Machine Learning

یادگیری ماشین
(۱۳-۱۱-۸۰۵-۰۱)
فصل دوم

دانشگاه شهید بهشتی

دانشکده‌ی مهندسی برق و کامپیوتر

پاییز 1393

احمد محمودی ازناوه

فهرست مطالب

- یادگیری کلاس
- ظرفیت یادگیری
- تعداد نمونه‌های آموزشی مورد نیاز
- دسته‌بندی چندکلاسی
- مقدمه‌ای بر رگرسیون
  - رگرسیون خطی تک‌متغیره
- انتخاب مدل

یادگیری کلاس

***Class learning** is finding a description that is shared by all positive examples and none of the negative examples.*

- بحث را با داده‌هایی که از دو کلاس مجزا تشکیل شده‌اند، آغاز می‌کنیم.
  - مثال مورد دسته‌بندی خودروها به دو دسته‌ی «خودروهای خانوادگی» و «سایر خودروها» ست.
  - با این کار می‌توان

Prediction

  - کلاس یک نمونه‌ی نامشخص را پیش‌بینی کرد.
  - یا این که دریافت مردم چه خودرویی را خودروی خانوادگی می‌دانند.

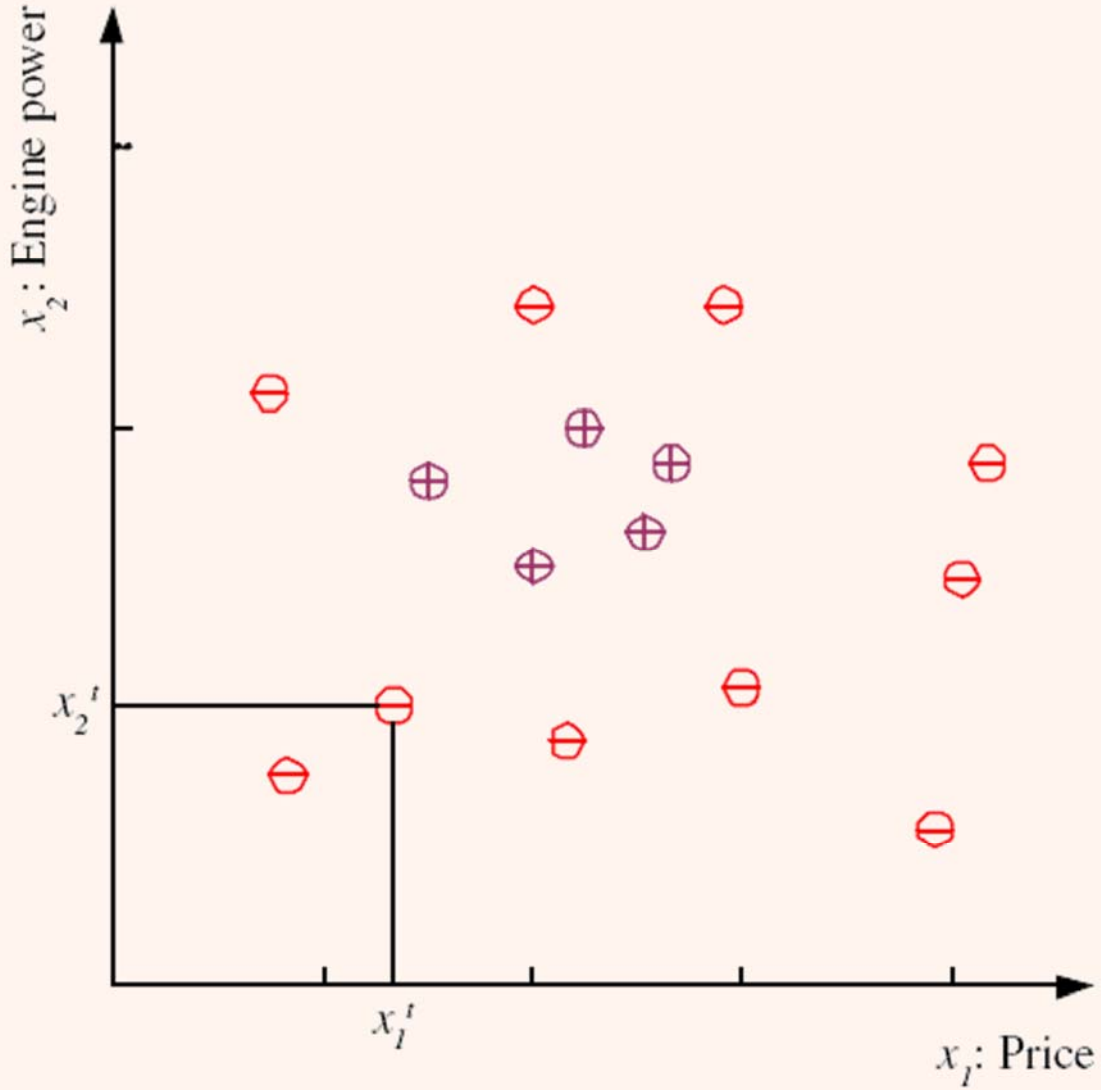
Knowledge extraction

  - نحوه‌ی نمایش ورودی‌ها:
    - x_1: price, x_2 : engine power

Training set

مجموعه‌ی آموزشی

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \mathcal{X} = \{\mathbf{x}^t, r^t\}_{t=1}^{N} \qquad r = \begin{cases} 1 \text{ if } \mathbf{x} \text{ is positive} \\ 0 \text{ if } \mathbf{x} \text{ is negative} \end{cases}$$

