

مقدمه‌ای بر یادگیری ماشین

دانلود

کارگاه یادگیری ماشین در فیزیک



دانشگاه شهید بهشتی
پژوهشگده فضای مجازی
بهار ۱۳۹۸
امد م Hammond ازناوه

فهرست مطالب

- یادگیری ماشین و تمولات اخیر
 - انقلاب صنعتی چهارم و نقش هوش مصنوعی
- یادگیری چیست؟
 - کاربردهای یادگیری ماشین
 - انواع شیوه‌های یادگیری
- تحلیل خطا
 - انتخاب مدل و تعمیمه‌پذیری
 - بیش‌بازش و کم‌بازش
 - مسئله سوگیری و واریانس
- Bayesian decision theory
 - آشنایی با softmax
- نمونه‌هایی از الگوریتم‌های یادگیری ماشینی
 - ترکیب یادگیرنده‌ها

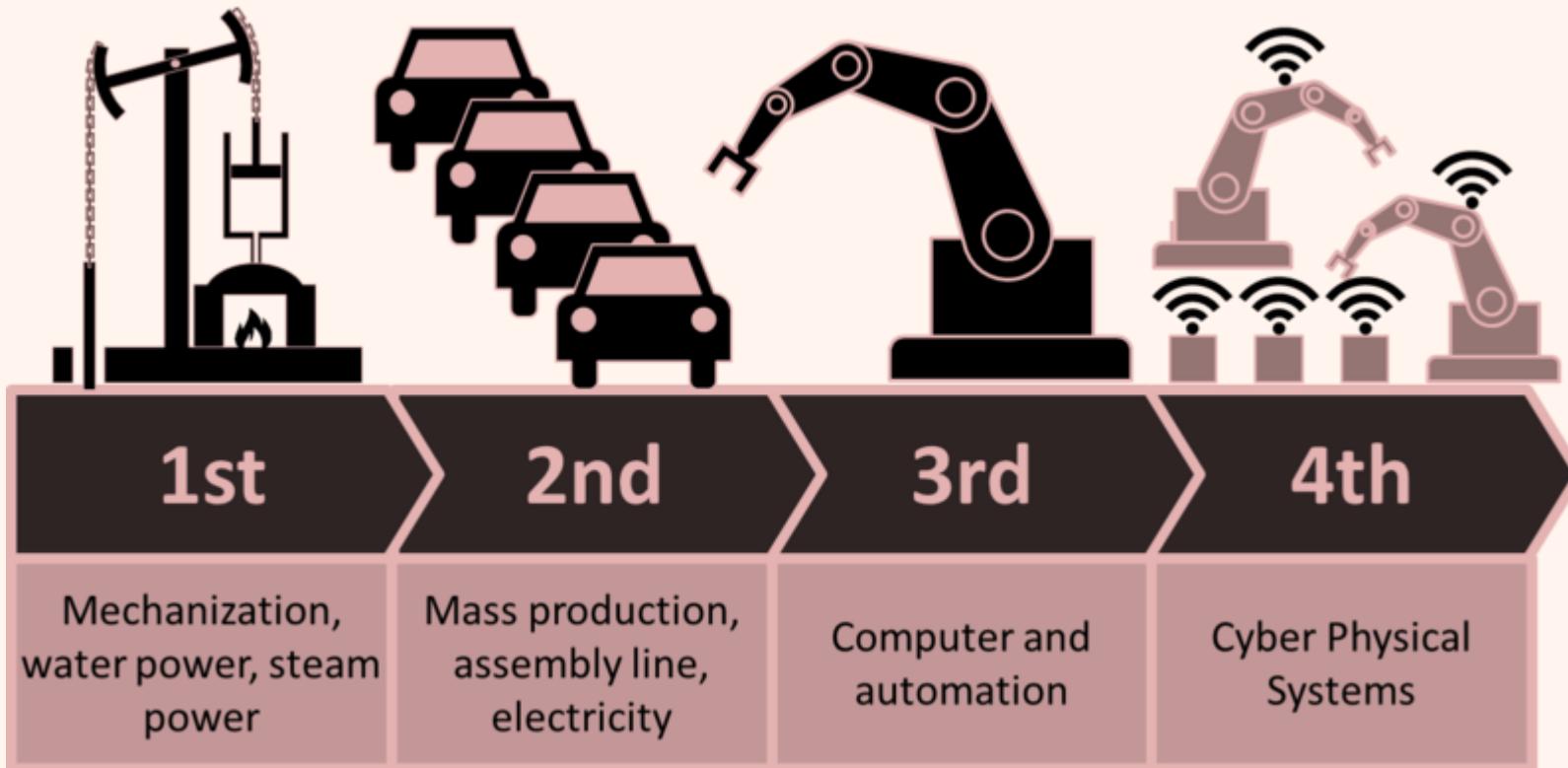


دانشکده
سینما
بهمیتی

نقش یادگیری ماشین



انقلاب صنعتی پهلو



دانشکده
بصیرتی

One Hundred Reasons to be a Scientist

GROWING UP IN ‘SCIENCE’

John J. Hopfield

Princeton University
Princeton, USA



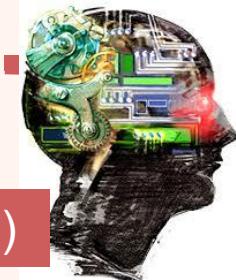
Children are naturally inquisitive, and will poke bugs to see how they respond, toss twigs in a stream to see how far they will go before they get stuck, will take apart a toy to see what its pieces are like, wonder where the water disappears to when it goes down the drain. I grew up in a household in which exploration was not just tolerated but encouraged. Activities that I can remember began on the kitchen floor playing with pots and pans, removing all part that could be unscrewed. My father repaired everything—the roof, the radio,

Physics is an exploration of what is not understood about the way things are, in search of essential principles, facts, and quantitative description. Some fall in love with the mysteries of the origin of the universe, or the nature of the world at distance scales that are unbelievably small. For me, having been brought up curious about the world around me, and fascinated to

My present enthusiasm in science might be described as ‘how do we think?’ It is the type of question I have always pursued, though with age the questions have gotten harder. Is it biology or physics? It doesn’t matter. Perhaps physics is best defined simply as ‘what those trained in physics do’.



هوش مصنوعی



Artificial intelligence (AI, also machine intelligence, MI)

- ای امکان تحقق کامل اهداف هوش مصنوعی

۵۵۵ دارد

It has long been believed, especially by older members of the scientific community, that for machines to be as intelligent as us, that is, for artificial intelligence to be a reality, our current knowledge in general, or computer science in particular, is not sufficient. People largely are of the opinion that we need a new technology, a new type of material, a new type of computational mechanism or a new programming methodology, and that, until then, we can only “simulate” some aspects of human intelligence and only in a limited way but can never fully attain it.

I believe that we will soon prove them wrong.

مقدمه‌ی چاپ دوم

دانشگاه
سینمای
بهریتی

Introduction to machine learning, Ethem Alpaydin

یادگیری ماشین در فریزک

ImageNet Competition

Progress of object recognition (1k ImageNet)



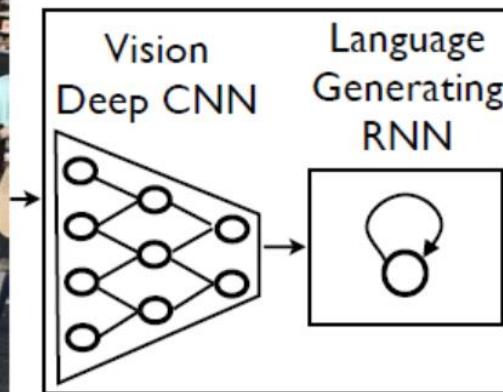
Announced
Dec. 2015



very deep: 152 layers



پیش‌رفت‌های افیدر



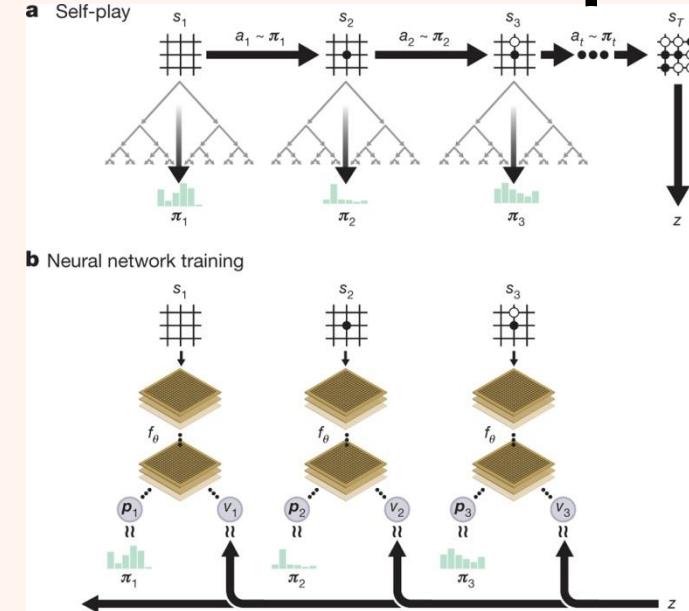
A group of people
shopping at an
outdoor market.

There are many
vegetables at the
fruit stand.



دانشکده
سینمای
بهریتی

پیش‌رفتهای افید



*Silver, D., et al. (2017). "Mastering the game of Go without human knowledge." *Nature* 550: 354.*



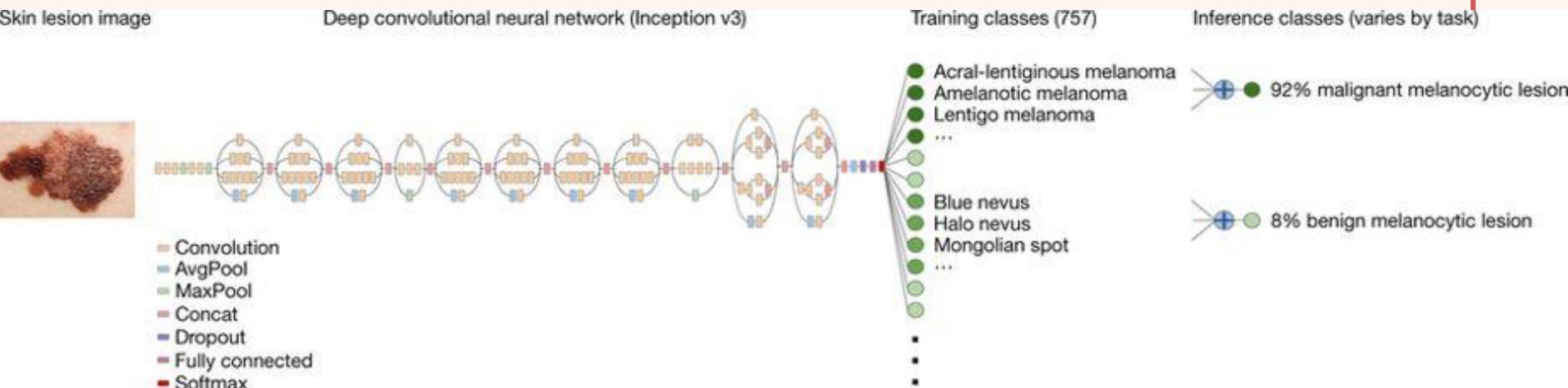
D Silver et al. *Nature* 550, 354–359 (2017) doi:10.1038/nature24270



nature

پیش‌رفت‌های افیز

A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, p. 115, 01/25/online 2017.



یادگیری ماشین در فیزیک

Long Programs

[Programs](#) > [Long Programs](#) > Machine Learning for Physics and the Physics of Learning



Machine Learning for Physics and the Physics of Learning

SEPTEMBER 4 - DECEMBER 8, 2019

OVERVIEW

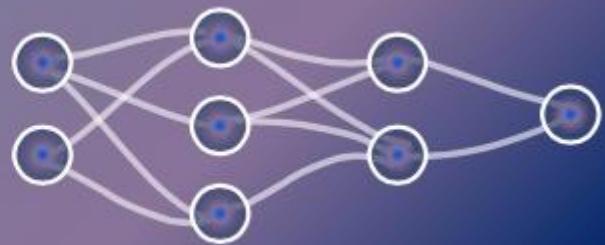
PARTICIPANT LIST

ACTIVITIES

APPLICATION

Machine Learning for Physicists

```
model.add(Dense(n_hiddenz, activation='relu'))  
model.add(Dense(3,activation='sigmoid'))  
model.compile(optimizer='Adam',  
              loss='categorical_crossentropy',  
              metrics=['accuracy'])  
  
inRange = range(len(data))  
for i in inRange:  
    all_batches = data[i].batch_size  
    samples,targets = myNet.produce_samples() # get the samples
```



سینمای
بصیرتی

Is Another AI Winter Coming?



<https://hackernoon.com/is-another-ai-winter-coming-ac552669e58c>



مقدمه‌ای یادگیری ماشین



- علوه مختلف از روش‌های مطرح شده در «**یادگیری ماشین**» استفاده می‌کنند.
- نقش «**یادگیری ماشین**» در زندگی تا په مد است؟
- کاربردهای **یادگیری ماشین** در زندگی (وزمراه اه یافته است:

– تشخیص دستنوشته

– خودرو بدون راننده

– تشخیص چهره

– تشخیص هرزناهه

– سیستم‌های توصیه‌گر



دانشکده
سینمایی
بهشتی

یادگیری چیست؟

- «یادگیری» عبارتست از تغییر نسبتاً پایدار در احساس، تفکر و رفتار فرد که بر اساس تجربه ایجاد شده باشد.

به نقل از ویکی‌پدیا

Learning is the act of acquiring new, or modifying and reinforcing existing knowledge, behaviors, skills, values, or preferences.

The ability to learn is possessed by humans, animals and some machines.



یادگیری ماشین چیست؟

Machine Learning

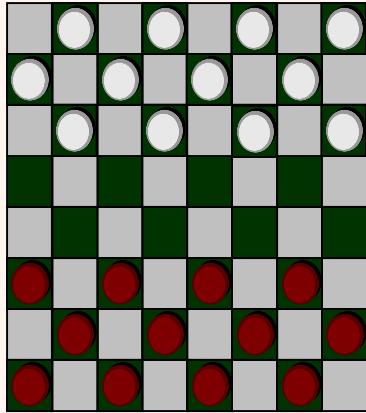
- «**یادگیری ماشین**» برنامه‌نویسی برای بهینه‌سازی یک عملکرد با استفاده از داده‌ها و تجربیات گذشته است.

Machine learning is programming computers to optimize a performance criterion using example data or past experience.

- «**یادگیری ماشین**» در پی راهی برای ایجاد برنامه‌ای است که عملکرد را به صورت خودکار و با توجه به تجربیات ارتقا دهد. (Tom. M. Mitchell)



Field of study that gives computers the ability to learn without being explicitly programmed.



Arthur Samuel (1959)



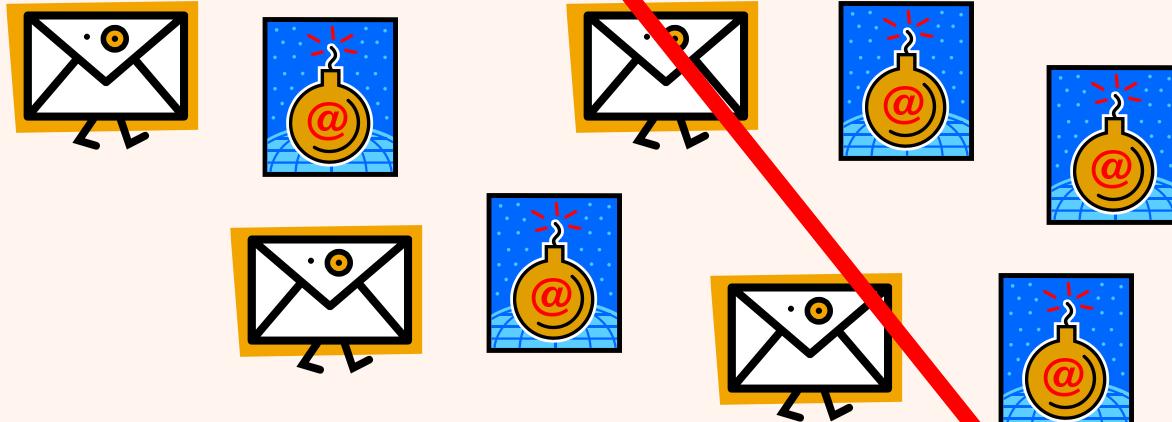
Well-posed Learning Problem: A computer program is said to learn from experience **E** with respect to some task **T** and some performance measure **P**, if its performance on **T**, as measured by **P**, improves with experience **E**.

Tom Mitchell (1998)



سایر تعاریف - مثال

- به عنوان مثال یک برنامه‌ی **تشخیص هرزنامه** را در نظر بگیرید که با توجه به ایمیل‌هایی که کاربر به عنوان spam اعلان می‌کند، سعی دو بهبود کارایی فود دارد.
- در این صورت T - عمل دسته‌بندی ایمیل‌ها به دو گروه spam/not spam - پی‌گیری ایمیل‌هایی که کاربر به عنوان spam اعلان می‌کند. E - تعداد ایمیل‌هایی که به درستی به عنوان spam دسته‌بندی شده‌اند. P -



دانشکده
سیستمی
بهسیانی

پرا یادگیری؟

- برای حل یک مسئله بر روی کامپیوتر به یک «الگوریتم» احتیاج داریم.
- برای برخی مسائل نمی‌توان یک الگوریتم نوشت، مانند تشخیص هرزنامه‌ها. حتی ممکن است برخی از ایمیل‌ها بسته به کاربر هرزنامه تلقی شوند یا نه، اما نمونه‌های زیادی از داده در اختیار داریم.
- در جاهایی که نمی‌توانیم مستقیماً برنامه‌یی مورد نظر را بنویسیم، به یادگیری احتیاج داریم، که با کمک یک سری **داده‌ی آموختی یا تجربیات** صورت می‌پذیرد.
 - به عنوان مثال برای مهاسبه‌ی محقق پرسنل نیازی به یادگیری وجود ندارد.



دانشکده
سینمایی
بهشتی

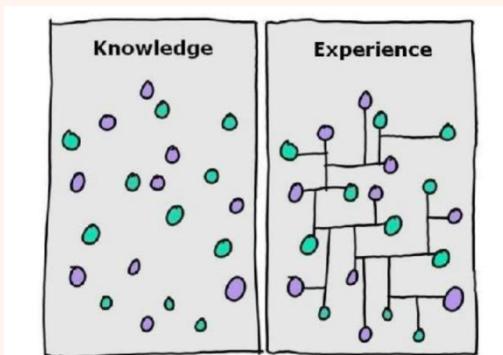
منظور از یادگیری

- یادگیری ← استخراج یک مدل کلی از روابط داده‌های
- بیشتر داده‌های اخذ شده توسط وسایل مختلف به صورت دیجیتال هستند.
- در واقع با محض انبوهی از داده‌ها مواجه هستیم که ارزان به دست نمی‌آیند، با این حال دانش در خصوص آنها به سادگی و با هزینه‌ی پایین حاصل نمی‌شود.

We are drowning in information and starving for knowledge. John Naisbitt.



دانشگاه
سینمایی



منظور از یادگیری (ادامه...)

