جلسه اول:

الگوریتم یادگیری میاد به فضای فرضیه نگاه میکند و یکی از آنها را که سازگار با داده های آموزشی هست انتخاب میکند. فرضیه همان فانکشن هدف ما هست که ورودی را به خروجی نگاشت میکند.

پرسپترون یک رابطه خطی مدل میکند. پرسپترون به مقادیر ورودی یک وزنی اختصاص میدهد و میگوید از این آستانه بالاتر بود فلان اگر نبود فلان. مثلا وام بدهیم یا ندهیم. آستانه عضو فضای فرضیه هست.

اگر داده های ما جدایی پذیر خطی باشند پرسپترون در تعداد متنهای میتواند این داده ها را از هم جدا کند.

در یادگیری غیر نظارتی به صورت ذاتی میتوانیم داده ها را با خطوط به خوشه ها تبدیل کنیم که تجمع داده در آنجا زیاد است.

موارد قبلی همه به صورت اپیزودیک بود و مقطعی بود یعنی این مشتری خوب هست یا نیست، اما در یادگیری تقویتی ما یک سری گام داریم و خروجی ما ترتیبی و بر اساس یک سلسله تصمیمات هستش. اینجا مسئله تصمیم گیری دنباله دار هست و تصمیمات فعلی در آینده هم تاثیر گذار است. و یک میزان مطلوبیت برای خروجی خودمان داریم و هدف یک پالیسی هستش که به ازای هر ورودی چه اکشنی را انجام بدهیم تا به مطلوبیت برسیم.

گاهی اوقات ممکن است یادگیری امکان پذیر نباشد و برای وقتی هست که ما اطلاعات کمی از تابع ناشناخته داریم و در این حالت یادگیری غیر ممکن است و برای این مواقع ما فضای فرضیه را سعی میکنیم محدود کنیم به آن مواردی که میشناسیم تا یادگیری امکان پذیر شود.

پرسپترون اون بردار وزن را میچرخاند. آستانه یک هایپر پارامتر هست. دقت کن آستانه هم عین وزن ها یاد میگیریم که همان bias ما هست که در 1 ضرب میشود و میگیم از اون مقدار کمتر یا بیشتر باشد جدا میکنیم.

چرا آمار و احتمال؟ چون از یک سری داده میخواهیم یک قاعده کلی استخراج بکنیم. و تعمیم بدهیم به کل دیتا داشتیم . اون مواردی که از روش یاد گرفتیم داده آموزشی ما هستش.

جلسه دوم:

یادگیری زمانی استفاده میشود که 1. الگو وجود داشته باشد 2. نتوانیم به صورت ریاضی بنویسیم 3. داده برای اون الگو داریم. یادگیری نظارتی میگه قرار است ما ورودی را به خروجی نگاشت کنیم که این تابع f ما هست پس ما باید از فضای فرضیه یک تابعی به نام g را انتخاب کنیم که به اندازه کافی به تابع f نزدیک باشد. تابع ناشناخته همان f ما هست. learning میگوید حتما باید فضای فرضیه را کامل بگردی نه اینکه خود فرضیه را ثابت فرض کنی و به شما بدهند.

تابع g را learning پیدا کرده است. دقت کن که بر اساس اصولی که بهش رسیدیم اگر m خیلی بزرگ باشد یعنی اون فضای فرضیه ما اون عدد 1 میشود و یادگیری امکان پذیر نمیشود چون اون مقدار باید به صفر میل بکند تا ما خطا نداشته باشیم و حدسی که میزنیم با داخل کوزه یکسان باشد پس اگر تعداد n که داده های ما هست کم باشد به همان اندازه فضای m هم باید محدود در نظر بگیریم تا یادگیری ممکن شود.

در پرسترون فضای فرضیه نامحدود است و عملا این قاعده ممکن نیست.

جلسه سوم:

هر چه قدر فضای فرضیه بزرگتر نزدیک تر کردن Ein , Eout سخت میشود. Ein میشود ورودی و برداشتی من. Eout میشود اون تخمین واقعی از کل مسئله.