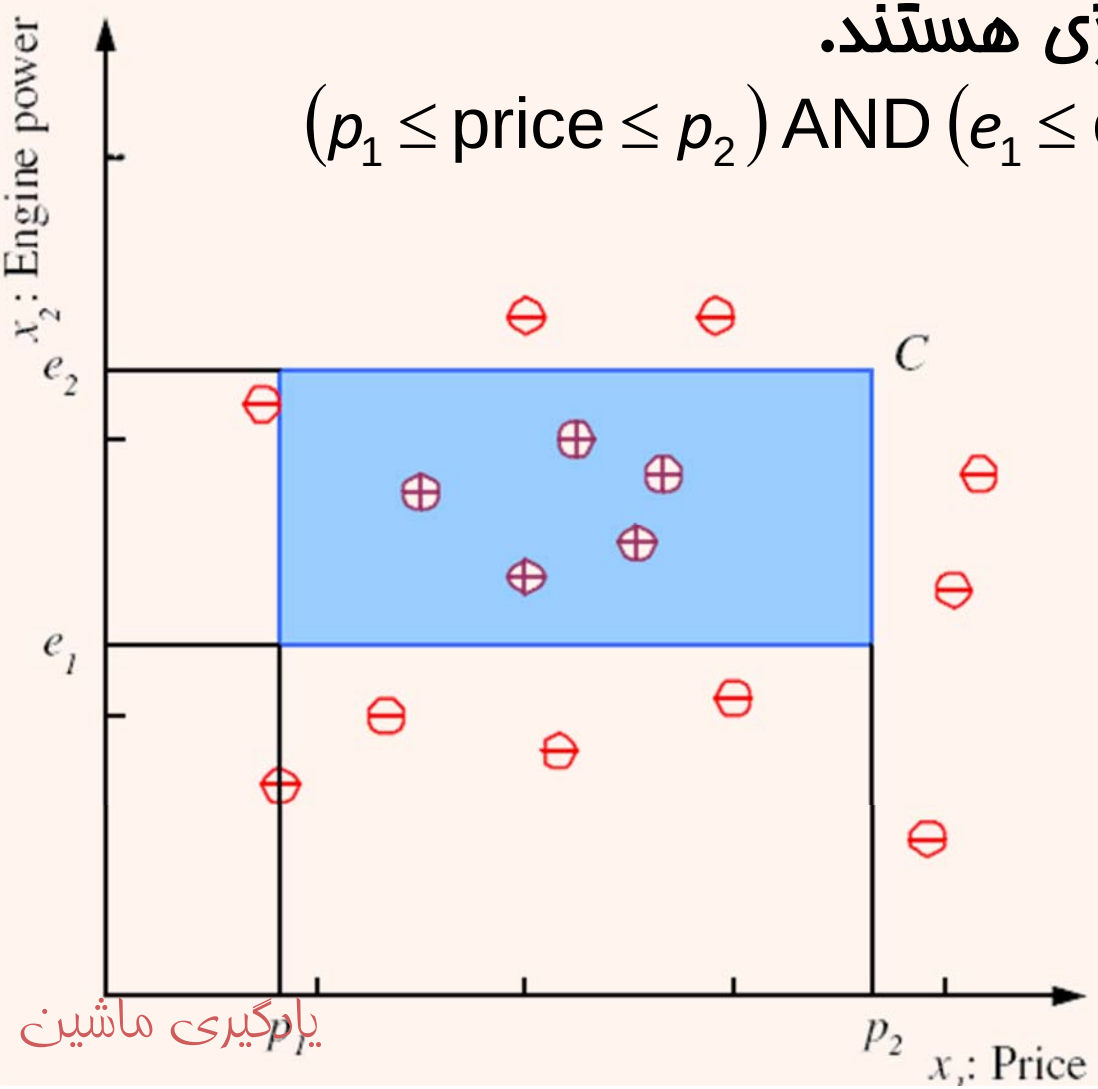


Hypothesis class $\mathcal{H}$

Hypothesis Class

- «کلاس فرضیه» مدلی است که برای دسته‌بندی مورد استفاده قرار می‌گیرد.
- در این مثال شامل همه‌ی مستطیل‌هایی است که با محورهای مختصات موازی هستند.

$$(p_1 \le \text{price} \le p_2) \text{ AND } (e_1 \le \text{engine power} \le e_2)$$



Hypothesis

$$h \in \mathcal{H}$$

- از بین کلاس فرضیه یک «فرضیه» با کم‌ترین خطا جستجو می‌شود.
- هر چهارتایی مرتب یک فرضیه را مشخص می‌کند.

فرضیه

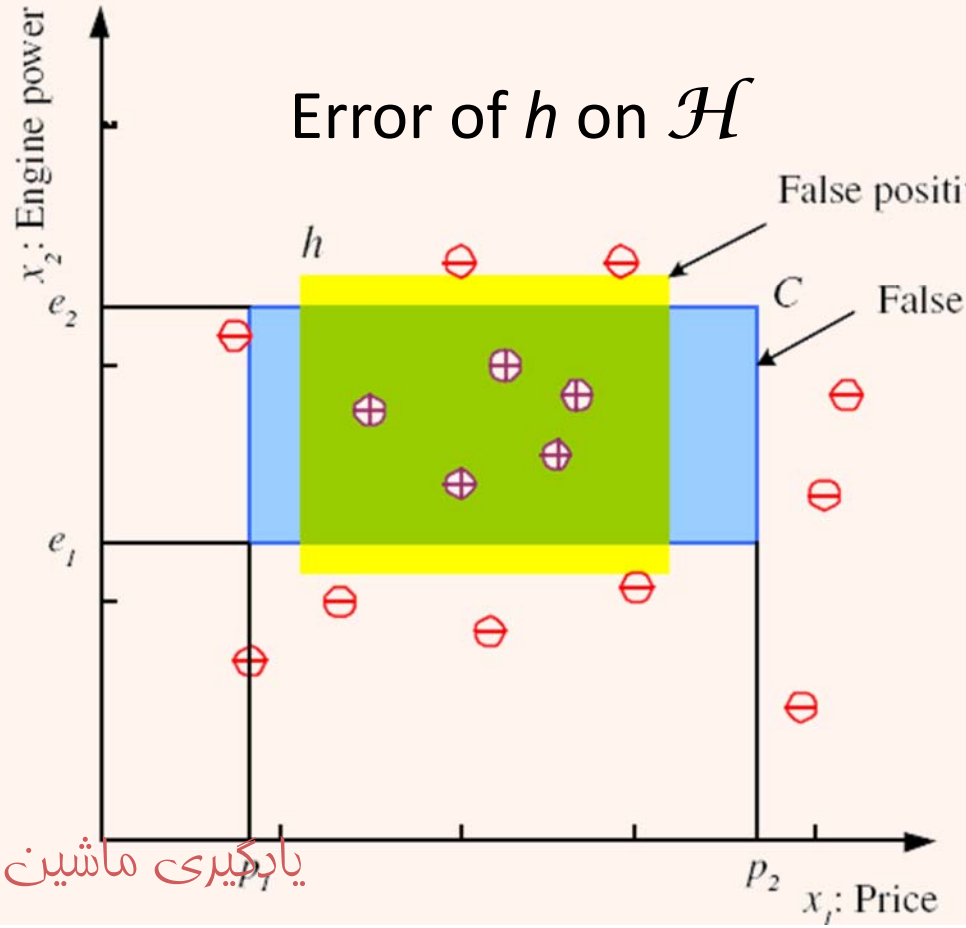
- هدف یافتن h به نحوی است که حداکثر شباهت به C را داشته باشد.

$$h(\mathbf{x}) = \begin{cases} 1 \text{ if } h \text{ says } \mathbf{x} \text{ is positive} \\ 0 \text{ if } h \text{ says } \mathbf{x} \text{ is negative} \end{cases}$$

Error of h on $\mathcal{H}$

$$E(h \mid \mathcal{X}) = \sum_{t=1}^{N} 1\left(h\left(\mathbf{x}^t\right) \neq r^t\right)$$

empirical error (training error)



- خطای آموزشی، میزان نمونه‌های آموزشی است که توسط h به درستی پیش‌بینی نمی‌شوند.

