دانشکده‌ی مهندسی کامپیوتر

درس‌نامه یکم

یادگیری ماشین

منابع درس

- An Introduction to Statistical Learning
- The Elements of Statistical Learning

تفاوت یادگیری ماشین با منطق

تمایز یادگیری ماشین با منطق در این است که در یادگیری ماشین داده‌ها می‌توانند برنامه را عوض نمایند. مثلا فرض نمایید که یک برنامه کامپیوتری نوشته شده است تا دو عدد را با یکدیگر جمع نماید

برنامه فوق یک برنامه مشخصی است که ورودیها را به خروجی انتقال می‌دهد. حال یک ماشین را در نظر بگیرید که یک برنامه دارد و در ضمن می‌تواند برنامه خود را نیز عوض کند.

برنامه متغیر معمولا یک سری عدد می‌باشند که برنامه ثابت از آنها استفاده می‌نماید. این اعداد را معمولا با β نمایش می‌دهیم.

مثال ۱. فرض کنید یک برنامه ثابت داریم که رابطه بین ورودی و خروجی آن به شکل زیر است:

$$y = \beta x$$

که در آن y ، x ، β همگی متعلق به $\mathbb{R}$ می‌باشند.در یک مدل یادگیری مقدار β ثابت نمی‌باشد و با توجه به ورودیها و خروجیها که به آن داده گفته می‌شود می‌توانند تغیر نمایند. ممکن است پس از مدتی β ثابت گردد و دیگر آنرا تغییر ندهیم. در این صورت دیگر برنامه ثابت شده و یادگیری صورت نمی‌پذیرد.
بنابراین هدف یادگیری استفاده از داده‌ها و رسیدن به روشی است که بتوان β را تنظیم نمود.

حفظ کردن یا تعمیم دادن

فرض نمایید داده‌ها به صورت زیر داده شده است:

$$X = \begin{array}{r} \\ \text{نمونه ۱} \\ \text{نمونه ۲} \\ \\ \text{نمونه } n \end{array}\begin{array}{c} \text{سن} \\ \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array}\right] \end{array} \qquad Y = \begin{array}{r} \\ \text{نمونه ۱} \\ \text{نمونه ۲} \\ \\ \text{نمونه } n \end{array}\begin{array}{c} \text{درآمد} \\ \left[\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array}\right] \end{array}$$

می‌خواهیم این داده‌ها را ذخیره نماییم و سپس هرگاه ورودی x_i را مشاهده کردیم y_i را در خروجی قرار دهیم. این عمل به سادگی قابل انجام است تنها یک مشکلی دارد و آن اینکه ممکن است برای یک سن خاص چندین درآمد وجود داشته باشد که می‌توان همه آنها را در خروجی ایجاد نمود.

اگر ورودی در داده‌ها وجود نداشت چه می‌توان کرد؟

به این مساله تعمیم می‌گوییم. چگونه می‌توانیم از مابقی داده‌ها به خروجی معقول برسیم و چگونه می‌توانیم نشان دهیم آنچه در خروجی ایجاد می‌شود مناسب است.

راه حل اول : درون یابی

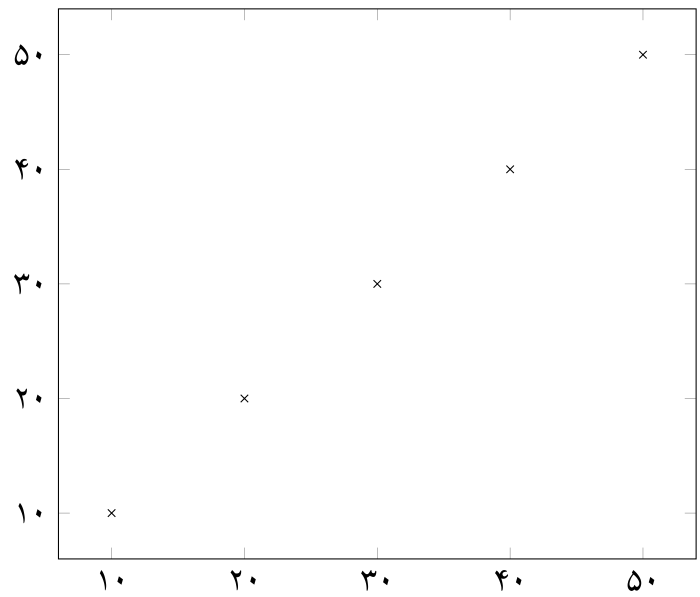
رابطه بین x و y را با یک تابع با پارامتر β نمایش می دهیم :

