Machine Learning

یادگیری با نظارت

دسته‌بندی

رگرسیون خطی

انتخاب مدل

یادگیری ماشین
بخش دوم

دانشگاه شهید بهشتی
پژوهشکده‌ی فضای مجازی
پاییز ۱۴۰۱
احمد محمودی ازناوه

فهرست مطالب

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

یادگیری دسته

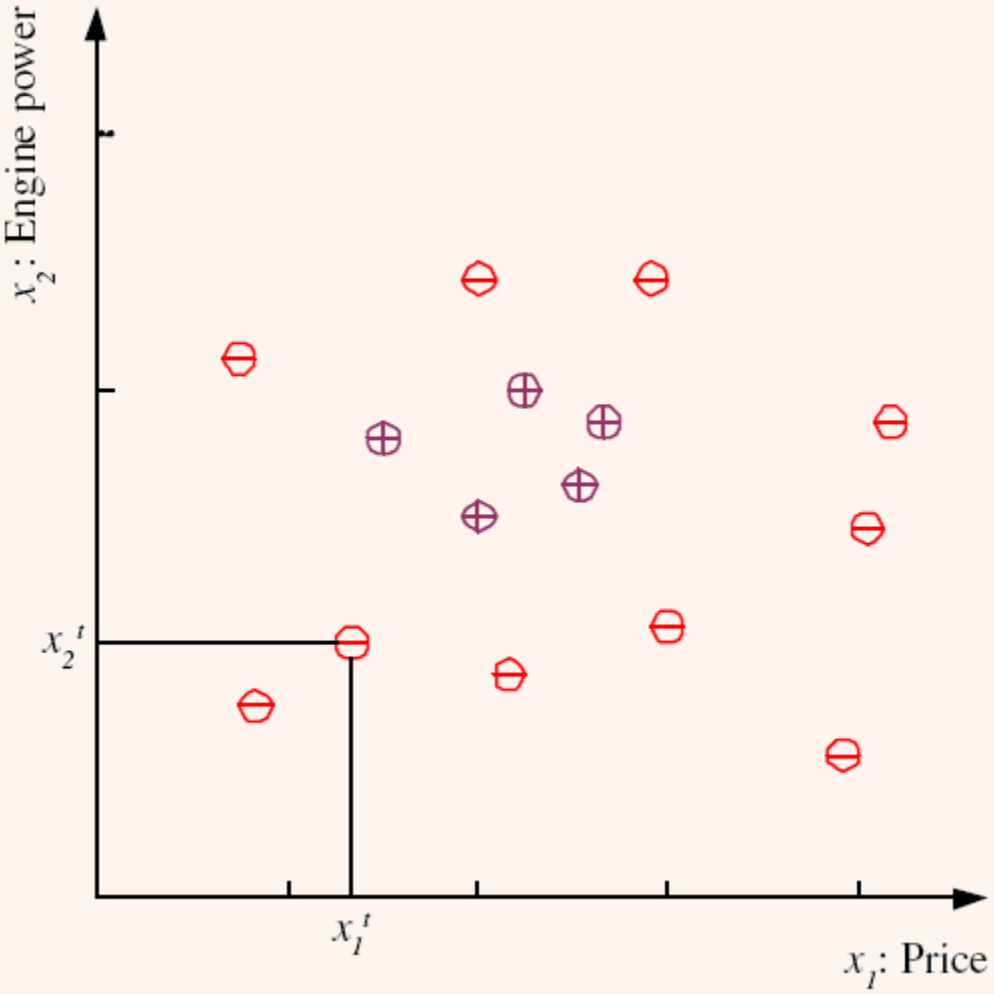
***Class learning** is finding a description that is shared by all positive examples and none of the negative examples.*