Generalization

تعمیم‌پذیری

- بی‌شمار فرضیه می‌توان یافت که دارای خطای صفر باشند.
  - یادگیری، را می‌توان جستجو برای یافتن بهترین پارامترها دانست.
- از بین تمامی فرضیه‌های درست، مناسب‌ترین فرضیه آن است که برای نمونه‌های جدیدی که در آینده دیده می‌شود، بهترین پاسخ را عرضه کند.

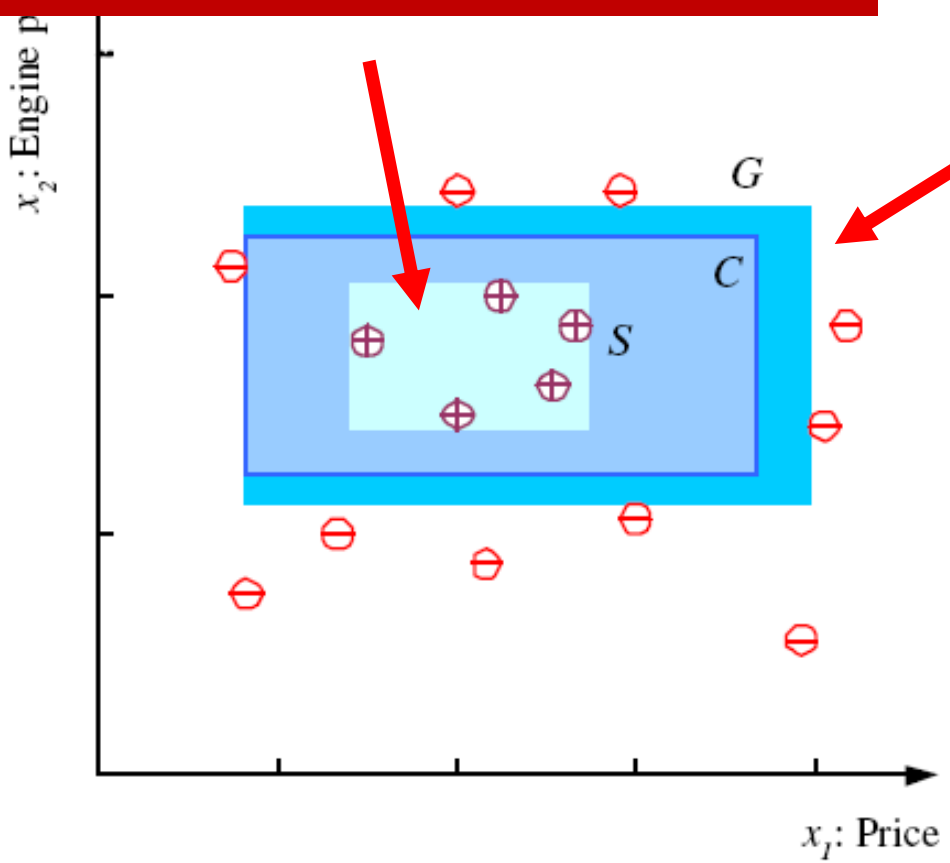
Version space

اختصاصی‌ترین فرضیه

most specific hypothesis, *S*

most general hypothesis, *G*

عمومی‌ترین فرضیه

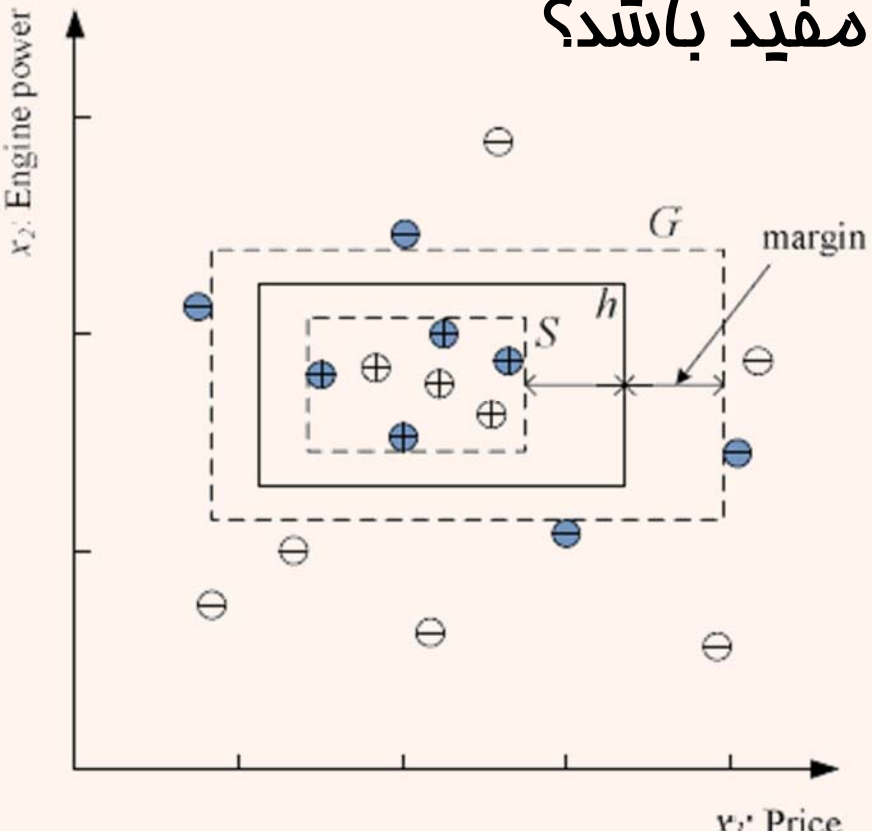


$h \in$ H, between *S* and *G* is consistent and make up the version space (Mitchell, 1997)

$$Consistent(h, D) \equiv (\forall \langle x, c(x) \rangle \in D)\ h(x) = c(x)$$

انتخاب بهترین فرضیه

- یک انتخاب مناسب برای فرضیه، در نظر گرفتن بیشترین حاشیه از اطراف است.
- بدین‌گونه، قابلیت تعمیم‌پذیری افزایش می‌یابد.
  - برای یافتن چنین فرضیه‌ای آیا تابع خطای مطرح شده می‌تواند مفید باشد؟



ظرفیت

- در برخی کاربردها هزینه‌ی اشتباه بسیار بالاست، از این رو در صورتی که نمونه در ناحیه‌ی حاشیه‌ای باشد، به جای دسته‌بندی برچسب «مشکوک» خواهد خورد و بررسی بیشتر آن به یک متخصص سپرده خواهد شد.

doubt

- تا این‌جا مطرح شد، هدف از یادگیری یافتن فرضیه‌ای با خطای صفر است.

$$E(h|\mathcal{X}) = 0$$

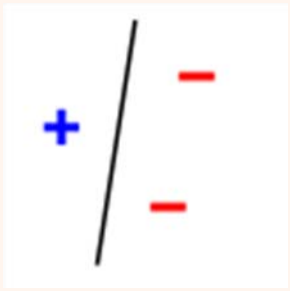
- اما همیشه، یافتن چنین فرضیه‌ای امکان‌پذیر نیست، در واقع برای هر کاربرد، باید مطمئن باشیم که کلاس فرضیه «ظرفیت کافی» برای آموختن C را دارد.

