دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین | بهار ۱۴۰۰

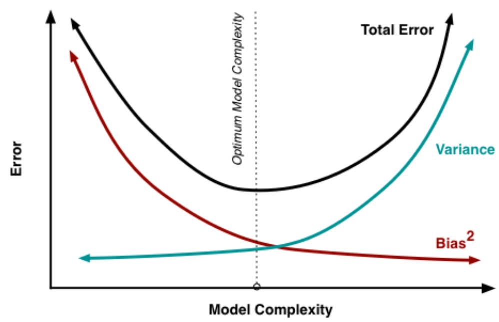
پاسخ تمرین سری سوم

مدرس: دکتر محمّدحسین رهبان

پاسخ سوال ۱

آ) به اسلایدهای درس مراجعه کنید. هم‌چنین در فصل ۳ کتاب $Bishop$ نیز این تجزیه با توضیحات کامل آورده شده است.

ب)



پ) استفاده از تعداد بیشتری $feature$ یا ویژگی - تنظیم پارامترهای مدل به شکلی که مدل پیچیده‌تر شود / در نظر گرفتن مدل پیچیده‌تر - تنظیم داده‌ها. در بسیاری از مسائل دسته‌های متنوعی وجود دارد ولی تعداد آن‌ها با یکدیگر برابر نیست. مثلا تشخیص سرطان. در چنین مواردی مدل به این سمت جهت‌دهی می‌شود که نمونه‌ی تست به احتمال بالایی سرطان ندارد در حالیکه اگر کیس مثبت، منفی اعلام شود هزینه‌ی زیادی برای ما خواهد داشت. در چنین مواردی از $oversampling$ یا $undersampling$ استفاده می‌شود و یا وزن داده‌هایی که کمتر ظاهر شده‌اند را بیشتر می‌گیریم.

ت) پاسخ برای حالت کلی این مسئله نمی‌توان ارائه داد و حالت‌های خاص متفاوتی پیدا می‌شود. عوامل مختلفی که باعث می‌شود برای حالت کلی نتوانیم پاسخی ارائه دهیم، میزان هم‌بستگی بین فیچرهایی که حذف می‌شوند و فیچرهایی که باقی می‌مانند، نسبت تعداد آن‌ها به کل فیچرها، میزان هم‌بستگی برچسب‌ها و این فیچرها و ... هستند.

اما اگر از موارد خاص بگذریم به صورت کلی کاهش فیچر به معنای از دست دادن $information$ از فضای داده است و این یعنی مدل‌مان ساده‌تر خواهد بود و به سمت خطای بیشتر روی داده‌های آموزش حرکت می‌کنیم که این به معنای کاهش واریانس و افزایش بایاس است.

هر یک از دو پاسخ بالا قابل قبول هستند.

ث) شناسایی و حذف فیچرهای غیراصلی (غیراصلی بودن تعاریف متفاوت می‌تواند داشته باشد، مثلا میزان هم‌بستگی با برچسب‌ها، یا میزان تنک بودن آن) - افزایش اندازه‌ی داده‌ی آموزش (تعداد نمونه‌ها/سمپل‌های بیشتر، یا معادلا سطرهای بیشتر در ماتریس داده‌مان) - تنظیم پارامترهای مدل برای جلوگیری از پیچیدگی زیاد

ج) فیچرهای هم‌بسته به صورت کلی باعث افزایش واریانس می‌شوند و این یعنی مدل ما از قدرت $generalization$ کمتری برخوردار خواهد بود. این به معنی بایاس کمتر نیز هست. با حذف فیچرهای هم‌بسته واریانس کم می‌شود و میزان بایاس رشد می‌کند.