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

دسته‌ی آموزشی

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

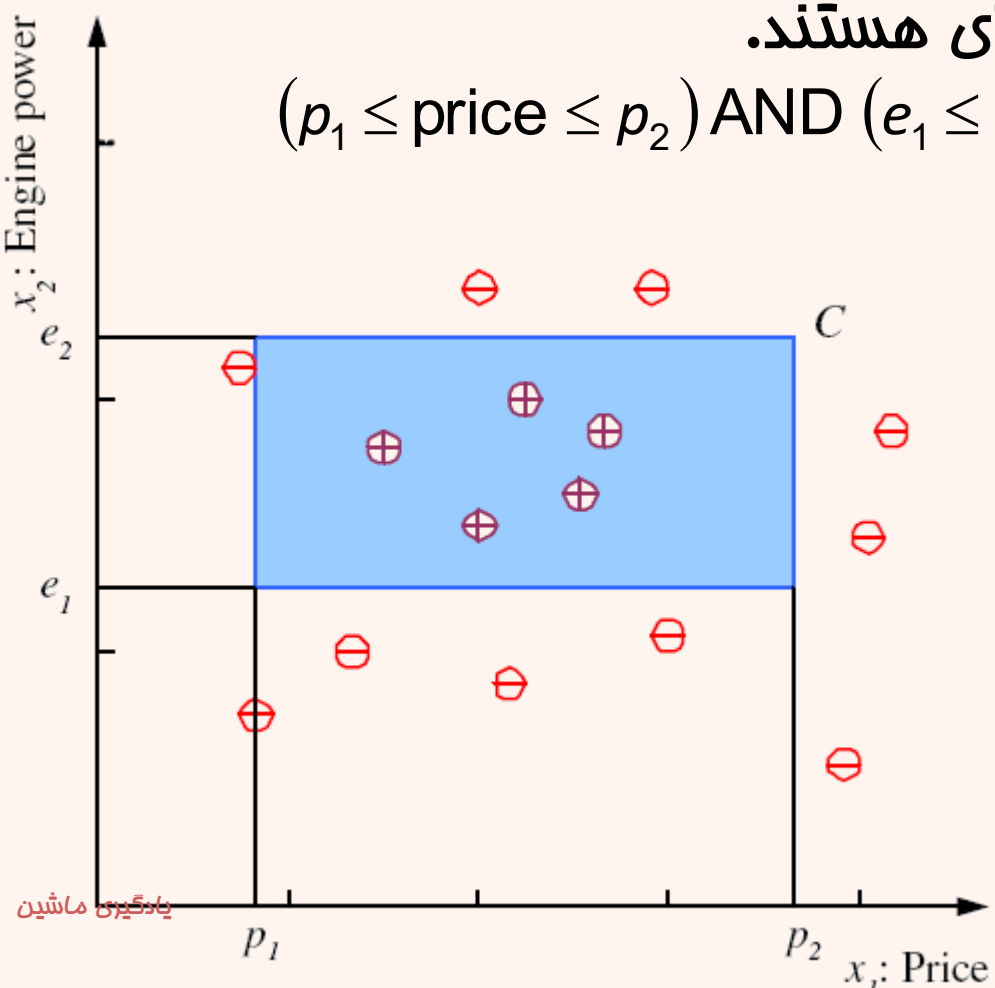


Hypothesis class $\mathcal{H}$

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

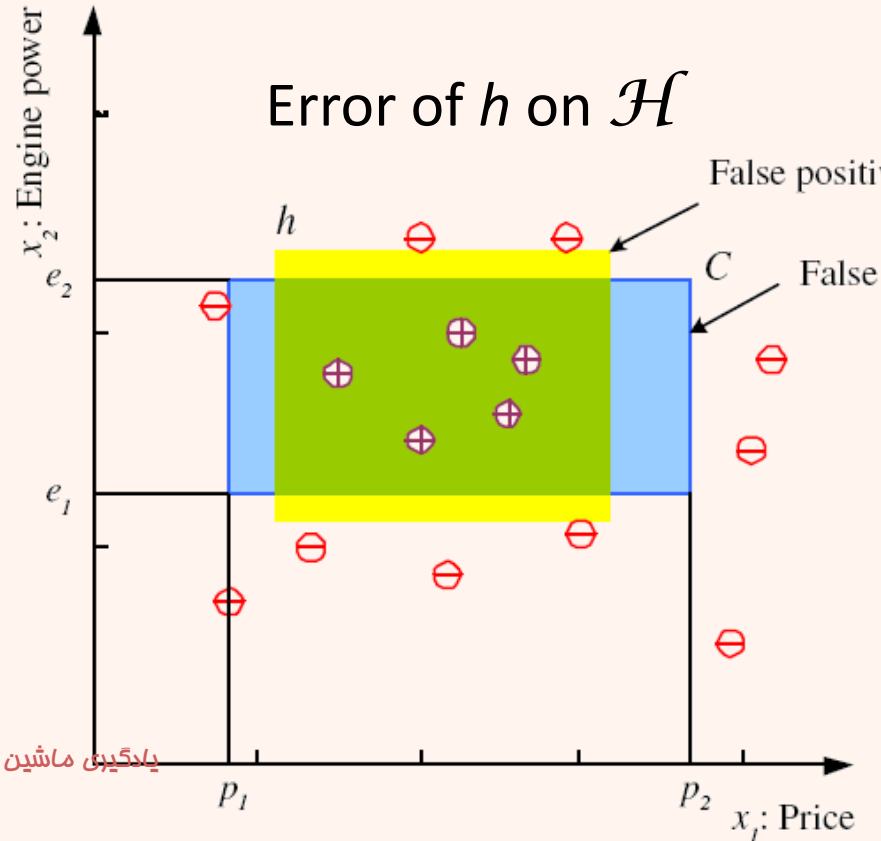
- هدف یافتن h به نحوی است که حداکثر شباهت به C را داشته باشد.