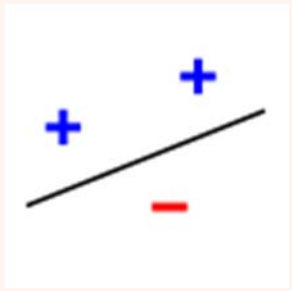
capacity

VC Dimension

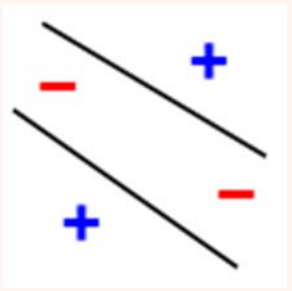
- در صورتی که یک پایگاه‌داده شامل N نمونه باشد، می‌توان آن را به 2^N شیوه‌ی متفاوت برچسب زد (دسته‌بندی دوکلاسه).
- در صورتی که بتوان برای همه‌ی این 2^N حالت، یک فرضیه $h \in \mathcal{H}$ یافت که کلاس‌ها قابل جداسازی باشند، گفته می‌شود که H، N نمونه(نقطه) را shatter می‌کند. به بیان دیگر، این فرضیه قابلیت آموختن N نمونه را بدون خطا دارد.
- به بیشترین مقدار N، **Vapnik-Chervonenkis (VC) Dimension** گفته می‌شود و به صورت $VC(\mathcal{H}) = N$ نمایش داده می‌شود.

VC Dimension

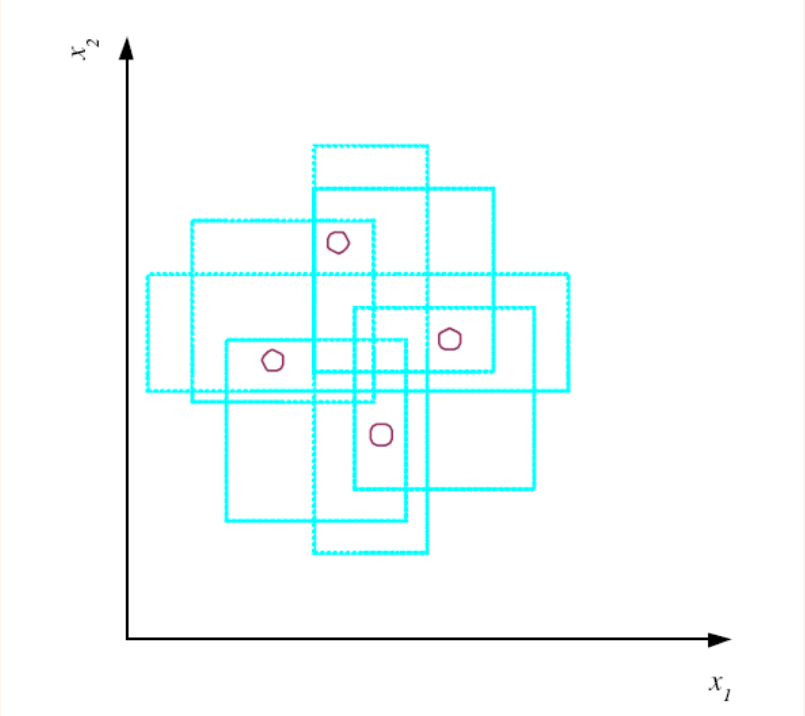




3 points shattered



4 points impossible



Probably Approximately Correct (PAC) Learning

- هر چه تعداد نمونه‌های آموزشی بیشتری در اختیار داشته باشیم، فرضیه‌ی به دست آمده، پاسخ دقیق‌تری خواهد داشت.
  - کلاس مفروض C که در آن نمونه‌ها از یک توزیع ثابت استخراج شده‌اند، **PAC-learnable** نامیده می‌شود چنانچه با احتمال $1 - \delta$ **بتوان فرضیه‌ای یافت که خطای آن کم‌تر از** ε باشد.
- $\delta \leq ½ , \varepsilon > 0$
  - با فرض این که اختصاصی‌ترین فرضیه در نظر گرفته شود، چه تعداد نمونه آموزشی (N) مورد نیاز است، تا با حداقل احتمال $1 - \delta$ میزان خطا حداکثر ε باشد؟

می‌خواهیم فرضیه‌ی مورد استفاده «بیشتر موارد»، «تقریبا درست» باشد

Probably Approximately Correct (PAC) Learning

Each strip is at most $\varepsilon/4$

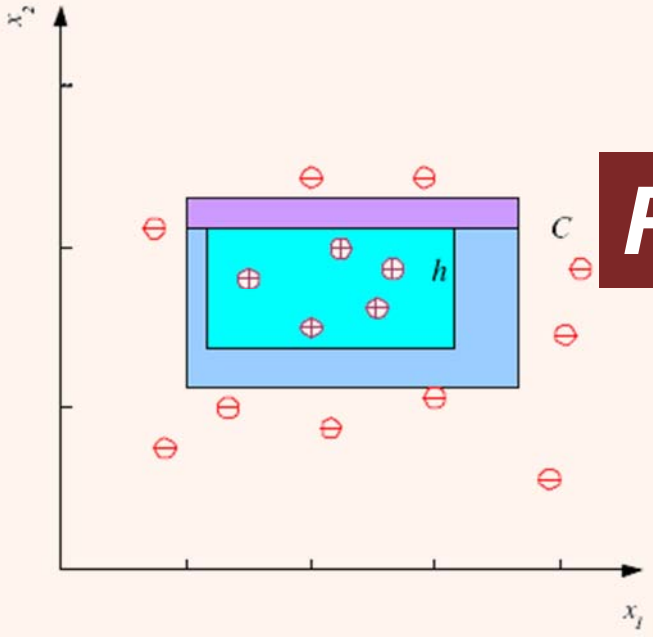
Pr that we miss a strip $1-\varepsilon/4$

