به نام خالق یکتا

تمرین دوم مبانی یادگیری‌ماشین

پاییز 1402

رگرسیون منطقی

1- فرض کنید به شما یک مجموعه آموزشی D شامل n نمونه به صورت $D=\{(\mathbf{x}^{(1)}, t^{(1)}), \dots, (\mathbf{x}^{(n)}, t^{(n)})\}$ داده شده است، که هر ورودی $\mathbf{x}^{(i)}$ یک بردار d بعدی $\mathbf{x}^{(i)} = (\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_d^{(i)})^{\mathrm{T}}$ بوده و خروجی هدف به صورت دودویی باشد، یعنی هر خروجی هدف $t^{(i)} \in \{0,1\}$ باشد. مدلی را در نظر بگیرید که برای هر نمونه $p(t=1|\mathbf{x}^{(i)}, \mathbf{w}, b)$ شکل یک تابع منطقی را به خود می‌گیرد:

$$p(t=1|\mathbf{x}^{(i)}, \mathbf{w}, b) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b) = \frac{1}{1+\exp(-\sum_{j=1}^{d} \mathbf{w}_d \mathbf{x}_d^{(i)} - b)}$$

تابع درستنمایی یا شباهت[1] به صورت $p(t^{(1)}, t^{(2)}, \dots, t^{(N)}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}, \mathbf{w}, b)$ تعریف شده است.

علاوه بر این، فرض کنید برای منظم سازی[2]، یک توزیع پیشین گاسی بر روی وزن‌ها $\mathbf{w} = \{\mathbf{w}_1, \dots, \mathbf{w}_d\}$ قرار داده شده به طوریکه $p(\mathbf{w}) = N(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$ است.

الف) تابع زیان[3] را طوری تعریف کنید که منفی احتمال پسین خطا[4] نسبت به وزن‌ها باشد. تابع زیان را تا جایی که می‌شود ساده بنویسید (تمام مشتق‌ها را نشان دهید). برای این که این کار را کنید، فرض کنید داده‌ها $i.i.d.$[5] هستند (یعنی، همه داده‌ها از هم مستقل هستند و از یک توزیع یکسان تولید شده‌اند).

ب) نشان دهید که مشتق تابع زیان نسبت به وزن‌های $b, \mathbf{w}_i$ چه خواهد بود، یعنی $\frac{\partial \text{Loss}}{\partial b}$ و $\frac{\partial \text{Loss}}{\partial \mathbf{w}_i}$ را حساب کنید.

پ) در نهایت، شبه کد برای نزول گرادیان را با استفاده از مشتق‌های محاسبه شده در قسمت قبل بنویسید.

2- فرض کنید ما مدل‌های رگرسیون منطقی زیر را برای دسته‌بندی دودویی با تابع $\text{sigmoid}(z) = \frac{1}{1+e^{-z}}$، در نظر داریم:

- مدل اول: $P(y = 1|\mathbf{x}, w_1, w_2) = \text{sigmoid}(w_1 x_1 + w_2 x_2)$
- مدل دوم: $P(y = 1|\mathbf{x}, w_1, w_2) = \text{sigmoid}(w_0 + w_1 x_1 + w_2 x_2)$

ما نمونه های آموزشی زیر را داریم:

$$\mathbf{x}^{(1)} = [1,1]^{\mathrm{T}} \qquad \mathbf{x}^{(2)} = [1,0]^{\mathrm{T}} \qquad \mathbf{x}^{(3)} = [0,0]^{\mathrm{T}}$$

$$y^{(1)} = 1 \qquad y^{(2)} = -1 \qquad y^{(3)} = 1$$

الف) اگر از مدل اول استفاده کنیم فرقی می‌کند که برچسب نمونه سوم چه عددی باشد؟

ب) اگر برچسب سومین نمونه آموزشی را به 1- تغییر دهیم، مقدار وزن‌های w_2 و w_1 بعد از آموزش تغییری می‌کند؟

---

[1] Likelihood Function
[2] Regularization
[3] Loss Function
[4] Loss
[5] Independent and Identically distributed

ج) اگر از مدل اول استفاده کنید چه تغییری برای مقادیر وزن‌ها اتفاق می‌افتد؟

د) اگر از مدل دوم استفاده کنید چه تغییری برای مقادیر وزن‌ها اتفاق می‌افتد؟

پ) فرض کنید مدل رگرسیون منطقی دوم را برای n داده $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}$ با برچسب‌های $y^{(1)}, y^{(2)}, \dots, y^{(n)}$ با استفاده از بیشینه کردن احتمال لگاریتم شباهت یا درستنمایی[6] برچسب‌ها آموزش دهیم.