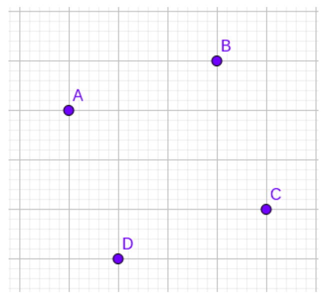
پاسخ سوال ۲ توضیحات کامل در این سند آمده است. قسمت $Corollary\ 5.9$ مشخصا به سؤال تمرین مربوط است. در واقع این قضیه با استفاده از نامساویِ $Jensen$ ثابت می‌کند که زمانی که تابع پیش‌بینی‌مان روی یک متغیر تصادفی حاشیه‌سازی شده باشد، واریانس کمتر و بایاس برابر نسبت به زمانی که این کار را انجام نمی‌دهیم خواهد داشت (نکته‌ی بسیار مهم در نتیجه‌ی این قضیه این است که در این حالت خطا بیشتر نمی‌شود و این یعنی عملکرد حالت سوم از دو حالت قبل بهتر است).

پاسخ سوال ۳

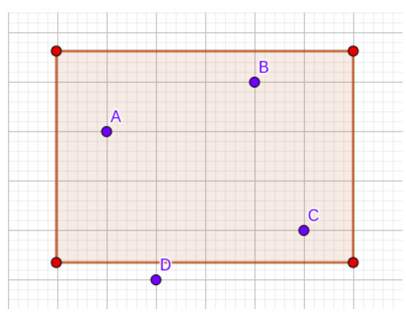
(آ) بخش ۱ و ۲ شبیه سوال ۴ از تمرین سری ۲ است.

۱. $VC\text{-}dimension = 4$

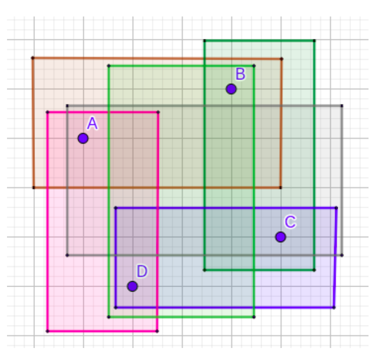
۴ نقطه را مانند تصویر زیر در نظر بگیرید:



حالا مشخص است که اگر ۱ نقطه مثبت باشد و ۳ نقطه منفی، به راحتی می‌توان دور نقطه‌ی مثبت را با یک مستطیل جدا کرد. همینطور اگر ۳ نقطه مثبت باشند و یک نقطه منفی. به شکل زیر توجه کنید:



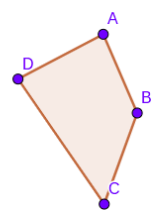
برای دو نقطه مثبت و دو نقطه منفی نیز می‌توان هر جایگشتی را مشابه شکل زیر پوشش داد:



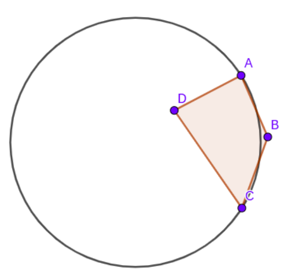
بنابراین حالتی از قرارگیری ۴ نقطه در فضا وجود داشت که *shatter* می‌شد. حالا اثبات می‌کنیم هیچ حالتی از قرار گیری ۵ نقطه در فضا وجود ندارد که بتوان آن ها را *shatter* کرد. برای ۵ نقطه کوچکترین مستطیلی که اضلاع آن موازی محور های مختصات باشد و همه‌ی ۵ نقطه را شامل شود در نظر بگیرید. قاعدتا بر روی هر ضلع این مستطیل حداقل یک نقطه قرار دارد. زیرا در غیر اینصورت می‌توان آن ضلع را تا جایی که با نقطه ای برخورد نداشته باشد حرکت داد تا مستطیل کوچکتری ایجاد شود. حالا اگر ۴ نقطه روی اضلاع باشند و یک نقطه داخل مستطیل، می‌توان آن ۴ نقطه را مثبت و نقطه‌ی داخل را منفی فرض کرد. بنابراین هر مستطیلی که شامل آن ۴ نقطه باشد باید نقطه‌ی داخل را هم دربرگیرد، پس نمی‌توان این ترکیب را *shatter* کرد. اگر هیچ نقطه ای داخل مستطیل نباشد هم حتما ضلعی وجود دارد که روی آن دو نقطه قرار دارد (بنابر اصل لانه‌ی کبوتری) حالا ۳ نقطه‌ی دیگر را به همراه یکی از این دو نقطه مثبت و نقطه‌ی دیگر را منفی در نظر بگیرید. مجددا هر مستطیلی که ۴ نقطه‌ی مثبت را شامل شود باید نقطه‌ی منفی را هم شامل شود. بنابراین هیچ ۵ نقطه ای را نمی‌توان *shatter* کرد. پس بعد VC برای این دسته بند ۴ است.