Pr that N instances miss a strip $(1-\varepsilon/4)^N$

Pr that N instances miss 4 strips is at most $4(1-\varepsilon/4)^N$

$4(1-\varepsilon/4)^N \le \delta$ and $(1-x) \le \exp(-x)$

$4\exp(-\varepsilon N/4) \le \delta$ and $N \ge (4/\varepsilon)\log(4/\delta)$



$$P\{C\Delta h \le \varepsilon\} \ge 1-\delta$$

نویز

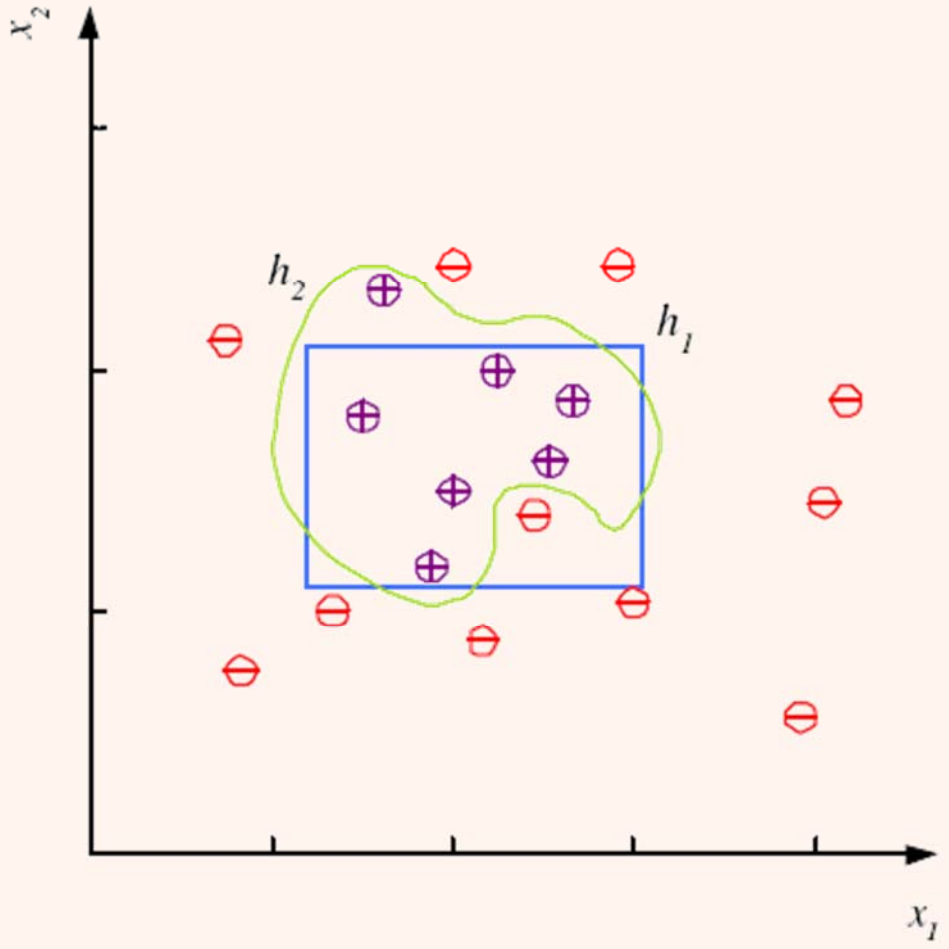
- «نویز»، ناهنجاری‌های ناخواسته در داده‌هاست.
- بر اثر نویز، دسته‌بندی دشوارتر خواهد بود و ممکن است دست‌یابی به خطای صفر امکان‌پذیر نباشد.
  - عدم دقت در وسایل اخذ داده
  - خطا در برچسب‌گذاری داده

Teacher noise

  - ممکن است برخی ویژگی در نظر گرفته نشده است و یا ویژگی‌هایی قابل مشاهده نبوده‌اند.

نویز و پیچیدگی مدل

- هنگامی که نویز وجود دارد، مدل به دست آمده **پیچیده‌تر** خواهد شد.



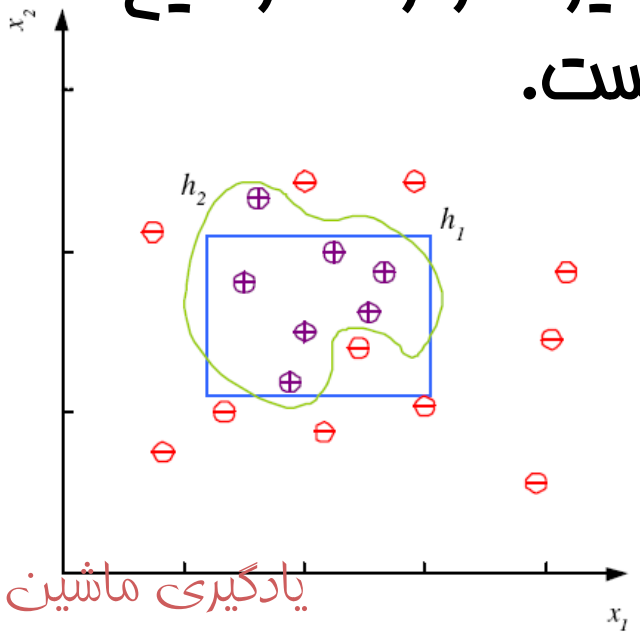
مزایای انتخاب مدل ساده‌تر

- استفاده از آن ساده‌تر است؛ پیچیدگی محاسباتی کم‌تری دارد.
- فرآیند آموزش آن، ساده‌تر است.
  - در صورت کم بودن داده‌های آموزشی، انتظار داریم با تغییر داده‌های آموزشی، مدل ساده تغییرات کم‌تری داشته باشد.
  - از سوی دیگر اگر مدل خیلی ساده باشد، با توجه به انعطاف کم آن، دارای bias بیشتر خواهد بود.
  - برای انتخاب مدل مناسب، باید هر دو این عامل‌ها را کمینه کنیم.
  - استخراج دانش از مدل ساده، به راحتی صورت می‌پذیرد.
- به ویژه در مواردی که با نویز مواجه هستیم، مدل‌های ساده‌تر کم‌تر از یک نمونه تاثیر می‌پذیرند، در این حالت هرچند دارای خطای بیشتری روی داده‌های آموزشی خواهند بود، ولی «تعمیم‌پذیری» بهتری خواهند داشت.

Less variance

تیغ Occam

- تیغ Occam اصلی منسوب به William of Ockham
- منطق‌دان و فیلسوف انگلیسی است. در قرن ۱۴ میلادی ویلیام اوکام اصلی را مطرح کرد که به نام اصل «تیغ Occam» شناخته شد. طبق این اصل، هر گاه درباره علت بروز پدیده‌ای دو توضیح مختلف ارائه شود، در آن توضیحی که **پیچیده‌تر** باشد احتمال بروز اشتباه بیشتر است و بنابراین، در شرایط مساوی بودن سایر موارد، توضیح **ساده‌تر**، احتمال صحیح بودنش بیشتر است.