چرا Ein و Eout فاصله میگیرند گاهی اوقات؟ (نمودار رنگ صورتی و سبز منظورمه) چون داده جدید میبینیم نسبت به قبلی ها در ترین و امکان misclassify شدن هست در اون خطی که پرسپترون میکشد و تا اپدیت بکند در این نتیجه داده ها به خوبی جدایی پذیر خطی نیستند.

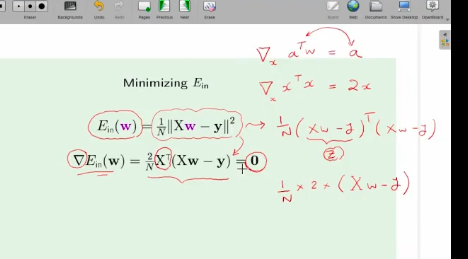
چرا پرسپترون خوب جدا نکرده؟ چون ممکن است در جای بدی متوقف شده باشد یا early stopping کرده باشد و تو جای بدی متوقف شده باشد و اگر چند تا iteration بیشتر میزد خطا کمتر بود.

تغییر در الگوریتم پرسپترون: بخاطر وجود این مشکل میایم یک تغییری ایجاد میکنیم، و میگیم کدوم یک از مدل ها کمترین Ein داشته و اون رو به عنوان جواب در نظر میگیریم. که به این الگوریتم میگوییم pocket که یعنی اون مدل بهتر را در جیب خودمان گذاشتیم و جلو میرویم و در نهایت از اون بهترین استفاده میکنیم.

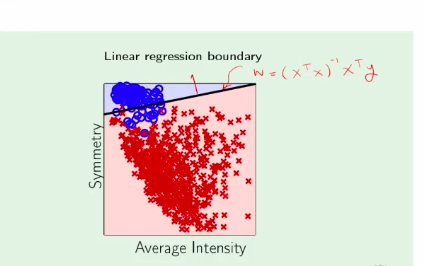
Regression: خروجی یک کلاس نیست یک مقدار حقیقی است. مثلا مقدار وامی که قرار است به آن شخص بدهیم. اون توان 2 علامت را خنثی میکند حالا چه h(x) بزرگتر از f(x) باشد یا نباشد. در این دست مسائل اگر Ein را کوچک بدست آوردی تقریبا میتوانی مطمئن شوی که Eout هم کوچک است. در دسته بند های خطی به جز وقتی که x خیلی بزرگ باشد بعد خیلی بزرگ باشد. ولی اگر کوچکتر از N باشد جای نگرانی نیست.

اگر ضرب داخلی اون بردار در وزن مثبت شد میشود جز کلاس 1 و اگر نشد میشود کلاس -1. X مثلا چون زاویه آن کمتر از 90 درجه هست میشود مال -1. اگر بیشتر از 90 درجه باشد 1. بردار وزن متعامد هست به زیر فضای خطی اون زیر فضای خطی ابر صفحه هست و مقدار ضرب داخلی آنها صفر میشود. تنها زیر فضای خطی که بعدش یکی کمتر از فضا هست میشود ابر صفحه.

نرم 2 به توان 2 تک به تک درایه ها را به توان 2 میرساند و با هم جمع میزد. که میشود همان 2 تا خط اطراف همون وزن منهای خروجی واقعی. برای کم کردن هم مشتق میگیریم.



