

## فصل دهم و یازدهم

شیوه به کارگیری و نکات کاربردی در طراحی الگوریتم ها و سیستم های یادگیری ماشین

# انتخاب الگوریتم یادگیری ماشین

- اگر در بحث پیش بینی یا دسته بندی خطا از حد انتظار بیشتر شد چه باید کرد؟
- انتخاب کورکورانه یا انتخاب آگاهانه
- جمع آوری داده های بیشتر (همیشه مفید نیست)
- کاهش یا افزایش ویژگی ها (زمان بر است باید مطلوبیت این کار مشخص گردد)
- اضافه کردن ویژگی های ترکیبی
- تنظیم ضرایب آلفا و لامبدا
- برخی از این سعی و خطاها هزینه بر و ممکن است ماه ها طول بکشد!!
- راه حل چیست؟؟

معاینه یادگیری ماشین

## معاینه یادگیری ماشین

■ معاینه یادگیری ماشین آزمونی است که از طریق آن می توان فهمید که چه اقداماتی عملکرد الگوریتم یادگیری را بهبود می بخشد و یا در افزایش عملکرد آن بی تاثیر است.

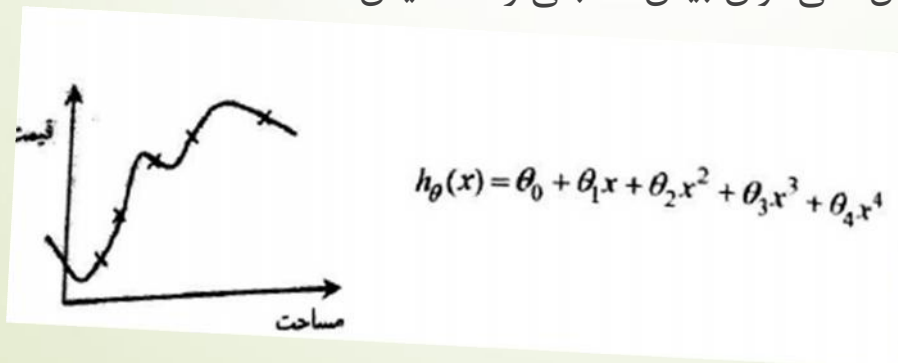
■ زمانگیر است ولی از روش سعی و خطا بهتر است.

■ ارزیابی فرضیه ها

■ بیش تطابقی و کم تطابقی

■ به دنبال کاهش خطا هستیم اما گاهی این کاهش به دلیل بیش تطابقی است.

■ اگر ویژگی ها زیاد باشد از طریق ترسیم شکل نمی توان بیش تطابقی را تشخیص داد.



## تقسیم داده ها به داده های آموزشی و آزمون

اندازه	قیمت	نوع داده	بیان پارامتریک نمونه ها
۲۰۱۴	۴۰۰	مجموعه آموزشی	$(x^{(1)}, y^{(1)})$
۱۶۰۰	۳۳۰		$(x^{(2)}, y^{(2)})$
۲۴۰۰	۳۶۹		$(x^{(3)}, y^{(3)})$
۱۴۱۶	۲۳۲		.
۳۰۰۰	۶۴۰		.
۱۹۸۵	۳۰۰		.
۱۵۳۴	۳۱۵		$(x^{(7)}, y^{(7)})$
۱۴۲۷	۱۹۹	مجموعه آزمون	$(x_{test}^{(1)}, y_{test}^{(1)})$
۱۳۸۰	۲۱۲		$(x_{test}^{(2)}, y_{test}^{(2)})$
۱۴۹۴	۲۴۳		$(x_{test}^{(3)}, y_{test}^{(3)})$

ترتیب داده ها بهم می ریخته می شود.  
 ۷۰٪ داده های آموزشی M-train  
 ۳۰٪ داده های آزمون M-test

■ ابتدا با داده های آموزشی مقدار بهینه پارمترهای به دست می آید.

■ و با داده های آزمون تابع مناسب انتخاب می شود.

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} h_{\theta}(x_{test}^i) - y_{test}^i)^2$$

$$J_{test}(\theta) = - \left[ \frac{1}{m_{test}} \sum_{i=1}^{m_{test}} y_{test}^i \log(\sigma_{\theta}(x_{test}^i)) + (1 - y_{test}^i) \log(1 - \sigma_{\theta}(x_{test}^i)) \right]$$

$$Test\ error = \frac{1}{m_{test}} \sum_{i=1}^{m_{test}} err(h_{\theta}(x_{test}^i) - y_{test}^i)$$

انتخاب تابع فرضیه مناسب از بین توابع مختلف:

تابعی که کمترین خطا آزمون را دارد.