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

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

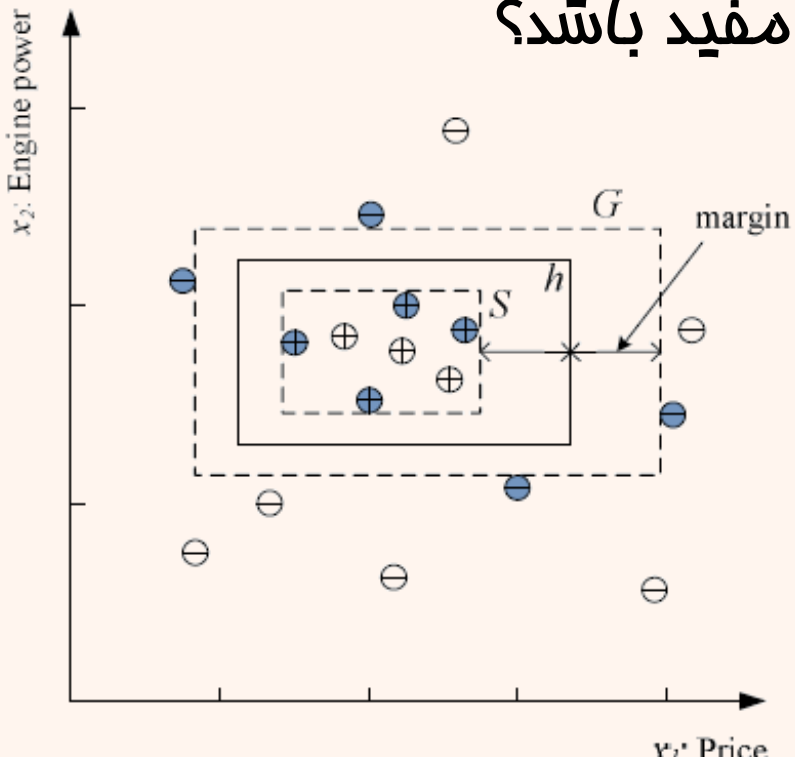
- بی‌شمار «فرضیه» می‌توان یافت که دارای خطای صفر باشند.
  - یادگیری، را می‌توان جستجو برای یافتن بهترین پارامترها دانست.
- از بین تمامی فرضیه‌های درست، مناسب‌ترین فرضیه آن است که برای نمونه‌های جدیدی که در آینده دیده می‌شود، بهترین پاسخ را عرضه کند.

A machine learning hypothesis is a candidate model that approximates a target function for mapping inputs to outputs