- هرچند ممکن است قادر به ارائه یک مدل دقیق نباشیم، اما می‌توان یک تقریب موب و مفید به دست آورد.
- مدل به دست آمده می‌تواند برای پیش‌بینی مورد استفاده قرار گیرد (predictive) و یا به منظور استخراج دانش از داده‌ها به کار آید (descriptive).
- در موارد زیر به یادگیری احتیاج داریم:
 - در مواردی که انسان خیلی در دسترس نیست،
 - زمانی که انسان قادر به توضیح مهارت خود نیست، مانند تشخیص صوت
 - زمانی که مساله مورد نظر در طول زمان تغییر می‌کند؛ به شرایط محیط وابسته است
- حالاتی که به حل مساله به تطبیق با شرایط خاصی وابسته است،



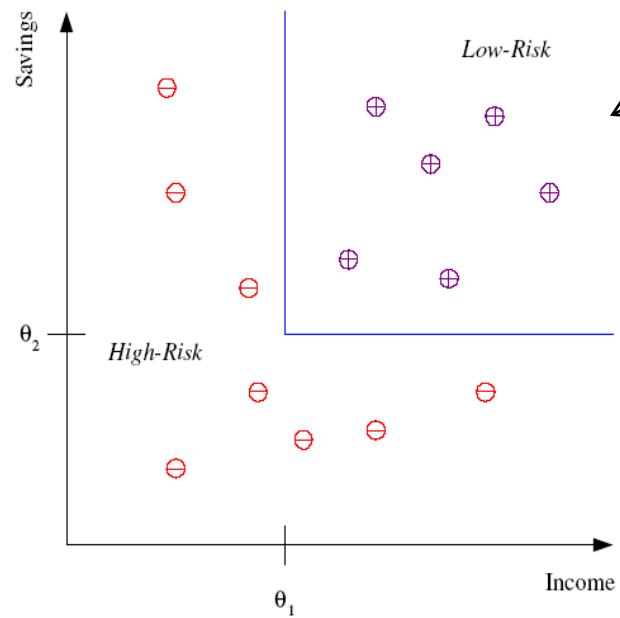
دانشکده
سینماسازی
بهشتی

• ارزیابی اعیان (credit scoring)

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_{t=1}^N \quad r = \begin{cases} 1 & \text{if } x \text{ is positive} \\ 0 & \text{if } x \text{ is negative} \end{cases}$$

Discriminant



Discriminant: IF *income* > θ_1 AND *savings* > θ_2

THEN low-risk ELSE high-risk



دسته‌بندی (ادامه...)

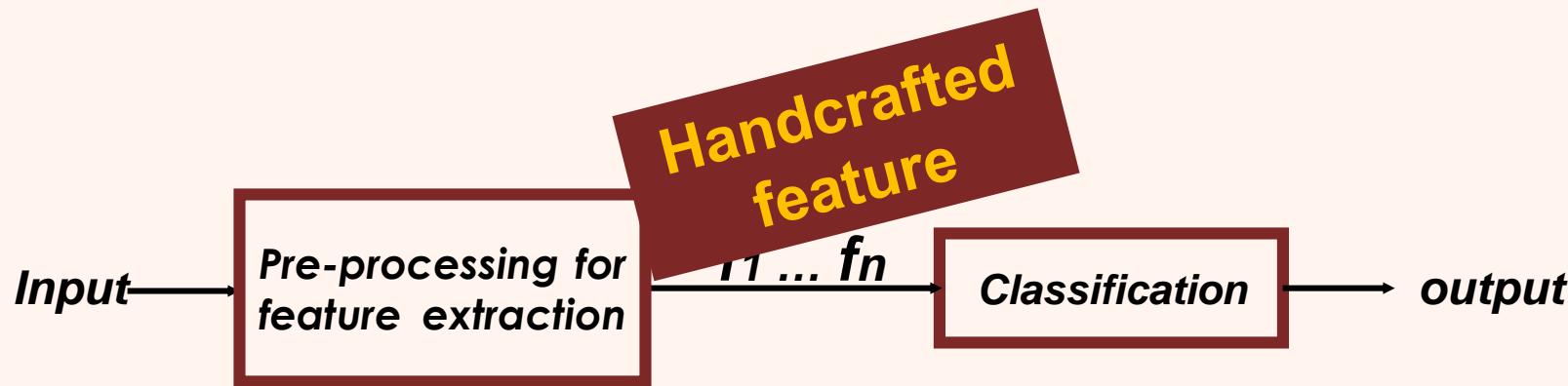
- «بازشناصی الگو» هم نامیده می‌شود.
- تشخیص کاراکتر (OCR)
 - تشخیص کاراکترهای دستنویس
 - یک کلمه دنباله‌ای از کاراکترهای است
 - t?e •
- تشخیص هویت با استفاده از دستخط

0	4	1	9	2	1	3	1	4	3
5	3	6	1	7	2	8	6	9	4
0	9	1	1	2	4	3	2	7	3
8	6	9	0	5	6	0	7	6	1
8	7	9	3	9	8	5	9	3	3
0	7	4	9	8	0	9	4	1	4
4	6	0	4	5	6	1	0	0	1
7	1	6	3	0	2	1	1	7	9
0	2	6	7	8	3	9	0	4	6
7	4	6	8	0	7	8	3	1	5



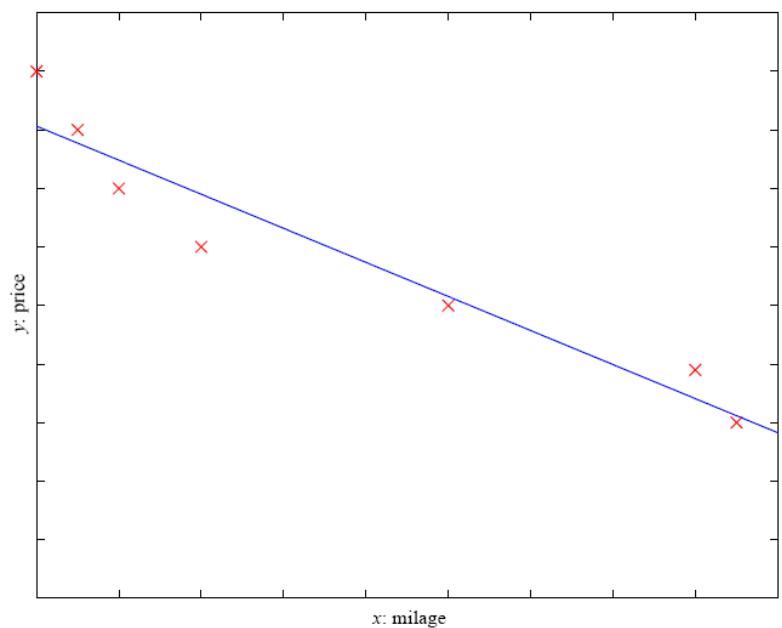
دانشکده
سینما و
بهاشتی

اسندخراج فصیحه در روش‌های کلاسیک



دانشکده
سینمایی

رگرسیون



$$y = w_1 x + w_0$$

پ

$$y = w_2 x^2 + w_1 x + w_0$$

یادگیری ماشین در فریزک

- دسته‌بندی و رگرسیون هر دو از نمونه‌های یادگیری با ناظارت (supervised) هستند.

- مثال: قیمت اتومبیل دسته‌بندی

x : car attributes

y : price

$$y = g(x | \theta)$$

$g(\cdot)$ model,

θ parameters



دانشکده
سینمایی

شیوه‌های یادگیری

Supervised learning

Unsupervised learning

Semi-supervised learning

Active learning

Reinforcement learning

• یادگیری بانظارت

• یادگیری بینظارت

• یادگیری نیمه‌نظارتی

- یادگیری فعال

• یادگیری تقویتی



دانشکده
سینماسازی
بهسیانی

- در این شیوه همراه با نمونه‌های آموزشی، پاسخ مطلوب هم وجود دارد.
 - پیش‌بینی نمونه‌های جدید
 - استخراج دانش
 - فشرده‌سازی



دانشکده
سینما
بهره‌برداری

- در این حالت تنها داده های وجودی وجود دارند، بدون این که ناظر مقدار مطلوب را مشخص کند.
- هدف پیدا کردن «نظم» (regularity) موجود در داده است، آنچه معمول و طبیعی است.

Density estimation

- خوشه بندی (clustering): گروه بندی

- مدیریت ارتباط با مشتری
- تشخیص نمونه های غیرنرمال؛ تشخیص تقلب و سوءاستفاده
- فشرده سازی تصویر (پندی سازی (نگ))
- بیوانفورماتیک (Learning motifs)



دانشگاه
بهشتی



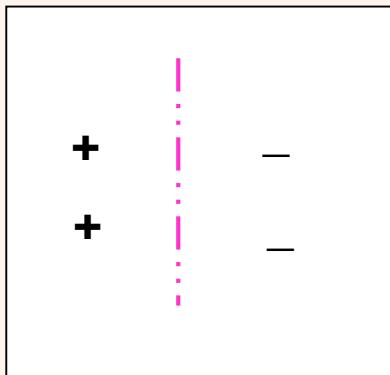
Datasets—not algorithms—might be the key limiting factor to development of human-level artificial intelligence.

(Alexander Wissner-Gross)

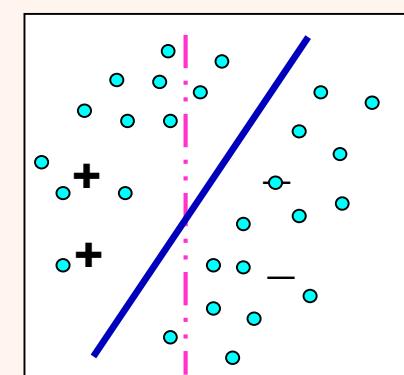
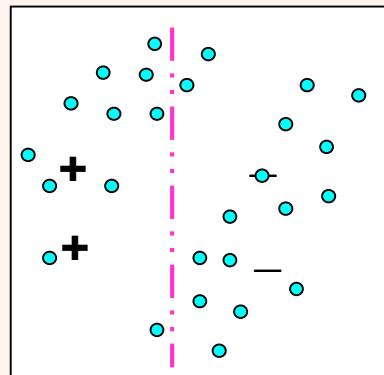


دانشکده
سینمای
بهریتی

- تنها بخشی از داده‌ها برچسب فورده‌اند، و مجمل زیادی از آن بدون برچسب هستند.
- برچسب زدن داده‌ها کار پرهزینه‌ای است.
- از طرفی، داده‌های برچسب نفوردیدی زیادی در اختیار داریم.



یادگیری بانظارت



یادگیری نیمه‌ناظاری



- در برخی موارد خروجی یک سیستم، دنباله‌ای از «گنش»‌هاست. به گونه‌ای که یک مرکت اهمیت ندارد، بلکه سیاستی است که باعث می‌شود مجموع مرکات، به هدف مناسب برسند.
- یک عمل مناسب است در صورتی که در مجموع و در کنار سایر اعمال مناسب باشد. در این حالت الگوریتم یادگیری باید قادر به انتخاب سیاست مناسب باشد.

Game playing

Robot in a maze

Multiple agents, partial observability, ...



ارزیابی الگوریتم‌های یادگیری

- بسته به کاربرد، برای ارزیابی الگوریتم‌های یادگیری، دقیقت دسته‌بندی، مجمم محاسبات و حافظه‌ی مورد نیاز در نظر گرفته می‌شود.
- الگوریتم‌های یادگیری متفاوتی وجود دارند؛ بسته به شرایط کاربرد مورد نظر، الگوریتم‌های متفاوتی را می‌توان مورد استفاده قرار داد.
- مجمم مورد نیاز داده‌های آموزشی، پیمیدگی الگوریتم‌های مورد استفاده و قابلیت تعمیم مسائلی است که باید مورد بررسی قرار گیرند.



دانشکده
سینمایی
بهشتی

مقدمه‌ای بر اگررسیون

- در اگررسیون، بخلاف دسته‌بندی با یک تابع پیوسته

مواجه هستید:

$$\mathcal{X} = \left\{ x^t, r^t \right\}_{t=1}^N \quad r^t \in \mathcal{R}$$

- برخلاف درون‌یابی، در اگررسیون وجود نویز در فروجی

$r^t = f(x^t) + \varepsilon$ را هم باید در نظر گرفت.

- وجود نویز را می‌توان به مربوط به متغیرهای مخفی (غیرقابل مشاهده) دانست.

$$r^t = f^*(x^t, z^t)$$

- هدف، تخمین فروجی با استفاده از مدل پیشنهادی $(g(x))$ است.



دانشکده
سینمایی

انتفاب مدل ۹

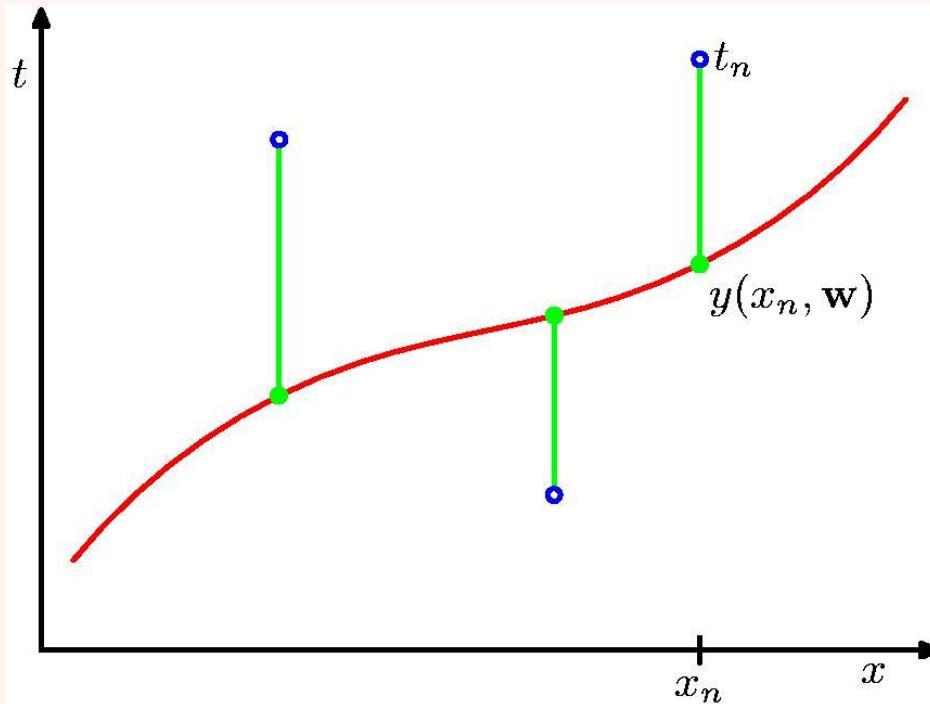
تعمید پذیری



مددهای بررسیون (داده...)

- خطا داده‌های آموزشی را می‌توان به صورت زیر تعریف کرد:

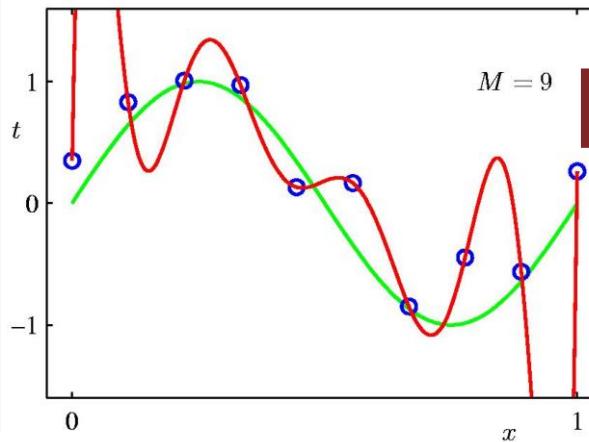
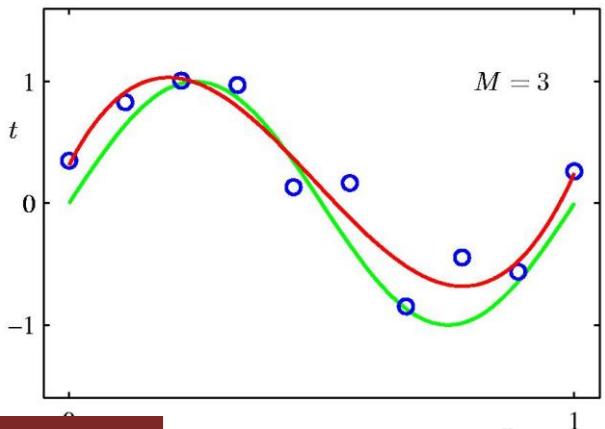
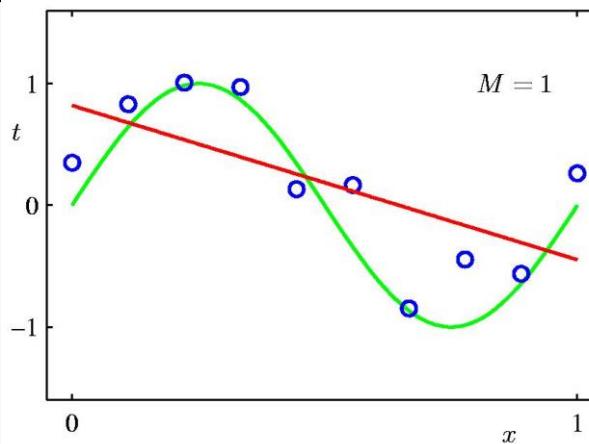
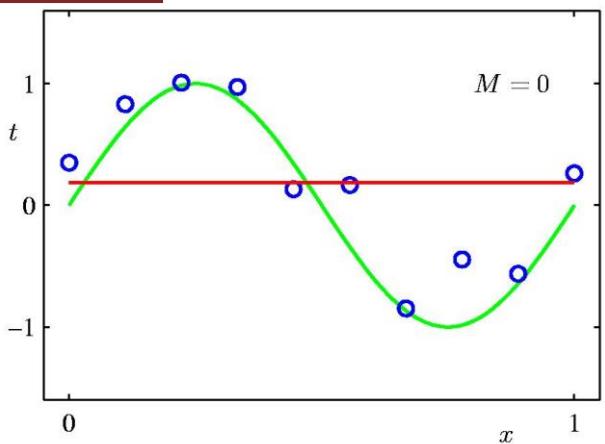
$$E(g | \mathcal{X}) = \frac{1}{N} \sum_{t=1}^N [r^t - g(x^t)]^2$$



دانشکده
سینمایی

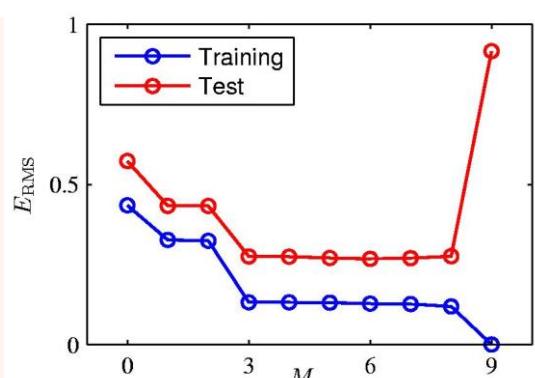
Underfitting

چند داده‌ها را با داده‌های بالاتر پیش‌ساخت



Overfitting

Best fit

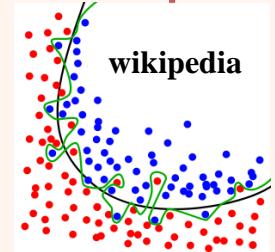


دانشکده
سینمایی

- در نتیجه، علاوه بر داده‌ها باید مفروضات دیگری در نظر گرفت که پاسخ یکتاًی به دست آید. این پیش‌فرض‌ها «**inductive bias**» نامیده می‌شود.
 - کلاس فرضیه، پیش‌فرض مذکور تلقی می‌شود.
- هر چه ظرفیت فرضیه افزایش یابد، پیمایدگی آن نیز بیشتر خواهد شد.
- دو مستطیل ناهمپوشان در مقابل یک مستطیل و انتخاب مستطیل به بیشترین حاشیه
- یا انتخاب خط برای رگرسیون و یا معیار مینیمم خطا
- در «**انتخاب مدل**» باید تعمیم‌پذیری را در نظر داشت.