۲. $VC\text{-}dimension = 3$

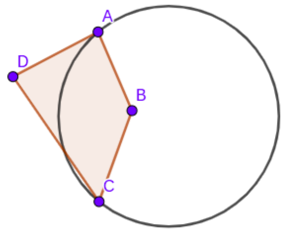
برای ۳ نقطه به طور مثال می‌توان آن ها را رئوس یک مثلث متساوی الاضلاع در نظر گرفت و به راحتی می‌توان هر ترکیبی از برچسب گذاری برای این نقاط را پوشش داد. حالا اثبات می‌کنیم که هیچ ۴ نقطه ای را نمی‌توان *shatter* کرد. دو حالت در نظر بگیرید؛ ۱) حالتی که این ۴ نقطه تشکیل یک چهارضلعی مقعر بدهند و ۲) حالتی که این چهار نقطه تشکیل یک چهارضلعی محدب بدهند. در حالت اول حتما نقطه ای وجود دارد که داخل پوش محدب ۳ نقطه‌ی دیگر است. اگر این ۳ نقطه را + در نظر بگیریم و نقطه‌ی داخل را -، آنگاه هیچ دایره ای وجود ندارد که تنها ۳ نقطه‌ی مثبت را شامل شود. زیرا هر دایره ای که آن ۳ نقطه را شامل شود، پوش محدب آن ها را هم شامل خواهد شد. در حالت دوم چهار ضلعی ایجاد شده با این ۴ نقطه را به طور مثال مانند شکل زیر در نظر بگیرید (این یک شکل عمومی است و هیچ فرض خاصی در آن نیست):



در این چهارضلعی حتما دو زاویه‌ی روبرو به هم وجود دارد که جمع آن‌ها کمتر یا مساوی ۱۸۰ درجه باشد. (در اینجا A و C) دو راس مربوط به این دو زاویه را + و دو راس دیگر را منفی در نظر بگیرید. حالا اثبات می‌کنیم هیچ دایره ای وجود ندارد که این دو راس را شامل شود و هیچ کدام از دو راس دیگر را شامل نشود. اگر دایره بخواهد راس B را شامل نشود در این صورت کمانی که روبروی وتر AC در دایره‌ی رسم شده قرار می‌گیرد حتما زاویه‌ای کمتر از $180 - B$ خواهد داشت. بنابراین تمام نقاط خارج از این دایره که رو به روی وتر AC قرار بگیرند، باید زاویه‌ای کمتر از $180 - B$ داشته باشند و تمام نقاطی که داخل دایره قرار بگیرند زاویه ی بیشتری خواهند داشت. ما فرض کرده بودیم که $A + C$ کمتر از ۱۸۰ می باشد، بنابراین $B + D$ بیشتر از ۱۸۰ خواهد بود و می توان نتیجه گرفت که D از $180 - B$ بزرگتر است. بنابر این D باید حتما داخل این دایره باشد.(مانند شکل زیر)



به طور عکس اگر قصد داشته باشیم دایره راس D را شامل نشود، باید حتما راس B را شامل شود. (مانند شکل زیر)



پس در نهایت VC برابر با ۳ خواهد بود.