https://machinelearningmastery.com/what-is-a-hypothesis-in-machine-learning/

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؟



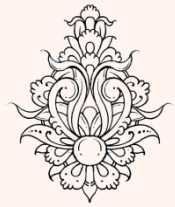
نویز

- «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

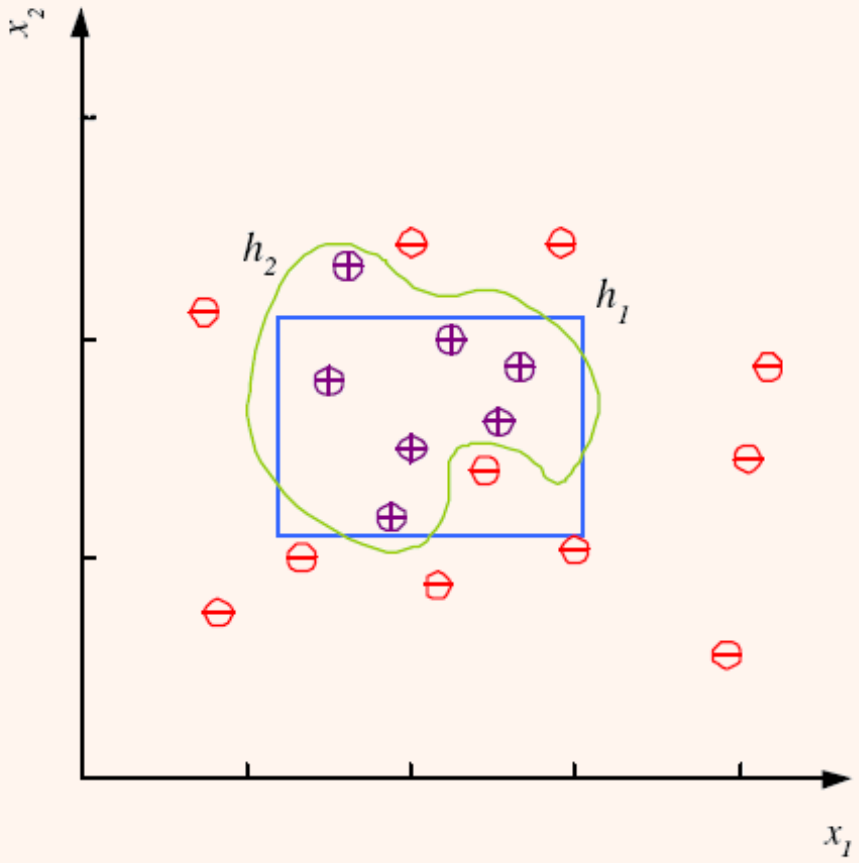
  - ممکن است برخی ویژگی در نظر گرفته نشده است و یا ویژگی‌هایی قابل مشاهده نبوده‌اند.

Hidden or latent attribute



نویز و پیچیدگی مدل

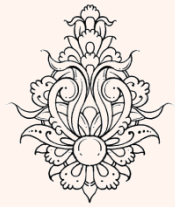
- هنگامی که نویز وجود دارد، مدل به دست آمده **پیچیده‌تر** خواهد شد.



مزایای انتخاب مدل ساده‌تر

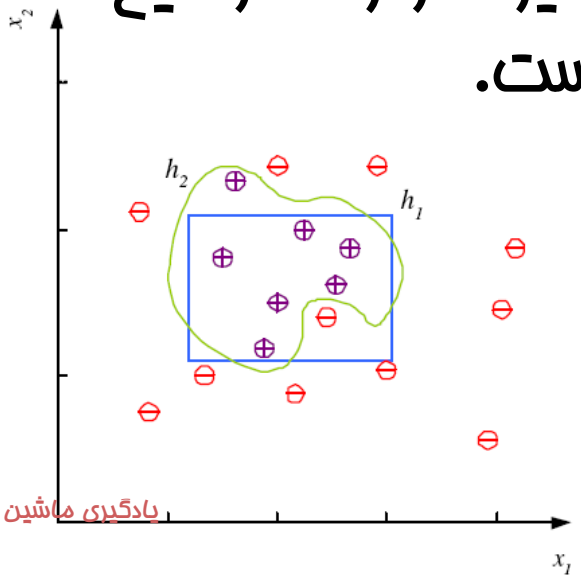
Variance v.s. Bias

- استفاده از آن ساده‌تر است؛ پیچیدگی محاسباتی کم‌تری دارد. (نمونه‌های جدید به راحتی آموزش می‌بینند).
- 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 سوگیری(bias) بیشتری 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(واریانس کم‌تری دارند)، در این حالت هرچند دارای خطای بیشتری روی داده‌های آموزشی خواهند بود، ولی «تعمیم‌پذیری» بهتری خواهند داشت.