آیا بهترین انتخاب است؟

	پارامترهای	خطای	درجه
	بهینه	تابع هزینه	چندجمله‌ای
$h_{\theta}(x) = \theta_0 + \theta_1 x$	$\rightarrow \theta^{(1)}$	$J_{\text{test}} \theta^{(1)}$	$\rightarrow d_1$
$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$	$\rightarrow \theta^{(2)}$	$J_{\text{test}} \theta^{(2)}$	$\rightarrow d_2$
$h_{\theta}(x) = \theta_0 + \theta_1 x + \dots + \theta_r x^r$	$\rightarrow \theta^{(r)}$	$J_{\text{test}} \theta^{(r)}$	$\rightarrow d_r$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$h_{\theta}(x) = \theta_0 + \theta_1 x + \dots + \theta_{10} x^{10}$	$\rightarrow \theta^{(10)}$	$J_{\text{test}} \theta^{(10)}$	$\rightarrow d_{10}$

چون انتخاب با داده های آزمون انجام شده است پس بررسی خطای مدل با این داده ها مناسب نیست.

داده های آموزشی

داده های صحت سنجی

داده های آزمونی



بیان پارامتریک نمونه‌ها	نوع داده	اندازه	قیمت
$(x^{(1)}, y^{(1)})$	مجموعه آموزشی	۲۰۱۴	۴۰۰
$(x^{(2)}, y^{(2)})$		۱۶۰۰	۳۳۰
$(x^{(3)}, y^{(3)})$		۲۴۰۰	۳۶۹
$\vdots$		۱۴۱۶	۲۳۲
$\vdots$		۳۰۰۰	۶۴۰
$(x^{(m)}, y^{(m)})$		۱۹۸۵	۳۰۰
$(x_{cv}^{(1)}, y_{cv}^{(1)})$	مجموعه روایی سنجی ۱	۱۵۳۴	۳۱۵
$\vdots$		۱۴۲۷	۱۹۹
$(x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})$			
$(x_{test}^{(r)}, y_{test}^{(r)})$	مجموعه آزمون	۱۳۸۰	۲۱۲
$\vdots$		۱۴۹۴	۲۴۳
$(x_{test}^{(m_{test})}, y_{test}^{(m_{test})})$			

ترتیب داده ها بهم می ریخته می شود.

۶۰٪ داده های آموزشی M-train

۲۰٪ داده های صحت سنجی M-cv

۲۰٪ داده های آزمون M-test


$$J_{train}(\theta) = \frac{1}{2m_{train}} \sum_{i=1}^{m_{train}} h_{\theta}(x_{train}^i) - y_{train}^i)^2$$

$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} h_{\theta}(x_{cv}^i) - y_{cv}^i)^2$$

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} h_{\theta}(x_{test}^i) - y_{test}^i)^2$$