$$y = f_\beta(x)$$

حال می گوییم β را طوری بدست می آوریم که

$$y_i = f_\beta(x_i) \qquad \forall i \in \{1, \cdots, n\}$$

در این حالت از روی n معادله بایستی β را به دست آوریم . حال ممکن است β وجود نداشته باشد ، برای آنکه جواب β را امکان پذیر نماییم بایستی همزمان به داده ها و تابع توجه کنیم .



در شرایط بالا هرچند به یک دستگاه با چهار معادله روبه‌رو می‌باشیم اما یک تابع خطی با یک پارامتر می‌تواند از تمامی نقاط عبور کند . بنابراین کافیست درنظر بگیریم :

$$y = \beta x$$

و β را از روی یک نقطه بدست آوریم مثلا

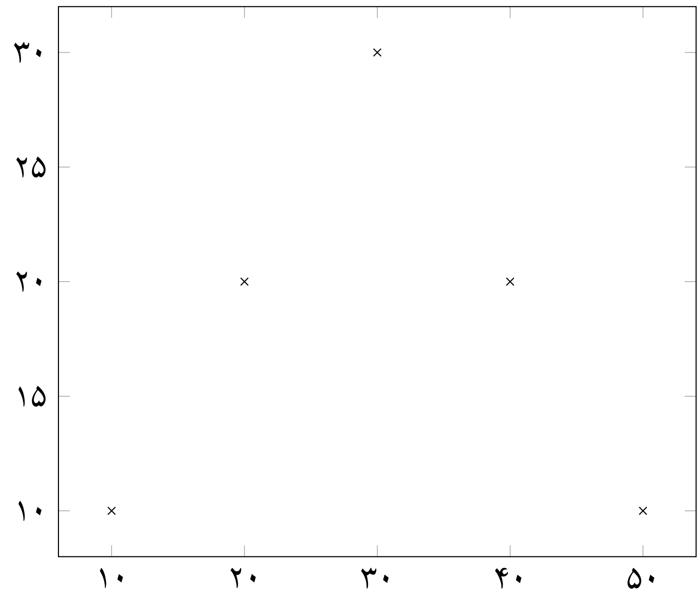
$$\beta = \frac{y_1}{x_1}$$

حال می توانیم برای یک x_{in} خروجی را به‌صورت

$$y_{out} = \beta x_{in}$$

در نظر بگیریم.

رابطه فوق را ممکن است به شکل زیر باشد :



ولیکن همچنان بتوانیم با تعداد پارامترهای کم تابعی بدست آوریم که از تمام نقاط بگذرد.مثلا در مثال فوق می‌توانیم یک درجه دوم درنظر بگیریم که از تمام داده ها عبور می‌کند .

$$y = \beta_1 x + \beta_2 x^2$$

اگر دو نقطه از داده را انتخاب نماییم ، می‌توانیم β_1 و β_2 را از یک دستگاه معادله بدست آوریم :