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)

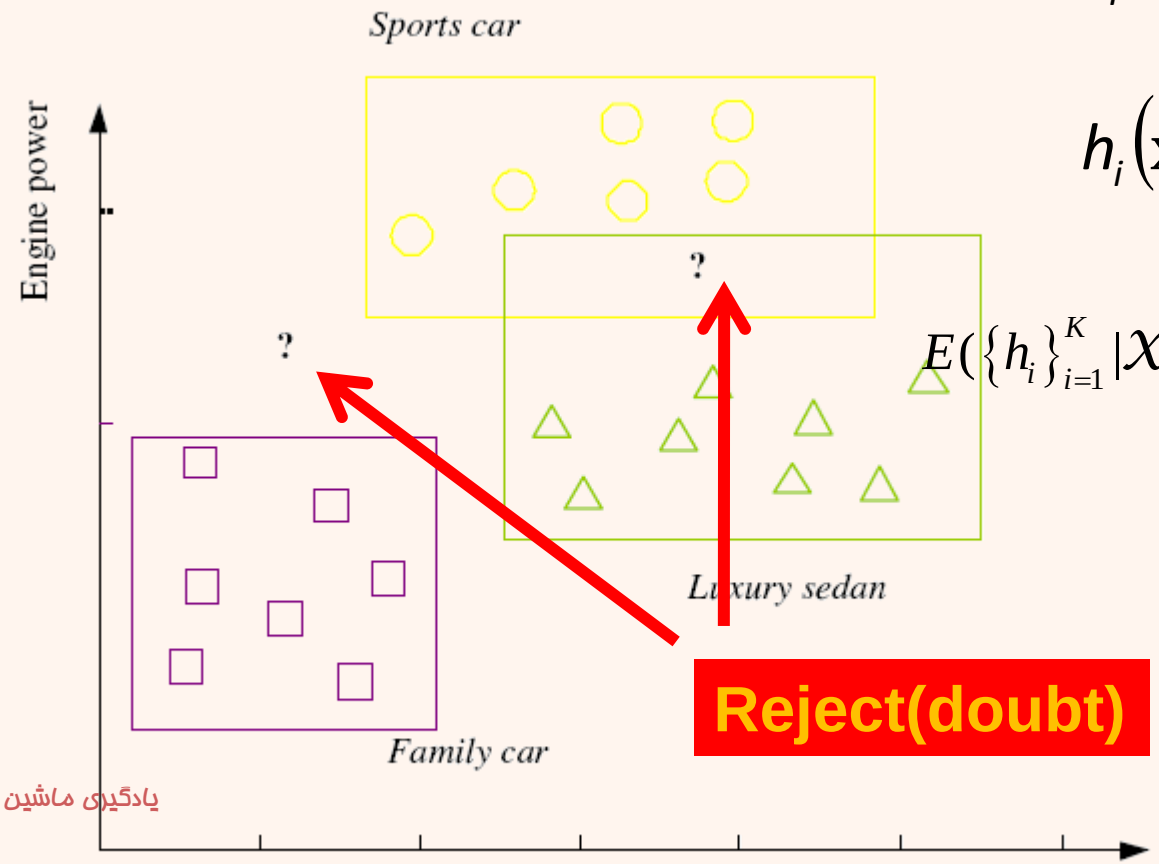
می‌توان مسأله را به صورت K دسته‌بندی دو کلاسه در نظر گرفت.

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 $h_i(\mathbf{x})$, $i=1,\ldots,K$:

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

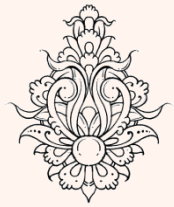


$$E(\{h_i\}_{i=1}^K|\mathcal{X})=\sum_{t=1}^N\sum_{i=1}^K 1\left(h_i\left(\mathbf{x}^t\right)\neq r_i^t\right)$$

Reject(doubt)

دسته‌بندی چندبرچسبی

- در این حالت به هر داده می‌تواند به صورت همزمان به چند دسته متعلق باشد.



مقدمه‌ای بر رگرسیون(پس‌نمایی)

- 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.