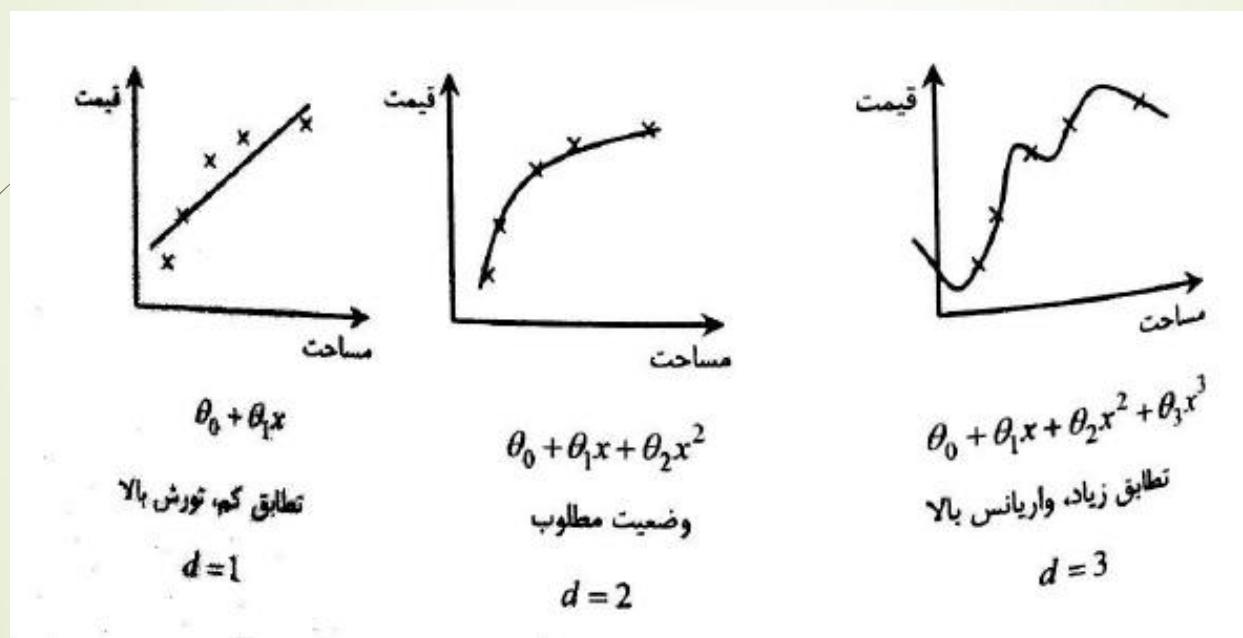


$h_{\theta}(x) = \theta_0 + \theta_1 x$	$\xrightarrow{\min_{\theta} J(\theta)}$	$\theta^{(1)}$	$\rightarrow$	$J_{cv} \theta^{(1)}$
$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$	$\xrightarrow{\quad}$	$\theta^{(2)}$	$\rightarrow$	$J_{cv} \theta^{(2)}$
$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$	$\xrightarrow{\quad}$	$\theta^{(3)}$	$\rightarrow$	$J_{cv} \theta^{(3)}$
$\vdots$		$\vdots$		$\vdots$
$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_{n-1} x^{n-1}$	$\xrightarrow{\quad}$	$\theta^{(n)}$	$\rightarrow$	$J_{cv} \theta^{(n)}$

انتخاب از طریق خطای صحت سنجی و محاسبه خطا از طریق داده های آزمون

## مقایسه تورش و واریانس

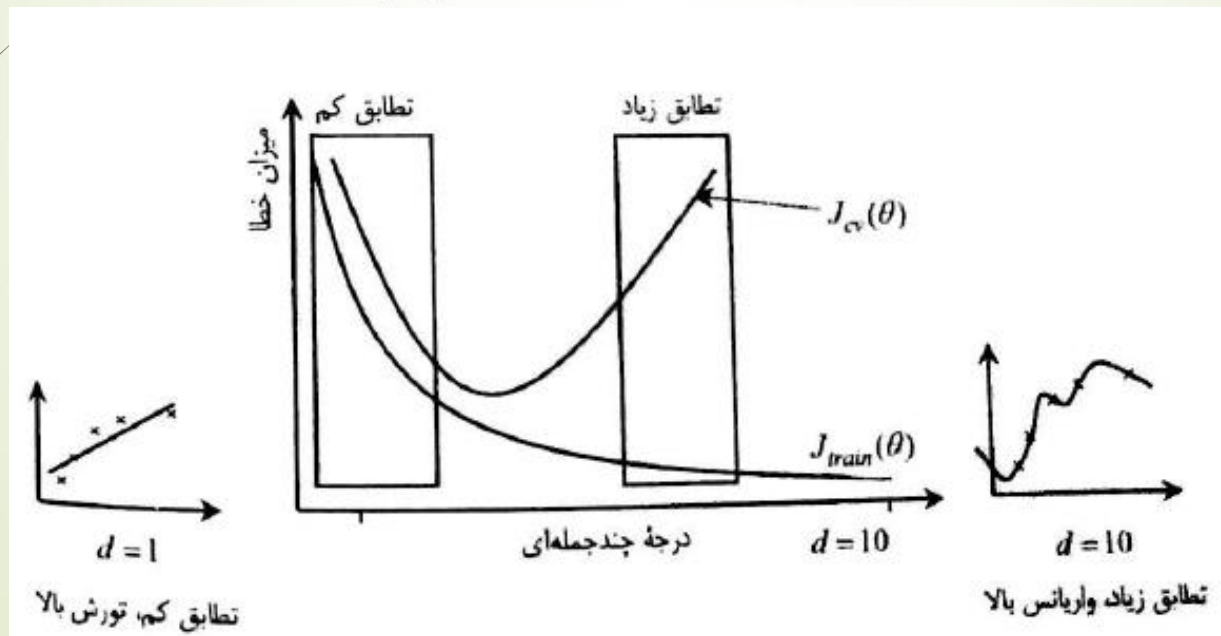
کم تطابقی باعث بایاس بالا و بیش تطابقی باعث واریانس بالا می شود.  
چگونه می توان از هر دو حالت اجتناب کرد؟



$$J_{train}(\theta) = \frac{1}{2m_{train}} \sum_{i=1}^{m_{train}} h_{\theta}(x_{train}^i) - y_{train}^i)^2$$