انتخاب مدل و تعمیم‌پذیری



برای افزایش «قابلیت تعمیم» باید پیچیدگی مدل مناسب با پیچیدگی داده‌ها انتخاب شود.

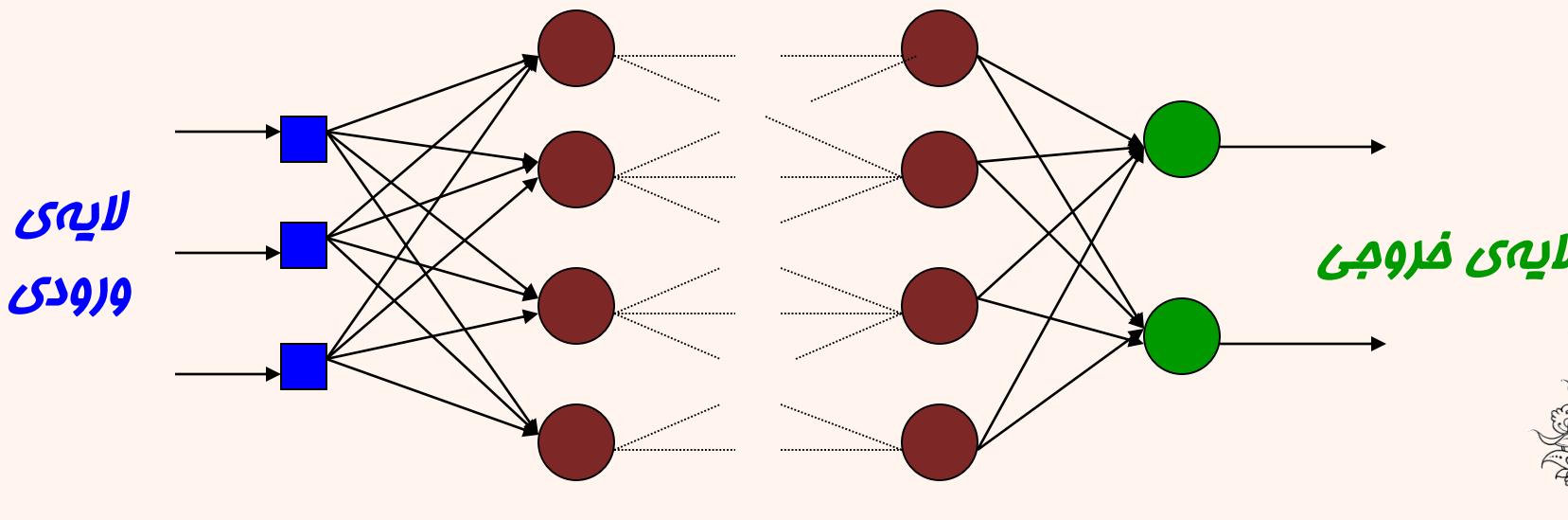
- در صورتی که پیچیدگی مدل کمتر از داده باشد، اصطلاحاً گفته می‌شود **underfitting** (خ داده است).
 - مانند زمانی که یک منحنی درجه‌ی سه با یک خط تقریب زده شود.
 - در چنین حالتی خطای آموختشی و خطای مرحله‌ی validation (validation error) هر دو بالا خواهند بود.
- در صورتی که مدل پیچیده‌تر انتخاب شود،
overfitting (خ می‌دهد).
 - با افزایش داده‌های آموختشی می‌توان اثر آن را تا حدی کاهش داد.



شبکه عصبی پنل لای

Multilayer Neural Network

Multilayer Neural Perceptron(MLP)



لایه‌های مخفی



Structure	Types of Decision Regions	Exclusive-OR Problem	Classes with Meshed regions	Most General Region Shapes
Single-Layer 	Half Plane Bounded By Hyperplane			
Two-Layer 	Convex Open Or Closed Regions			
Three-Layer 	Arbitrary (Complexity Limited by No. of Nodes)			

دانشکده
سینمایی
بهریتی

گانه Tradeoff

- بین عوامل زیر tradeoff وجود دارد:
 - پیچیدگی کلاس فرضیه \mathcal{H} ,
 - اندازه مجموعه آموزشی N ,
 - خطا تعمیمی E .

$N, E \downarrow$
 $c(\mathcal{H})$, first $E \downarrow$ and then E



دانشکده
سینمایی

- برای بررسی تعمیم‌پذیری، بخشی از داده‌ها را در آموزش مورد استفاده قرار نمی‌دهیم (validation set)، و تنها برای بررسی تعمیم‌پذیری از آن‌ها استفاده می‌شود.
- در نتیجه، فرضیه‌ای که با داده‌های validation بهترین پاسخ را دارند، به عنوان فرضیه‌ی مناسب انتخاب می‌شود.
- بعد از آموزش، برای مقایسه (وش) مورد استفاده، داده‌های آزمایش که باید متفاوت از داده‌های آموزشی و داده‌های validation هستند، مورد استفاده قرار گیرد.

Test set (publication set)



دانشکده
سینما
بهره‌بری

ابعاد متفاوت الگوریتم‌های یادگیری ماشین

$$g(\mathbf{x} | \theta)$$

- مدل (cost or loss function)

$$E(\theta | \mathcal{X}) = \sum_t L(r^t, g(\mathbf{x}^t | \theta))$$

- فرآیند بهینه‌سازی

$$\theta^* = \arg \min_{\theta} E(\theta | \mathcal{X})$$

- در صورتی که مدل پیمایده‌تر شود، به روش‌های پیمایده‌تری برای یافتن پارامترهای بهینه احتیاج خواهیم داشت.

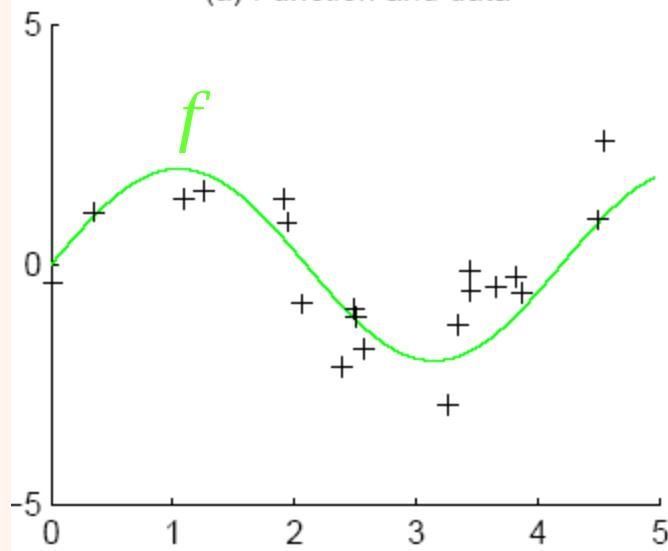
- برای انجام مناسب آموزش به مدلی با ظرفیت مناسب، تعداد نمونه‌های آموزشی مناسب و یک فرآیند بهینه‌سازی خوب احتیاج داریم.



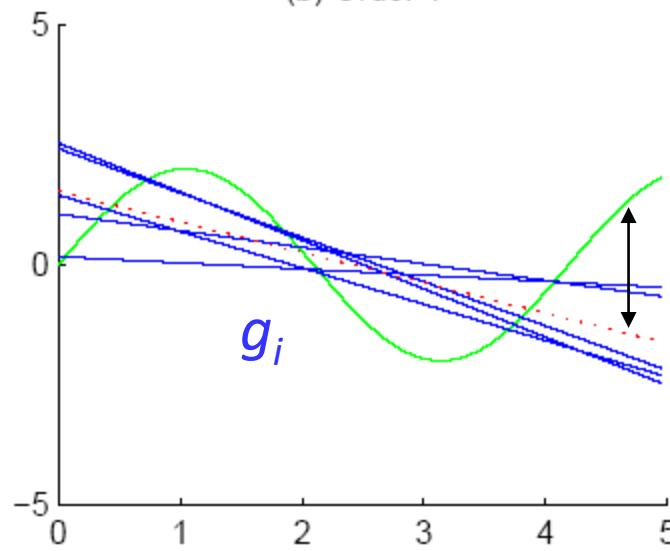
دانشکده
سینمایی
بهینه‌سازی



(a) Function and data

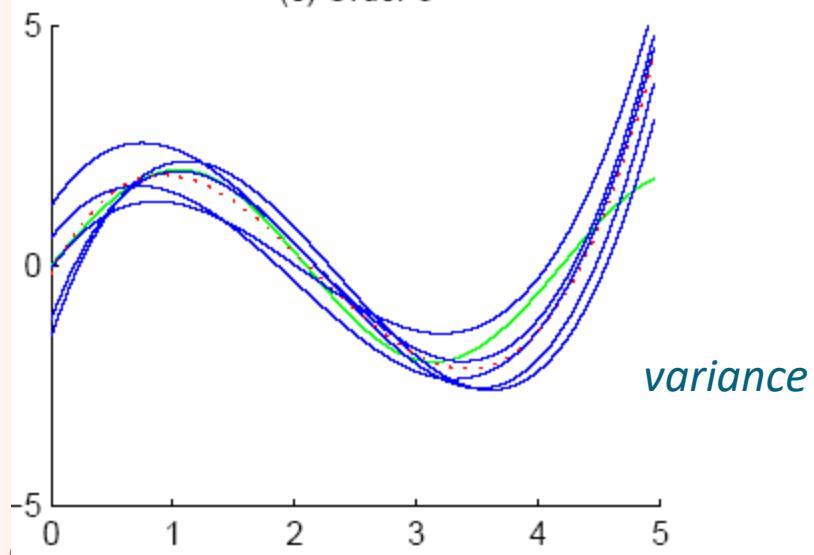


(b) Order 1

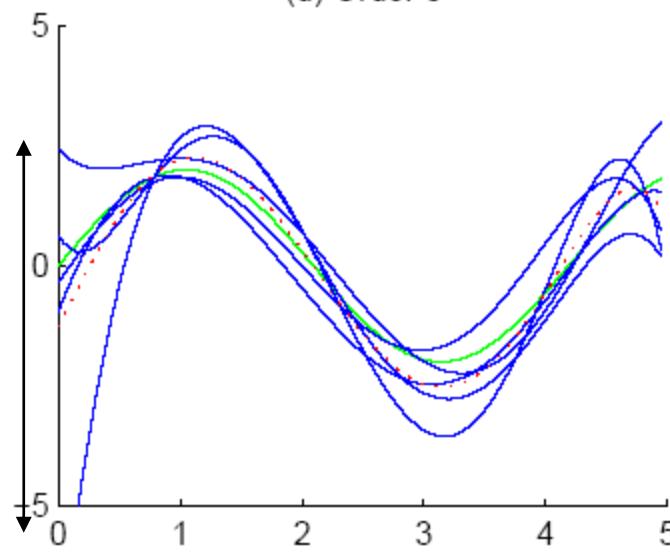


f
bias
 \bar{g}

(c) Order 3



(d) Order 5



Bias/Variance Dilemma

مثال

- M samples $X_i = \{x^t_i, r^t_i\}, i=1, \dots, M$ are used to fit $g_i(x), i=1, \dots, M$

$$\text{Bias}^2(g) = \frac{1}{N} \sum_t [\bar{g}(x^t) - f(x^t)]^2$$

$$\text{Variance}(g) = \frac{1}{NM} \sum_t \sum_i [g_i(x^t) - \bar{g}(x^t)]^2$$

$$\bar{g}(x) = \frac{1}{M} \sum_i g_i(x)$$

$$g_i(x) = 2$$

واریانس صفر است، اما بایاس بالایی دارد

$$g_i(x) = \sum_t r^t / N$$

یا و گیری ماشین در فریبک

بایاس کاهش می‌یابد، اما واریانس افزایش می‌یابد



دانشگاه
سلام

Bias and Variance

Expected square error at x

noise

squared error

$$E[(r - g(x))^2 | x] = E[(r - E[r | x])^2 | x] + (E[r | x] - g(x))^2$$

به مدل بستگی ندارد، واریانس
نویز است؛ در واقع بخشی از
خطای است که قابل مذکو نیست

میزان خطای وابسته به
داده‌های آموزشی و مدل
است

$$E_x[(E[r | x] - g(x))^2 | x] = (E[r | x] - E_x[g(x)])^2 + E_x[(g(x) - E_x[g(x)])^2]$$

bias

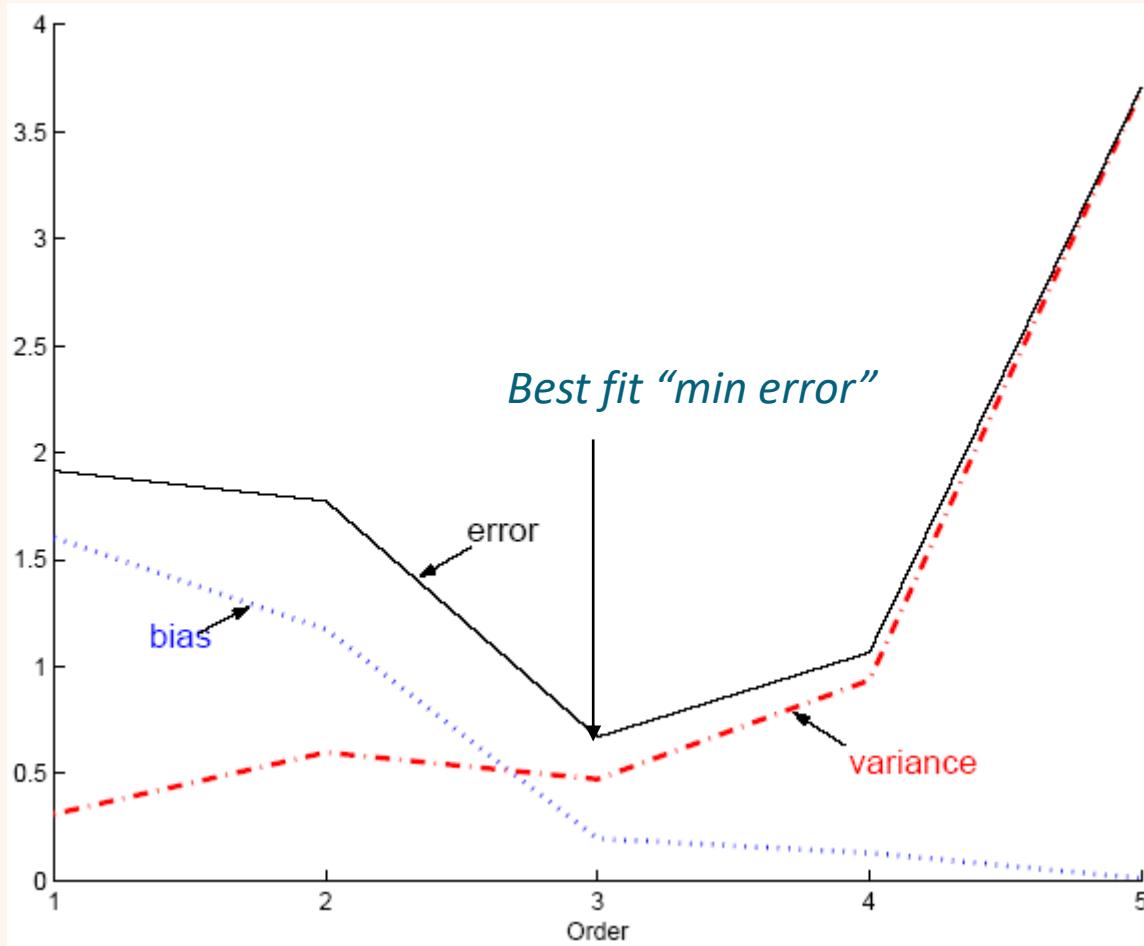
variance



محیا ری است که میزان خطای را
صرف نظر از نمونه‌های آموزشی
نشان می‌دهد

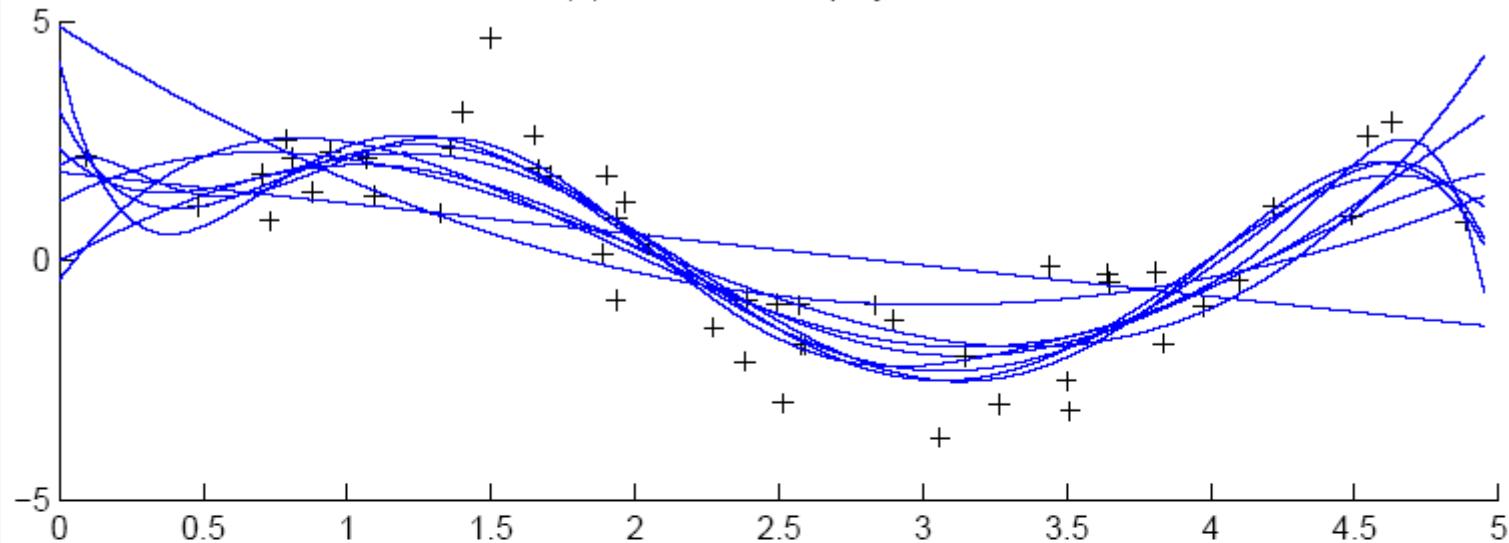
آنرا با تغییرات داده‌های آموزشی،
بهم مقدار (x) به په میزان تغییر
می‌کند.