$$y_1 = \beta_1 x_1 + \beta_2 x_1^2$$

$$y_2 = \beta_1 x_2 + \beta_2 x_2^2$$

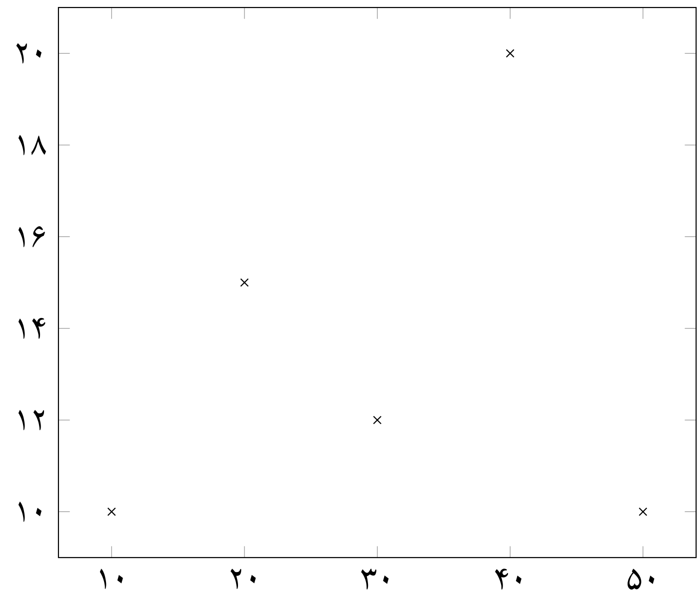
$$\Rightarrow \quad \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 & x_1^2 \\ x_2 & x_2^2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

بنابراین خواهیم داشت:

$$\Rightarrow \quad \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} x_1 & x_1^2 \\ x_2 & x_2^2 \end{pmatrix}^{-1} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

مثال های فوق نشان می‌دهند که در حالتیکه شر ایط مناسب باشد با تعداد کمی نمونه می توانیم رابطه بین ورودی و خروجی را بدست آورم .

در بسیاری از مواقع به این میزان خوش شانس نمی‌باشیم و بایستی از تمام نمونه ها برای بدست آوردن پارامتر ها استفاده نمود .یعنی تعداد پارامتر ها با تعداد نمونه ها بایستی برابر باشد .



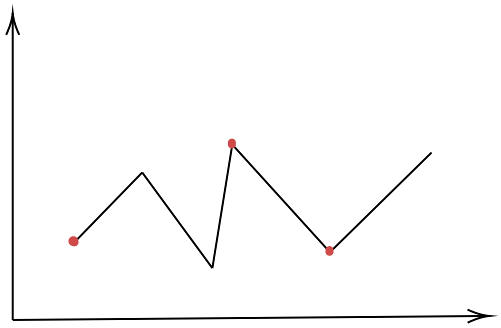
برای حل مسئله فوق دو راه پیش‌رو داریم :

۱. تابعی را برای کل R در نظر بگیریم و سعی می‌کنیم نقاط را fit نماییم ، مثلا اگر یک تابع درجه $n-1$ در نظر بگیریم که چند جمله ای به صورت زیر باشد :

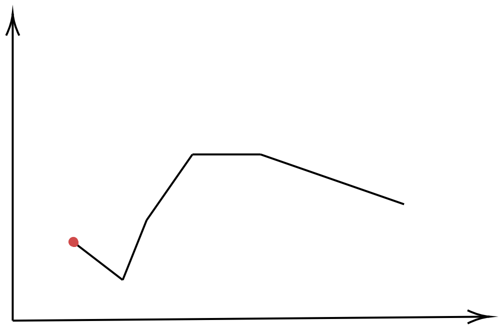
$$f(x) = \sum_{i=0}^{n-1} \beta_i x^i$$

آنگاه می‌توانیم با داشتن n معادله و n مجهول سعی نماییم β_i ها را بدست آوریم . (دیدگاه جهانی)

۲. تعداد نقاط کمی در کنارهم را در نظر گرفته و یک تابع ساده را روی آنها برازش نماییم و سپس آنها را به یکدیگر وصل کنیم . (دیدگاه محلی)



اگر دقت نماییم درون‌یابی خاصیت ذخیره سازی را نیز دارد، یعنی به ازای X_i همان Y_i را در خروجی قرار می‌دهد. سوال اصلی را البته بایستی جواب دهیم، اینکه برای یک نقطه جدید X_{in} خروجی Y_{out} چقدر معتبر می‌باشد؟ جواب این مسئله را می‌توان اینگونه داد که تعدادی از نقاط را در مسئله درون‌یابی مورد استفاده قرار نمی‌دهیم و پس از مشخص کردن پارامترهای مسئله، روی نقاط کنار گذاشته شده چک می‌کنیم که خروجی مدل به ازای این نقاط چه میزان به آنچه نیاز داریم نزدیک است.
در مثال زیر نقاط استفاده شده در فرایند درون‌یابی با رنگ قرمز و نقاط آزمون با رنگ آبی مشخص شده‌اند که مشاهده می‌شود که نقاط آزمون فاصله زیادی تا خروجی بدست آمده دارند.