$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} h_{\theta}(x_{cv}^i) - y_{cv}^i)^2$$

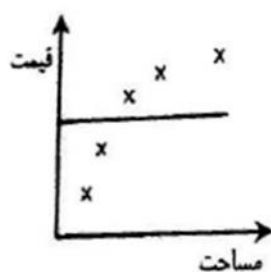
- در کم تطابقی هر دو خط بالاست.
- در بیش تطابقی خطای خطای آموزش کم ولی خطای صحت سنجی بالاست.
- درجه چند جمله ای  $d=4$



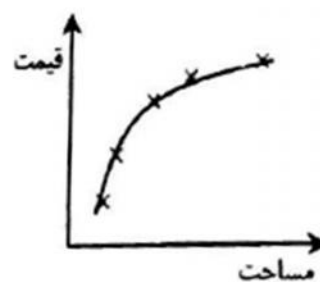
## اثر ضریب تنظیم لامبدا در بایاس و واریانس بالا

$$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m} \sum_{j=1}^m \theta_j^2$$



ضریب  $\lambda$  بزرگ  
نورس بالا (تطابق کم)  
 $\lambda = 10,000$   
 $\theta_1 \approx 0, \theta_2 \approx 0, h_{\theta}(x) \approx 0$



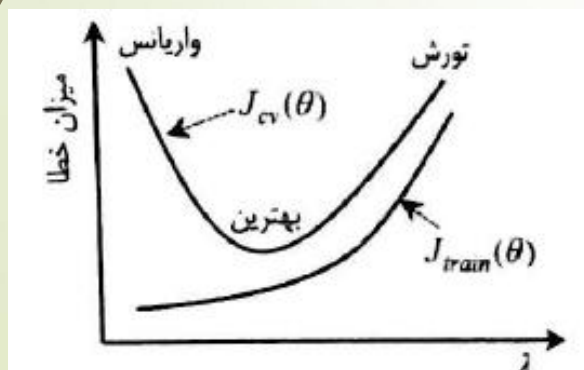
ضریب  $\lambda$  متوسط  
وضعیت درست



ضریب  $\lambda$  کوچک  
واریانس بالا (تطابق زیاد)

توسط داده های آموزشی چند جمله را انتخاب و سپس با توجه به مقادیر مختلف  $\lambda$  و داده های صحت سنجی پارامتر بهینه انتخاب می شود.

۱	$\lambda = 0$	$\min_{\theta} J(\theta)$	$\theta^{(1)}$	$\rightarrow$	$J_{cv} \theta^{(1)}$
۲	$\lambda = 0.1$		$\theta^{(2)}$	$\rightarrow$	$J_{cv} \theta^{(2)}$
۳	$\lambda = 0.2$		$\theta^{(3)}$	$\rightarrow$	$J_{cv} \theta^{(3)}$
۴	$\lambda = 0.5$		$\theta^{(4)}$	$\rightarrow$	$J_{cv} \theta^{(4)}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
۱۰	$\lambda = 1.0$		$\theta^{(10)}$	$\rightarrow$	$J_{cv} \theta^{(10)}$



اگر  $\lambda$  کوچک باشد خطای آموزش کم ولی خطای صحت سنجی زیاد شده و بیش تطابقی (واریانس بالا) داریم.

