**فصل 5: مدیریت داده‌های دسته‌ای[[1]](#footnote-1)**

**5. مقدمه**

اغلب مفید است که اشیاء را نه از نظر کمیت، بلکه از نظر کیفیت اندازه‌گیری کنیم. ما اغلب این اطلاعات کیفی را به عنوان عضویت یک نمونه در یک دسته‌ی مجزا مانند جنسیت، رنگ یا برند خودرو نشان می‌دهیم. با این حال، همه‌ی داده‌های دسته‌ای یکسان نیستند. مجموعه‌ای از دسته‌ها که بدون ترتیب ذاتی هستند، اسمی[[2]](#footnote-2) نامیده می‌شوند. نمونه‌هایی از دسته‌های اسمی عبارتند از:

* آبی، قرمز، سبز
* مرد، زن
* موز، توت فرنگی، سیب

در مقابل، وقتی مجموعه‌ای از مقوله‌ها دارای ترتیب طبیعی هستند، از آن به عنوان ترتیبی[[3]](#footnote-3) یاد می‌کنیم. مثلا:

* کم، متوسط، ​​زیاد
* جوان، پیر
* موافق، خنثی، مخالف

علاوه بر این، اطلاعات طبقه‌بندی اغلب در داده‌ها به‌عنوان بردار یا ستونی از رشته‌ها نشان داده می‌شوند (مانند «Maine»، «Texas»، «Delaware»). مشکل این است که بیشتر الگوریتم‌های یادگیری ماشین به ورودی‌هایی نیاز دارند که مقادیر آن‌ها عددی باشند. الگوریتم k نزدیکترین همسایه، یک مثال ساده ارائه می‌دهدکه یک مرحله در الگوریتم محاسبه فواصل بین مشاهدات است و اغلب از فاصله اقلیدسی استفاده می‌کند:

که در آن x و y دو مشاهده‌های ما هستند و زیرمجموعه‌ی i، مقدار ویژگی مشاهدات را نشان می‌دهد. با این حال، اگر مقدار x یک رشته باشد محاسبه فاصله‌ی آشکارا غیرممکن است(به عنوان مثال، " Texas"). در عوض، باید رشته را به فرمت عددی تبدیل کنیم تا بتوان آن را در معادله‌ی فاصله اقلیدسی وارد کرد. هدف ما ایجاد تحولی است که به درستی، اطلاعات موجود در دسته‌ها را منتقل کند(ترتیبی بودن، فواصل نسبی بین دسته‌ها و غیره). دراین فصل، تکنیک‌هایی را برای ایجاد این تحول و همچنین غلبه بر چالش‌هایی که اغلب هنگام مدیریت داده‌های طبقه‌بندی با آن مواجه می‌شوند، پوشش خواهیم داد.

**1.5رمزگذاری ویژگی­های طبقه‌بندی اسمی**

**مسئله**

شما یک ویژگی با کلاس­های اسمی دارید که ترتیب ذاتی ندارد (به عنوان مثال، سیب، گلابی، موز).

**راه حل**

با استفاده ازکتابخانه scikit-learn’s ، ویژگی را یک‌بار OHE با LabelBinarizer می­کنیم.



ما می‌توانیم از متد classes برای خروجی کلاس‌ها استفاده کنیم:

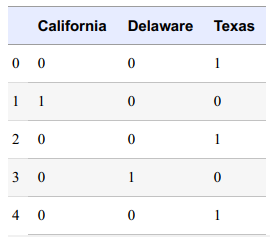


اگر بخواهیم رمزگذاری OHE را معکوس کنیم، می‌توانیم از inverse\_transform استفاده کنیم:



ما حتی می‌توانیم از pandas برای رمزگذاری یکباره این ویژگی استفاده کنیم.