انتفاب مدل

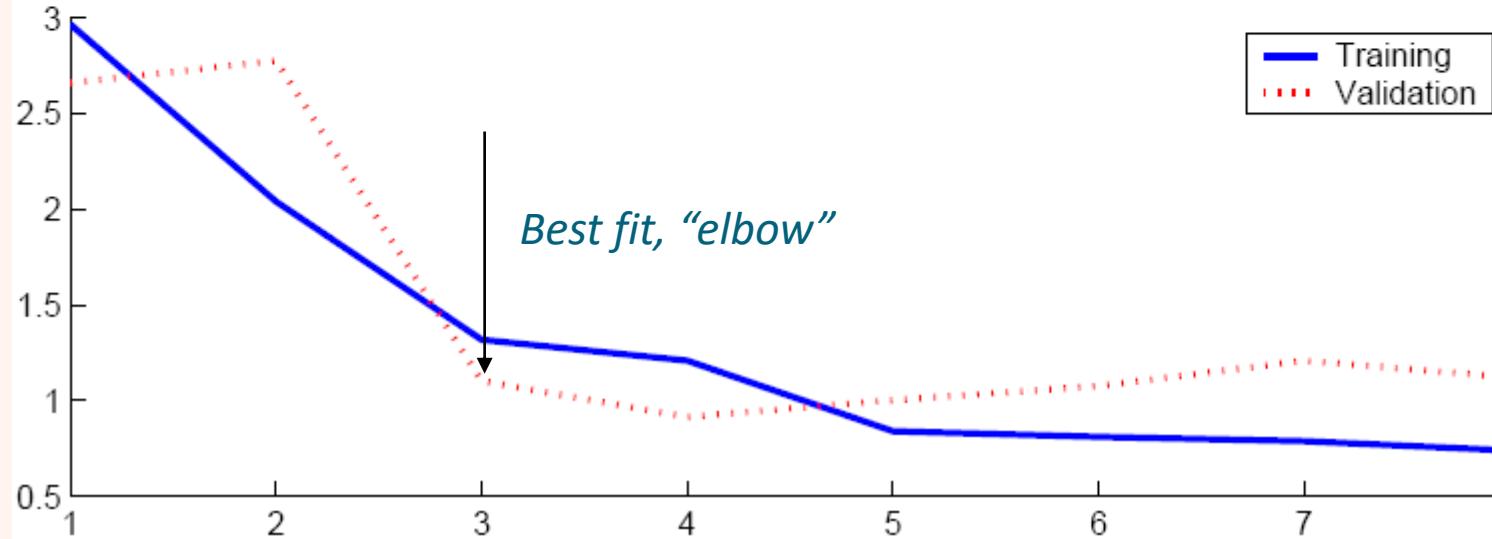


Cross validation

(a) Data and fitted polynomials



(b) Error vs polynomial order

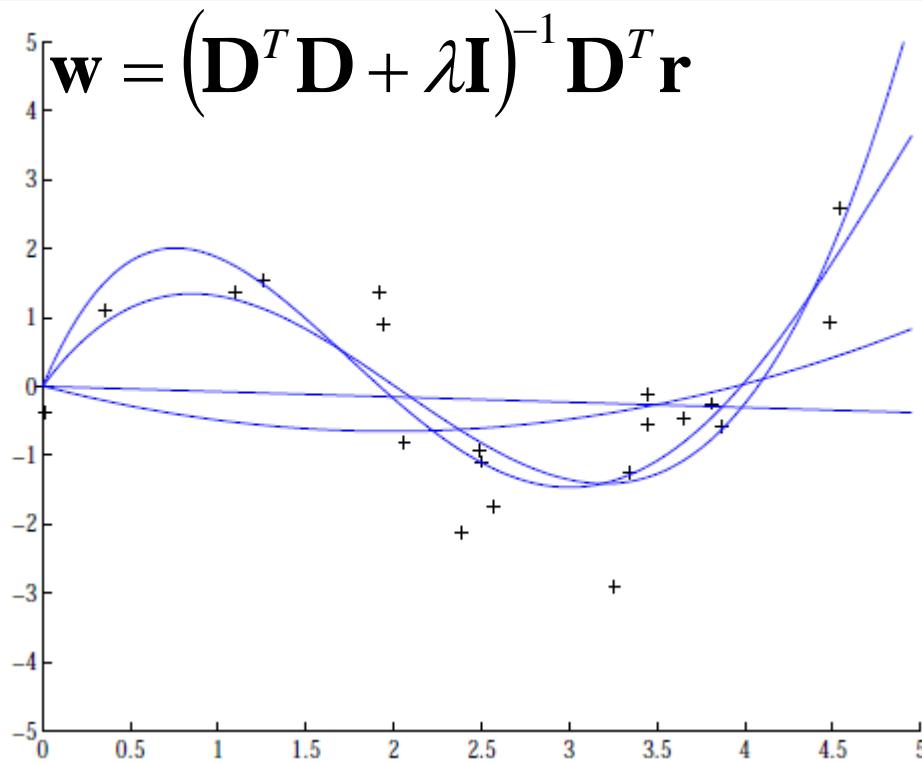


دانشکده
سینما
بیوپیشید

Regularization

Penalize complex models

E' = error on data + λ model complexity



Coefficients increase in magnitude as order increases:

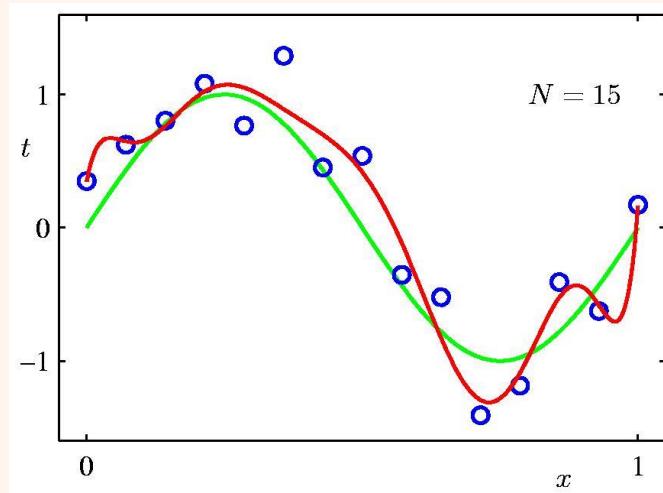
- 1: [-0.0769, 0.0016]
- 2: [0.1682, -0.6657, 0.0080]
- 3: [0.4238, -2.5778, 3.4675, -0.0002]
- 4: [-0.1093, 1.4356, -5.5007, 6.0454, -0.0019]



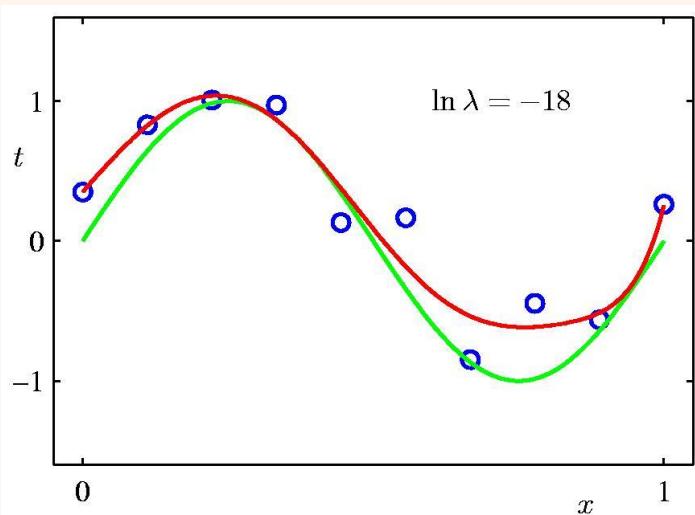
۱۵۹

$$\text{regularization: } E(\mathbf{w} | \mathcal{X}) = \frac{1}{2} \sum_{t=1}^N [r^t - g(x^t | \mathbf{w})]^2 + \lambda \sum_i w_i^2$$

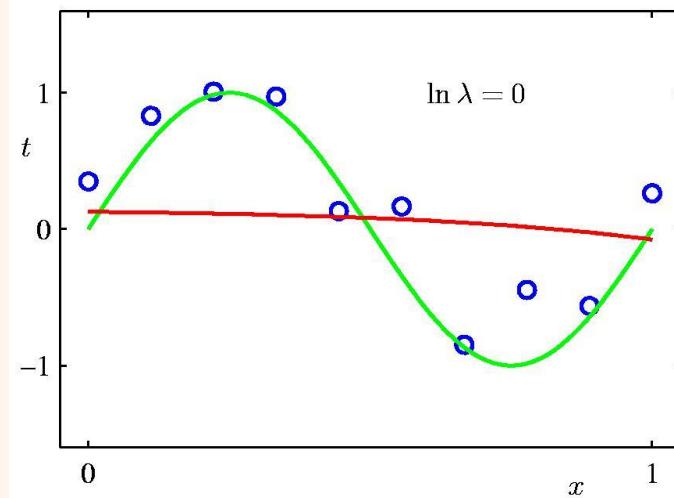
Regularization



9th Order Polynomial



$\ln \lambda = -18$



$\ln \lambda = 0$



دانشکده
سینمایی

روش‌های پارامتری



- مسئله‌ی دسته‌بندی اعتبار مشتریان:

- ۹۰٪ی: درآمد و پس‌انداز

- خروجی: مشتری High risk و low risk

- Input: $x = [x_1, x_2]^T$, Output: $C \in \{0, 1\}$

- پیش‌بینی:

- high risk($C=1$) or low risk($C=0$)

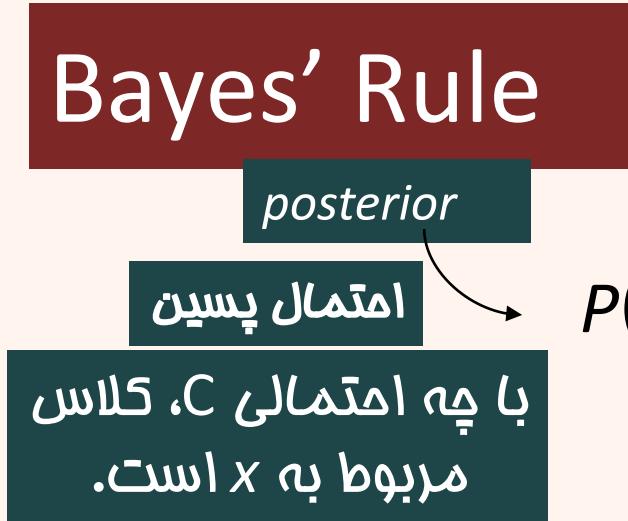
choose $\begin{cases} C = 1 & \text{if } P(C=1 | x_1, x_2) > 0.5 \\ C = 0 & \text{otherwise} \end{cases}$

or

choose $\begin{cases} C = 1 & \text{if } P(C=1 | x_1, x_2) > P(C=0 | x_1, x_2) \\ C = 0 & \text{otherwise} \end{cases}$

دسته‌بندی (ادامه...)

- با فرض این وجودی x , متغیر مشاهده شده است، مسئله یافتن احتمال $P(C|x)$ است.



$$P(C=0) + P(C=1) = 1$$

$$p(x) = p(x|C=1)P(C=1) + p(x|C=0)P(C=0)$$

امتمال پیشین

prior

درست‌نمایی کلاس

Class likelihood

با په احتمالی x توسط کلاس C تولید می‌شود.



- با توجه به این فرض که توزیع داده‌ها، از توزیعی خاص پیدوی می‌گند، این روش‌ها را «روش‌های پارامتری» می‌نامند.

- $\mathcal{X} = \{x^t\}_{t=1}^N$ where $x^t \sim p(x)$

- تخمین پارامتر:

- تخمین پارامترهای θ از داده‌های آموزشی X
- برای داده‌ها یک مدل به صورت $(x | \theta) p$ در نظر گرفته می‌شود («آماری بسنده» است؛ تمایل اطلاعات در مورد توزیع را در بر دارد)



- «تابع درستنما^ی»، تابعی از پارامترهای مدل آماری است.
- درستنما^ی یک مجموعه از پارامترها، θ ، برای مقادیری معین (X)؛ برابرست با احتمال (福德اد X به ازای مجموعه پارامترها (احتمال درستی θ آن به شرط (X))

$$- l(\theta | X) \equiv p(X | \theta)$$

• X ثابت است و θ را تغییر داده می‌شود.

• این تابع در «استنباط آماری» نقش اساسی دارد.



Statistical inference

دانشکده
سینمایی
بهره‌بری

برآورد درستنمایی بیشینه

Maximum Likelihood Estimation

Make sampling x^t from $p(x^t|\theta)$ as likely as possible

- در صورتی که نمونه‌ها، $\mathcal{X} = \{x^t\}$ ، «متغیرهای مستقل با توزیع یکسان (i.i.d.)» باشد:

independent and identically distributed

- $l(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = \prod_t p(x^t|\theta)$
- در برآورد درستنمایی بیشینه در پی یافتن θ هستیم به کونهای که احتمال تعلق X به p مدهاکثر شود؛ درستنمایی بیشینه شود.
- برای سادگی محاسبات، به جای درستنمایی، از لگاریتم آن استفاده می‌شود:

$$\mathcal{L}(\theta|\mathcal{X}) = \log l(\theta|\mathcal{X}) = \sum_t \log p(x^t|\theta)$$

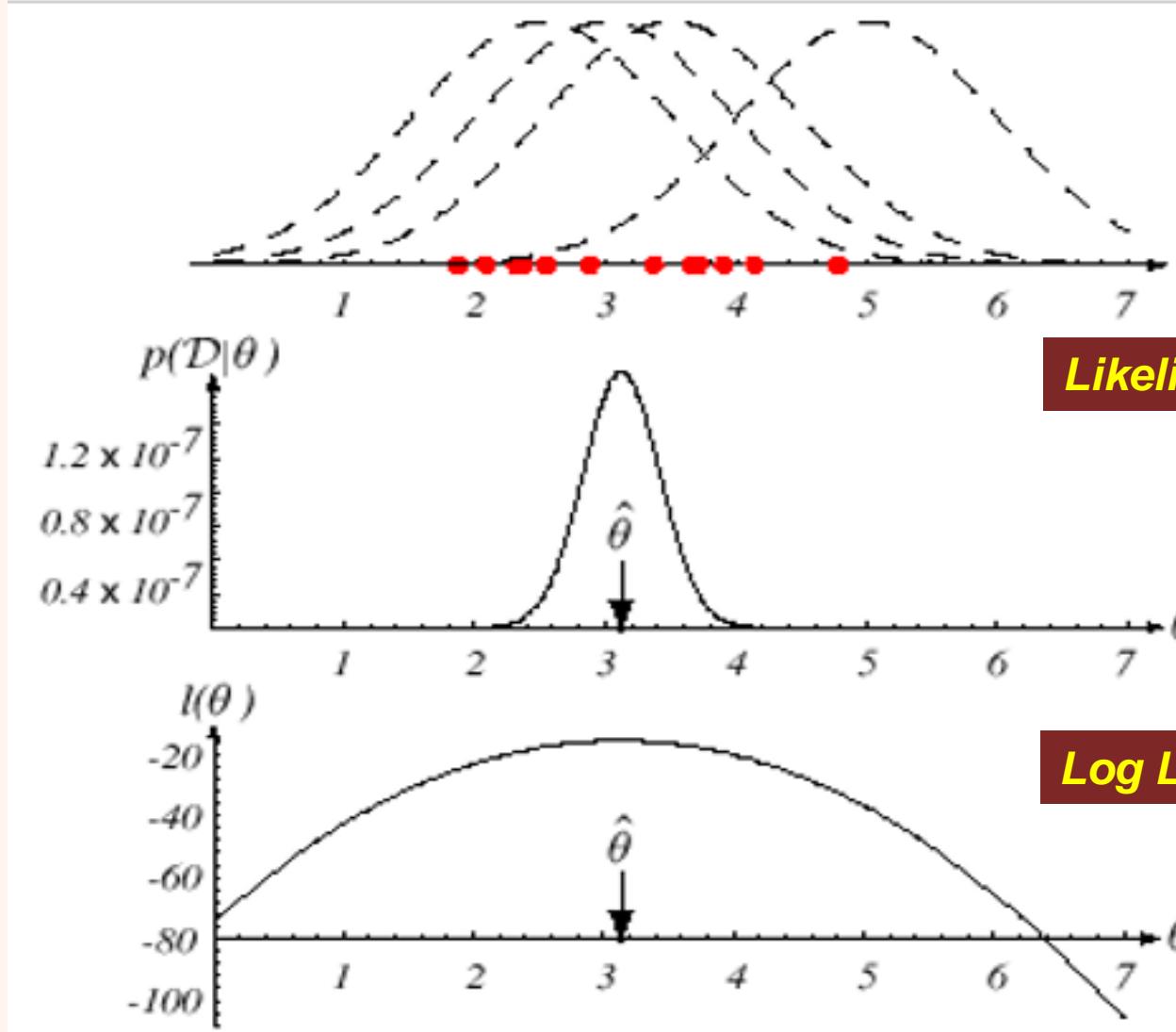
Log likelihood

$$\theta^* = \operatorname{argmax}_{\theta} \mathcal{L}(\theta|\mathcal{X})$$



دانشکده
سینمایی
بهره‌وری

برآورد درست نمایی پیشینه



Pattern Classification, Chapter 3

۵۷



دانشکده
سینمایی

Generative models v.s. Discriminative models



آشنایی با
softmax

دسته‌بندی بر پایه‌ی تابع درست‌نمایی

Likelihood-based classification (Generative models)

- برای دسته‌بندی یک سری «تابع جداساز ($g_i(x)$)» مماسبه می‌شود:

$$g_i(x) = \max_{j=1}^K g_j(x)$$

اگر

کلاس C_i انتخاب می‌شود

- ابتدا احتمال پیشین ($P(C_i)$) و تابع پگالی احتمال داده‌ها در هر کلاس ($P(x|C_i)$) بر اساس تابع درست‌نمایی مماسبه شده، سپس بر اساس قانون Bayes، احتمال پسین ($P(C_i|x)$) به دست آمده و بر اساس آن تابع جداساز تعریف می‌شود:

$$g_i(x) = \log P(C_i|x)$$

- این شیوه به «دسته‌بندی بر پایه‌ی درست‌نمایی» موسوم است. در واقع بر اساس مدلی که برای داده‌ها تخمین زده می‌شود، جداساز به دست می‌آید.

- در روش‌های پارامتری، نیمه‌پارامتری و ناپارامتری از این شیوه استفاده شد.



دانشگاه
سینه‌پیشی

دسته‌بندی بر پایه‌ی جداساز

discriminant-based classification (discriminative models)

- در این (وش)، بدون تفمین توزیع داده‌ها، به صورت مسقیم جداساز تفمین زده می‌شود.
 - در این حالت یک مدل برای جداساز تعریف می‌شود:

$$g_i(\mathbf{x}|\Phi_i)$$

- پارامترهای جداساز به صورت «صریح» مشخص شده‌اند برای خلاف (وش‌های مبتنی بر درستنمایی که به صورت «ضمنی» و براساس توزیع داده‌ها به دست می‌آیند.



دسته‌بندی بر پایه‌ی جداساز(ادامه...)

