دانشکده مهندسی کامپیوتر

دانشگاه صنعتی شریف

استاد درس: دکتر محمدحسین رهبان

بهار ۱۴۰۰

تمرین ششم

درس یادگیری ماشین

نام و نام خانوادگی: امیر پورمند

شماره دانشجویی: ۹۹۲۱۰۲۵۹

آدرس ایمیل **pourmand1376@gmail.com**

۱ سوال ۱

۱.۱ الف

سه اصل یادگیری عبارتند از:

۱. تیغ آقای اوکم یا Occam's razor

دروافع دانشمندی به نام اوکم در قرن ۱۴ بیان کرده که یک مدل هر چه ساده تر باشد بهتر است. البته سادگی مدل را با مفهوم سادگی فضای فرضیه و سادگی تک تک فرضیه ها پیوند میزنند که این دو مفهوم در بسیاری از موارد با یکدیگر ارتباط یک به یک دارند. این مفهوم در ساده ترین حالت ان بیان میکند که ساده ترین مدل در اکثر موارد منطقی ترین آنها است البته مدل ساده را نباید ساده تر از چیزی کرد که لازم است و سادگی بیش از حد نیز نمیتواند داده را به خوبی توضیح دهد.

۲. بایاس نمونه گیری یا sampling bias

دومین اصل مهم یادگیری بیان میکند که در نمونه برداری از جمعیت باید همواره دقت کنیم که اگر از جمعیت با ویژگی خاصی نمونه برداری میکنیم پیش بینی ما نیز روی همان قسمتی از جمعیت که ان ویژگی خاص را دارند درست خواهد بود. به طور خاص وقتی که توزیع داده تست با توزیع داده آموزش یکی نباشد میگوییم بایاس نمونه برداری رخ داده است. مثال آن نیز مشخص است که استاد در سرکلاس توضیح دادند که چگونه با انتخاب اشتباه تلفن به عنوان ابزاری ارتباطی در سال ۱۴۰۰ دچار خطای نمونه برداری شده بودند.

۳. مشاهده دزدکی داده یا data snooping

این مفهوم اساسا به این معناست که مشاهده داده های اصلی قبل از انتخاب مدل و بعضا انتخاب مدل بر اساس مشاهدات میتواند گمراه کننده و در بسیاری از مواقع عامل اصلی overfitting باشد علت این امر نیز ان است که بعد vc این انتخاب شهودی در نظر گرفته نمیشود و به نوعی یک جستجو داخل ذهن دیتاساینتیست انجام میشود. یک دلیل دیگر نیز آن است که در طی فرایند ممکن است از دیتاست تست استفاده شود که هرگز اینکار نباید انجام میشد. به طور خاص اقای ابومصطفی اشاره میکند که اگر کل دیتاست در یادگیری تاثیر بگذارد کل فرایند یادگیری زیرسوال میرود که اهمیت استفاده نکردن از داده تست را بیان میکند.

۲.۱ ب

خیر. این حرف لزوما درست نیست. مثال آن هم الگوریتم SVM با کرنل RBF است که در درس نیز این مثال زده شد. با این که مرز تصمیم گیری پیچیده بود ولی باز میتوان مدل را به نوعی ساده در نظر گرفت زیرا تعداد پارامترهای آزاد مدل به تعداد بردارهای پشتیبان است.

۳.۱ پ

sampling bias در واقع نحوه نمونه برداری و جمع آوری دیتا است در حالی که data snooping میگوید که چگونه داده بر روی فرایند یادگیری تاثیر گذاشته است در حالی که ویژگی های داده باید بر فرایند یادگیری تاثیر بگذارند نه خود داده و نه مشخصا کل داده!

در بعضی موارد بایاس داده به این خاطر ایجاد میشود که به چیزی نگاه کردیم که نباید نگاه میکردیم!

مثلا فرض کنیم میخواهیم خرید و نگه داشتن سهام شرکت را پیش بینی کنیم و شرکت هایی که الان وجود دارند را انتخاب میکنیم و روی ۵۰ سال گذشته پیش بینی میکنیم که آیا خرید و نگهداری سهم های انها تاثیر مثبت خواهد داشت یا نه. به طرز شگفت انگیزی میبینیم که سود خیلی زیادی نصیبمان میشود اما اشکال کار در کجاست؟

در اینجاست که ما انگار از قبل داده های اینده را نگاه کردیم و میدانیم که کدامین شرکت ها موفق خواهند بود و اصلا وجود خواهند داشت. همین مسئله باعث بایاس کردن ما به سمت شرکت هایی شده است که سهام بهتری دارند در حالی که این مسئله در واقعیت اصلا وجود ندارد.