یکی از توانایی‌های مفید scikit-learn موقعیتی است که در آن هر یک از آنها چندین کلاس را فهرست می‌کند:



یک بار دیگر می‌توانیم کلاس‌ها را با متد classes ببینیم:



**بحث**

ممکن است فکر کنیم استراتژی مناسب این است که به هر کلاس یک مقدار عددی اختصاص دهیم (به عنوان مثال، Texas = 1, California = 2)

با این حال، وقتی کلاس‌های ما ترتیب ذاتی ندارند (به عنوان مثال، Texas «کم‌تر از California » نیست)، مقادیر عددی ما به اشتباه، ترتیبی را ایجاد می‌کنند که وجود ندارد. استراتژی مناسب، ایجاد یک ویژگی باینری برای هر کلاس در ویژگی اصلی است. این مورد، اغلب رمزگذاری یک‌طرفه(در ادبیات یادگیری ماشین) یا ساختگی(در ادبیات آماری و تحقیقاتی) نامیده می‌شود. ویژگی راه حل ما یک بردار حاوی سه کلاس (یعنی 'California', 'Florida', 'Texas') بود.

در رمزگذاری OHE، هر کلاس به ویژگی خاص خود تبدیل می‌شود که در صورت ظاهر شدن کلاس، 1 و در غیر این صورت 0 می‌شود. از آنجا که ویژگی ما دارای سه کلاس بود، OHE سه ویژگی باینری(یکی برای هر کلاس) را برگرداند. با استفاده از OHE، می‌توانیم عضویت یک مشاهده را در یک کلاس ثبت کنیم و در عین حال این مفهوم را حفظ کنیم که کلاس فاقد هر نوع سلسله مراتبی است. در نهایت، شایان ذکر است که اغلب، توصیه می‌شود که پس از رمزگذاری یکباره‌ی یک ویژگی، یکی از ویژگی‌های OHE را در ماتریس حاصل رها کنیم تا از وابستگی خطی جلوگیری کنیم.

همچنین ببینید:

* [Dummy Variable Trap، Algosome](https://www.algosome.com/articles/dummy-variable-trap-regression.html)
* [رها کردن یکی از ستون‌ها هنگام استفاده از OHE](https://stats.stackexchange.com/questions/231285/dropping-one-of-the-columns-when-using-one-hot-encoding)

**2.5 رمزگذاری طبقه‌بندی ترتیبی**

**مسئله**

شما یک ویژگی طبقه‌بندی ترتیبی دارید(به عنوان مثال، زیاد، متوسط، کم).

**راه حل**

برای تبدیل برچسب‌های رشته‌ها به معادل‌های عددی از روش replace در کلاس DataFrame در کتابخانه‌ی Pandas استفاده کنید.



**بحث**

اغلب ما یک ویژگی با کلاس‌هایی داریم که دارای نوعی طبیعی هستند. یک مثال معروف مقیاس لیکرت است:

* کاملا موافقم
* موافق
* خنثی
* مخالف بودن
* به شدت مخالف

هنگام رمزگذاری ویژگی‌ها برای استفاده در یادگیری ماشین، باید کلاس‌های ترتیبی را به مقادیر عددی تبدیل کنیم که مفهوم مرتب‌سازی را حفظ کنند. متداول‌ترین روش، ایجاد دیکشنری‌ای است که برچسب رشته‌ی کلاس را به یک عدد ترسیم می‌کند و سپس آن نقشه را روی ویژگی اعمال می‌کند. مهم است که انتخاب مقادیر عددی ما بر اساس اطلاعات قبلی ما در مورد کلاس‌های ترتیبی باشند. در راه حل ما، بالا به معنای واقعی کلمه سه برابر بزرگتر از پایین است. این در هر موردی خوب است، اما اگر فواصل فرضی بین کلاس‌ها برابر نباشد، می‌تواند خراب شود:



در این مثال، فاصله بین Low و Medium برابر با فاصله بین Medium و Barely More Than Medium است که تقریباً دقیق نیست. بهترین رویکرد این است که نسبت به مقادیر عددی نگاشت شده به کلاس‌ها آگاه باشید:



**3.5 رمزگذاری دیکشنری ویژگی‌ها**

**مسئله**

شما یک فرهنگ لغت(دیکشنری) دارید و می‌خواهید آن را به یک ماتریس ویژگی تبدیل کنید.