- فرآیند آموزش یافتن(بهینهسازی) پارامترهای جداساز بر اساس یک مجموعه‌ی آموزشی و با هدف افزایش درستی دسته‌بندی است. در این حالت به جای تفمین درست توزیع داده‌ها هر کلاس، هدف تفمین درست مرزهای بین دسته‌هاست.
- از نظر طرفداران این (ویکرد برآورد توزیع داده‌های یک کلاس از تفمین جداساز، مساله‌ی دشوارتری است.
 - برای یک حل مسئله، محققول نیست آن را به مسائل دشوارتر تقسیم کردا
 - البته زمانی این گفته درست است که جداساز با یک مدل ساده تفمین زده شود.



دانشکده
سینمایی

جداساز خطي

- ساده‌ترین جداسازی که می‌توان در نظر گرفت، «جداساز خطي» است:

$$g_i(\mathbf{x} | \mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} = \sum_{j=1}^d w_{ij} x_j + w_{i0}$$

- در بيشتر موارد استفاده از جداساز خطي ترجيع داده می‌شود، به دلائل زير:

- سادگي: داراي پيچيدگي از مرتبه $O(d)$
- تفسيرپذيری: به راحتی می‌توان بر اساس آن به استخراج دانش پرداخت؛ فروجي هم‌مجموع وزن‌دار و وودی است. وزن هر بعد اهمیت و علامت آن اثر آن خصیصه را نشان می‌دهد.
- در بسیاری موارد جداساز خطي بهینه است: توزیع داده‌های کلاس گاوی با ماتریس کواریانس یکسان



دانشگاه
سینمایی
بهشتی

- چنان‌که مدل خطي پیمیدگي لازم را نداشته باشد، می‌توان سراغ جداسازهاي پیمیده‌تری (فت):

$$g_i(\mathbf{x} | \mathbf{W}_i, \mathbf{w}_i, w_{i0}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

$$g_i(\mathbf{x}) = w_0 + \sum_{i=1}^k w_i x_i + \sum_{i=1}^k \sum_{j=1}^k w_{ij} x_i x_j$$

جداساز درجهی دو

- دارای پیمیدگی از مرتبه‌ی $O(d^2)$
- نیاز به داده‌های آموختنی بیشتر
- احتمال بروز overfitting بیشتر

- یک راه معادل استفاده از جملات با مرتبه بالاتر (higher order terms) است.

$$z_1 = x_1, z_2 = x_2, z_3 = x_1^2, z_4 = x_2^2, z_5 = x_1 x_2$$

- به جای جداساز پیمیده نگاشت غير خطي به فضای با جداساز خطي

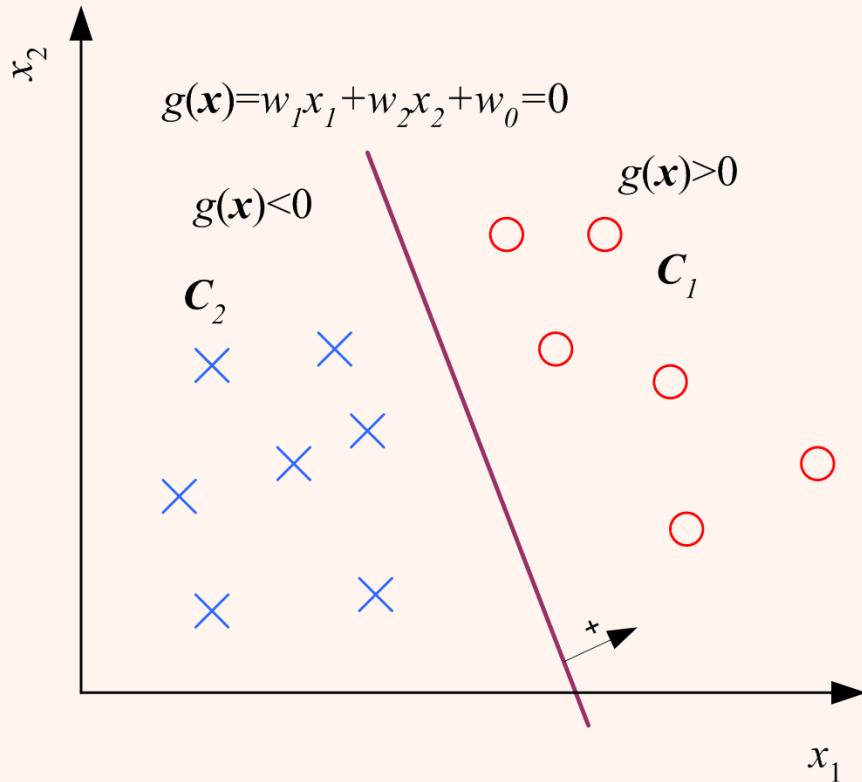
$$g_i(\mathbf{x}) = \sum_{j=1} w_{ij} \phi_j(\mathbf{x})$$

Potential functions(1964)



دانشکده
سینمایی
بهره‌بری

- در این حالت یک تابع مدارساز (ابزار سطح جداکننده) کافیست:



choose $\begin{cases} C_1 & \text{if } g(\mathbf{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}$



دانشکده
سینمایی

- در این حالت به k تابع جداساز نیاز است:

$$g_i(\mathbf{x} | \mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

در نظر گرفتن پنین جداسازی به معنای این است که همهی دسته‌ها جدایی‌پذیر فقط در نظر گرفته شوند.

$$g_i(\mathbf{x} | \mathbf{w}_i, w_{i0}) = \begin{cases} > 0 & \text{if } \mathbf{x} \in C_i \\ \leq 0 & \text{otherwise} \end{cases}$$

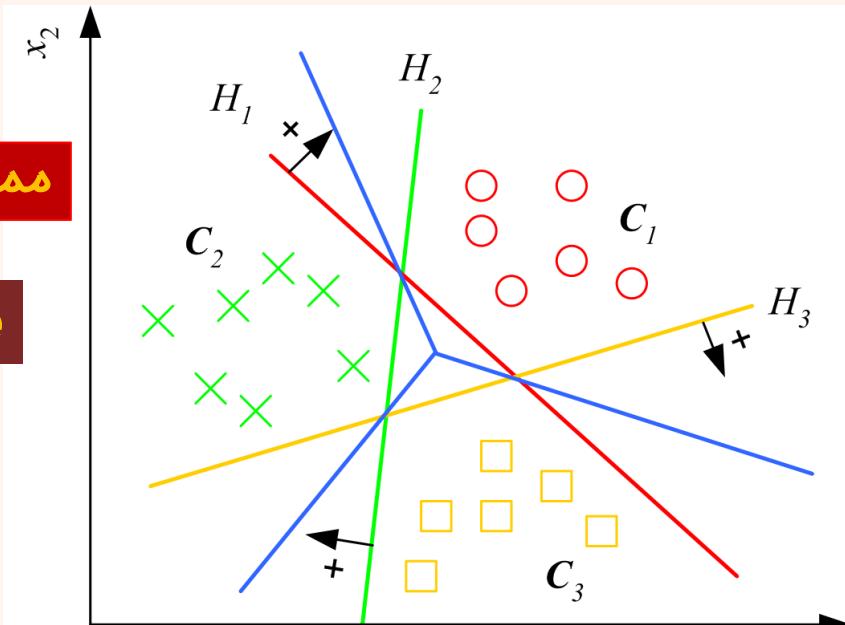
Linearly separable

ممکن است همهی دسته‌ها جدایی‌پذیر فقط نباشند:

Choose C_i if

$$g_i(\mathbf{x}) = \max_{j=1}^K g_j(\mathbf{x})$$

Linear machine



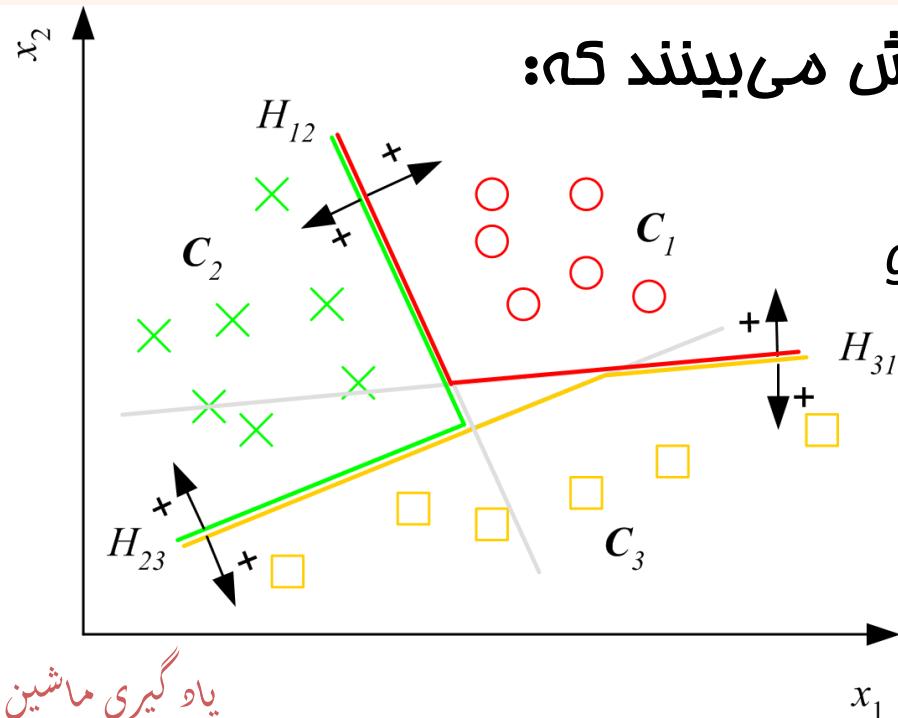
دانشکده
سینمایی

۴۵

در این حالت فضای k ناممی‌یار ممکن تفکیم می‌شود.

- اگر همهی کلاس‌ها جدایی‌پذیر خطي نباشند، يك (ويکرد مناسب تقسيم مسئله به چند جدایی‌ساز خطي است.
 - برای هر دو کلاس يك جداساز تعریف شود.
- در این صورت $k(k-1)/2$ جداساز مورد نیاز است.

$$g_{ij}(\mathbf{x} | \mathbf{w}_{ij}, w_{ij0}) = \mathbf{w}_{ij}^T \mathbf{x} + w_{ij0}$$



- پارامترها به گونه‌ای آموخته می‌بینند که:

$$g_{ij}(\mathbf{x}) = \begin{cases} > 0 & \text{if } \mathbf{x} \in C_i \\ \leq 0 & \text{if } \mathbf{x} \in C_j \\ \text{don't care} & \text{otherwise} \end{cases}$$

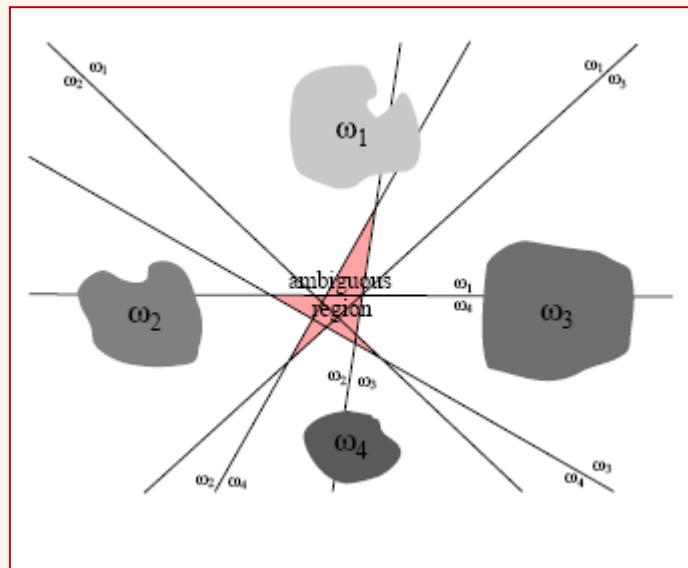


- در زمان تست:
choose C_i if
 $\forall j \neq i, g_{ij}(\mathbf{x}) > 0$



- در این حالت ممکن است برقی نواحی (ابطه‌ی پیش برقار) نباشد.
- در این حالت معیار دیگری را می‌توان جایگزین نمود:

$$g_i(x) = \sum_{j \neq i} g_{ij}(x)$$



Duda et. al.

این شیوه نیز نمونه‌ی دیگری از تبدیل یک مسئله پیچیده به چند مسئله ساده‌تر است

هزاری بر جداسازی پارامتری

- در صورتی که داده‌ها از توزیع گاوسی تبعیت کنند و ماتریس کواریانس یکسانی داشته باشند، جداساز بینه، فطی خواهد بود:

$$p(x | C_i) \sim N(\mu_i, \Sigma)$$

$$g_i(\mathbf{x} | \mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

$$\mathbf{w}_i = \Sigma^{-1} \mu_i, \quad w_{i0} = -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log P(C_i)$$

- برای یک مجموعه‌ی آموزشی ابتدا مقادیر میانگین و ماتریس کواریانس هر کلاس محاسبه می‌شود.
- برای هالت دو کلاسه خواهیم داشت:

$$y \equiv P(C_1 | \mathbf{x}) \text{ and } P(C_2 | \mathbf{x}) = 1 - y$$

choose C_1 if $\begin{cases} y > 0.5 \\ y / (1 - y) > 1 \quad \text{and } C_2 \text{ otherwise} \\ \log [y / (1 - y)] > 0 \end{cases}$

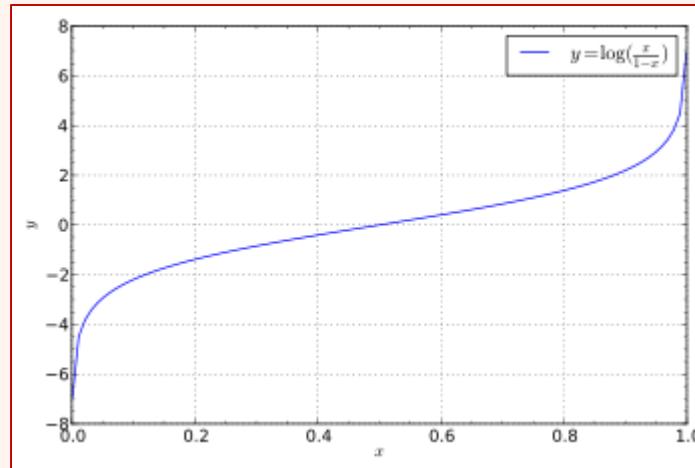


دانشکده
سینمایی

مروری بر جداسازی پارامتری (ادامه...)

• LOGIT به صورت زیر تعریف می‌شود:

$$\text{logit}(y) = \log \frac{y}{1-y}$$



در نتیجه داده به کلاس یک تعلق دارد اگر و تنها اگر

$$\text{logit}(P(C_1|x)) > 0$$

وگرنه متعلق به کلاس دو فواهد بود.



دانشکده
سینمای
بصیری

مودعی بر جداسازی پارامتری (ادامه...)

$$\begin{aligned}
 \text{logit}(P(C_1|\boldsymbol{x})) &= \log \frac{P(C_1|\boldsymbol{x})}{1-P(C_1|\boldsymbol{x})} = \log \frac{P(C_1|\boldsymbol{x})}{P(C_2|\boldsymbol{x})} \\
 &= \log \frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x}|C_2)} + \log \frac{P(C_1)}{P(C_2)} \\
 &= \log \frac{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left[-(1/2)(\boldsymbol{x}-\mu_1)^T \Sigma^{-1} (\boldsymbol{x}-\mu_1)\right]}{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left[-(1/2)(\boldsymbol{x}-\mu_2)^T \Sigma^{-1} (\boldsymbol{x}-\mu_2)\right]} + \log \frac{P(C_1)}{P(C_2)} \\
 &= \boldsymbol{w}^T \boldsymbol{x} + w_0
 \end{aligned}$$

where $\boldsymbol{w} = \Sigma^{-1}(\mu_1 - \mu_2)$ $w_0 = -\frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) + \log \frac{P(C_1)}{P(C_2)}$

The inverse of logit

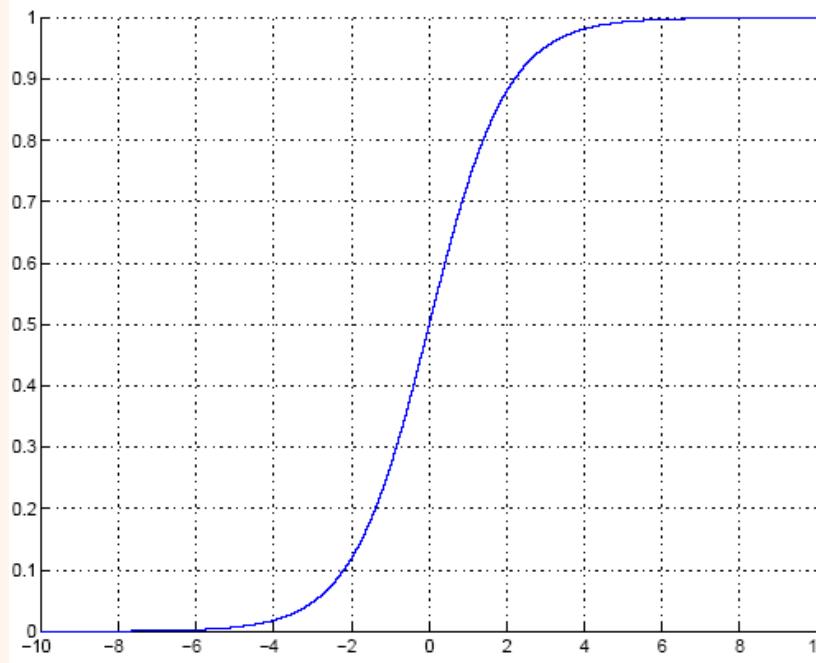
$$\log \frac{P(C_1|\boldsymbol{x})}{1-P(C_1|\boldsymbol{x})} = \boldsymbol{w}^T \boldsymbol{x} + w_0$$

$$P(C_1|\boldsymbol{x}) = \text{sigmoid}(\boldsymbol{w}^T \boldsymbol{x} + w_0) = \frac{1}{1 + \exp\left[-(\boldsymbol{w}^T \boldsymbol{x} + w_0)\right]}$$



دانشکده
سینمایی

تابع sigmoid



Calculate $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0$ and choose C_1 if $g(\mathbf{x}) > 0$, or

Calculate $y = \text{sigmoid}(\mathbf{w}^T \mathbf{x} + w_0)$ and choose C_1 if $y > 0.5$

در مالت دو^ه تابع sigmoid ارتباط مقدار تابع
جداساز و احتمال پسین را نشان می‌دهد.