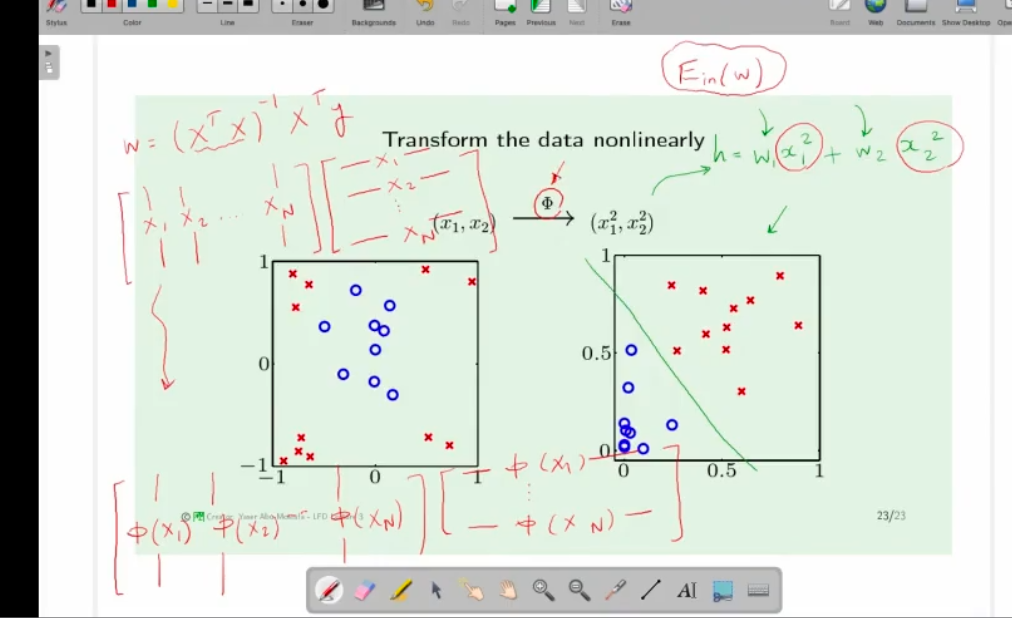
رنک ماتریس میشود تعداد ستون هایی که از همدیگر مستقل خطی هستند. برای وقتی هست که تعداد داده ها خیلی بزرگتر از بعد هست و فیچر ها باید مستقل از هم باشند. فیچر ها نباید به همدیگه وابستگی خطی داشته باشند اگر این را تعریف کنیم میتوانیم بگوییم رنک ماتریس حداکثر d+1 شود که به معنی full rank بودن هست پس معکوس پذیر هست.



از regression برای کلاس بندی استفاده کردیم اما اشکال کجاست؟ چون در پایین قرمز ها مثلا خروجی تابع وزن در ورودی میشود -3 ولی این تابع میخواست -1 اتلاق کند به این نقاط و MSE زیاد میشود این باعث میشود دچار خطا شود و اذیت شود برای همین به یک جواب مطلوب نمیرسیم و معمولا از این برای وزن اولیه یا جواب اولیه برای الگوریتم پرسپترون استفاده میکنند.

مسائل در خطی با SINGLE SHOT حل میشوند یعنی کلا ورودی ضربدر وزن میکنی و دیگر iteration ندارد. مثلا برای مقدار ماندن در خانه طرف بعد 5 سال فرار نمیکند ولی این مقدارش با 20 سال در حالت خطی فرقی نمیکند یعنی اهمیتی ندارد 5 سال تا 20 سال ولی هر چه قدر بیشتر باشد این یعنی ماندگاری بالاتر است و برای ما مهم است پس میایم از مسائل غیر خطی استفاده میکنیم و مثلا میگیم طرف آیا کمتر از 1 سال و بیشتر از 5 سال ساکن بوده است؟

وقتی یک regression خطی هست نسبت به چی هست؟ نسبت به وزن ها یا ورودی؟ نسبت به وزن ها خطی هست دیگه چون داریم نسبت به وزن مشتق میگیریم. برای کلاس بندی هم همین است. حالا ورودی را میخواهیم ثابت فرض میکنیم و یک transformation غیر خطی بزنیم روی داده ها (اینم مشکلی نداره چون خطی بودن نسبت به وزن ها هست نه نسبت به ورودی ها مثلا فاصله داده از مراکز) و به فضای جدیدی برویم که و درفضای جدیدی جدای پذیر خطی هستند.



جلسه چهارم:

در regression خطی با یک گام میتوانیم مسائل را حل بکنیم که میشود وزن برابر با pseudo inverse ضربدر y که همان ورودی ما هستش. خطی بودن نسبت به وزن ها هست و نه نسبت به ورودی ها و مزیتی ندارد پس برای غیر خطی کردن تابع دسته بند یا classifier این است که بر روی ورودی یک تابع غیر خطی اعمال کنیم به فضای جدید برویم و اون ورودی را وارد کنیم و میتوانیم مسائل و الگو های پیچیده تر راه حل بکنیم نسبت به حالت صرفا خطی.

تمام مشتق گیری نسبت به وزن ها هستند.