۳. *VC-dimension = d*

برای d نقطه در فضا آن ها را برابر با بردار یکه‌های فضا در نظر بگیرید. برای یک برچسب گذاری دلخواه از این نقاط، بردار W را به این شکل در نظر میگیریم که هر درایه‌ی آن ۱ است اگر برچسب بردار یکه‌ی متناظر با آن + باشد و ۱- است اگر برچسب بردار یکه‌ی متناظر با آن - باشد. در اینصورت بدیهی است که حاصلضرب داخلی W با هر یک از نقاط برابر با ۱ می‌شود اگر آن نقطه برچسب + داشته باشد و برابر با ۱- می‌شود اگر برچسب منفی داشته باشد. این بردار دقیقا همان بردار نرمال ابرصفحه‌ی مدنظر است. حالا برای $d+1$ نقطه، از آنجایی که تعداد آن ها از بعد فضا بیشتر است، حتما نقطه ای وجود دارد که به صورت جمع وزن دار باقی نقاط قابل نوشتن است. این نقطه را P_j در نظر بگیرید. بنابراین داریم:

$$P_j = \sum a_i P_i$$

که در این رابطه a_i ضریب مربوط به هر نقطه‌ی P_i است. حالا اگر تمام a_i ها برابر با صفر باشند، پس نقطه‌ی P_j روی مبدا است و چون ابر صفحه از مبدا می‌گذرد پس به هیچ وجه نمی‌تواند این نقطه را دسته بندی کند. پس فرض می‌گیریم تمام a_i ها صفر نباشند. برچسب گذاری نقاط را اینگونه در نظر بگیرید که اگر برای نقطه‌ای a_i مثبت بود، برچسب آن نقطه نیز مثبت و اگر برای نقطه ای a_i منفی بود، برچسب آن نیز منفی باشد. برچسب نقطه‌ی P_j را نیز منفی در نظر بگیرید. فرض کنید بردار W وجود داشته باشد به نحوی که دسته بندی را به درستی انجام دهد. بنابراین داریم:

$$W^\top P_i = \sum a_i W^T P_i$$

با فرض آنکه دسته بندی به درستی انجام می‌شود، حاصل عبارت سمت چپ باید منفی باشد در صورتی که عبارت سمت راست مجموع چند جمله‌ی مثبت است و در نهایت مثبت خواهد شد. (زیرا هر جمله حاصل ضرب یک a_i در $W^T P_i$ است و برچسب گذاری ها به گونه‌ای انتخاب شدند که این دو هم علامت باشند.) پس در نهایت سمت چپ منفی شد در حالی که سمت راست مثبت است. این تناقض نشان می‌دهد هیچ Wای وجود ندارد که دسته بندی را به درستی انجام دهد. بنابراین بعد VC برابر با d است.

(ب) باند زیر را در نظر بگیرید. و می دانیم $VC(H) = 4, 3$، $\delta = 0.1$ و $\epsilon = 0.05$

$$m \geq \frac{1}{\epsilon}(4\log_2(\frac{2}{\delta}) + 8VC(H)\log_2(\frac{13}{\epsilon}))$$

در نتیجه قرار دادن این مقادیر در باند بالا برای قسمت الف-۱ $m \geq 5480$ و برای قسمت الف-۲ $m \geq 4197$.

(پ) یک توزیع $P(X)$ و تابع هدف f را انتخاب کنید. حال به طور تکرار شونده مجموعه نمونه های آموزش با اندازه ۴۱۹۷ بر اساس $P(X)$ و f نمونه برمیداریم. برای هر مجموعه یک فرضیه *h (hypothesis)* را در H_c پیدا کنید که به طور کاملا درست مجموعه آموزش را دسته بندی می کند و سپس *true accuracy* فرضیه h را محاسبه کنید. در حداقل ۰.۹ از این آزمایشات *true accuracy* حداقل برابر ۰.۹۵ است.

پاینده باشید