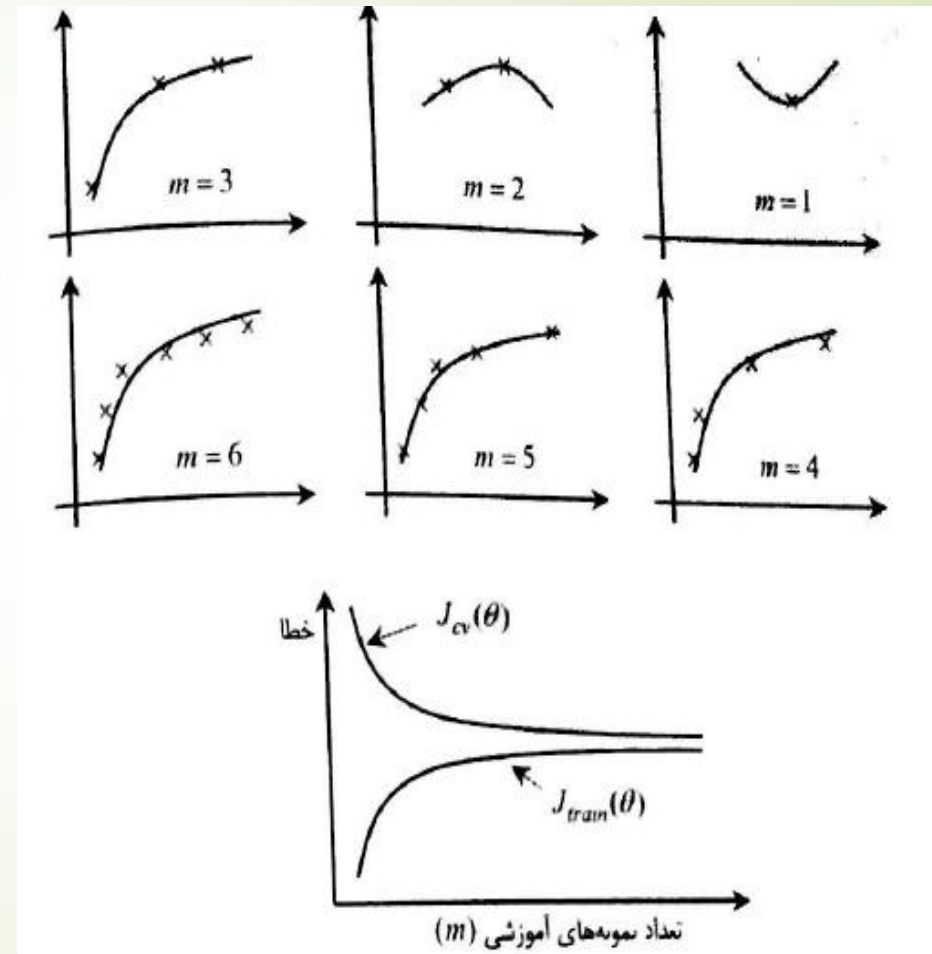
اگر  $\lambda$  بزرگ باشد خطای آموزش زیاد و صحت سنجی زیاد شده و کم تطابقی (بایاس بالا) داریم.

## تأثیر افزایش داده ها: منحنی یادگیری

$$\hat{x}(x) = \theta_0 x_0 + \theta_1 x + \theta_2 x^2$$

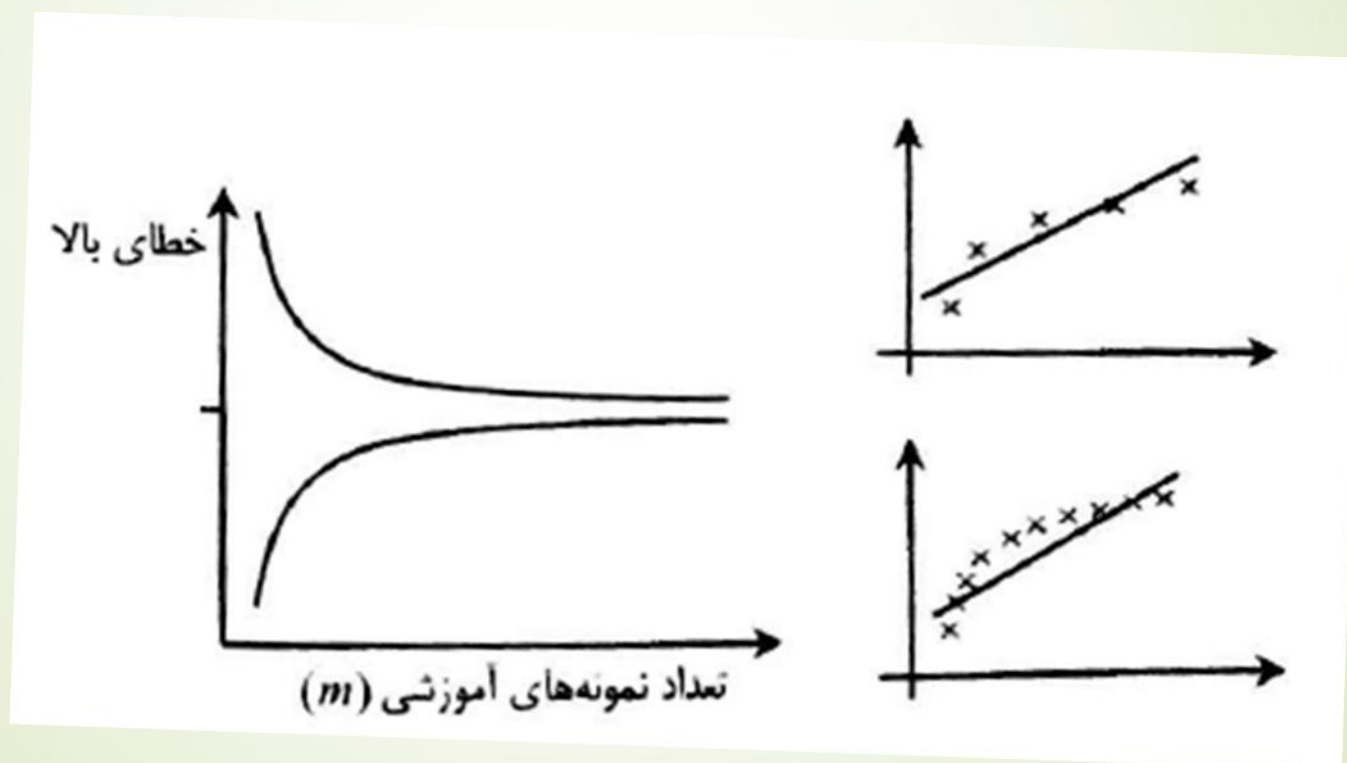
$$J_{train}(\theta) = \frac{1}{2m_{train}} \sum_{i=1}^{m_{train}} h_{\theta}(x_{train}^i) - y_{train}^i)^2$$