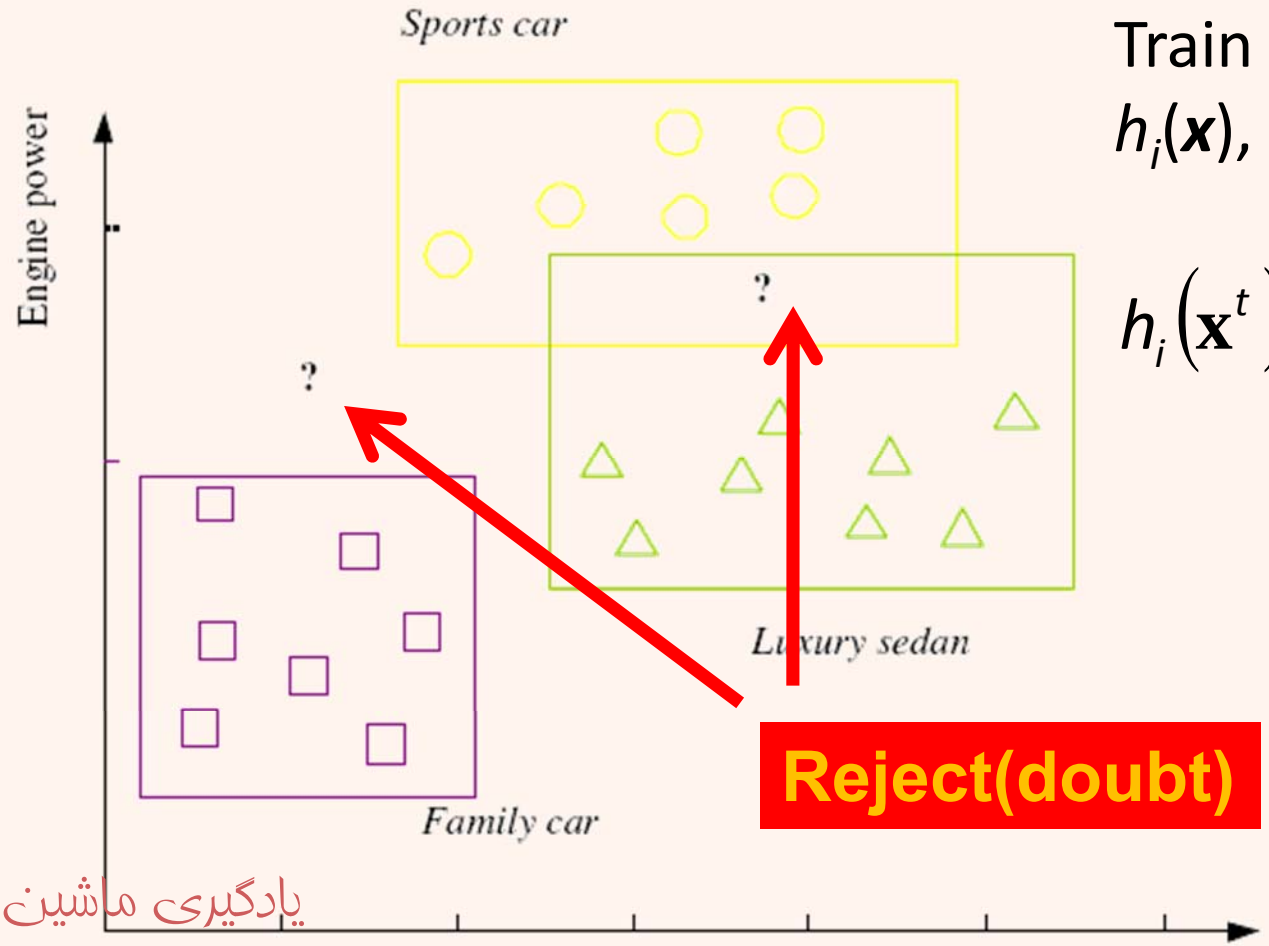


دسته‌بندی چندکلاسی(k)

می‌توان مساله را به صورت K دسته‌بندی دو کلاسه در نظر گرفت.

r، k بعدی است.

$$\mathcal{X}=\{\mathbf{x}^t,r^t\}_{t=1}^N \qquad r_i^t=\begin{cases}1 \text{ if } \mathbf{x}^t\in C_i\\ 0 \text{ if } \mathbf{x}^t\in C_j,\ j\neq i\end{cases}$$



Train hypotheses
$h_i(\boldsymbol{x})$, $i=1,\ldots,K$:

$$h_i\left(\mathbf{x}^t\right)=\begin{cases}1 \text{ if } \mathbf{x}^t\in C_i\\ 0 \text{ if } \mathbf{x}^t\in C_j,\ j\neq i\end{cases}$$

Reject(doubt)

چند نکته در مورد دسته‌بندی

- در برخی کاربردها به جای یک فرضیه برای موارد مثبت، می‌توان دو فرضیه یکی برای موارد مثبت و دیگری برای موارد منفی در نظر گرفت، در این حالت اگر نمونه‌ای توسط هر دو کلاس تشخیص داده نشود، رد خواهد شد.
- توزیع دو کلاس همیشه مانند هم نیست، در نتیجه نمی‌توان همیشه فرضیه‌ی یکسانی را در نظر گرفت، به عنوان مثال کلاس افراد بیمار و سالم
  - افراد سالم خصوصیات مشابهی دارند اما بیماران بسته به نوع بیماری علائم و در نتیجه نشانه‌های متفاوتی خواهند داشت.

مقدمه‌ای بر رگرسیون

- در رگرسیون، برخلاف دسته‌بندی با یک تابع پیوسته مواجه هستیم:

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

- بر خلاف درون‌یابی، در رگرسیون وجود نویز در خروجی را هم باید در نظر گرفت.

$$r^t = f\left(x^t\right) + \varepsilon$$

– وجود نویز را می‌توان به مربوط به متغیرهای مخفی(غیرقابل‌مشاهده) دانست.