$$\sum_i \log \mathrm{P}\left(y^{(i)} \middle| \mathbf{x}^{(i)}, \mathbf{w}\right) - \frac{\lambda}{2}\ ||\mathbf{w}||^2 = \sum_i \log \mathrm{sigmoid}(y^{(i)}\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)}) - \frac{\lambda}{2}\ ||\mathbf{w}||^2 \quad (1)$$

برای λهای بزرگ (منظم‌سازی قوی[7])، لگاریتم شباهت به صورت تابع خطی از $\mathbf{w}$ عمل خواهد کرد:

$$\log \mathrm{sigmoid}\left(y^{(i)}\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)}\right) \approx \frac{1}{2}\ y^{(i)}\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} \quad (2)$$

لگاریتم شباهت نهایی را با استفاده از این تقریب بیان کنید (با مدل 1) و از عبارت بیشینه شباهت نسبت به $\mathbf{w}$ مشتق بگیرید و جواب را بر حسب λ و مجموعه آموزشی $\{\mathbf{x}^{(i)}, y^{(i)}\}$ بیان کنید. هم چنین بگویید رفتار $\mathbf{w}$ با افزایش λ چگونه تغییر می‌کند (فرض کنید هر $\mathbf{x}^{(i)} = (\mathrm{x}_1^{(i)}, \mathrm{x}_2^{(i)})$ باشد و هر $y^{(i)}$، 1- یا 1 باشد).

رگرسیون خطی

3- دو متغیر حقیقی x و y که x مشروط به y به شکل زیر تولید می‌شود را در نظر بگیرید:

$$\varepsilon \sim N(0, \sigma^2)$$

$$y = ax^2 + bx + c + \varepsilon$$

که ε یک متغیر است که یک نویز گاسی با توزیع نرمال با میانگین 0 و واریانس σ را نمایندگی می‌کند. یک رگرسیون خطی دارای پارامترهای a, b و c هم در فوق مشاهده شده است. احتمال شرطی y دارای رابطه

$$\mathrm{P}(y|x, a, b, c) \sim N(ax^2 + bx + c, \sigma^2)$$

است که به شکل زیر نوشته می‌شود:

$$\mathrm{P}(y|x, a, b, c) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y - ax^2 + bx + c)^2\right)$$

الف) فرض کنید n زوج آموزشی $(x^{(i)}, y^{(i)})$ داریم و واریانس σ مجهول است. با استفاده از بیشینه شباهت هر یک از پارامتر های این مدل را تخمین بزنید.

ب) اگر مجموعه داده ما چهار زوج آموزشی به شکل D={(5,8),(1,4),(3,-2),(0,-8)} داشته باشد، با استفاده از بیشینه شباهت هر یک از پارامترهای این مدل را محاسبه کنید.

---

[6] Log-Likelihood
[7] Strong regularization

4- اگر داده ما $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ ، $\mathbf{x}_i \in \mathbb{R}^m$ و $y_i \in \mathbb{R}^n$ باشد و مدل زیر را داشته باشیم:

$$y_i \sim N(\mathbf{w}^{\mathrm{T}}\mathbf{x}, \sigma^2)$$

شباهت به صورت زیر تعریف خواهد شد:

$$\mathrm{P}(y|\mathbf{x}, \mathbf{w}) = N(\mathbf{x}\mathbf{w}, \sigma^2\mathbf{I}) = \prod_{i=1}^{|\mathrm{D}|} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i)^2}{2\sigma^2}\right)$$

توزیع پیشین برای وزن‌ها را به صورت زیر در نظر بگیرید:

$$\mathrm{P}(\mathbf{w}) = N(0, \sigma_0^2\mathbf{I}) = \prod_{j=1}^{|\mathbf{w}|} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\mathbf{w}_j)^2}{2\sigma_0^2}\right)$$

تخمین MAP را برای وزن‌ها به دست آورید.

5- درست یا غلط بودن عبارات زیر را با دلیل مشخص کنید.

الف) بیشینه شباهت پارامتر a در مدل می‌تواند با استفاده از رگرسیون خطی یاد گرفته شود اگر مدل به این شکل باشد: $y_i = \log\left(x_1^{a_1} \mathrm{e}^{a_2}\right) + \varepsilon_i$ و $\varepsilon_i \sim N(0, \sigma^2)$ نویز باشد.