$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} h_{\theta}(x_{cv}^i) - y_{cv}^i)^2$$



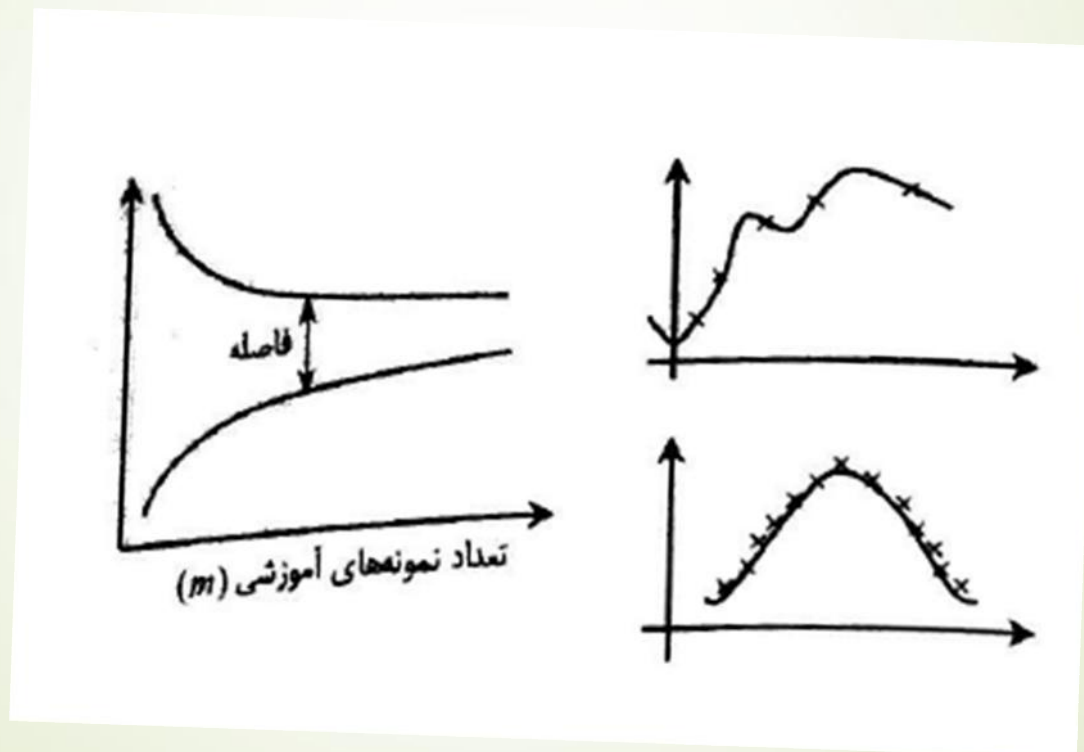
در کم تطابقی (بایاس بالا) افزایش داده باعث بهبود عملکرد نمی شود.

در نمودار خطا فاصله کم نمودار آموزش و صحت سنجی نشان دهنده کم تطابقی است.