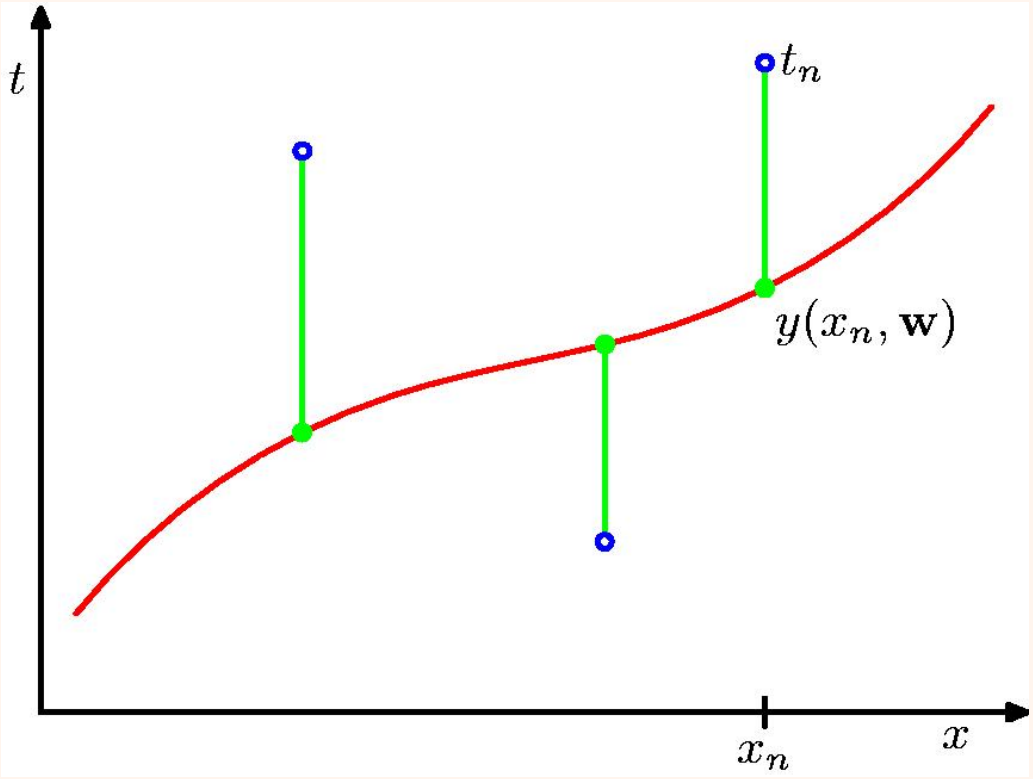
$$r^t = f^*\left(x^t, z^t\right)$$

– هدف تخمین خروجی با .استفاده از مدل پیشنهادی (g(x)) است.

مقدمه‌ای بر رگرسیون (ادامه...)

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

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



رگرسیون خطی تک‌متغیره

- با فرض این که g(x) خطی است:

$$g(\mathbf{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_0 = \sum_{j=1}^{d} w_j x_j + w_0$$

- فرض می‌کنیم که مثال مورد نظر دوبعدی باشد، در نتیجه تابع خطا به صورت زیر خواهد شد:

$$E(w_1, w_0 \mid \mathcal{X}) = \frac{1}{N} \sum_{t=1}^{N} \left[r^t - \left(w_1 x^t + w_0 \right) \right]^2$$

رگرسیون خطی تک‌متغیره(ادامه...)

- برای پیدا کردن پارامترهای بهینه باید مشتق گرفته و آن را مسوی صفر قرار دهیم:

$$\frac{\partial E}{\partial w_0} = \frac{2}{N}\sum_{t=1}^{N}\left[r^t - \left(w_1 x^t + w_0\right)\right](-1)$$

$$\frac{\partial E}{\partial w_1} = \frac{2}{N}\sum_{t=1}^{N}\left[r^t - \left(w_1 x^t + w_0\right)\right](-x^t)$$

- دو معادله و دو مجهول

$$w_0 = \bar{r} - w_1 \bar{x}$$

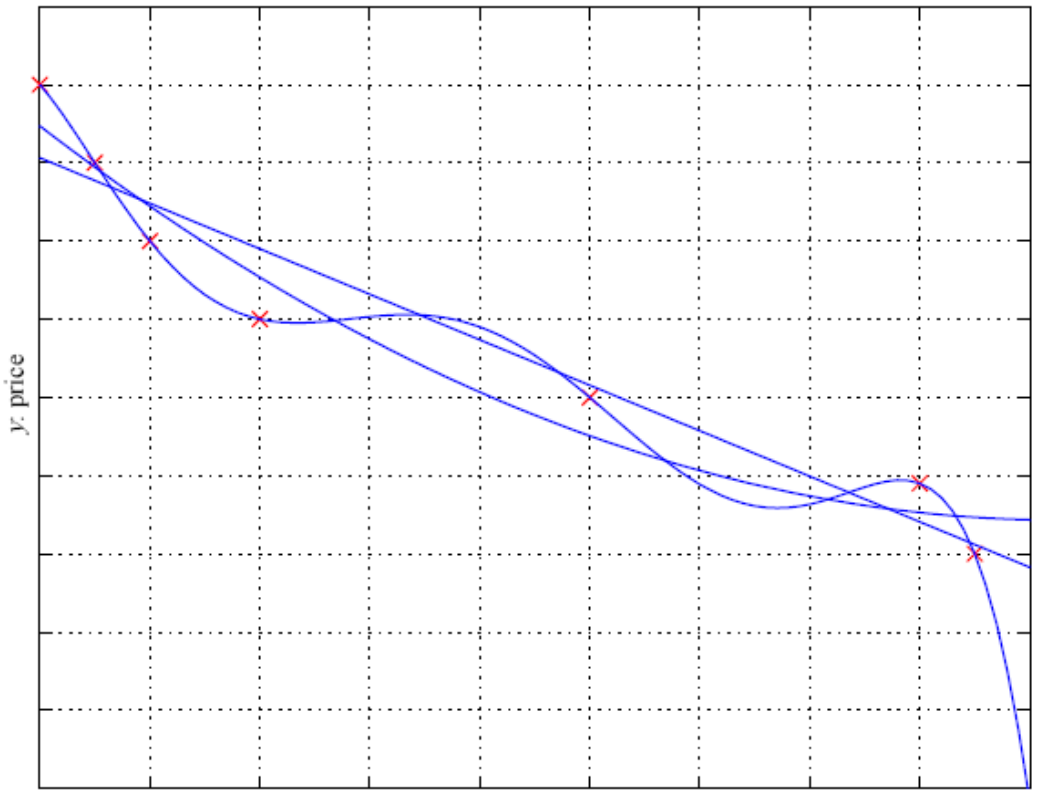
$$w_1 = \frac{\sum_t x^t r^t - \bar{x}\bar{r}N}{\sum_t \left(x^t\right)^2 - N\left(\bar{x}\right)^2}$$

