**فصل 16- رگرسیون لجستیک**

**16.0 مقدمه**

علی‌رغم اینکه رگرسیون لجستیک، رگرسیون نامیده می‌شود، در واقع یک تکنیک دسته‌بندی نظارت شده پرکاربرد است. رگرسیون لجستیک و بسط‌های آن، مانند رگرسیون لجستیک چند جمله‌ای، به ما این امکان را می‌دهند که با استفاده از یک رویکرد ساده و کاملاً قابل درک، احتمال اینکه یک مشاهده از یک کلاس خاص است را پیش‌بینی کنیم. در این فصل، آموزش انواع دسته‌بندها را با استفاده از scikit-learn پوشش خواهیم داد.

**16.1 آموزش یک دسته‌بند دوکلاسه**

**مسئله**

شما باید یک مدل طبقه‌بندی کننده ساده را آموزش دهید.

**راه‌حل**

آموزش یک رگرسیون لجستیک در scikit-learn با استفاده از LogisticRegression:



**بحث**

علیرغم وجود «رگرسیون» در نام خود، رگرسیون لجستیک در واقع یک طبقه‌بندی کننده‌ی دودویی پرکاربرد است (یعنی بردار هدف فقط می‌تواند دو مقدار بگیرد). در یک رگرسیون لجستیک، یک مدل خطی (به عنوان مثال، ) در یک تابع لجستیک (که سیگموئید نیز نامیده می‌شود)، گنجانده می‌شود، به طوری که:

که در آن احتمال کلاس 1 بودن مقدار هدف () در مشاهده iم است، بردار داده آموزشی است، و پارامترهایی هستند که باید یاد گرفته شوند و عدد اویلر است. اثر تابع لجستیک این است که مقدار خروجی تابع را بین 0 و 1 محدود می‌کند تا بتوان آن را به عنوان یک احتمال تفسیر کرد. اگر بزرگتر از 0.5 باشد، کلاس 1 پیش بینی می‌شود. در غیر این صورت کلاس 0 پیش بینی می‌شود.

در scikit-learn، ما می‌توانیم یک مدل رگرسیون لجستیک را با استفاده از LogisticRegression یاد بگیریم. پس از آموزش، می‌توانیم از مدل برای پیش‌بینی کلاس مشاهدات جدید استفاده کنیم:





در این مثال، مقدار پیش‌بینی شده برای مشاهده نمونه کلاس 1 است. علاوه بر این، می‌توانیم این‌که یک نمونه مشاهده شده متعلق به چه کلاسی است را ببینیم:





نمونه مشاهده شده ما با احتمال 18.8 درصد متعلق به کلاس 0 و با احتمال 81.1 درصد متعلق به کلاس 1 است.

**16.2 آموزش یک طبقه‌بندی کننده چند کلاسه**

**مسئله**

داده‌های مورد نظر شمال بیش از دو کلاس است و شما باید یک مدل طبقه‌بندی کننده را آموزش دهید.

**راه‌حل**

یک رگرسیون لجستیک را در scikit\_learn با LogisticRegression با استفاده از روش یکی در مقابل بقیه[[1]](#footnote-1) یا روش‌های چندجمله‌ای آموزش دهید:



**بحث**

رگرسیون‌های لجستیک به خودی خود فقط طبقه‌بندی‌کننده‌های دوکلاسه هستند، به این معنی که نمی‌توانند بردارهای هدف با بیش از دو کلاس را مدیریت کنند. با این حال، با دو روش هوشمندانه برای رگرسیون لجستیک می‌توان کاری کرد که از این روش برای طبقه‌بندی داده‌هایی که چند کلاسه هستند نیز استفاده کرد. در روش اول، در رگرسیون لجستیکِ یکی در مقابل بقیه (OVR) یک مدل جداگانه برای هر کلاس آموزش داده می‌شود که پیش‌بینی می‌کند آیا یک مشاهده، از آن کلاس است یا خیر (در نتیجه آن را به یک مسئله طبقه‌بندی دوکلاسه تبدیل می‌کند). فرض می‌کند که هر مسئله طبقه‌بندی (به عنوان مثال، کلاس 0 یا نه) مستقل است.

روش دیگر، در رگرسیون لجستیک چندجمله‌ای (MLR) تابع لجستیکی که در دستور العمل 15.1 دیدیم با یک تابع softmax جایگزین می‌شود:

که در آن احتمال کلاس بودن مقدار هدف () در مشاهده iم است و تعداد کل کلاس‌ها است. یکی از مزیت‌های عملی MLR این است که احتمالات پیش‌بینی شده آن‌ها با استفاده از روش predict\_proba قابل اطمینان‌تر هستند (یعنی کالیبره بودنِ بهتر).

هنگام استفاده از LogisticRegression، می‌توانیم با تنظیم مقدار multi\_class انتخاب کنیم که کدام یک از دو تکنیک را می‌خواهیم که البته پارامتر ovr به صورت پیش‌فرض استفاده می‌شود. ما می‌توانیم با تنظیم آرگومان بر روی multinomial روش را به MNL تغییر دهیم.

**16.3 کاهش واریانس از طریق تنظیم‌ساز**

**مسئله**

شما باید واریانس مدل رگرسیون لجستیک خود را کاهش دهید.