در بیش تطابقی (واریانس بالا) افزایش داده باعث بهبود عملکرد می شود.

در نمودار خطا فاصله زیاد نمودار آموزش و صحت سنجی نشان دهنده بیش تطابقی است.





## جمع بندی

- در کم تطابقی (بایاس بالا) افزایش داده باعث بهبود عملکرد نمی شود
- در بیش تطابقی (واریانس بالا) افزایش داده باعث بهبود عملکرد می شود.
- در بیش تطابقی (واریانس بالا) کاهش تعداد ویژگی ها کمک کننده است ولی در کم تطابقی (بایاس بالا) تاثیری ندارد.
- در کم تطابقی (بایاس بالا) اضافه کردن ویژگی های ترکیبی کمک کننده است ولی در بیش تطابقی (واریانس بالا) تاثیری ندارد.
- کاهش مقدار  $\lambda$  در کم تطابقی (بایاس بالا) و افزایش  $\lambda$  بیش تطابقی (واریانس بالا) کمک کننده است.

From:  
cheapsales@buystufffromme.com  
To: ang@cs.stanford.edu  
Subject: Buy now!

Deal of the week! Buy now!  
Rolex w4tchs - \$100  
Medicine (any kind) - \$50  
Also low cost Mortgages available.

From: Alfred Ng  
To: ang@cs.stanford.edu  
Subject: Christmas dates?

Hey Andrew,  
Was talking to Mom about plans  
for Xmas. When do you get off  
work? Meet Dec 22?

ایمیل های عادی و اسپم

$$x = \begin{bmatrix} \cdot & \text{andrew} \\ 1 & \text{buy} \\ 1 & \text{deal} \\ \cdot & \text{discount} \\ \vdots & \vdots \\ 1 & \text{now} \\ \vdots & \vdots \end{bmatrix} \quad x \in \mathbb{R}^n$$

برای ساخت مدل ویژگی ها و خروجی ها مشخص می شود.

معمولاً از تکنیک دانه باشی استفاده می شود.

ساخت تدریجی مدل

ساخت مدل ساده سپس اضافه کردن ویژگی ها و داده ها با تحلیل مدل

با استفاده از نمودار خطا و تحلیل کم و بیش برآزشی

## تحلیل خطا

- منظور از تحلیل خطا، بررسی دقیق نمونه هایی است که الگوریتم به درستی دسته بندی نکرده است تا علت آن مشخص گردد.
- افزایش یا کاهش ویژگی ها مشخص می شود.
- ضعف مدل مشخص می شود.
- مانند اسپم هایی که درست تشخیص داده نشده اند.
- وجود کلمه خاصی
- وجود علامت خاصی
- شیوه نوشتن متفاوت حروف
- مدل ساده اولیه کمک می کند که نمونه های دشوار به موقع تشخیص و تحلیل شوند.

## محاسبه شاخص کمی در داده های دارای چولگی

➤ محاسبه شاخص کمی وقتی نمونه ها دارای چولگی باشد دارای اشکال است.

➤ شاخص کمی در بیماران سرطانی

➤ نسبت بیماران به افراد سالم کمتر است.

➤ اگر تمام بیماران را به اشتباه سالم تشخیص دهد عدد ناچیزی خواهد شد. (۹۹٪ نمونه های سالم و بیمار را درست تشخیص می دهد و خطا برای هر کدام ۱٪ است)

➤ آیا خطای ۱٪ مطلوب است؟؟؟

➤ در داده های دارای چولگی شاخص خطای کلی، شاخص مناسبی نیست.

### Confusion Matrix

		دسته واقعی	
		1	0
دسته پیش‌بینی شده	1	به درستی مثبت TP	به اشتباه مثبت FP
	0	به اشتباه منفی FN	به درستی منفی TN

$$\text{دقت} = \frac{\text{به درستی مثبت}}{\text{پیش‌بینی شده مثبت}} = \frac{\text{به درستی مثبت}}{\text{به اشتباه مثبت} + \text{به درستی مثبت}}$$

$$\text{بازخوانی} = \frac{\text{به درستی مثبت}}{\text{واقعا مثبت}} = \frac{\text{به درستی مثبت}}{\text{به اشتباه منفی} + \text{به درستی مثبت}}$$

➤ دقت (precision): نشان می‌دهد که چند درصد مواردی که به درستی مثبت ( $y=1$ ) پیش‌بینی کرده است، واقعاً مثبت بوده‌اند.