راه حل دوم: یادگیری آماری

در یادگیری آماری رابطه بین ورودی و خروجی را یک توزیع مشترک در نظر میگیریم. به‌عنوان مثال در مسئله سن و درآمد، یک تویع مشترک می‌تواند به شکل زیر باشد:

$$f_{XY}(x,y) = \frac{۱}{\sqrt{det(۲\pi Q)}} e^{-\frac{\begin{pmatrix} x & y \end{pmatrix} Q^{-۱} \begin{pmatrix} x \\ y \end{pmatrix}}{۲}}$$

که یک توزیع گوسی می‌باشد و می‌توان آن را به صورت $(X,Y) \sim N(۰,Q)$ نشان داد. از آنجاییکه هدف، پیدا کردن خروجی Y بر پایه ورودی X می‌باشد، می‌توانیم از روی توزیع مشترک مقدار Y را بدست آوریم. می‌توان نشان داد در شرایط خاص بهترین خروجی به صورت زیر است:

$$E[Y|X] = f(X)$$

که در صورت وجود توزیع مشترک می‌توان آن را محاسبه نمود. به عنوان مثال در رابطه فوق داریم:

$$f(X) = \beta X$$

که در آن

$$\beta = \frac{E[XY]}{E[Y^{۲}]} = \frac{Q_{۱۲}}{Q_{۲۲}}$$

که ماتریس Q را به صورت زیر فرض کرده‌ایم:

$$Q = \begin{pmatrix} Q_{۱۱} & Q_{۱۲} \\ Q_{۲۱} & Q_{۲۲} \end{pmatrix} = \begin{pmatrix} E[X^{۲}] & E[XY] \\ E[XY] & E[Y^{۲}] \end{pmatrix}$$

داشتن توزیع مشترک عملا هم آنچه بدان نیاز داریم را در اختیار قرار می‌دهد و هم نیازی به داده برای تعمیم وجود ندارد. در بسیاری از مواقع توزیع مشترک وجود ندارد و تنها داده‌های ورودی و خروجی دردسترس است. دو کار می‌توان انجام داد:

۱. توزیع مشترک را بیابیم.

۲. مستقیما $E[Y|X]$ را پیدا کنیم.

روش دوم را فعلا انتخاب می‌کنیم و فرض می‌کنیم که $E[Y|X]$ از یک دسته تابع مشخص با پارامتر β درست شده است:

$$E[Y|X] = f_{\beta}(X)$$

که در این شرایط رابطه بین X و Y را می‌توان به صورت زیر نوشت:

$$Y = f_{\beta}(X) + \epsilon$$

که در آن ϵ خود یک متغیر تصادفی است که دارای توزیع مشترک با X بوده و

$$E[\epsilon|X] = ۰.$$

در اغلب موارد و برای سادگی فرض می‌گردد که X و ϵ مستقل از هم بوده و درنتیجه:

$$E[\epsilon|X] = E[\epsilon] = ۰$$

در نظر گرفته می‌شود.

نکته ۱. اگر Y گسسته باشد، آنگاه فرض استقلال نمی‌تواند برقرار باشد. جز در شرایط خاص که مثلا تابع f را خود گسسته بدانیم.

روش دوم را فعلا انتخاب میکنیم و فرض میکنیم که $\mathbb{E}[Y|X]$ از یک دسته تابع مشخص با پارامتر β درست شده است.

