر گرفتن احتمال شرطی زیر به مسیر خود ادامه میدهیم: فرض کنید برای داده آموزشی (x(i) | y(i)) میخواهم فرایند ماکسیمم سازی را اجرا کنیم ...

$$Pr(y_i = k|\mathbf{x}_i) = \frac{\exp(\mathbf{w}_k^T\mathbf{x}_i)}{\sum_{j=1}^{K}\exp(\mathbf{w}_j^T\mathbf{x}_i)}$$

صورت و مخرج کسر بالا را به صورت کسر تقسیم میکنیم و از طرفین منفی لگاریتم میگیریم که بدین ترتیب داریم:

$$-\log Pr(y_i = k|\mathbf{x}_i) = \log(1 + \sum_{j\neq k}\exp(\mathbf{w}_j^T\mathbf{x})/\exp(\mathbf{w}_k^T\mathbf{x}))$$

عبارت بدست آمده همان مقدار logistic loss برای حالت چند کلاسه میباشد:

$$L(x_i, y_i) = \log(1 + \sum_{j\neq y_i}\exp(\mathbf{w}_j^T\mathbf{x} - \mathbf{w}_{y_i}^T\mathbf{x}))$$

سمت راست عبارت بدست آمده مجموع تمام احتمال های دو به دو بر پایه کلاس y(i) میباشد. بدین ترتیب احتمال برای کلاس های درست و نادرست به ترتیب زیر حاصل میشود.

- کلاس نادرست (j برابر نیست با y(i)):

$$\frac{Pr(y = j|\mathbf{x})}{Pr(y = y_i|\mathbf{x})} = \exp(\mathbf{w}_j^T\mathbf{x} - \mathbf{w}_{y_i}^T\mathbf{x}_i)$$

- کلاس درست (j برابربرست با y(i)) :

$$\frac{Pr(y = j|\mathbf{x})}{Pr(y = y_i|\mathbf{x})} = 1$$

برای train روش لاجستیکی رگرسیون ما نیازمند یافتن w ای هستیم که به طور میانگین loss مارا مینیموم کند.

$$\boldsymbol{min\ j(w) = \left(\frac{1}{m}\right) * \sum_{i=1}^{m} \log(1 + \exp(-z))}$$

عبارت بدست آمده یک تابع پیوسته بوده ولی به دلیل اینکه غیرخطی میباشد یافتن مقدار بهینه در آن کار ساده ای نیست و باید سرچ incremental برای یافتن آن صورت پذیرد.

قسمت الف:

$$\text{if } f_i(\mathbf{x}) \begin{cases} > 0 \\ < 0 \end{cases}, \text{ then } \begin{cases} \mathbf{x} \in C_+ \\ \mathbf{x} \in C_- \end{cases}$$

$$f_k(\mathbf{x}) = \max\{f_1(\mathbf{x}), \cdots, f_K(\mathbf{x})\}, \qquad (k = 1, \cdots, K)$$

$$p(y = i|\mathbf{x}) = \phi_i(\mathbf{w}_i^T\mathbf{x}) = \frac{\exp(\mathbf{w}_i^T\mathbf{x})}{\sum_{k=1}^K \exp(\mathbf{w}_k^T\mathbf{x})}, \qquad \sum_{i=1}^K \phi_i = 1$$

$$\phi_0(\mathbf{w}_0^T\mathbf{x}) = \frac{\exp(\mathbf{w}_0^T\mathbf{x})}{\exp(\mathbf{w}_0^T\mathbf{x}) + \exp(\mathbf{w}_1^T\mathbf{x})} = \frac{1}{1 + \exp(-(\mathbf{w}_0 - \mathbf{w}_1)^T\mathbf{x})} = \sigma(-\mathbf{w}^T\mathbf{x})$$

$$\phi_1(\mathbf{w}_1^T\mathbf{x}) = \frac{\exp(\mathbf{w}_1^T\mathbf{x})}{\exp(\mathbf{w}_0^T\mathbf{x}) + \exp(\mathbf{w}_1^T\mathbf{x})} = \frac{1}{1 + \exp((\mathbf{w}_0 - \mathbf{w}_1)^T\mathbf{x})} = \sigma(-\mathbf{w}^T\mathbf{x})$$

$$\mathbf{h}_\mathbf{w}(\mathbf{x}) = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_K \end{bmatrix} = \frac{1}{\sum_{k=1}^K \exp(\mathbf{w}_k^T\mathbf{x})} \begin{bmatrix} \exp(\mathbf{w}_1^T\mathbf{x}) \\ \vdots \\ \exp(\mathbf{w}_K^T\mathbf{x}) \end{bmatrix}$$

$$\boldsymbol{L\,(D\mid W)} = \prod_{n=1}^N p(y_n|\mathbf{x}_n) = \prod_{n=1}^N \prod_{i=1}^K p(y_n = i|\mathbf{x}_n)^{1\{y_n = i\}}$$

$$= \log L(\mathbf{w}|\mathcal{D}) = \sum_{n=1}^N \sum_{i=1}^K 1\{y_n = i\} \log p(y_n = i|\mathbf{x}_n)$$

$$= \sum_{n=1}^N \sum_{i=1}^K 1\{y_n = i\} \log\left(\frac{\exp(\mathbf{w}_i^T\mathbf{x}_n)}{\sum_{k=1}^K \exp(\mathbf{w}_k^T\mathbf{x}_n)}\right)$$

$$= \sum_{n=1}^N \sum_{i=1}^K 1\{y_n = i\} \left(\mathbf{w}_i^T\mathbf{x}_n - \log\sum_{k=1}^K \exp(\mathbf{w}_k^T\mathbf{x}_n)\right)$$