**راه حل**

از DictVectorizer استفاده کنید.

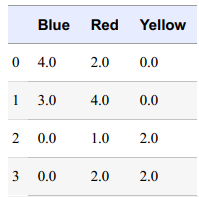


به طور پیش فرض، DictVectorizer یک ماتریس پراکنده را خروجی می‌دهد که فقط عناصرِ با مقداری غیر از 0 را ذخیره می‌کند. هنگامی که ماتریس‌های عظیمی داریم(که اغلب در پردازش زبان طبیعی با آن مواجه می‌شویم) و می‌خواهیم نیازهای حافظه را به حداقل برسانیم، این مورد می‌تواند بسیار مفید باشد. می‌توانیم DictVectorizer را مجبور کنیم که یک ماتریس متراکم را با استفاده از sparse=False تولید کند. با استفاده از متد get\_feature\_names می‌توانیم نام هر ویژگی تولید شده را بدست آوریم:



در حالی که ضروری نیست، برای این مثال می‌توانیم یک DataFrame ایجاد کنیم تا خروجی را بهتر ببینیم:





**بحث**

دیکشنری یک ساختار داده‌ای محبوب است که توسط بسیاری از زبان‌های برنامه نویسی استفاده می‌شود. با این حال، الگوریتم‌های یادگیری ماشین انتظار دارند داده‌ها در قالب یک ماتریس باشند. ما می‌توانیم این کار را با استفاده از dictvectorizer انجام دهیم. این یک وضعیت رایج هنگام کار با مدل‌های پردازش زبان طبیعی است. برای مثال، ممکن است مجموعه‌ای از اسناد داشته باشیم و برای هر سند یک فرهنگ لغت داشته باشیم که حاوی تعداد دفعاتی است که هر کلمه در سند ظاهر می‌شود. با استفاده از dictvectorizer، ما به راحتی می‌توانیم یک ماتریس ویژگی ایجاد کنیم که در آن هر ویژگی تعداد دفعاتی است که یک کلمه در هر سند ظاهر می‌شود:



در مثال اسباب بازی ما، تنها سه کلمه منحصر به فرد (Red, Yellow, Blue) وجود دارد؛ بنابراین تنها سه ویژگی در ماتریس ما وجود دارد. با این حال، می‌توانید تصور کنید که اگر هر سند واقعاً یک کتاب در کتابخانه دانشگاه باشد، ماتریس ویژگی‌های ما بسیار بزرگ خواهد بود(و سپس می‌خواهیم sparse را روی True تنظیم کنیم).

همچنین ببینید:

* [نحوه استفاده از دیکشنری‌ها در پایتون](https://www.pythonforbeginners.com/dictionary/how-to-use-dictionaries-in-python)
* [ماتریس‌های پراکنده SciPy](https://docs.scipy.org/doc/scipy/reference/sparse.html.)

**4.5 درج مقادیر گمشده کلاس**

**مسئله**

شما یک ویژگی طبقه‌بندی دارید که حاوی مقادیر گم شده است و می‌خواهید آن را با مقادیر پیش‌بینی شده جایگزین کنید.

**راه حل**

راه‌حل ایده‌آل، آموزش یک الگوریتم طبقه‌بندی کننده یادگیری ماشین برای پیش‌بینی مقادیر از دست رفته است که معمولاً یک طبقه‌بندی کننده k نزدیک‌ترین همسایه(KNN) استفاده می‌شود:



یک راه حل جایگزین، پر کردن مقادیر از دست رفته با متداول‌ترین مقدار ویژگی است:



**بحث**

وقتی مقادیر گمشده در یک ویژگی طبقه‌بندی شده داریم، بهترین راه حل ما این است که جعبه ابزار الگوریتم‌های یادگیری ماشین را برای پیش‌بینی مقادیر مشاهدات از دست رفته باز کنیم. ما می‌توانیم این کار را با در نظر گرفتن ویژگی با مقادیر از دست رفته به عنوان بردار هدف و سایر ویژگی‌ها به عنوان ماتریس ویژگی انجام دهیم. الگوریتمی که معمولاً مورد استفاده قرار می‌گیرد KNN است (که بعداً در این کتاب به طور عمیق مورد بحث قرار می‌گیرد) که به مقدار گمشده‌، کلاس میانه‌ی k نزدیکترین مشاهدات همسایه را نسبت می‌دهد. از طرف دیگر، می‌توانیم مقادیر از دست رفته را با متداول‌ترین کلاس ویژگی پر کنیم. اگرچه نسبت به KNN پیچیده‌تر است، اما برای داده‌های بزرگتر، مقیاس پذیرتر است. در هر صورت، توصیه می‌شود که یک ویژگی باینری را شامل شود که نشان می‌دهد کدام مشاهدات حاوی مقادیر منتسب هستند.

**همچنین**

* [غلبه بر مقادیر از دست رفته در یک طبقه‌بندی جنگل تصادفی](http://bit.ly/2HSsNBF)
* [مطالعه K -نزدیکترین همسایه به عنوان یک روش انتساب](http://bit.ly/2HS9sAT)

**5.5 مدیریت کلاس‌های نامتعادل**

**مسئله**

شما یک بردار هدف با کلاس‌های بسیار نامتعادل دارید.

**راه حل**

داده‌های بیشتری جمع آوری کنید. اگر این راه برای شما امکان پذیر نیست، معیارهای مورد استفاده برای ارزیابی مدل خود را تغییر دهید. اگر این مورد نیز جواب نداد، از پارامترهای وزن کلاس داخلی مدل (در صورت وجود) و نمونه‌برداری پایین یا نمونه‌برداری استفاده کنید. ما معیارهای ارزیابی را در فصل بعدی پوشش می‌دهیم، بنابراین فعلاً اجازه دهید روی پارامترهای وزن کلاس، نمونه‌برداری پایین و نمونه‌برداری بالا تمرکز کنیم. برای نشان دادن راه حل‌های خود، باید داده‌هایی با کلاس‌های نامتعادل ایجاد کنیم.

مجموعه داده‌ی Fisher’s شامل سه کلاس متعادل از 50 مشاهدات است که هر کدام گونه‌های گل را نشان می‌دهند(Iris setosa، Iris virginica و Iris versicolor). برای عدم تعادل مجموعه داده، 40 مورد از 50 مشاهدات Iris setosa را حذف می‌کنیم و سپس کلاس Iris virginica و Iris versicolor را ادغام می‌کنیم. نتیجه نهایی یک بردار هدف باینری است که نشان می‌دهد آیا مشاهده گل زنبق ستوسا است یا خیر. نتیجه 10 مشاهده از Iris setosa (کلاس 0) و 100 مشاهده از Iris setosa ( کلاس 1) است:



بسیاری از الگوریتم‌ها در Sikit-Learn پارامتری را به کلاس‌های وزنی در طول تمرین ارائه می‌کنند تا تأثیر عدم تعادل آنها را خنثی کند؛ در حالی که ما هنوز آن را پوشش نداده‌ایم. RandomForestClassifier یک الگوریتم طبقه‌بندی محبوب است و شامل یک پارامتر class\_weight است. شما می‌توانید یک آرگومان ارسال کنید که وزن کلاس مورد نظر را به صراحت مشخص کند:



یا می‌توانید به صورت متعادل پاس دهید که به طور خودکار وزن‌هایی را به طور معکوس با فرکانس‌های کلاس ایجاد می‌کند:



از طرف دیگر، می‌توانیم کلاس اکثریت را نمونه‌سازی پایین انجام دهیم یا کلاس اقلیت را نمونه‌برداری بالا کنیم. در نمونه‌گیری پایین، ما به‌طور تصادفی، بدون جایگزینی از کلاس اکثریت(یعنی کلاس با مشاهدات بیشتر) نمونه‌برداری می‌کنیم تا یک زیرمجموعه جدید از مشاهدات برابر با کلاس اقلیت ایجاد کنیم. به عنوان مثال، اگر کلاس اقلیت 10 مشاهده داشته باشد، به طور تصادفی 10 مشاهده را از طبقه اکثریت انتخاب می‌کنیم و از آن 20 مشاهده به عنوان داده استفاده می‌کنیم. در اینجا ما دقیقاً این کار را با استفاده از داده‌های Iris نامتعادل خود انجام می‌دهیم:



گزینه دیگر ما این است که از کلاس اقلیت نمونه‌برداری کنیم. در این نمونه‌برداری، برای هر مشاهده در کلاس اکثریت، به طور تصادفی یک مشاهده از کلاس اقلیت را با جایگزینی انتخاب می‌کنیم. نتیجه نهایی همان تعداد مشاهدات طبقات اقلیت و اکثریت است.

Upsampling بسیار شبیه به downsampling اجرا می‌شود، فقط به صورت معکوس:



**بحث**

در دنیای واقعی هم، کلاس‌های نامتعادل در همه جا وجود دارند. بیشتر بازدیدکنندگان روی دکمه خرید کلیک نمی‌کنند و بسیاری از انواع سرطان‌ها خوشبختانه نادر هستند. به همین دلیل، مدیریت کلاس­های نامتعادل یک فعالیت رایج در یادگیری ماشین است. بهترین استراتژی ما صرفاً جمع آوری مشاهدات بیشتر به ویژه مشاهدات طبقه اقلیت است. ولی این اغلب ممکن نیست، بنابراین ما باید به گزینه‌های دیگر متوسل شویم.

راهبرد دوم، استفاده از معیار ارزیابی مدل است که برای کلاس‌های نامتعادل مناسب‌تر است. دقت، اغلب به‌عنوان معیاری برای ارزیابی عملکرد یک مدل استفاده می‌شود، اما وقتی کلاس‌های نامتعادل وجود دارند، دقت می‌تواند مناسب باشد. به عنوان مثال، اگر فقط 0.5٪ از مشاهدات سرطان از نوع نادر داشته باشند، حتی یک مدل ساده که پیش‌بینی می­کند هیچ کس سرطان ندارد، 99.5٪ دقیق خواهد بود. واضح است که این ایده آل نیست. برخی از معیارهای دیگر مانند ماتریس‌های درهم ریختگی، precision، recall، امتیازات F1 و منحنی‌های ROC را در فصل‌های بعدی مورد بحث قرار می‌دهیم.

استراتژی سوم، استفاده از پارامترهای وزن کلاس موجود است. پیاده سازی برخی از مدل‌ها به ما این امکان را می­دهد که الگوریتم را برای کلاس­های نامتعادل تنظیم کنیم. خوشبختانه، بسیاری از طبقه‌بندی‌کننده‌هایscikitlearn دارای پارامتر class\_weight هستند که آن را به گزینه خوبی تبدیل کرده است.

راهبردهای چهارم و پنجم یعنی نمونه‌برداری پایین و نمونه‌برداری بالا به یکدیگر مرتبط هستند. در نمونه‌برداری پایین، یک زیرمجموعه تصادفی از کلاس اکثریت با اندازه مساوی با کلاس اقلیت ایجاد می­کنیم. در نمونه‌برداری بالا، ما به طور مکرر با جایگزینی از کلاس اقلیت نمونه‌برداری می‌کنیم تا اندازه آن با کلاس اکثریت برابر باشد. تصمیم بین استفاده از نمونه‌برداری پایین و نمونه‌برداری بالا به یک زمینه‌ی خاص بستگی دارد، و به طور کلی باید هر دو را امتحان کنیم تا ببینیم کدامیک نتایج بهتری ایجاد می‌کند.

1. Categorical data [↑](#footnote-ref-1)
2. nominal [↑](#footnote-ref-2)
3. ordinal [↑](#footnote-ref-3)