$$\bar{x} = \frac{\sum_t x^t}{N}$$

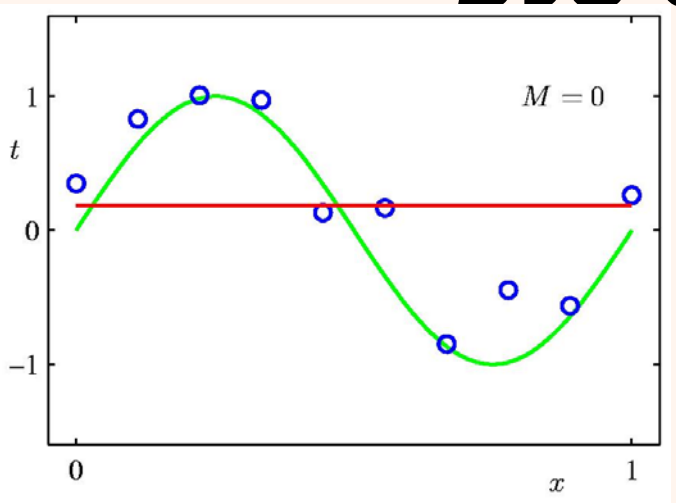
$$\bar{r} = \frac{\sum_t r^t}{N}$$

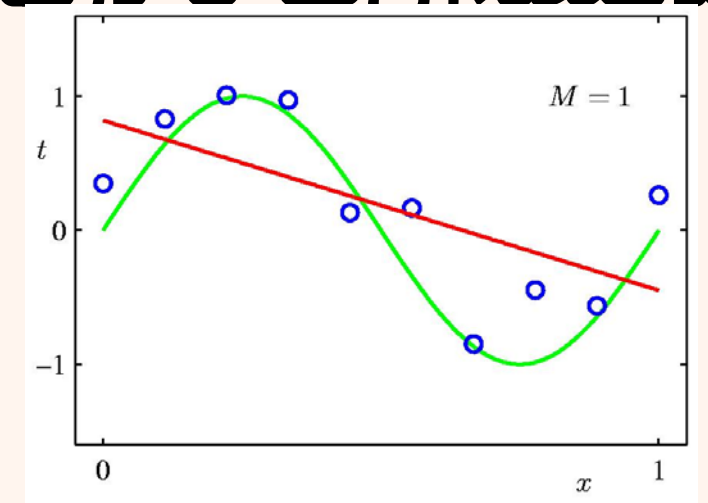
رگرسیون

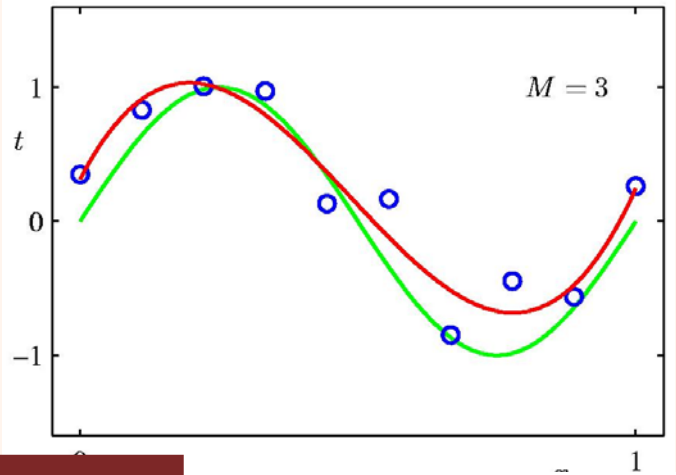
- در صورتی که مدل خطی برای داده‌ها ساده باشد، می‌توان از تابع درجه‌ی دو و یا درجات بالاتر استفاده کرد:

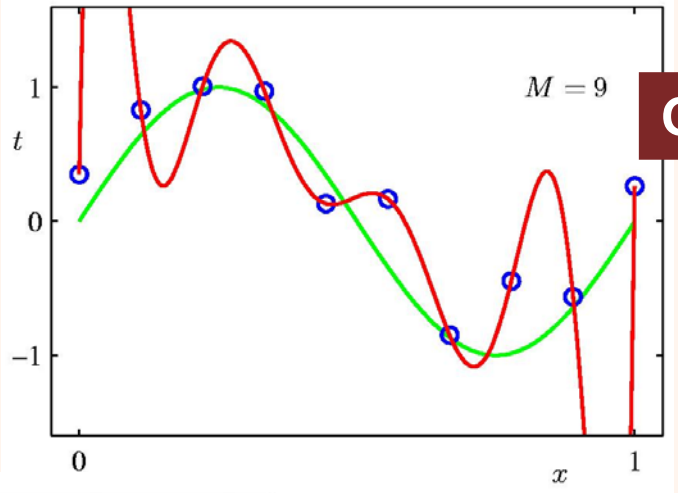


چندجمله‌ای‌ها با درجات بالاتر



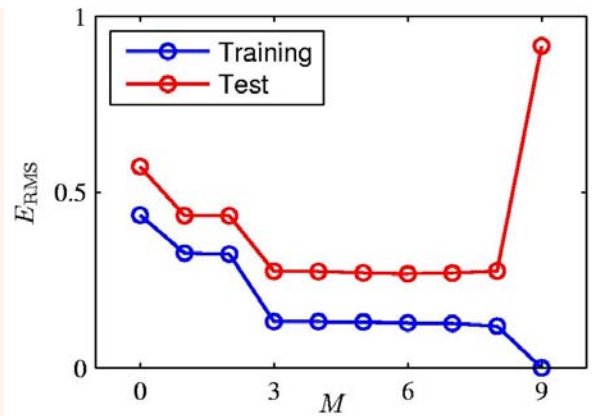






Over fitting

Best fit



انتخاب مدل و تعمیم‌پذیری

- یادگیری، یک مساله‌ی **ill-posed** است؛ داده‌های آموزشی به تنهایی برای یافتن یک راه حل یکتا، کافی نیستند.

x_1	x_2	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}	h_{13}	h_{14}	h_{15}	h_{16}
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

- در یادگیری استقرایی، افزون بر داده‌ها باید مفروضاتی را نیز در نظر گرفت. بدون پیش‌فرض قادر به حل مساله نخواهیم بود.

انتخاب مدل

- در نتیجه، علاوه بر داده‌ها باید مفروضات دیگری در نظر گرفت که پاسخ یکتایی به دست آید. این پیش‌فرض‌ها «inductive bias» نامیده می‌شود.
  - کلاس فرضیه، پیش‌فرض مذکور تلقی می‌شود.
- هر چه ظرفیت فرضیه افزایش یابد، پیچیدگی آن نیز بیشتر خواهد شد.
  - دو مستطیل ناهمپوشان در مقابل یک مستطیل
- در «انتخاب مدل» باید تعمیم‌پذیری را در نظر داشت.

انتخاب مدل و تعمیم‌پذیری

- برای افزایش «قابلیت تعمیم» باید پیچیدگی مدل متناسب با پیچیدگی داده‌ها انتخاب شود.
- در صورتی که پیچیدگی مدل کم‌تر از داده باشد، اصطلاحا گفته می‌شود underfitting رخ داده است.
  - مانند زمانی که یک منحنی درجه‌ی سه با یک خط تقریب زده شود.
  - در چنین حالتی خطای آموزشی و خطای مرحله‌ی validation (validation error) هر دو بالا خواهند بود.
- در صورتی که مدل پیچیده‌تر انتخاب شود، overfitting رخ می‌دهد.
  - با افزایش داده‌های آموزشی می‌توان اثر آن را تا حدی کاهش داد.

Tradeoff سه‌گانه

- بین عوامل زیر tradeoff وجود دارد:
  - پیچیدگی کلاس فرضیه $\mathcal{H}$، $c(\mathcal{H})$
  - اندازه‌ی مجموعه‌ی آموزشی
  - خطای تعمیم

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

Cross-Validation

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

Test set (publication set)

ابعاد متفاوت الگوریتم‌های یادگیری ماشین

- مدل $g(\mathbf{x}|\theta)$
- تابع هزینه (cost or loss function)

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

- فرآیند بهینه‌سازی

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

  - در صورتی که مدل پیچیده‌تر شود، به روش‌های پیچیده‌تری 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.