پس چیشد در فرمولی که برای آلفا ضربدر بردار ستونی که میزان شباهت ورودی با تک به تک داده ها هست چیکار میکند؟ ضرب داخلی میدهد و مشخص میکند و حاصل میاد ورودی را با تک به تک آموزشی شباهت را میسنجد و به هر شباهت یک وزنی اختصاص میدهد و میاد وزن ها را جمع میزند تا score اون داده را بدست بیاورد و مشخص کند به کدام کلاس تعلق دارد اگر مثبت بود یک کلاس اگر منفی بود کلاس دیگر.

البته میزان تاثیری داریم یعنی اگر شباهت بیشتر بود ضریب بیشتری باشد و تاثیر اون برچسب زیاد تر باشد تا جمع مثبت شود و برود به کلاس مثبت ها و اگر منفی بود ضریب کمتر و میرود یک کلاس دیگر و کلاس بند انجام میشود. الفا برچسب داده i ام خواهد بود. این ضرب از regression شروع شد بعد الفا بدست اومد.

ضرب ها رو دقت کن اگر بردار سمت چپ ماتریس بود روی سطر کار میکند اگر سمت راست ماتریس بود دارد روی ستون کار میکند و تاثیر دارد. این y که در فرمول آلفا هست میخواهد بگوید که برای اینکه برچسب نقطه j را پیدا بکنی بیا تمام نقاط آموزشی را در نظر بگیر و ببین شباهت اون نقطه به نقطه j که همان ضرب داخلی هست چه قدر هست مقدار را ضربدر آلفا بکن پس اگر داده قرار هست مثبت باشد برای کلاس مثبت باشد آلفا آن باید مثبت باشد یعنی به داده هایی شبیه هست که آلفا های آنها مثبت هست و اگر نقاط مشابه را الفا آنها را جمع بزنی به برچسب اون نقطه برسی. یعنی اگه برچسب به مثبت ها نزدیک بود یعنی مثبت هست دیگه.

حالا اگر y منفی بود بیا داده هایی که منفی هستند و آلفا آنها منفی هست را جمع بزن و k های آن باید بزرگ باشند دیگه.

وقتی داده ورودی را غیر خطی میکنی به یک نمایش جدید از ورودی میرسی. ماتریس k درایه های آن ضرب داخلی داده ها هستند. و ضرب داخلی نمایش جدید میشود kernel که ماتریس کرنل هم گفته میشود. از فی برای غیر خطی کردن بدست میاد و حاصل فقط به ضرب داخلی وابسته میشود نه خود فی و ضرب داخلی آنها مهم است واسه همین میشود کرنل که یک بازنمایی جدید از همین regression ها خواهد بود.

دقت کن در یادگیری ماشین دنبال یک تابع فرضی هستیم که دقیقا اون تخمین مپ کردن ورودی به خروجی را انجام دهد حالا خطا چیست؟

Error: دنبال خطای فرضیه ما هستیم. و اکثر اوقات از یک تعریف point wise استفاده میکنیم و میایم نگاه را به جای اینکه به کل داده ها نگاه بکنیم به یک داده نگاه میکنیم و بر اساس اون میبینیم چه قدر فرضیه ما خطا دارد مثلا اگر مسئله regression باشد از square error استفاده میکنیم اگر classification باشد اگر عبارت داخل براکت صحیح بود 1 میشود یعنی برابر بود و گرنه صفر و دسته بندی انجام میشود.

میانگین خطا هم میایم متوسط گیری میکنیم و خطا ما روی داده های آموزشی ما هست. حالا اگر روی جمعیت کل داده ها به جای آموزشی استفاده بکنیم e out را بیرون خواهد داد. Pointwise کجا باید باشد؟ در قسمت ارزیابی فرضیه نهایی ما خواهد بود.

معیار error: معیار خطا بر اساس مسائل مختلف، مختلف باید باشد.

Friendly measure: اگر جزئیات ریز ندانیم و میایم صرفا برای راه حل ساده تر استفاده میکردیم مثلا از square loss برای کلاس بندی استفاده میکردیم و اگر استفاده میکردیم در یک گام میتوانست w0 خوب به ما بدهد. مثلا در مسئله دسته بندی خطی اگر از entropy cross استفاده بکنیم بهینه سازی را ساده تر میکند و بهینه سازی محدب میشود و این بهینه سازی به شکل عددی پیدا کردن اون واقعا کار راحتی هست یا در حالت بسته برای regression مناسب است.