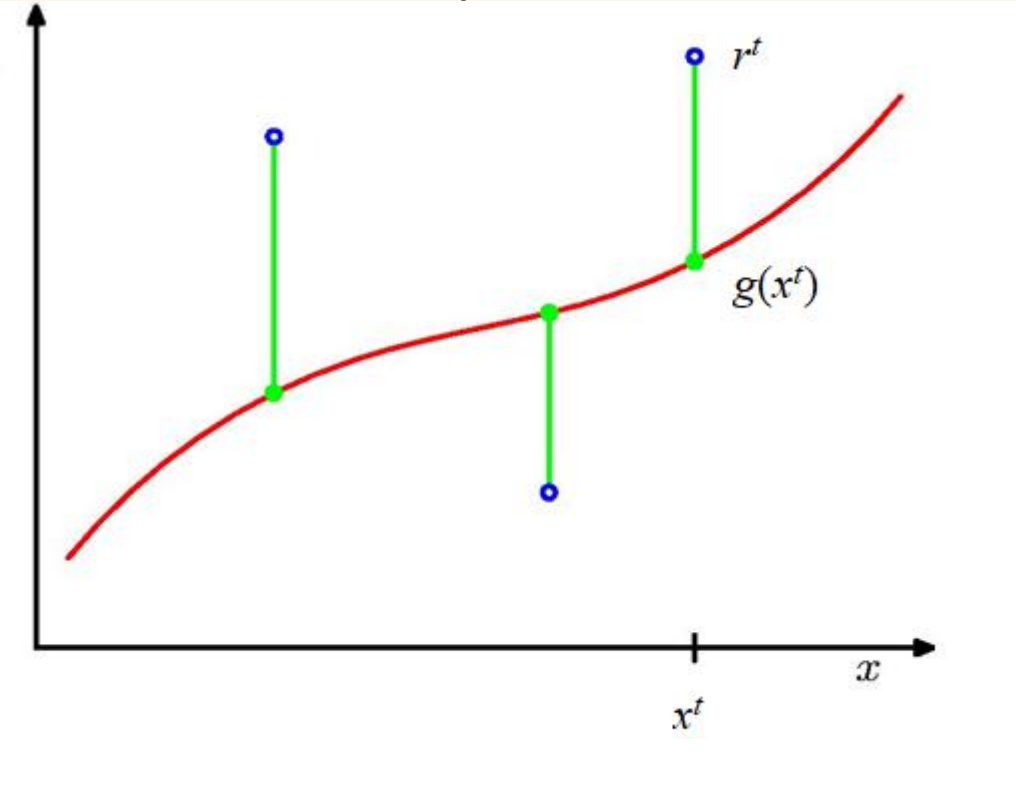
$$r^t = f^*\left(x^t, z^t\right)$$

  - 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:

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



Pattern Recognition and Machine Learning (Bishop)

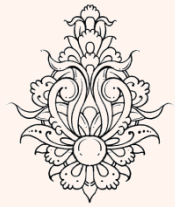
رگرسیون خطی تک‌متغیره

- 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...)

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

- برای پیدا کردن پارامترهای بهینه باید مشتق گرفته و آن را مساوی 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