$$\mathbb{E}[Y|X] = f_\beta(x)$$

در این شرایط رابطه بین X و Y را میتوان به صورت زیر نوشت:

$$y = f_\beta(x) + \epsilon$$

که در آن ϵ خود یک متغیر تصادفی است که دارای توزیع مشترک با X بوده و

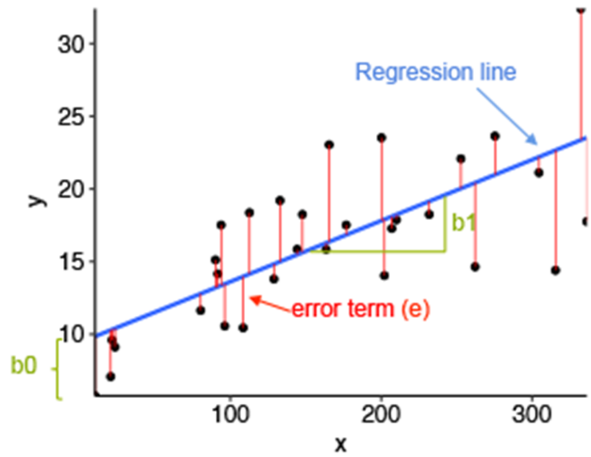
$$\mathbb{E}[\epsilon|X] = ۰$$

در اغلب موارد و برای سادگی فرض میشود که X و ϵ از هم مستقل بوده و در نتیجه:

$$\mathbb{E}[\epsilon|X] = \mathbb{E}[\epsilon] = ۰$$

در دیدگاه آماری چون هدف دیگر برازش نیست آنگاه میتوانیم تابع f را با آزادی بیشتری انتخاب کنیم. به عنوان مثال در داده های زیر تابع خطی را در نظر میگیریم و خواهیم داشت:

$$\mathbb{E}[Y|X] = \beta X$$



شکل ۱: $f_\beta(x) = \beta x$

حال باید از روی داده مقدار β را محاسبه کنیم:

$$y_1 = \beta x_1 + \epsilon_1$$

$$y_2 = \beta x_2 + \epsilon_2$$

$$\dots$$

$$y_n = \beta x_n + \epsilon_n$$

که به دلیل وجود $\epsilon_1, ..., \epsilon_n$ عملا میتوانیم بی نهایت جواب پیدا کنیم. برای رهایی از این مشکل باید فرضیاتی داشته باشیم که برای رسیدن به این فرضیات از نظریه احتمالات بهره میبریم

قضیه‌ی ۱ (قانون اعداد بزرگ). اگر متغیرهای تصادفی $X_1, ..., X_n$ متغیرهای تصادفی $i.i.d$ باشند، داریم:

$$\frac{\sum\limits_{i=0}^{n} X_i}{n} \longrightarrow \mathbb{E}[X]$$

با بهره گیری از قانون اعداد بزرگ و فرض $\mathbb{E}[\epsilon] = 0$ خواهیم داشت:

$$\sum_{i=0}^{n} y_i = \beta \sum_{i=0}^{n} x_i + \sum_{i=0}^{n} \epsilon_i$$

بنابراین:

$$\hat{\beta} = \frac{\sum\limits_{i=0}^{n} y_i}{\sum\limits_{i=0}^{n} x_i}$$

میتواند تقریب خوبی از β اصلی باشد. حال اگر $\sum x_i = 0$ باشد، $b\hat{e}ta$ قابل محاسبه نیست. فرض بعدی این است که واریانس متغیر تصادفی ϵ کمینه باشد یعنی β طوری انتخاب شود که σ_ϵ^2 کمینه شود.

به این ترتیب، طبق قانون اعداد بزرگ داریم:

$$\sigma_\epsilon^2 \approx \frac{\sum\limits_{i=0}^{n} \epsilon_i^2}{n} = \frac{1}{n} \sum_{i=0}^{n} (y_i - \beta x_i)^2$$

در این حالت β که تقریب اعداد بزرگ مذکور را کمینه میکند به صورت زیر خواهد بود:

$$\hat{\beta} = \frac{\sum\limits_{i=0}^{n} x_i y_i}{\sum\limits_{i=0}^{n} x_i^2}$$

حال مشابه قبل این سوال مطرح است که مدل بدست آمده را چطور ارزیابی کنیم؟

راه حل پیشنهادی قبل را میتوانیم دوباره مدنظر قرار دهیم: تعدادی از داده ها را کنار گذاشته و در فرآیند یادگیری از آنها استفاده نکنیم. سپس در مرحله ازمون کیفیت مدل را بر حسب آنچه خروجی داده میشود ارزیابی کنیم.