پس برای دسته بندی از cross entropy برای تابع loss استفاده میکنیم.

Error measure هم برای ارزیابی سیستم مهم است هم در قسمت یادگیری تا به یک فرضیه خاص برسیم در یادگیری و بعد اون فرضیه را ارزیابی کنیم.

Noise: گاهی اوقات ممکن است که target function تابع نباشد، یا الگو تابع نباشد. مثلا پیشبینی بر اساس آب و هوا مثلا برای ساعت 10 شب ممکن است آب و هوا یا ابری یا بارونی باشد که این در تضاد با مفهوم تابع است که میگفت به ازای یک ورودی مشخص یک مقدار مشخص برگردانده میشود. پس میایم به صورت احتمالی نگاه میکنیم به تارگت فانکشن که تابع نیست و میایم توزیع نگاه میکنی یعنی یک مشتری 90 درصد خوب هست و 10 درصد بد هست بعد میتوانیم تصمیم گیری بکنیم. به این توزیع تارگت distribution گفته میشود. در نهایت noise میشود اون انحرافی که از مقدار متوسط داریم. هر چه قدر عدد بالا تر نویز بیشتر هست چه منفی چه مثبت. اگر تابع بود ارتباط خروجی با x میتوانیم بگیم در همه مقادیر صفر است به جز جایی که برابر با y است. الان هدف ما این است که توزیع تارگت را یاد بگیریم که p y | x را یاد بگیریم که target distribution ما خواهد بود و تا الان نیازی نداریم در یادگیری نظارت شده بیایم و p x را یاد بگیریم که توزیع داده های ورودی ما هستند.

کلا هدف چی هست برای امکان پذیر بودن یادگیری؟ Ein خطای اون فرضیه ما به ازای Eout آن یکسان باشد به ازای همان فرضیه به این generalization bound گفته میشود تا فرضیه مطمئن باشیم روی داده های جدید هم درست کار بکند. در واقع به نوعی دنبال این هستیم که خطا صفر شود خطای چی ؟ Eout به ازای اون فرضیه ای که ما استفاده کردیم. اندازه فضای فرضیه ما مهم هست و اگر پیچیده باشد امکان پذیر نخواهد بود یادگیری. اون کرنل برای این بود که خطا صفر شود و اون generalization برای این بود که خطای فرضیه نمونه ما به ازای کل یکی شود. بعضی ها هم از هر دو روش به صورت توام استفاده میکنند. در پرسپترون بی نهایت وزن ممکن داریم و تعداد هایپر پلین ها به صورت بی نهایت هست در صورتی که قرار بود فضای فرضیه ما محدود باشد. VC dimension: پیچیدگی فضای فرضیه را مدل میکند برای وقتی که فضای فرضیه به سمت بینهایت میرود و حالت صعودی دارد و در ادامه Ein کم و Eout به سمت زیاد شدن حرکت میکند. Generalization gap مقدار بین Ein و نمودار Eout خواهد بود. هم Ein باید کم شود هم فاصله تعمیم پذیری تا Eout مناسبی داشته باشیم. اگر پیچیدگی فرضیه کم باشد اصلا نمیتواند مدل کند پس Ein زیاد هست هر چه قدر پیچیده تر میشود این مقدار کمتر میشود و بهتر مدل میکند از یه جایی به بعد اینقدر با سرعت پیچیده میشود که خطای out بیشتر میشود. گفتیم پیچیدگی بالا برود خطا کم میشود دیگه حالا فرض کن فی ما اینقدر بتواند بعد بالا ببرد و پیچیدگی ایجاد بکند تا بینهایت که خطای ما صفر شود ولی از طرف دیگر ممکن است overfit شود و generalization bound ما خراب شود. kij یعنی شباهت بین iوj چون ضرب داخلی آنها هست.

**جلسه پنجم:**

Point wise میاد و برای مقدار Ein در یک نقطه مشخص یعنی ورودی x خطا را میسنجد که میزان واقعی اختلاف خروجی واقعی نسبت به فرضیه ما هست. و اگر این میانگین را روی کل داده های جمعیت خودمان بگیریم امید ریاضی نسبت به x را Eout یا خطای کلی نسبت به تمام جمعیت بدست خواهد آمد.