قسمت ب:

فرایند لگاریتم گیری برای تبدیل ضرب به جمع (با توجه به اینکه رویکرد این تابع با احتساب عملگر لگاریتمی درمورد مینیموم و ماکسیمم تغیری نمیکند) در قسمت الف انجام شد و حال با گرادیان کاهشی به دنبال نقطه مورد نظر میگردیم.

$$\begin{aligned}
\mathbf{g}_l(\mathbf{w}_j) &= \frac{d}{d\mathbf{w}_j} l(\mathbf{W}|\mathcal{D}) = \frac{d}{d\mathbf{w}_j} l(\mathbf{W}|\mathcal{D}) \left[\sum_{n=1}^{N} \sum_{i=1}^{K} 1\{y_n = i\} \left(\mathbf{w}_i^T \mathbf{x}_n - \log \sum_{k=1}^{K} \exp(\mathbf{w}_k^T \mathbf{x}_n) \right) \right] \\
&= \sum_{n=1}^{N} 1\{y_n = j\} \left(\mathbf{x}_n - \frac{\exp(\mathbf{w}_j^T \mathbf{x}_n)}{\sum_{k=1}^{K} \exp(\mathbf{w}_k^T \mathbf{x}_n)} \mathbf{x}_n \right) \\
&= \sum_{n=1}^{N} 1\{y_n = j\} \left(1 - p(y_n = j|\mathbf{x}_n)\right) \mathbf{x}_n = \sum_{n=1}^{N} 1\{y_n = j\} (1 - \phi_i) \mathbf{x}_n
\end{aligned}$$

تحلیل الگوریتم ها:

استفاد از الگوریتم های دوکلاسه مانند on vs on & on vs all اگرچه برای داده های کوچک کارگشا هستند ولی به دلیل وجود عیبی بزرگ که همان ایجاد فضاهایی در مساله که نمیتوان برای دو کلاس مرز مشخصی عنوان کرد به گونه ای که طبق ارزیابی توسط یک حالت ممکن است متعلق به کلاس اول و طبق ارزیابی توسط حالت دیگری متعلق به کلاس دوم باشد نمیتواند عملکرد مناسبی ارایه کند. هرچقدر تعداد این فضاها که ابهام درمورد آنها وجود دارد افزایش میابد عملکرد الگوریتم نیز بدتر میشود. به طورکلی میتوان گفت که استفاده از الگوریتم های چندکلاسه مخصوصا برای داده های بزرگ انتخاب بهتری میباشد.

قسمت آخر:

همانطور که دیدیم لازم است این عبارت را در رگرسیون لجستیک کلاسیک ارزیابی کنیم.

$$\beta = (X^T * X)^{-1} * X^T * Y$$

این عبارت از سیستم معادلات خطی بدست آمده است. $Y = X * \beta$

به طور غیرمستقیم فرض کردیم که همه مشاهدات در همان اندازه مهم هستند و از این رو وزن یکسانی دارند و ما سعی کردیم مقدار باقی مانده های توان دو رساندن را به حداقل برسانیم. با این حال ، هنگامی که ما رگرسیون لجستیک وزن دار انجام می دهیم ، اهمیت مشاهدات خود را اهمیت می دهیم تا مشاهدات مختلف دارای وزن های مختلفی باشند که مربوط به آنها است.

بدین ترتیب ما یک ماتریس وزنی W خواهیم داشت که وزن مشاهده شده در i در درایه W_{ii} آن وجود دارد. حال به جای محاسبه و تخمین عبارت روبرو عبارت زیر را مورد محاسبه قرار میدهیم. $\beta = (X^T * X)^{-1} * X^T * Y$

$$\beta = (X^T * W * X)^{-1} * X^T * W * U$$

$$U_i = X_i^T * \beta + \frac{y_i - \mu_i}{W_{ii}}$$
$$W_{ii} = \mu_i * (1 - \mu_i)$$

و µi برآورد ما برای p است ، که قبلاً دیدیم می تواند به صورت زیر نوشته شود

$$\mu_i = \frac{1}{1 + e^{-X_i^T * \beta}}$$

$$Var(X) = n * p * (1 - p) \quad X \sim Bin(p, n)$$

حال از سویی دیگر با درنظر گرفتن تابع سیگموید برای تمام خطوط داریم:

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))}$$

از آنجایی که ما مدل خطی خود را در یک محدوده احتمال قرار داده ایم ، ضرایب (یا وزن) هر فیچر بر روی خروجی به تأثیر نمی گذارد. بنابراین ، برای تفسیر مدل های خود ، می توانیم با تبدیل احتمالات خود برای ورود به شانس ، عملکرد سیگموئید خود را از محدوده خروجی ۰–۱ ، به +/- بی نهایت ببریم.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$.

***p* احتمال تعلق داشتن به کلاس است.**

با رگرسیون خطی ، بهترین خط مناسب را پیدا می کنیم و mean squared error را مینیموم میکنیم.به همین دلیل تمام مشاهدات ما دارای مقادیر مثبت منفی بی نهایت هستند زیرا ما ۱۰۰٪ مطمئن هستیم که داده های برچسب ما به کدام طبقه تعلق دارد.

$$\begin{aligned} logit(1) &= log(\tfrac{1}{(1-1)}) \\ &= log(1) - log(0) \\ &= 0 - -infinity \\ &= +infinity \end{aligned}$$

برای داده های مورد بررسی با این روش چون تمامی حالت ها بررسی میشوند میتوان عمل کرد و از آن به جای ماکسیموم کردن احتمال لایکلی هود استفاده کرد.