درس‌نامه دهم

یکی از مباحث مهم در یادگیری ماشین، انتخاب مدل مناسب می‌باشد. به‌عنوان مثال در مساله دسته‌بندی دو روش مختلف را یاد گرفته‌ایم : LDA و Logistic Regression. حال سوال این است که بین این دو کدام را انتخاب نماییم.

معیارهای انتخاب مدل

دلایلی را که یک مدل را بر مدل دیگر ترجیح می‌دهیم می‌تواند متعدد باشد. در زیر تعدادی از آن‌ها را بیان می‌داریم.

۱. کیفیت تعمیم‌پذیری

شاید از مهم‌ترین دلایلی که یک مدل را بر مدل دیگر ترجیح می‌دهیم کیفیت آن در عمل می‌باشد. در این درس از این جهت به انتخاب مدل می‌پردازیم.

۲. پیچیدگی محاسباتی

ممکن است زمان یادگیری یک مدل بسیار طولانی باشد و یا از قدرت محاسباتی موجود بالاتر باشد. یا اینکه یک مدل را بایستی برای داده‌های جدید آموزش دهیم که بار محاسباتی کم مورد نظر است. همچنین در زمان اجرای مدل ممکن است برای سرویس دادن به تعداد زیادی کاربر مجبور باشیم تا از مدل های ساده تر استفاده نماییم. همچنین مدل‌های ساده‌تر قابلیت اجرا روی دستگاه‌های ساده‌تر را دارند و بنابراین ممکن ازست آن‌ها را ترجیح دهیم.

۳. تفسیرپذیری

در یک کاربرد عملی ممکن است کاربر به اینکه پیشگویی چگونه و با چه مولفه هایی انجام گرفته است اهمیت داشته باشد. به‌عنوان مثال در مسائل پزشکی یک متخصص نیاز دارد که کدام ژن‌ها در یک بیماری تاثیر دارند.

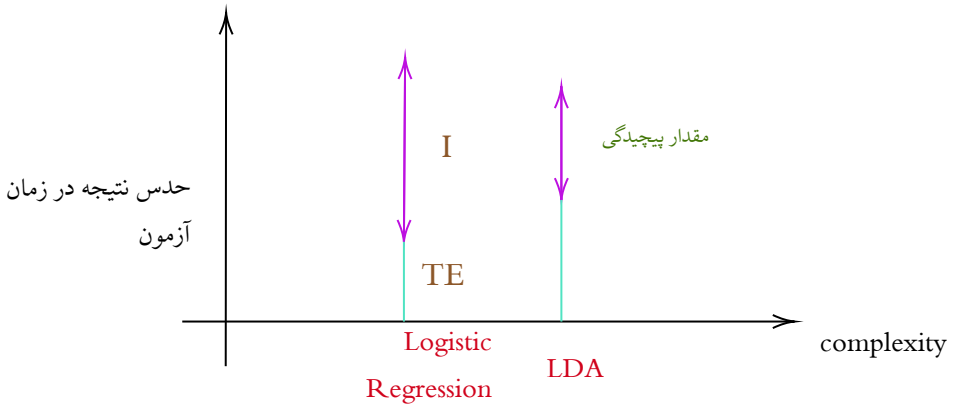
روش‌های انتخاب مدل

بدین منظور از داده‌ها بایستی استفاده کنیم قدرت تعمیم این دو روش را بسنجیم. بدین منظور دو راه وجود دارد.

۱. روش‌های تحلیلی

۲. روش‌های احتمالاتی

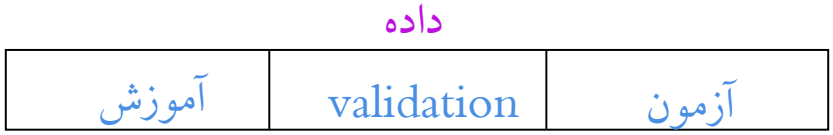
در روش تحلیلی از روی داده‌های آموزش یک حدسی برای نتیجه آزمون می‌زنیم و از روی آن یکی از روش‌ها را که بهترین نتیجه دارد را انتخاب می‌کنیم:



به عنوان مثال در شکل فوق هر چند LDA در زمان آموزش دارای ریسک بالاتری است ولیکن حدس ما از مقدار آزمون کوچکتر بوده و بنابراین آن را انتخاب می‌نماییم.

$$\widehat{\text{TEST ERROR}} = TE + I$$

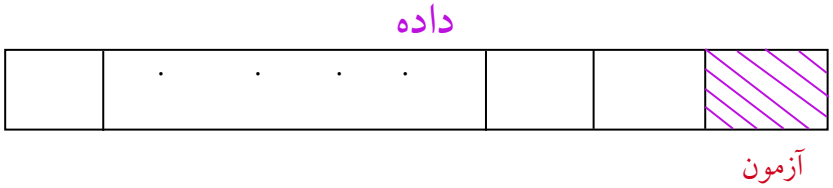
در روش احتمالاتی، داده‌ها را به سه قسمت تقسیم می‌نماییم.



با داده‌های آموزش پارامترهای مدل‌ها را یاد می‌گیریم و سپس با داده‌های $validation$ مقدار تعمیم هر کدام را تخمین می‌زنیم. سپس آن‌را که بهترین تعمیم را دارد انتخاب می‌کنیم. داده‌های آزمون برای ارزیابی بهترین مدل مورد استفاده قرار می‌گیرند.

نمونه برداری مجدد (Resampling)

در روش احتمالاتی تقسیم داده‌ها به سه قسمت آموزش، اعتبارسنجی و آزمون زمانی مناسب است که داده‌های فراوانی در اختیار داشته باشیم. اگر داده‌ها به میزان کافی نباشد بایستی از روش‌های نمونه‌برداری مجدد استفاده نماییم. یکی از روش‌های پرکاربرد روش اعتبارسنجی متقابل cross validation می‌باشد. در ادامه به این روش می‌پردازیم.



در این روش یک قسمت از داده‌ها را به عنوان آزمون جدا کرده و مابقی داده را به K قسمت مساوی تقسیم مینماییم. آنگاه داده‌های $K-1$ قسمت را برای یادگیری پارامترهای مدل‌ها مورد استفاده قرار می‌دهیم و اعتبار سنجی را روی آن قسمت که استفاده نشده است، انجام می‌دهیم. این عمل را K بار انجام می‌دهیم و برای هر مدل K مقدار از تخمین میزان ریسک بدست می‌آوریم، متوسط این اعداد برابر است با تخمین از میزان ریسک مدل‌ها:

$$CV_{(k)} = \frac{1}{K}\sum_{i=1}^{K} L_i$$

به طور خاص اگر تعداد نمونه‌های مورد استفاده n باشد و $K = n$ قرار دهیم، آنگاه اعتبار سنجی یک‌طرفه خواهیم داشت. Leave-one-out CV (LOOCV)