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

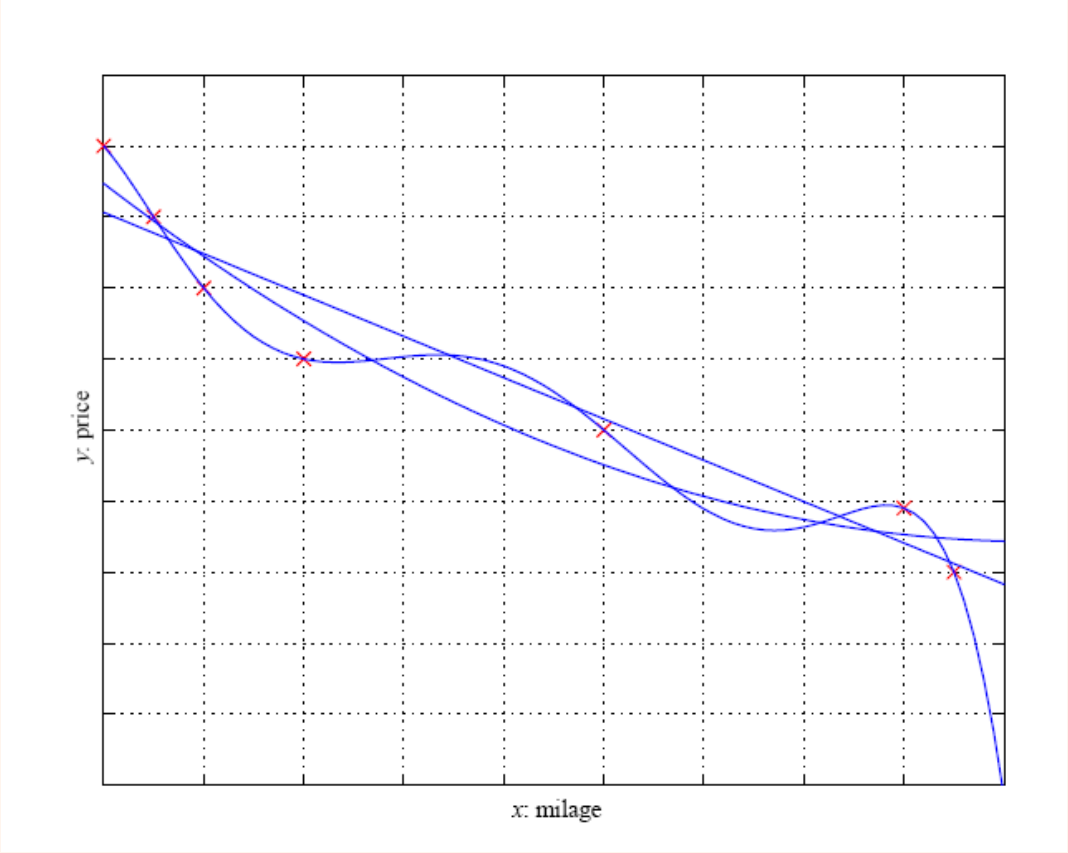
$$w_1 = \frac{\sum_t x^t r^t - \bar{x}\bar{r}N}{\sum_t \left(x^t\right)^2 - N\left(\bar{x}\right)^2}$$