دانشکده
سینمایی

Logistic Discrimination

مالت دوکلاس

Logistic Regression

- در دسته‌بندی مبتنی بر درست‌نمایی ابتدا، $P(C_1)$ و $P(x|C_1)$ مماسبه شده، سپس مقدار $P(x|C_1)$ به دست می‌آید.
- در logistic discrimination احتمال پسین به صورت مستقیم براورد می‌شود.
- در صورتی که لگاریتم نسبت درست‌نمایی دو کلاس خطی باشد:

$$\log \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} = \mathbf{w}^T \mathbf{x} + w_0^o$$

- با توجه به قانون Bayes :

$$\text{logit}(P(C_1|\mathbf{x})) = \log \frac{P(C_1|\mathbf{x})}{1 - P(C_1|\mathbf{x})} = \log \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} + \log \frac{P(C_1)}{P(C_2)}$$



دانشکده
سینمایی
بهشتی

حالت دوگلاس

$$\text{logit}(P(C_1|\boldsymbol{x})) = \log \frac{P(C_1|\boldsymbol{x})}{1 - P(C_1|\boldsymbol{x})} = \log \frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x}|C_2)} + \log \frac{P(C_1)}{P(C_2)}$$
$$= \boldsymbol{w}^T \boldsymbol{x} + w_0$$

where $w_0 = w_0^o + \log \frac{P(C_1)}{P(C_2)}$

$$y = \hat{P}(C_1|\boldsymbol{x}) = \frac{1}{1 + \exp[-(\boldsymbol{w}^T \boldsymbol{x} + w_0)]}$$

بدین ترتیب، تقریبی از احتمال پسین به دست می‌آید.

در واقع مسئله یافتن (آموفتن) \boldsymbol{w} و w_0 است.



دانشکده
سینمایی
بهشتی

آموزش (حالت دو کلاس)

$$y = \hat{P}(C_1 | \mathbf{x}) = \frac{1}{1 + \exp[-(\mathbf{w}^T \mathbf{x} + w_0)]}$$

$$\mathcal{X} = \left\{ \mathbf{x}^t, r^t \right\}_t \quad r^t | \mathcal{X}^t \sim \text{Bernoulli}(y^t)$$

در اینجا به صورت مسأله، مقدار y تفمین زده می‌شود.
درست‌نمایی \mathcal{X} به ازای پارامترهای مورد نظر محاسبه شده:

$$l(\mathbf{w}, w_0 | \mathcal{X}) = \prod_t \left(y^t \right)^{(r^t)} \left(1 - y^t \right)^{(1-r^t)}$$

بر اساس آن تابع خطای محاسبه می‌شود:

$$E = -\log l$$

$$E(\mathbf{w}, w_0 | \mathcal{X}) = -\sum_t r^t \log y^t + (1 - r^t) \log (1 - y^t)$$

هدف کاهش میزان خطای است این هیچ راه تمیلی برای حل این مسئله وجود ندارد.



- در دسته‌بندی مبتنى بر جداساز، پaramترها به گونه‌ای بهینه‌سازی مى‌شوند که فطاوی دسته‌بندی مداخلل شود:

$$w^* = \arg \min_w E(w | X)$$

- در بیشتر مواقع، اهمل تملیلی برای یافتن پaramترها وجود ندارد و پارهای جز استفاده از یک روش بهینه‌سازی تکرارشونده نفواید بود.
- استفاده از گرادیان نزولی یکی از پرکاربردترین راهکارهاست.
- ماتریس گردایان به صورت زیر تعریف می‌شود:

$$\nabla_w E = \left[\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \dots, \frac{\partial E}{\partial w_d} \right]^T$$



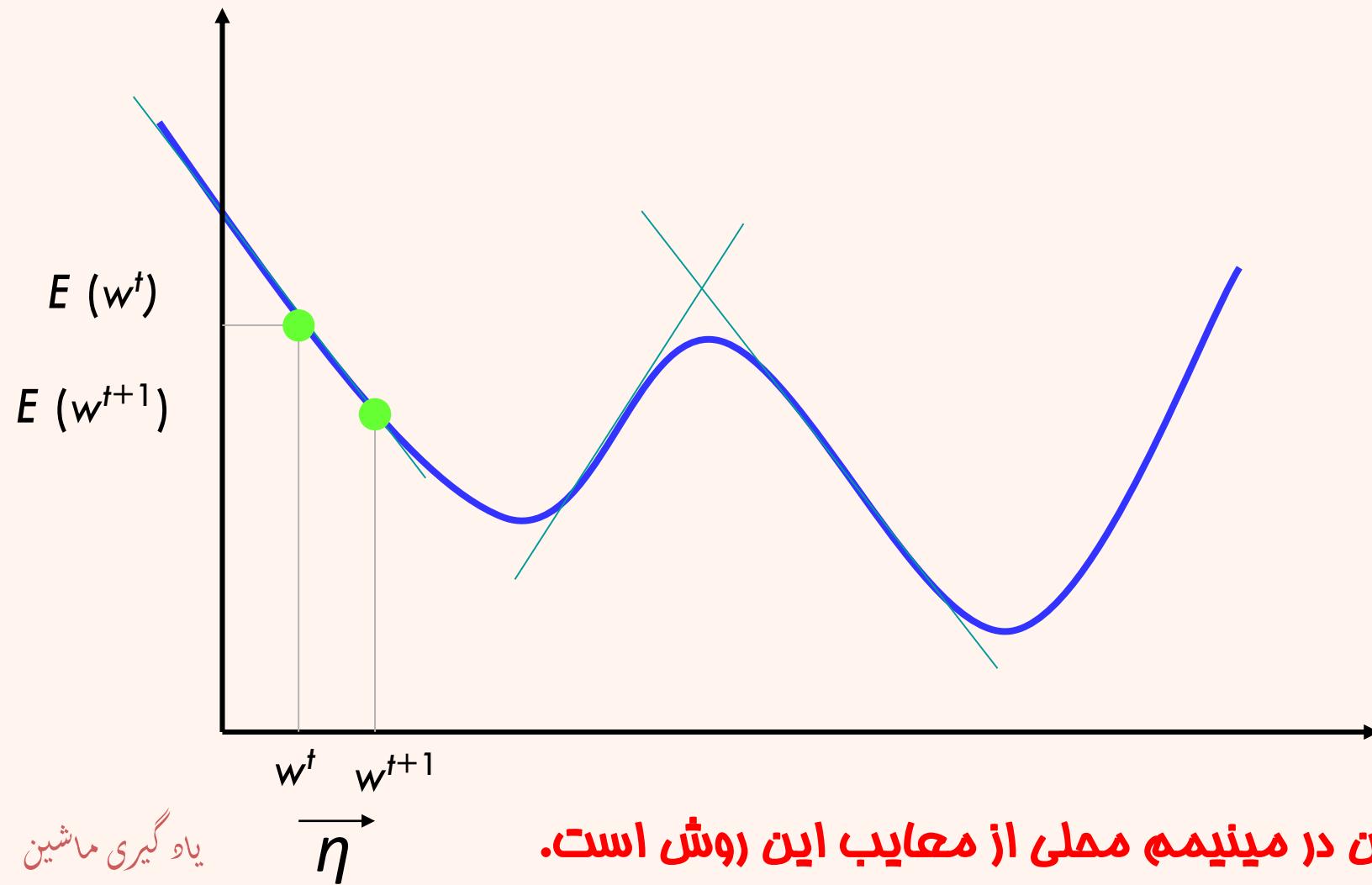
Gradient-Descent

$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i}, \forall i$$

$$w_i = w_i + \Delta w_i$$

پارامترها با یک مقدار تصادفی مقداردهی می‌شوند.

برخلاف جهت گرادیان مقدار پارامترها را به دو زمینه می‌شوند.

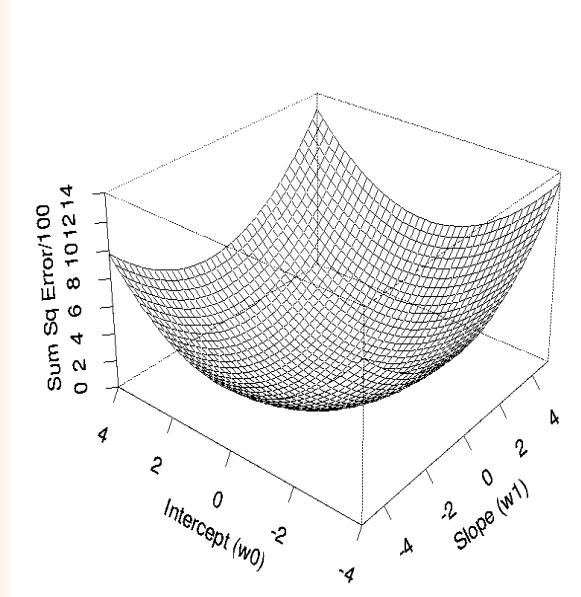
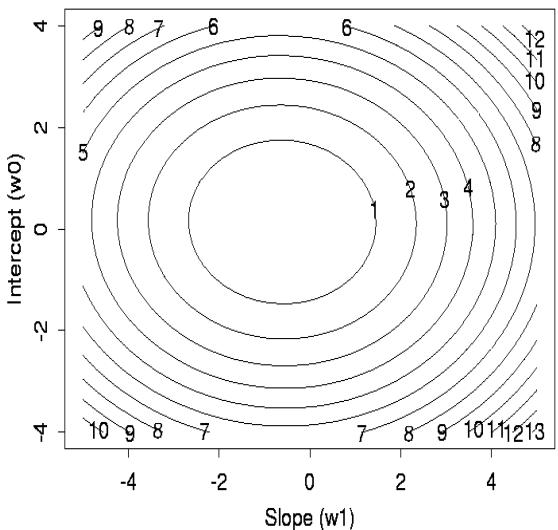


دانشکده
سینمایی

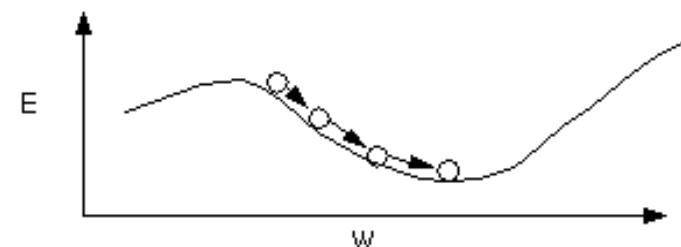
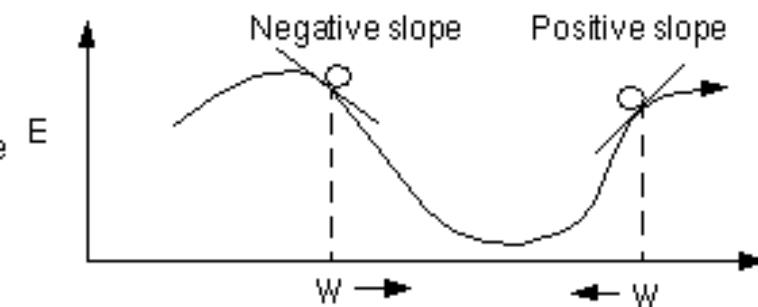
گیر کردن در مینیمم محلی از معایب این روش است.

کمینه کردن خط (ادا...)

rule



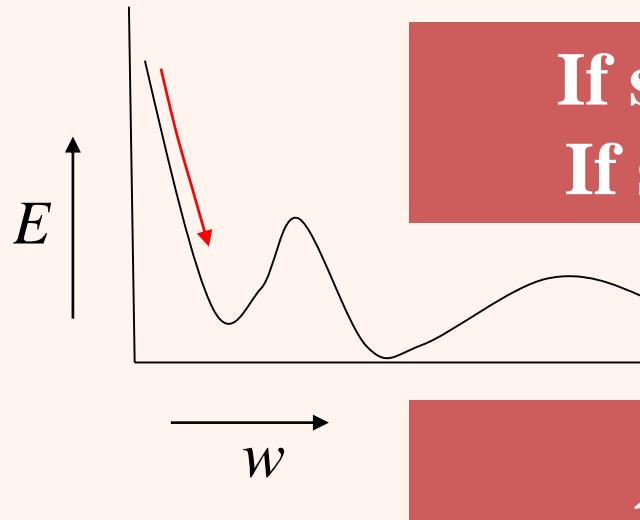
Slope of E positive
=> decrease W
Slope of E negative
=> increase W



دانشگاه
سینمایی

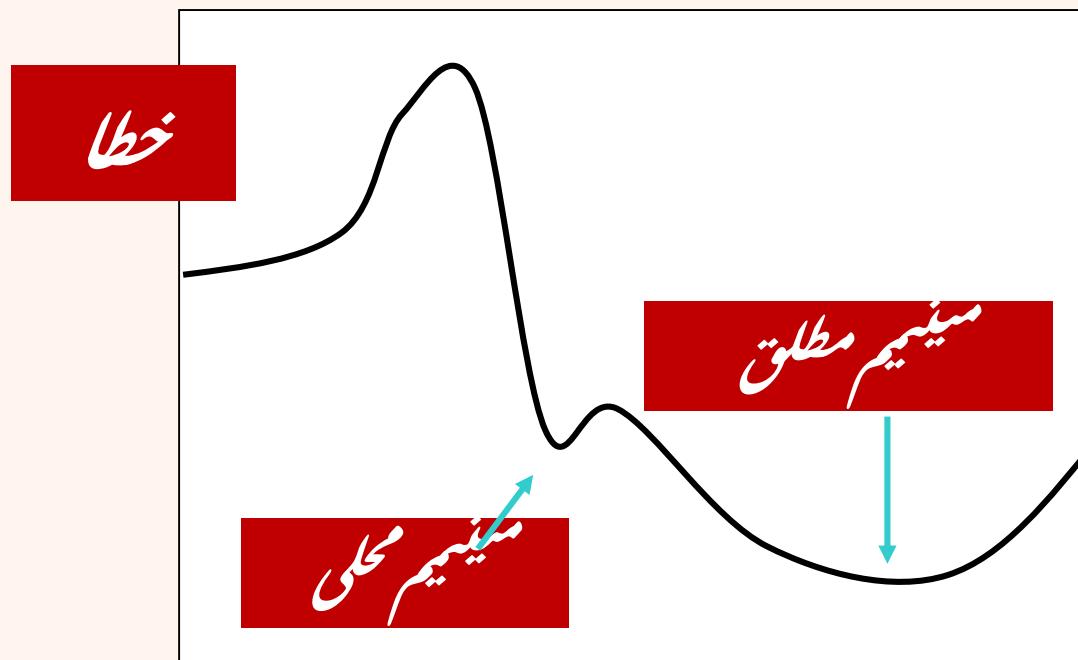
۷۷

Gradient Descent



If slope is negative \rightarrow increase w
If slope is positive \rightarrow decrease w

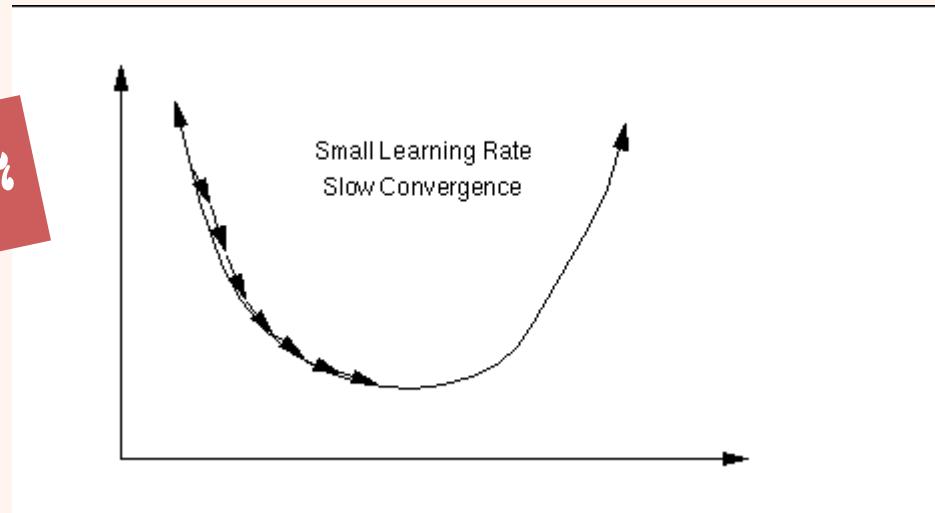
مینیمم محلی جایی است که مشتق صفر گردد



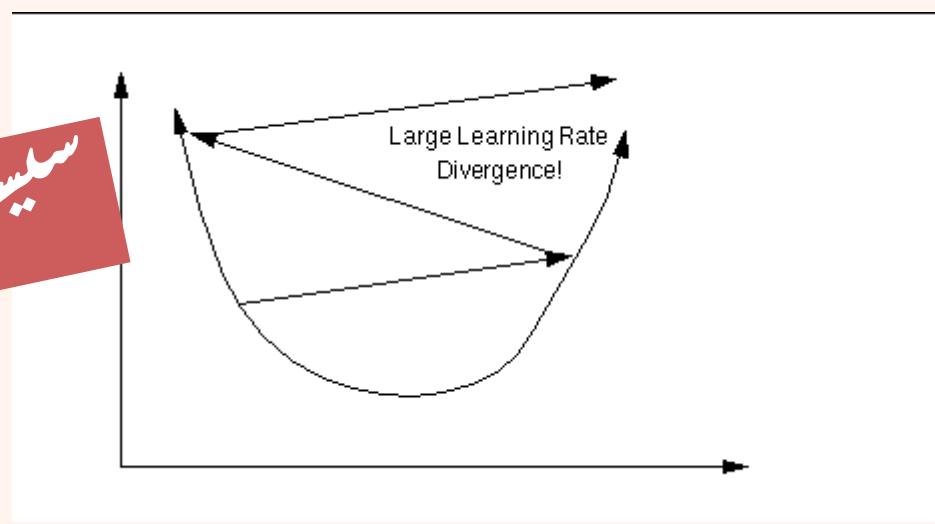
دانشکده
سینماسینما

تنظیم نرخ یادگیری

همگرایی کند است.



سیستم ناپایدار است.



$$\mathcal{X} = \left\{ \boldsymbol{x}^t, \boldsymbol{r}^t \right\}_t \quad r^t | \boldsymbol{x}^t \sim \text{Mult}_K(1, \boldsymbol{y}^t)$$

$$\log \frac{p(\boldsymbol{x}|C_i)}{p(\boldsymbol{x}|C_K)} = \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}^o$$

$$\frac{p(C_i|\boldsymbol{x})}{p(C_K|\boldsymbol{x})} = \exp \left[\boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}^o \right] \quad w_{i0} = w_{i0}^o + \log P(C_i)/P(C_k)$$

$$\sum_{i=1}^{K-1} \frac{p(C_i|\boldsymbol{x})}{p(C_K|\boldsymbol{x})} = \frac{1 - p(C_K|\boldsymbol{x})}{p(C_K|\boldsymbol{x})} = \sum_{i=1}^{K-1} \exp \left[\boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}^o \right]$$

$$p(C_K|\boldsymbol{x}) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp \left[\boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}^o \right]}$$



دانشکده
سینمایی

حالت پندهای

$$p(C_K|\boldsymbol{x}) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp[\boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}]} \quad \frac{p(C_i|\boldsymbol{x})}{p(C_K|\boldsymbol{x})} = \exp[\boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}]$$

$$p(C_i|\boldsymbol{x}) = \frac{\exp[\boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}]}{1 + \sum_{i=1}^{K-1} \exp[\boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}]}$$

برای این که با همهی کلاس‌ها یکسان بروزد شود:

$$y_i = \hat{P}(C_i|\boldsymbol{x}) = \frac{\exp[\boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}]}{\sum_{j=1}^K \exp[\boldsymbol{w}_j^T \boldsymbol{x} + w_{j0}]}, i = 1, \dots, K$$

softmax



دانشگاه
سینمایی

محاسبهی ماقریمه است، با این تفاوت که مشتق‌پذیر است.