$$p = \frac{TP}{TP + FP}$$

➤ بازخوانی (recall): نشان می‌دهد که چند درصد موارد واقعاً مثبت ( $y=1$ ) درست مثبت پیش‌بینی شده‌اند.

$$R = \frac{TP}{TP + FN}$$

		دسته واقعی	
		1	0
دسته پیش‌بینی‌شده	1	1558 TP	125 FP
	0	255 FN	2633 TN

ایمیل 4571

عادی 2758

اسپم 1813

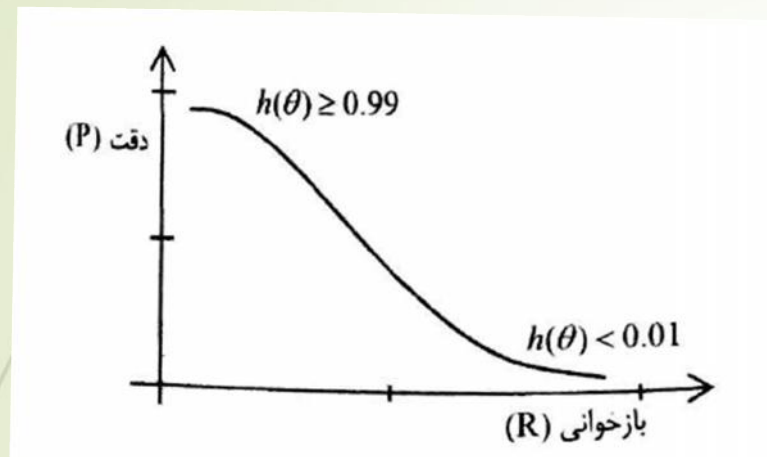
Confusion Matrix

$$p = \frac{TP}{TP + FP} = \frac{1558}{1558 + 125} = 92.5\%$$

$$R = \frac{TP}{TP + FN} = \frac{1558}{1558 + 255} = 86\%$$

افزایش یا کاهش شاخص های دقت و بازخوانی به چه معناست؟

در مواردی که پیش بینی دست های کوچک تر اهمیت بیشتری دارد (سرطانی) شاخص دقت و بازخوانی ما را به اشتباه می اندازد.



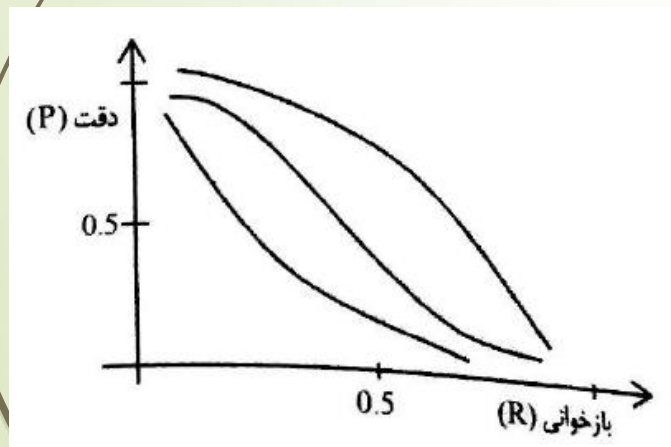
شاخص دقت بیشتر  
 اگر  $h(\theta) \geq 0.7$  آن گاه  $y = 1$

اگر  $h(\theta) < 0.7$  آن گاه  $y = 0$

شاخص بازخوانی بیشتر

اگر  $h(\theta) \geq 0.3$  آن گاه  $y = 1$

اگر  $h(\theta) < 0.3$  آن گاه  $y = 0$



تعادل مطلوب بین شاخص دقت و بازخوانی

## انتخاب الگوریتم براساس شاخص های دقت و بازخوانی

$$\text{شاخص} = \frac{P + R}{2}$$

شاخص دقت (P)	شاخص بازخوانی (R)
۰.۵	۰.۴
۰.۷	۰.۱
۰.۰۲	۱.۰

میانگین شاخص ها	شاخص بازخوانی (R)	شاخص دقت (P)
۰.۴۵	۰.۴	۰.۵
۰.۴	۰.۱	۰.۷
۰.۵۱	۱.۰	۰.۰۲

F <sub>۱</sub> Score	میانگین شاخص ها	شاخص بازخوانی (R)	شاخص دقت (P)
۰.۴۴۴	۰.۴۵	۰.۴	۰.۵
۰.۱۲۵	۰.۴	۰.۱	۰.۷
۰.۳۹۲	۰.۵۱	۱.۰	۰.۰۲

$$F_1\text{Score} = 2 \frac{PR}{P + R}$$