**راه‌حل**

تنظیم ابرپارامتر قدرت تنظیم[[2]](#footnote-2)، C:



**بحث**

منظم سازی، روشی برای جریمه کردن مدل‌های پیچیده برای کاهش واریانس آن‌‎ها است. به طور خاص، یک عبارتِ جریمه به تابع هزینه اضافه می‌شود که ما در تلاش برای به حداقل رساندن آن هستیم، معمولاً جریمه‌های L1 و L2. در پنالتی L1:

که در آن پارامترهای j ام از ویژگی‌های p است که آموخته می‌شوند و α یک ابرپارامتر است که قدرت تنظیم‌سازی را نشان می‌دهد. با پنالتی L2:

مقادیر بالاتر α باعث افزایش جریمه برای مقادیر پارامترهای بزرگتر (به عنوان مثال، مدل‌‎های پیچیده‌تر) می‌شود. scikit-learn از روش رایج استفاده از C به جای α پیروی می‌کند که در آن C معکوس قدرت تنظیم‌سازی است: . برای کاهش واریانس در حین استفاده از رگرسیون لجستیک، می‌توانیم C را به عنوان یک فراپارامتر در نظر بگیریم تا مقدار C را که بهترین مدل را ایجاد می‌کند، تنظیم کنیم. در scikit-learn می‌توانیم از کلاس LogisticRegressionCV برای تنظیم مؤثر C استفاده کنیم. پارامتر LogisticRegressionCV، Cs، می‌تواند محدوده‌ای از مقادیر را برای جستجوی C بپذیرد (اگر لیستی از اعداد اعشاری به عنوان یک آرگومان ارائه شود) یا اگر یک عدد صحیح ارائه شود، فهرستی از تعداد زیادی از مقادیر کاندید را که از مقیاس لگاریتمی‌بین 10،000- تا 10،000 ترسیم شده اند ایجاد می‌کند.

متأسفانه، LogisticRegressionCV به ما اجازه نمی‌دهد تا در مورد عبارات مختلف جریمه‌ها جستجو کنیم. برای انجام این کار، باید از تکنیک‌های انتخاب مدل کمتر کارآمدی که در فصل 12 بحث شده است استفاده کنیم.

**16.4 آموزش یک دسته‌بند بر روی داده‌های بسیار بزرگ**

**مسئله**

شما باید یک مدل دسته‌بند ساده را روی مجموعه بسیار بزرگی از داده‌ها آموزش دهید.

**راه‌حل**

با استفاده از حل‌کننده گرادیان میانگین تصادفی[[3]](#footnote-3) (SAG) یک رگرسیون لجستیک را در scikit-learn با LogisticRegression آموزش دهید:



**بحث**

رگرسیون لجستیک scikit-learn تعدادی تکنیک را برای آموزش یک رگرسیون لجستیک ارائه می‌دهد که حل‌کننده[[4]](#footnote-4) نامیده می‌شوند. اکثر اوقات scikit-learn بهترین حل‌کننده را به طور خودکار برای ما انتخاب می‌کند یا به ما هشدار می‌دهد که نمی‌توانیم کاری با آن حل کننده انجام دهیم. با این حال، یک مورد خاص وجود دارد که باید از آن آگاه باشیم.

در حالی که توضیح دقیق، فراتر از محدوده این کتاب است (برای اطلاعات بیشتر به اسلایدهای مارک اشمیت در «[همچنین ببینید](http://bit.ly/2GRrVw0)» مراجعه کنید)، نزول گرادیان میانگین تصادفی به ما امکان می‌دهد که زمانی که داده‌های ما بسیار بزرگ هستند، یک مدل را بسیار سریع‌تر از حل‌کننده‌های دیگر آموزش دهیم. با این حال، به مقیاس‌بندی ویژگی‌ها نیز بسیار حساس است. بنابراین استاندارد کردن ویژگی‌های ما اهمیت ویژه ای دارد. ما می‌توانیم الگوریتم یادگیری خود را برای استفاده از این حل کننده با تنظیم solver=’sag’ تنظیم کنیم.

**همچنین ببینید:**

* [به حداقل رساندن مجموع محدود با الگوریتم گرادیان میانگین تصادفی، مارک اشمیت](http://www.birs.ca/workshops/2014/14w5003/files/schmidt.pdf)

**16.5 مدیریت کلاس‌های نامتوازن**

**مسئله**

شما باید یک مدل طبقه‌بندی کننده ساده را آموزش دهید.

**راه‌حل**

آموزش یک رگرسیون لجستیک در scikit-learn با استفاده از LogisticRegression:



**بحث**

مانند بسیاری از الگوریتم‌های یادگیری دیگر در Sicit-Learn، LogisticRegression با یک روش داخلی برای مدیریت کلاس‌های نامتعادل همراه است. اگر کلاس‌های بسیار نامتعادل داریم و در طول پیش‌پردازش به آن توجه نکرده‌ایم، می‌توانیم از پارامتر class\_weight برای وزن‌دهی کلاس‌ها استفاده کنیم تا مطمئن شویم که ترکیب متعادلی از هر کلاس داریم. به طور خاص، آرگومان balanced، به طور خودکار، کلاس‌ها را به طور معکوس و متناسب با فرکانس آنها وزن‌دهی می‌کند:

که در آن وزن کلاس j، n تعداد مشاهدات، تعداد مشاهدات در کلاس j و k تعداد کل کلاس‌ها است.

1. . one-vs-rest [↑](#footnote-ref-1)
2. Regularization [↑](#footnote-ref-2)
3. Stochastic Average Gradient (SAG) [↑](#footnote-ref-3)
4. - solver [↑](#footnote-ref-4)