آشنایی با چند شیوه‌های متفاوت یادگیری ماشینی



(انتخاب - استخراج) خصیصه

Feature Selection vs Extraction

- انتخاب خصیصه:

- K خصیصهای مهم‌تر ($k < d$) انتخاب می‌شود.
- الگوریتم‌های انتخاب زیرمجموعه

- استخراج خصیصه:

- K خصیصهای جدید، استخراج می‌شود.
- نگاشت از فضای n-بعدی به فضای k-بعدی
- (وش‌های استخراج خصیصه نیز از دیدگاه‌های مختلف قابل طبقه‌بندی هستند، (وش‌های خطي در برابر (وش‌های غيرخطي و یا (وش‌های بی‌نظارت در برابر (وش‌های با‌نظارت



دانشکده
سینمایی
بهشتی

کاربرد PCA در شناسایی چهره



پایگاه داده‌ی
ORL



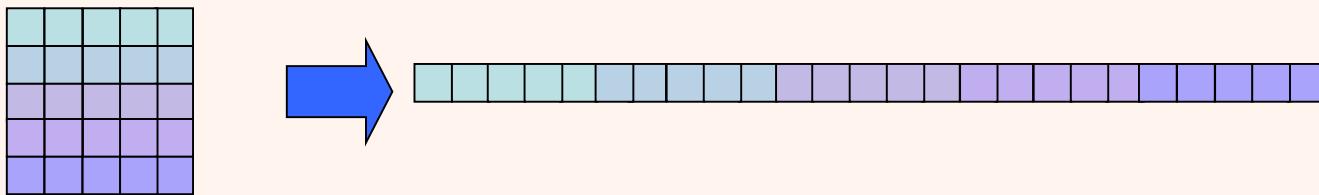
میانگین چهره‌ها



دانشکده
سینما و
بصیرتی

M. Turk, A. Pentland, "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.

کاربرد PCA در شناسایی چهره (ادامه...)



Eigenfaces



۴



۵



۱۵



۳۰



۲۹



۱



۱۵۰



۳۹



۳۸



۳۷



۳۶



۳۵



دانشکده
سینمایی
بهشتی

۸۵

کاربرد PCA در شناسایی چهره‌ها (...)

تصویر اصلی



تصویر میدانگین

=



اولین مؤلفه اساسی

+ (2.4



- 0.05

دومین مؤلفه اساسی



+ 1.9



سومین مؤلفه اساسی

+ ... - 0.012



) 10³

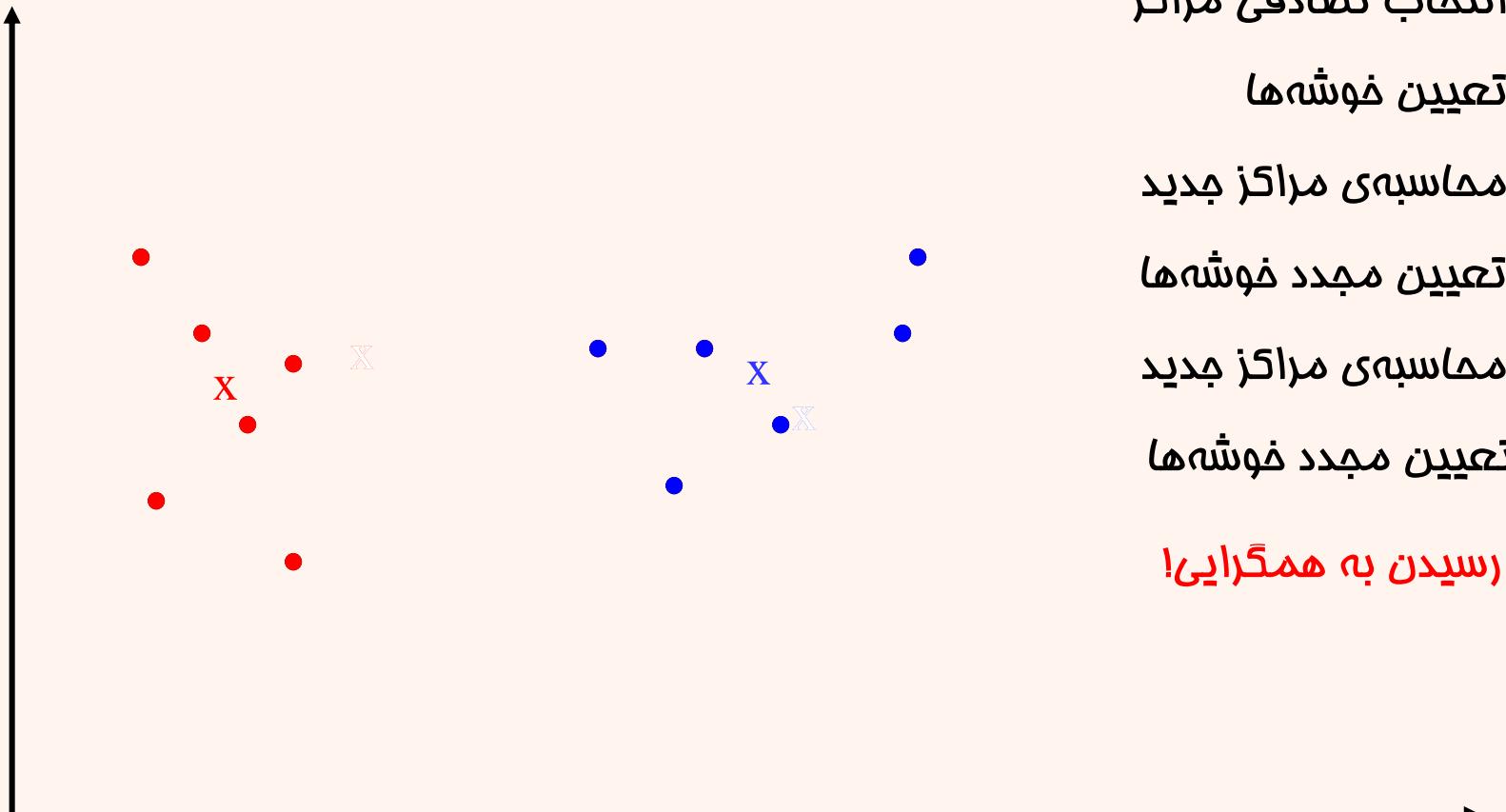
آفرين مؤلفه اساسی

دانشکده
سینمایی



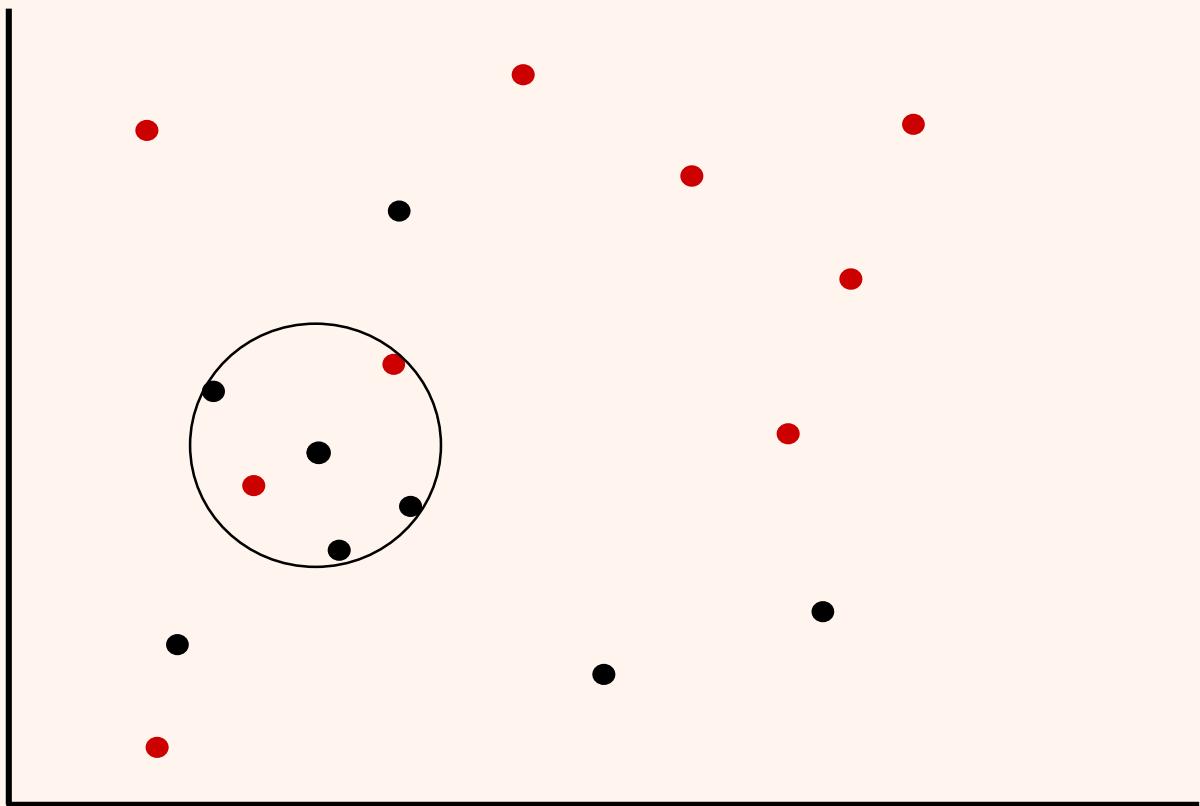
فوتشبندی

K=2



دانشکده
سینمایی

مثال

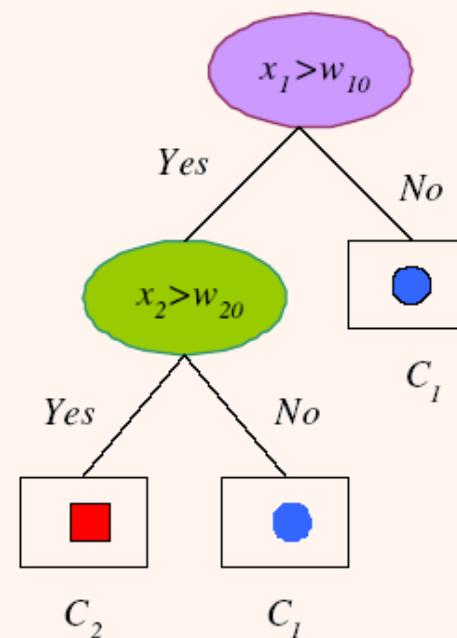
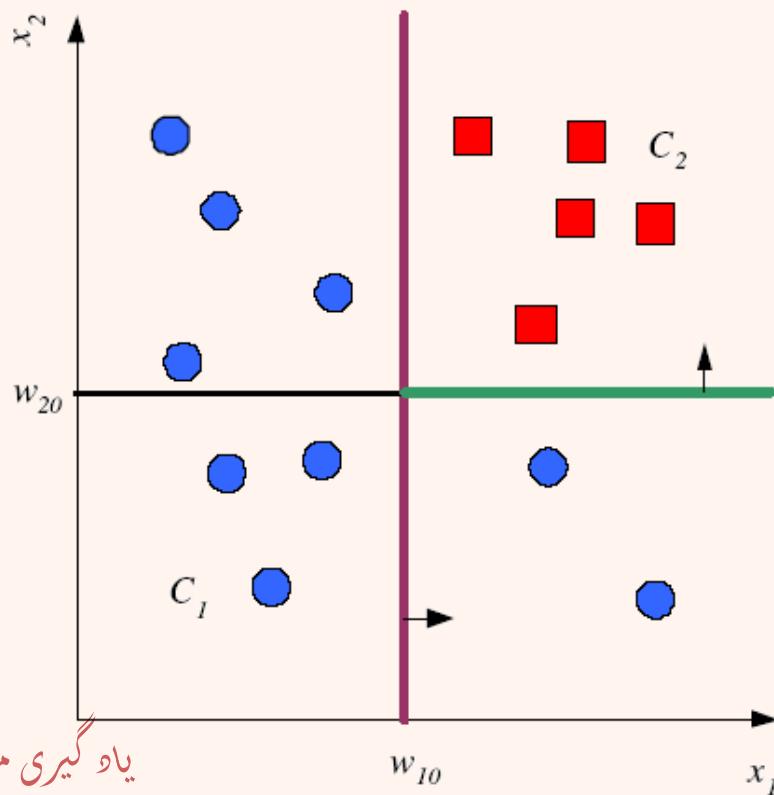


دانشکده
سینمایی

۸۸

درخت تصمیم

- هر «تابع آزمون ($f_m(x)$)» یک جداساز در فضای d -بعدی است، که فضای ۹۰۰ دی را به نواحی کوچک‌تر تقسیم می‌کند. هر f یک تابع ساده است که با ترکیب با یکدیگر توابعی پیچیده را فواهند ساخت.



ماشین بُردار پشتیبان

- نسخه‌ی اولیه‌ی SVM توسط آقای Vladimir Vapnik استادارد ارائه شد.
- با همکاری خانم Corinna Cortes Vapnik کنونی SVM را در سال ۱۹۹۳ پایه‌ریزی کرده و در سال ۱۹۹۵ منتشر نمودند.



Cortes, C. and V. Vapnik (1995). "Support-vector networks." Machine Learning 20(3): 273-297.

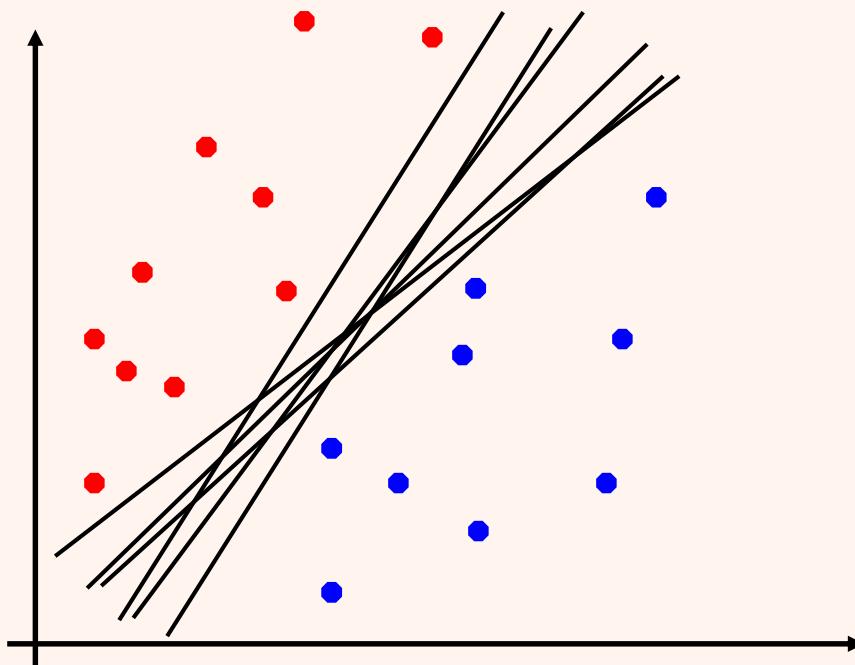


دانشگاه
بهشتی

مرز بهینه

• سوال

- گذاه یک از مرزها، مرزی بهینه برای جداسازی است؟

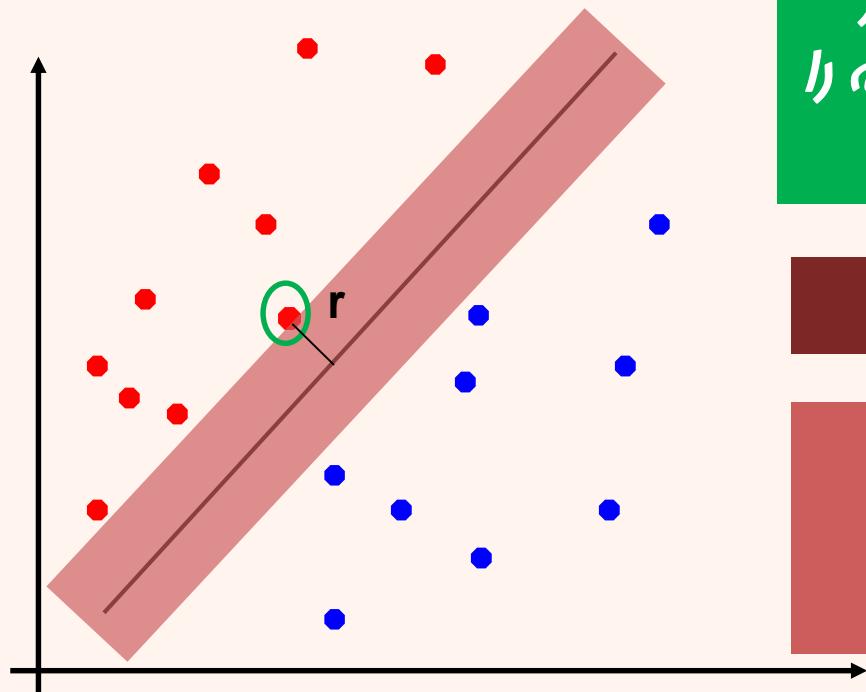


۹۱

مرز جداسازی

- هی فوایدیم به گونه‌ای بهترین مرز جداسازی را

Margin of separation



فرض کنیم نزدیک‌ترین نقطه به مرز جداسازی در نظر گرفته شده و فاصله را r بنامیم.

هدف ماکزیمم نمودن r است.

یک ماشین مشخص می‌کنیم هر مرزی که ماشینی پهن‌تری را تیجه دهد، بهتر است.

ماشینی ماکزیمم

- ماکزیمم نمودن ماشین (Margin) ایده‌ی خوبی است جهت چهارسازی خطا، این شیوه را **LSVM** یا **Linear SVM** می‌نامند.
- در این حالت نمونه‌هایی که به روی مرز ماشین هستند، از اهمیت ویژه‌ای برخوردارند.
- بدین‌وسیله می‌توان از نمونه‌های دیگر صرفنظر کرد و تنها به نمونه‌های مجهود روی مرز ماشین پرداخت.