۲ سوال ۲

۱.۲ آ

میدانیم رگرسیون خطی معمولی مشکلاتی دارد که در بعضی اوقات بهتر است از یک ورژن پیشرفته تر آن استفاده کنیم که مسائلی مانند ناهمگونی دیتا را را نیز در نظر میگیرد. میتوان رگرسیون معمولی را با ماتریس کوواریانس زیر در نظر گرفت:

$$C = E[ee^T] = \sigma^2 I$$

ابتدا مدل را تعریف کنیم: (البته n در اینجا ابعاد ویژگی هاست)

$$h(x) = \sum_{i=0}^{d} \theta_i x_i = \theta^T x$$

سپس برای تابع هزینه داریم: (در اینجا n تعداد داده های آموزش است)

$$\begin{aligned}
J(\theta) &= \frac{1}{2}\sum_{i=0}^{n} w^i (h(x^i) - y^i) \\
&= \frac{1}{2}(X_{n\times d}\theta_{d\times 1} - Y_{n\times 1})^T W_{n\times n}(X_{n\times d}\theta_{d\times 1} - Y_{n\times 1}) \\
&= \frac{1}{2}(X\theta - Y)^T W(X\theta - Y) \\
&= \frac{1}{2}(\theta^T X^T - Y^T)(WX\theta - WY) \\
&= \frac{1}{2}(\theta^T X^T WX\theta + Y^T WY - \theta^T X^T WY - Y^T WX\theta)
\end{aligned}$$

حال برای مینیمم کردن تابع هزینه داریم:

$$\begin{aligned}
\min J(\theta) &= \frac{\partial}{\partial\theta} J(\theta) \\
&= \frac{1}{2}(2X^T WX\theta + 0 - X^T WY - X^T WY) \\
&= (X^T WX\theta - X^T WY)
\end{aligned}$$

پس داریم:

$$\begin{aligned}
X^T WX\theta &= X^T WY \\
\theta &= (X^T WX)^{-1} X^T WY
\end{aligned}$$

۲.۲ ب

همان طور که در بالا نیز گفتم وقتی ناهمگونی یا heteroscedasticity در واریانس های ویژگی های دیتاست وجود داشته باشد و به بیان دیگر واریانس ویژگی ها یکسان نباشد بهتر است از این نسخه تعمیم یافته linear regression استفاده کنیم. البته رگرسیون خطی معمولی به هر حال بایاس ندارد اما معمولا جواب خوبی وقتی که واریانس متفاوت باشد نمیدهد.

این روش یک روش غیرپارامتریک است که به هر داده وزن خاصی را نسبت میدهد که این نیز در الگوریتم باید یاد گرفته شود.

اگر وزن ها گوسی باشند یعنی یک سری داده ها اهمیت خیلی بیشتری پیدا میکنند و مقدار کمی از داده ها با اهمیت خیلی کمتری خواهند بود حال اگر این واریانس خیلی زیاد باشد وزن ها میتوانند مقادیر خیلی بیشتری را اتخاذ کنند و اگر این واریانس کمتر باشند مقادیر وزن ها رنج کمتری از اعداد را شامل میشود و وزن ها به یکدیگر نزدیک تر خواهند بود.

و البته لازم نیست که واریانس این توزیع برای مجموعه دادگان مختلف یکی باشد چنین فرضی را ما جایی impose نکرده ایم و به هر حال متغیر بودن واریانس باعث flexible تر شدن مدل و fit شدن بهتر به داده ها خواهد شد که نتیجتا دقت بیشتری را به همراه خواهد داشت.

۳ سوال ۳

۱.۳ آ

در مسئله رگرسیون مراحل مختلف با استفاده از درخت های متعدد ساخته میشود و در هر مرحله در واقع یک درخت ساخته خواهد شد که به درخت های قبلی اضافه میشود و خطا را کم میکند. این کار در مسئله دسته بندی نیز به همین شکل انجام میشود.

در مسئله رگرسیون و دسته بندی هر بار residual حساب شده و سعی در کاهش این تابع در کل داریم و ایده کلی این است که در نهایت با توجه به سلسله ای از کم شدن ها با مقدار بهینه خواهیم رسید.

تفاوت این دو مسئله در استفاده از تابع سیگموید برای مسئله دسته بندی است که در واقع یک احتمال را برمیگرداند و سپس از یک حد آستانه برای پیش بینی کلاس مورد نظر استفاده میگردد.

۲.۳ ب

این سوال نیز دستی حل شده است و جواب آن در فایل Exercise 3 است.

۳.۳ پ

این سوال دستی حل شده است و جواب آن در فایل Exercise 3 است.

۴.۳ ت

خالی