Training and testing:

دقت کن حالا هدف این هست که برای تعداد بینهایت عضو بررسی میکند اون boundیی که قرار هست بدست بیاوریم.

مشکل چی بود؟ مشکل این بود در اون linear bound مقدار همپوشانی بسیار زیاد هست و حالتی نخواهد بود که این مجموعه های bad event جدا از هم دیگر باشند به همین دلیل اگر m ما به سمت بینهایت برود که تعداد فرضیه های ما هست خروجی بسیار بد خواهد بود و bound ما دارد با تعداد فرضیه ها تغییر میکند و linear bound راهکار خیلی شدیدی را در پیش میگیرد.

Ein اون خطای ما نسبت به داده ها خواهند بود که اشتباه مرز بندی شده اند و Eout میشود اون ناحیه ای که مرز ما اشتباه تصمیم گیری کرده و شده 2 تا مثلث چون اون ناحیه ای است که دارد خطا رخ میدهد Ein خطای sample ما هست و میخواهیم بدونیم اون داده هایی که اشتباه کلاس بندی شده اند. M سایز فضای فرضیه ما خواهد بود.

Dichotomy میاد تابع فرضیه را روی تعداد نقاط محدودی از ورودی اعمال میکند به جای اینکه مثل فرضیه روی تمام ورودی اعمال شود.

تابع رشد: بیا حداکثر dichotomy های رو n تا نقطه پیدا بکن یعنی این n تا نقطه را جوری انتخاب بکن که بیشترین pattern را از h استخراج بکنی یعنی ببین این n تا نقطه کدوم بیشترین dichotomy رو ایجاد میکند. یک انتخاب بد انتخاب خطی بودن هست چون اصلا باعث میشود همه مقدار یکسانی بگیرند و اشتباه هست و حداکثر الگو بدست نمیاد.

مجموعه محدب: اگر هر 2 تا نقطه را داخل آن در نظر بگیریم و خط بکشیم تمام اون خط در داخل اون هست. برای اینکه به مقدار رشد خیلی بالا مثل 2 به توان n برسیم لازم نیست که حتما فضای فرضیه کل توابع ممکن باشد فقط کافی است فضای فرضیه به اندازه کافی پیچیده باشد و شکل های پیچیده تولید کند به این نوع حالت با یک مجموعه محدب میگوییم داریم به کرانه بالا میرسیم به ازای n تا نقطه. غیر محدب ها هم سوپر ست محدب ها هستند پس چون بیشتر از 2 به توان n محدب ها نمیشوند برای غیر محدب هم در نهایت به این مقدار میرسیم. شمارش تمام pattern ها از یک محدب بسیار سخت خواهد بود بنابراین سعی میکنند به جای اینکه تمام فضای فرضیه را محاسبه بکنیم میایم یک ویژگی از فضای فرضیه انتزاع میکنیم تا کرانه بالا را بدست بیاوریم نه مقدار دقیق را پس میایم فقط ثابت میکنیم m چند جمله ای هست همینکه این را ثابت کنیم یادگیری ممکن خواهد بود. Shattered نشود یعنی چی؟ یعنی به ازای اون نقاطی که داریم بر اساس فرضیه که مد نظر داریم مثلا پرسپترون نتوانیم همه الگو ها را توسط فرضیه تولید بکنیم که بهش break point هم میگوییم اون نقطه ای که دیگر بیشتر از آن نمیتوانیم همه الگو ها را با این فرضیه تولید کنیم. دونستن break point کافی است چون میدانیم یادگیری ممکن است مقدارش ضروری نیست. در حالت مسائل محدب تضمین نمیتوانیم بکنیم یادگیری را چون به سمت بینهایت میرود و breakpoint لزوما شاید وجود نداشته باشد. اگر بدانیم breakpoint باشد اثبات میکند تابع رشد به صورت چند جمله ای خواهد بود. دقت کن ما کران بالا را نمیتوانیم رد بکنیم اگر رد کنیم یعنی کران بالا درست نبوده است.