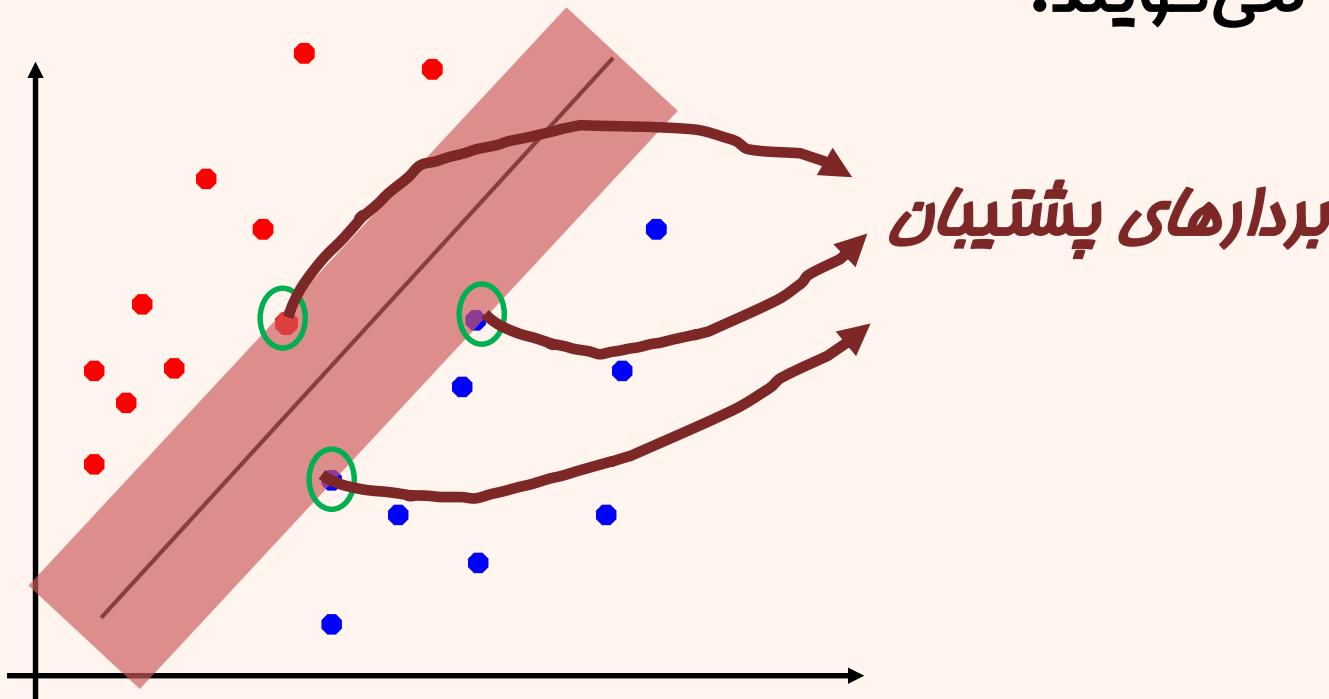


دانشکده
سینمایی
بهشتی

Support Vector

بردار پشتیبان

- به نمونه‌های (وی مرز حاشیه «بردار پشتیبان») می‌گویند.



Optimal hyperplane



روش‌های ترکیبی



بەھرین الگوریتم یادگیری

- یک (ا)ه، انتخاب بەھرین الگوریتم یادگیری بە پایهی یک «مجموعه‌ی اعتبار» است.

- هر الگوریتم یادگیری محدود به مدل خاصی است: در واقع «پیش‌فرضی» خاصی دارد که در صورتی که مفروضات برای داده‌ها معتبر نباشد، موجب ایجاد خطا می‌شود.

- یادگیری یک مسئله‌ی ill-posed مجموعه داده‌های محدود الگوریتم‌های مختلف به پاسخ‌های متفاوتی می‌رسند.

- در صورت تنظیم پارامترهای مدل برای یک مجموعه‌ی آموزشی، (fine-tuned) ممکن است مدل برای برخی داده‌ها مناسب نباشد و مدل دیگری برای آن‌ها پاسخ بەھری داشته باشد.

- با ترکیب چند الگوریتم یادگیری می‌توان کارایی بەھری بە دست آورد.



دانشکده
سینمایی
بهشتی

ترکیب الگوریتم‌های یادگیری (ادامه...)

Base-learner

- بدین ترتیب چندین «یادگیرنده‌ی پایه» برای به دست آوردن کارایی بهتر با هم ترکیب می‌شوند.
 - چگونه یادگیرنده‌های پایه‌ای انتخاب شوند که عملکرد یکدیگر را پوشت دهند؟
 - چگونه خروجی یادگیرنده‌های مختلف برای به دست آوردن بهترین نتیجه با هم ترکیب شوند؟
- ترکیب یادگیرنده‌های یکسان سودی ندارد، گاهی ممکن است این کار تزها منجر به افزایش هزینه‌ی محاسباتی شود.



دانشکده
سینما و
بصیرتی

انتخاب یادگیرنده

- استفاده از الگوریتم‌های یادگیری متفاوت:
 - هر الگوریتم یادگیری مفروضات خاصی دارد. با انتخاب تنها یک الگوریتم (وی فرضیات آن مسئله تأکید شده و سایر موارد محفول می‌مانند. به عنوان مثال ترکیب (وش‌های پارامتری و ناپارامتری
- استفاده از hyperparameter-های متفاوت:
 - برای یک شیوه‌ی یادگیری پارامترهای متفاوتی در نظر گرفت که در کارایی نهایی مؤثرند: تعداد گره‌ها در لایه‌ی مخفی شبکه‌ی عصبی، k در در $n-n$ ، حدآستانه‌ی خطای در درخت تصمیم، تابع کرنل در ماشین‌های بردار پشتیبان، استفاده از ماتریس کواریانس یکسان/متفاوت در (وش‌های پارامتری، انتخاب حالت اولیه در (وش‌های تکرار شونده(نزول گرادیان)
 - با آموزش با hyperparameter های متفاوت و میانگین‌گیری ساده واریانس و در نتیجه خطای کلی کاهش می‌یابد.



دانشکده
سینماسازی
بهشتی

انتخاب یادگیرنده(داده‌ها...)

- استفاده از نمایش متفاوت داده‌ها برای هر یادگیرنده:
 - به جای ترکیب داده‌های مختلف، می‌توان بردارهای خصیصه‌ی کوچک‌تری برای یادگیرنده‌های متفاوت در نظر گرفت.
- به عنوان مثال در شناسایی صوت؛ استفاده‌ی جدایانه ویدئوی حرکت لب‌ها و سیگنال صوتی
- استفاده از کلمات کلیدی در گزار تصویر در «بازیابی تصویر»
- در صورتی که داده‌ها دارای نمایش واحدی باشند، باز هم می‌توان از چنین شیوه‌ای بهره گرفت، باعث می‌شود هر یادگیرنده (وی زیرفضای خاصی از داده تمرکز کند، مانند sensor fusion
- random subspace (attribute bagging)



انتخاب یادگیرنده(ادامه...)

• استفاده از مجموعه‌های آموزشی متفاوت:

- با توجه به تأثیر داده‌هایی که برای آموزش به کار می‌روند، (و)یکرد دیگر استفاده از مجموعه‌های آموزشی متفاوت است.
- یک راه انتخاب تصادفی داده‌های آموزشی برای هر یادگیرنده است. bagging
- راه دیگر استفاده از یک یادگیرنده برای داده‌های داده‌های است که در یادگیرنده‌ها)ی مورد استفاده منجر به ارائه پاسخ مطلوب نشده‌اند. boosting
- هر یادگیرنده به نواحی خاصی از داده‌ها به صورت محلی افتقاضی یابد. cascading
- می‌توان کار اصلی را به پند کار جزئی‌تر تقسیم نمود. mixture of experts

error correcting output code(ECOC)

یادگیرنده‌ها صرفاً به فاطر کارایی مورد استفاده قرار نمی‌گیرند، سادگی و تفاوت آن‌ها اهمیت دارد. (Diversity vs. Accuracy)



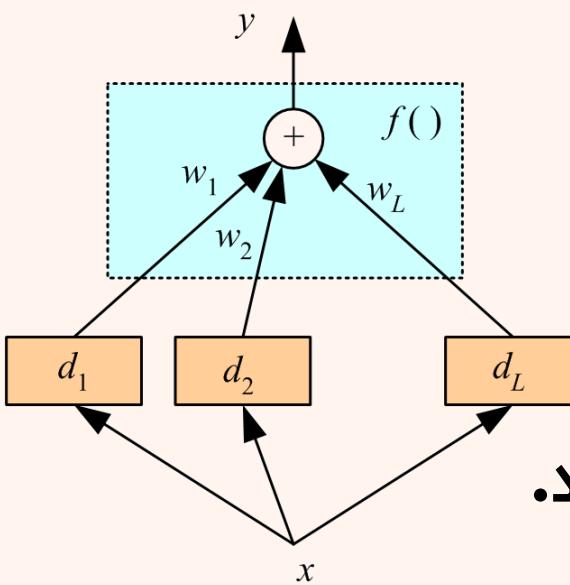
- ترکیب فطی خروجی یادگیرندها:

$$y = \sum_{j=1}^L w_j d_j$$

$$w_j \geq 0 \text{ and } \sum_{j=1}^L w_j = 1$$

ensembles -linear opinion pools

- برای دسته‌بندی هر یادگیرنده به هر کلاس یک رأی افتراض می‌دهد:



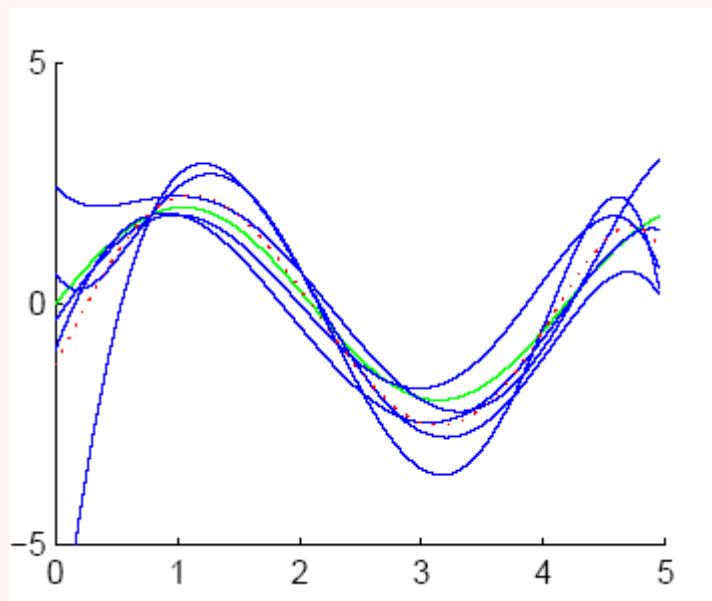
$$y_i = \sum_{j=1}^L w_j d_{ji}$$

- وزن هر یادگیرنده را می‌توان از روی صفت عملکرد آن تعیین کرد.



رأی‌گیری (ادامه...)

- می‌توان هر یادگیرنده را به صورت تابع درست به همراه نویز تصویر کرد. در نتیجه رأی‌گیری به نوعی فرآیند کاهش نویز فواهد بود.
 - استفاده از مدل‌هایی با بایاس کم و واریانس بالا



Bagging (Bootstrap aggregating)

- در این شیوه، یادگیرنده‌های یکسان از مجموعه‌های آموزشی متفاوت استفاده می‌کنند.
 - برای تولید مجموعه‌های آموزشی متفاوت از bootstrap استفاده می‌شود.
- از یک مجموعه‌ی آموزشی به طول N ، N داده به صورت **تصادفی** و همراه با جایگزینی استخراج می‌شود. فرآیند فوق ۱ بار تکرار می‌شود.
- از رأی‌گیری(میانگین‌گیری) برای ترکیب نتایج استفاده می‌شود.
- پایداری الگوریتم یادگیری را افزایش می‌دهد. (کاهش واپیانس)
 - یک الگوریتم «**نابایدار**» است، در صورتی که تغییر کوچکی در مجموعه‌ی آموزشی منجر به تغییر کلی یادگیرنده شود.



دانشکده
سینمایی
بهشتی

Boosting

- در این شیوه به جای انتخاب تصادفی داده‌های آموزشی، مجموعه‌ی آموزشی برای یک یادگیرنده بر اساس اشتباهات یادگیرنده‌ها)ی قبلی شکل می‌گیرد.
- در روش اولیه از ترکیب سه یادگیرنده‌ی ضعیف یک یادگیرنده‌ی قوی‌تر ایجاد می‌شود.
- یادگیری:
 - مجموعه‌ی آموزشی (\mathcal{X}) به سه بخش افزایش می‌شود ($\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$).
 - از \mathcal{X}_1 برای آموزش \mathcal{A}_1 استفاده می‌شود.
 - از تماه نمونه‌های اشتباه تشخیص داده شده \mathcal{X}_1 توسط \mathcal{A}_1 و \mathcal{X}_2 برای آموزش \mathcal{A}_2 استفاده می‌شود.
 - به همین ترتیب برای یادگیرنده‌ی سوم



دانشکده
سینمای
بهره‌بری

۱۰۴

Schapire, 1990

Boosting

- آزمون:

- در این مرحله نتیجه‌ی اعمال ۹۰٪ی به یادگیرنده‌ی اول و دوی برسی می‌شود. در صورت یکسان بودن نتیجه، تصمیم دو یا کیرنده‌ی اول پذیرفته می‌شود و گرنه رأی یادگیرنده‌ی سوم تعیین کننده خواهد بود.
- از معایب این روش نیاز به تعداد داده‌های آموختی بالاست.
- در روش از یک مجموعه‌ی آموختی بارها استفاده می‌شود.



دانشکده
سینما
بهریتی

AdaBoost.M1(Adaptive Boosting)

- در این روش تعداد یادگیرنده‌های پایه محدود نیست.
- در این شیوه احتمال انتخاب داده براساس خطای یادگیرنده‌های قبلی تعیین می‌شود.
- بعد از تکمیل آموخت، از رأی‌گیری ساده استفاده می‌گند.
- در صورتی که یادگیرنده‌ها پیچیده انتخاب شوند، در مراحل بعدی تنها نویز باقی خواهد ماند.
 - به عنوان مثال در حالتی که از درخت تصمیم استفاده می‌شود، برای عمق درخت محدودیت در نظر گرفته می‌شود.

Decision stumps

- نشان داده شده است که این شیوه منجر به افزایش حاشیه جداسازی می‌شود.



دانشکده
سینمایی
بهشتی

AdaBoost.M1(Adaptive Boosting)

Training:

For all $\{x^t, r^t\}_{t=1}^N \in \mathcal{X}$, initialize $p_1^t = 1/N$

For all base-learners $j = 1, \dots, L$

Randomly draw \mathcal{X}_j from \mathcal{X} with probabilities p_j^t

Train d_j using \mathcal{X}_j

For each (x^t, r^t) , calculate $y_j^t \leftarrow d_j(x^t)$

Calculate error rate: $\epsilon_j \leftarrow \sum_t p_j^t \cdot 1(y_j^t \neq r^t)$

If $\epsilon_j > 1/2$, then $L \leftarrow j - 1$; stop

$\beta_j \leftarrow \epsilon_j / (1 - \epsilon_j)$

For each (x^t, r^t) , decrease probabilities if correct:

If $y_j^t = r^t$ $p_{j+1}^t \leftarrow \beta_j p_j^t$ Else $p_{j+1}^t \leftarrow p_j^t$

Normalize probabilities:

$Z_j \leftarrow \sum_t p_{j+1}^t$; $p_{j+1}^t \leftarrow p_{j+1}^t / Z_j$

Testing:

Given x , calculate $d_j(x), j = 1, \dots, L$

Calculate class outputs, $i = 1, \dots, K$:

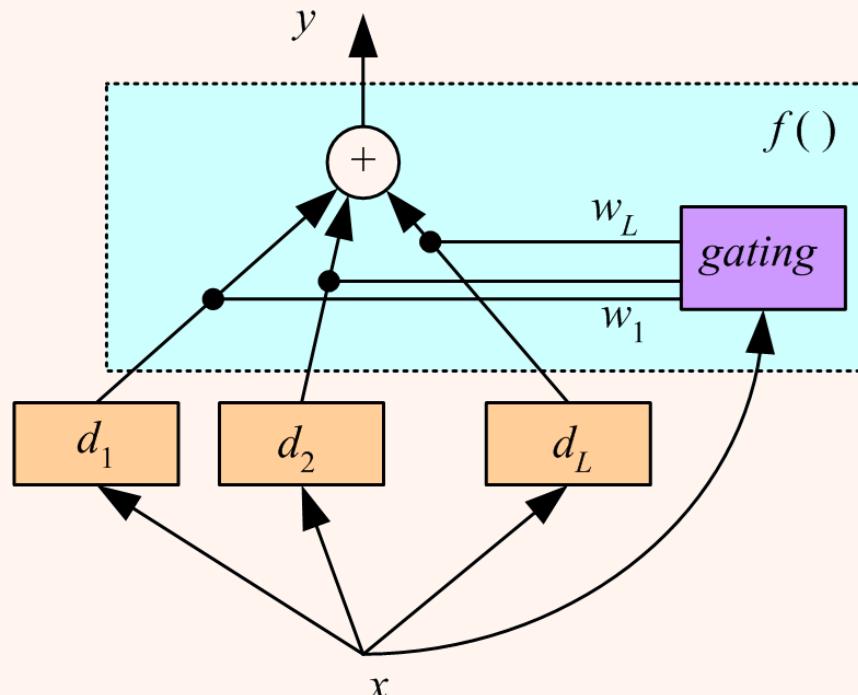
$$y_i = \sum_{j=1}^L \left(\log \frac{1}{\beta_j} \right) d_{ji}(x)$$



دانشکده
سینمایی

- مانند رأى گیری است با این تفاوت که وزن آرا به ۹۰٪ بستگی دارد.
- مانند الگوریتم‌های قابتی

$$y = \sum_{j=1}^L w_j(x) d_j$$



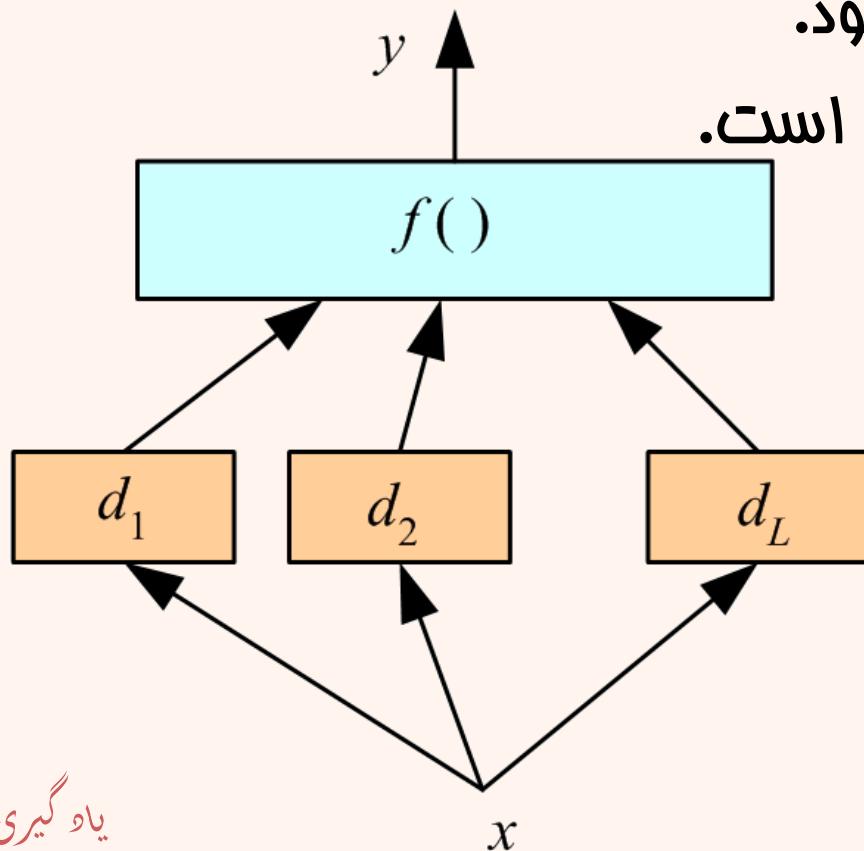
دانشکده
سینمایی
بهشتی

۱۰۸

(Jacobs et al., 1991)

Stacking

- در این شیوه، نمودهٔ ترکیب نتایج نیز توسط الگوریتم‌های یادگیری انجام می‌شود.
 - برای این منظور باید از مجموعه‌ای جدا از مجموعه‌ی آموزشی استفاده شود.
 - هدف کاهش باپاس است.



Cascading