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

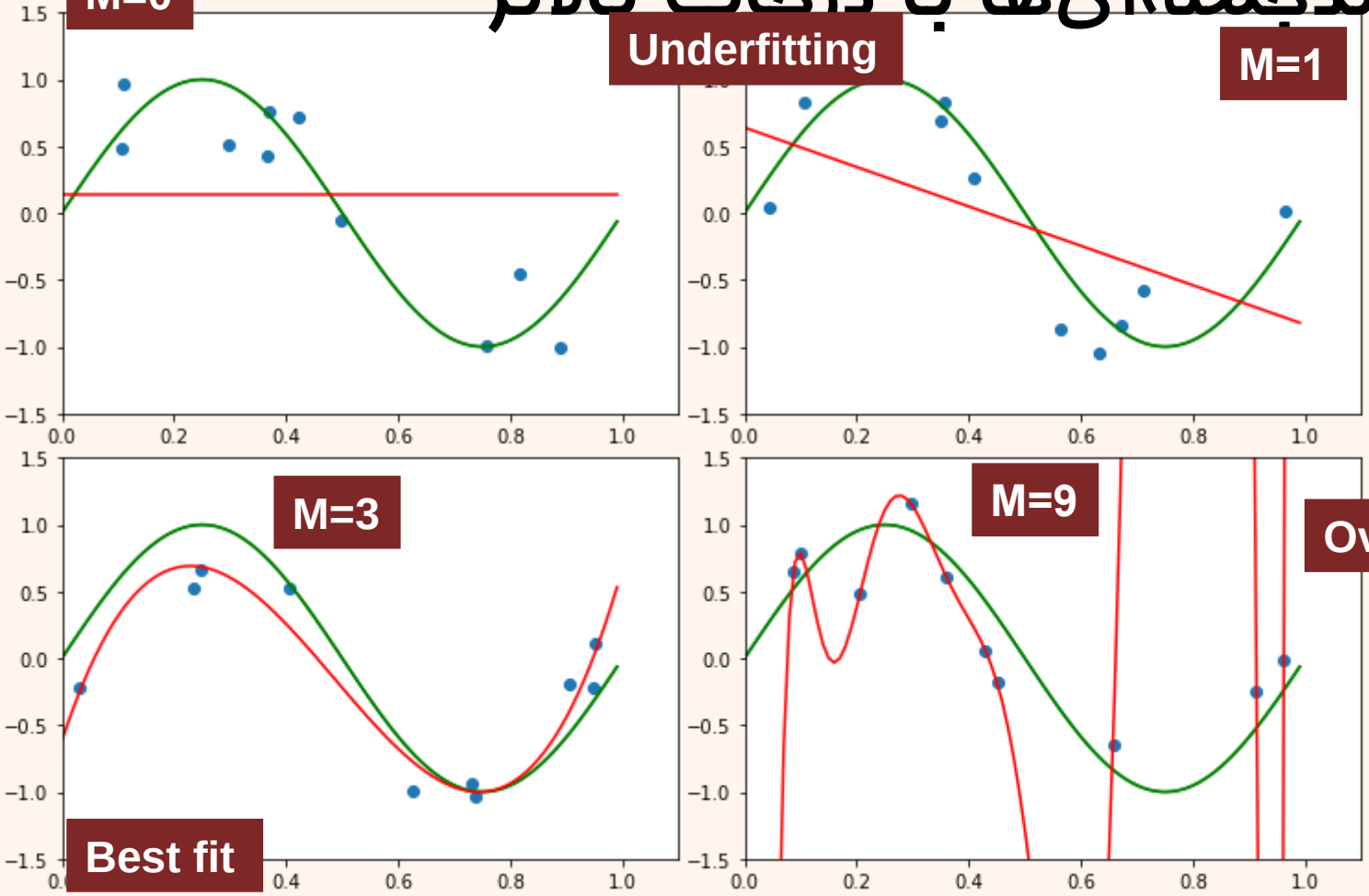
$$\bar{r} = \frac{\sum_t r^t}{N}$$

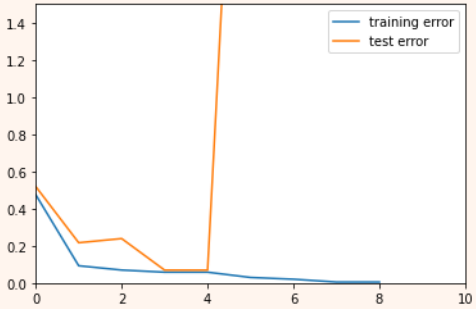
رگرسیون

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



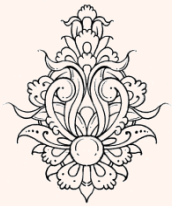
چندجمله‌ای‌ها با درجات بالاتر





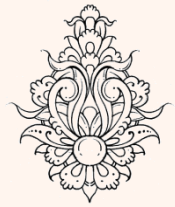
انتخاب مدل و تعمیم‌پذیری

- یادگیری، یک مساله‌ی **ill-posed** است؛ داده‌های آموزشی به تنهایی برای یافتن یک راه حل یکتا، کافی نیستند.
- در یادگیری استقرایی، افزون بر داده‌ها باید مفروضاتی را نیز در نظر گرفت. بدون پیش‌فرض قادر به حل مساله نخواهیم بود.

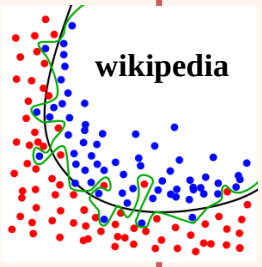


انتخاب مدل

- 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 و انتخاب مستطیل به بیشترین حاشیه
  - یا انتخاب خط و یا چندجمله‌ای مرتبه بالا برای رگرسیون
- در «انتخاب مدل» باید تصمیم‌پذیری 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.
  - در چنین حالتی خطای آموزشی پایین و خطای مرحله‌ی validation (validation error) بالا خواهد بود.

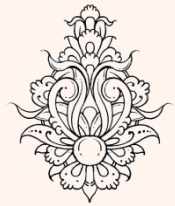
  - با افزایش داده‌های آموزشی می‌توان اثر آن را تا حدی کاهش داد.

Tradeoff سه‌گانه

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

$N\uparrow, E\downarrow$

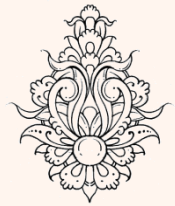
$c(\mathcal{H})\uparrow$, first $E\downarrow$ and then $E\uparrow$



اعتبارسنجی متقابل

- 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 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.