ب) بیشینه شباهت پارامتر a در مدل می‌تواند با استفاده از رگرسیون خطی یاد گرفته شود اگر مدل به این شکل باشد: $y_i = x_1^{a_1} \mathrm{e}^{a_2} + \varepsilon_i$ و $\varepsilon_i \sim N(0, \sigma^2)$ نویز باشد.

6- می‌خواهیم پارامترهای یک رگرسیون را برای یک مجموعه داده که می‌دانیم توسط یک تابع چند جمله‌ای تولید شده است اما درجه این چند جمله‌ای را نمی‌دانیم، یاد بگیریم. فرض کنید داده واقعا توسط یک چند جمله‌ای درجه 4 با یک نویز گاسی جمع شونده تولید شده باشد (یعنی به صورت $y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + \varepsilon$ با $\varepsilon \sim N(0,1)$ باشد). تعداد 100 زوج $\{\mathbf{x}, y\}$ برای آموزش و تعداد 100 زوج $\{\mathbf{x}, y\}$ دیگر برای آزمایش موجود است.

چون درجه چندجمله‌ای را نمی‌دانیم، پارامترهای دو مدل را برای داده یاد می‌گیریم، مدل A یک چندجمله‌ای درجه 2 و مدل B یک چندجمله‌ای درجه 5 است. کدام یک از این دو مدل بهتر بر داده آزمایشی منطبق می‌شوند؟ چرا؟

بخش پیاده سازی

سوال اول)

1. مقدمه

در این سوال، هدف پیاده‌سازی رگرسیون لاجستیک برای پیش‌بینی نجات یافتن یا نیافتن مسافران کشتی تایتانیک است. در این تمرین، با پلتفرم کگل نیز آشنا می‌شوید. کگل به عنوان منبع عظیمی از داده‌ها و پروژه‌های داده معتبر شناخته می‌شود و جامعه بزرگی از داده‌کاوان‌ها و تحلیل‌گران داده را به خود جذب کرده است. این پلتفرم به افراد امکان می‌دهد تا مهارت‌های خود را در حوزه داده و یادگیری ماشین بهبود بخشیده و در پروژه‌های واقعی مشارکت کنند و همچنین مسابقات بزرگی در حوزه یادگیری ماشین برگزار میکند.

2. معرفی مجموعه‌داده

این مجموعه‌داده، اطلاعات مربوط به مسافران کشتی تایتانیک را دارد. نام و توضیحات هر یک از ستون‌های این مجموعه‌داده به شرح زیر است:

- Survival: زنده ماندن یا نماندن مسافر که با 0 یا 1 نمایش داده می‌شود
- pclass: کلاس بلیط مسافر که می‌تواند کلاس 1، 2 و یا 3 باشد
- sex: جنسیت
- Age: سن مسافر
- sibsp: تعداد همسران یا خواهران و برادران همراه مسافر
- parch: تعداد فرزندان یا والدین همراه مسافر
- ticket: شماره بلیط مسافر
- cabin: شماره کابین مسافر
- embarked: کد بندر مربوطه که می‌تواند C، Q و یا S باشد. C = Cherbourg, Q = Queenstown, S = Southampton

3. ثبت پاسخ

برای ثبت پاسخ مساله از طریق لینک زیر پاسخ خود را در سایت کگل آپلود نمایید و اسکرین‌شات نتیجه‌ی آنرا به مستندات تحویلی خود اضافه نمایید.

https://www.kaggle.com/competitions/titanic/data

سوال دوم)

1. مقدمه

در این سوال، هدف پیاده‌سازی الگوریتم k نزدیکترین همسایه برای پیش‌بینی کلاس گل‌ها است. این پیش‌بینی به کمک اطلاعات مجموعه‌داده‌ی Iris صورت خواهد گرفت.

2. معرفی مجموعه داده

مجموعه‌داده‌ی انواع گل Iris، اطلاعات مربوط به کاسبرگ و گلبرگ گل و نوع آنها را دارد. نام و توضیحات مربوط به هر یک از ستون‌های این مجموعه‌داده به شرح زیر است:

- ID: آی دی منسوب به هر گل
- SepalLengthCm: طول کاسبرگ گل به سانتی متر
- SepalWidthCm: عرض کاسبرگ گل به سانتی متر
- PetalLengthCm: طول گلبرگ گل به سانتی متر
- PetalWidthCm: عرض گلبرگ گل به سانتی متر
- Species: نوع گل

3. مصورسازی

برای این سوال لازم است که نموداری مانند نمودار زیر رسم کنید و در آن نتایج پیشبینی مدل خود را نشان دهید.

4. ارزیابی

معیار accuracy را برای راه حل خود گزارش کنید.

5. آپلود

در انتها نتایج مدل خود را در فایلی به فرمت .csv ذخیره کنید. این فایل باید به تعداد سطرهای داده‌ی تست پیشبینی داشته باشد که در یک ستون با نام Species ذخیره شده است. فایل نتایج را به علاوه‌ی کد در پوشه‌ی مربوط به سوال در ریپازیتوری گیت‌هاب قرار دهید.

سوال سوم)

1. مقدمه

در این تمرین، هدف پیاده‌سازی رگرسیون لاجستیک برای پیش‌بینی اسپم بودن یا نبودن ایمیل‌ها است بر روی دیتاست Spambase است. در این سوال شما اجازه دارید تا از کتابخانه‌ی scikit learn استفاده کنید.

2. معرفی مجموعه‌داده

داده‌های مجموعه‌داده شامل اطلاعاتی عمدتا آماری از کلماتی است که در متن ایمیل وجود دارند. همچنین اسپم بودن یا نبودن ایمیل نیز در این ویژگی‌ها نمایش داده شده است. 58 ویژگی در این مجموعه‌داده وجود دارد که توضیحات مربوط به آنها به شرح زیر است:

- 48 ستون با مقادیر پیوسته با نام‌هایی به فرمت word_freq_WORD که درصد کلاماتی که با لغت WORD مطابقت دارند را نمایش می‌دهد.
- 6 ستون با مقادیر پیوسته با نام‌هایی به فرمت char_freq_CHAR که درصد کرکترهایی که با کرکتر CHAR تطابق دارند را نمایش می‌دهد.
- 1 ستون با نام capital_run_length_average که میانگین طول رشته‌های متشکل از حروف بزرگ (Capital) را نمایش می‌دهد.
- 1 ستون به نام capital_run_length_longest که طول بلندترین رشته‌ی متشکل از حروف بزرگ را نشان می‌دهد.
- 1 ستون به نام capital_run_length_total که تعداد کل حروف بزرگ به کاررفته در متن ایمیل را نمایش می‌دهد.
- 1 ستون به نام Class که در آن اسپم نبودن یا بودن ایمیل، به ترتیب با مقدار 0 یا 1 نمایش داده شده است.

این مجموعه داده یک مجموعه‌داده‌ی نامتعادل (imbalanced) است (تقریبا 39.4 درصد) که برای مدیریت آن باید از سه تکنیک oversampling،undersampling و تابع loss وزن‌دار استفاده نمایید.

همچنین در سند نهایی شما باید سه معیار precission recall, f1 وaccuracy را برای هر سه تکنیک گزارش کنید.

3. مصورسازی

برای این تمرین لازم است تا ماتریس درهم‌ریختگی (Confusion Matrix) مربوطه را رسم کنید و در سند نهایی تمرین قرار دهید.

4. آپلود

در انتها نتایج مدل خود را در فایلی به فرمت .csv ذخیره کنید. این فایل باید به تعداد سطرهای داده‌ی تست پیشبینی داشته باشد که در یک ستون با نام Class ذخیره شده است. فایل نتایج را به علاوه‌ی کد در پوشه‌ی مربوط به سوال در ریپازیتوری گیت‌هاب قرار دهید.

آنچه باید تحویل دهید:

لازم است که موارد زیر را به عنوان موارد مورد تحویل هر سوال در پوشه‌ی مربوط به آن سوال داخل ریپازیتوری گیت‌های قرار دهید:

- فایل ژوپیتر نوت بوک به فرمت .ipynb که شامل کد پاسخ و توضیحات مربوط به آن است.
- فایل پیش‌بینی‌های خروجی مدل که به فرمت Q#.csv خواهد بود و # شماره‌ی سوال می‌باشد.
- برای سوال اول، اسکرین شات نتیجه بارگذاری پیشبینی مدل در سایت کگل به شکلی که نام اکانت مشخص باشد.
- تحویل همه موارد فقط و فقط از گیت هاب صورت میگیرد. لینک تمرین:
- https://classroom.github.com/a/pBm5fJ4H
- مهلت تمرین تا سه شنبه 30 آبان، ساعت 3 بامداد می‌باشد.

موفق باشید