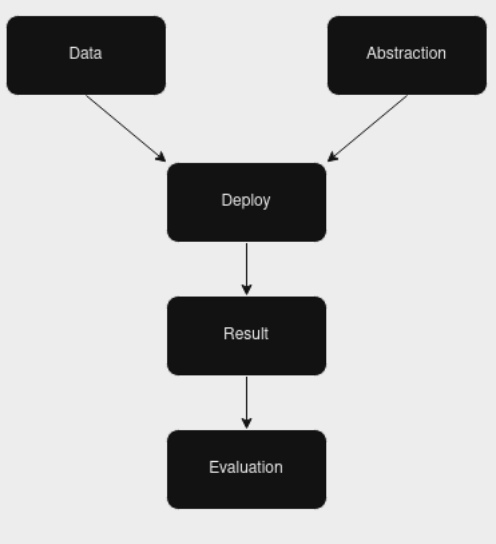
درس‌نامه یکم

فرآیند یادگیری ماشین

یادگیری ماشین با داده شروع می‌شود. سپس آن داده با استفاده از Abstraction که نشان دهنده طرز تفکر و چگونگی نگاه به داده است، به پیاده‌سازی یا Deployment می‌رسد. از این پیاده‌سازی نتایجی گرفته می‌شود و در نهایت این نتایج ارزیابی می‌شوند.

داده

تعدادی عدد است، مانند یک ماتریس. اگر عدد نباشد هم آن را به عدد تبدیل می‌کنیم.



شکل ۱: فرآیند یادگیری

اهداف یادگیری

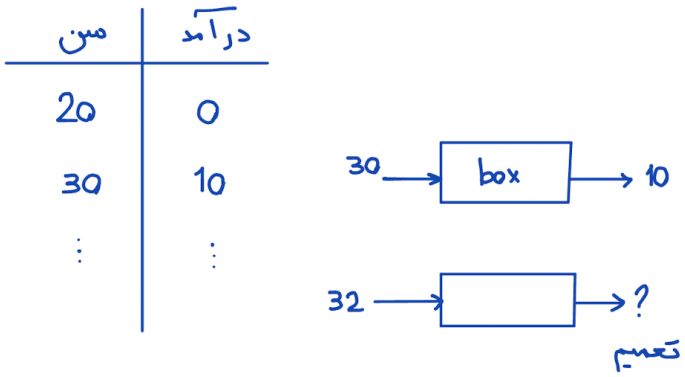
در یادگیری ماشین ما سعی می‌کنیم یک الگو یا pattern را در داده شناسایی کنیم. همچنین داده‌ها را بر اساس یک دسته از ویژگی‌ها دسته‌بندی کنیم. گاهی هم نیاز داریم که بفهمیم چگونه میتوانیم یک داده جدید تولید کنیم؟
بنابراین با توجه به نوع Abstraction ما، مدل‌ها به دو دسته Discriminative و Generative تقسیم‌بندی می‌شوند. در مدل‌های Generative برای تولید خروجی های مختلف نیاز به Random seed داریم.

Karyotyping

برای مثال، ما نمونه‌های تصاویر کروموزوم‌ها را از آزمایشگاه گرفته و به همراه label خود، در یک ماتریس قرار می‌دهیم. پس از دسته‌بندی ویژگی‌های داده ها یا همان pre-processing، اگر نمونه جدیدی به ما داده شود، می‌توانیم آن را دسته‌بندی کنیم و مدل ما Discriminative است. به عبارت دیگر، مدل Generative توزیع $\mathbb{P}(x)$ را مدل می‌کند و مدل Discriminative توزیع $\mathbb{P}(y|x)$.

حفظ کردن یا تعمیم دادن

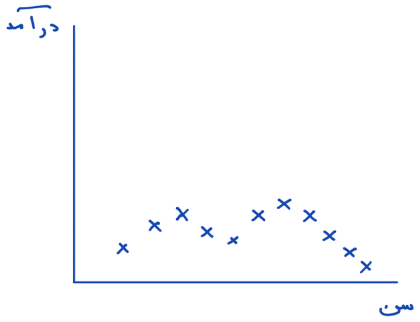
فرض کنید ما داده‌های سن و درآمد افراد را به صورت زیر داریم:



هدف ساختن جعبه‌ای است که با دریافت یک x_i از ورودی y_i صحیح را در خروجی بدهد. یک روش حفظ کردن داده‌ها مثل یک جدول است ولی با استفاده از آن با دو مشکل رو به رو هستیم، اول اینکه ممکن است یک داده ورودی چند خروجی در جدول ما داشته باشد و دوم اینکه اگر داده‌ی جدیدی به عنوان ورودی بیاید چه اتفاقی می‌افتد؟
بنابراین سعی می‌کنیم از ترکیب تعمیم دادن Generalization و حفظ کردن Memorization استفاده کنیم. راه حل ما برای ترکیب این دو، درون‌یابی است.

Interpolation یا درون‌یابی

فرض کنید داده‌ها را به صورت نقاطی در شکل زیر نشان داده‌ایم:



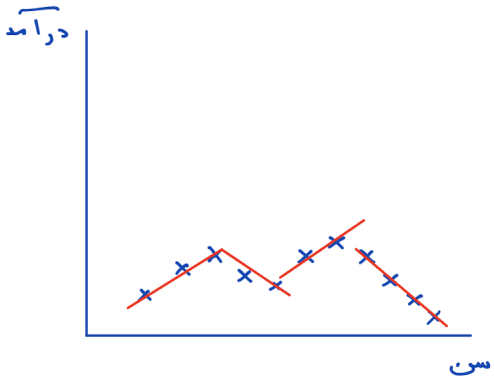
اگر داشته باشیم

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

سعی می‌کنیم پارامتر های β_i و تابع f را به گونه‌ای پیدا کنیم تا از تمام نقاط بگذرد. یعنی

$$f_\beta(x) = \sum_{i=0}^{5} (\beta_i x^i)$$

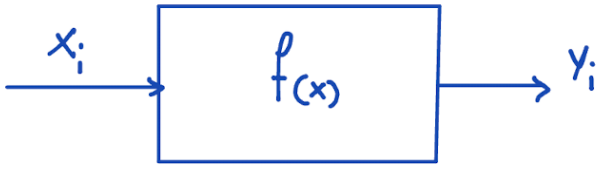
به این دیدگاه که پارامترها را برای کل دامنه پیدا کرده و ثابت در نظر بگیریم، درون‌یابی Global می‌گوییم. در مقابل دیدگاه گلوبال دیدگاه Local را داریم که مانند شکل زیر دامنه را بازه‌بندی کرده و سپس برای هر بازه پارامترهای جداگانه در نظر می‌گیریم.



تفاوت Regression و Classification و توصیف مدل و ارزیابی آن

اگر بتوانیم برای y_iها فاصله تعرف کنیم و از آن استفاده کنیم می‌توانیم از Regression استفاده کنیم و در غیر این صورت از classification باید استفاده کنیم.
X می تواند هر چیزی باشد و در نهایت می توانیم آن را به عدد تبدیل کنیم.

x	y
x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4
⋮	⋮



اگر $f(x)$ را $f_\beta(x)$ در نظر بگیریم و داده ها را بصورت $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ هدف ما این است که از روی نمونه های داده شده که D هستند β را بیابیم.

برای این کار سه رویکرد وجود دارد که شامل memorization وpredictionو inference می‌شود که در memorization تعمیم یا generalization وجود ندارد در صورتی که در prediction وجود دارد و تفاوت prediction و inference این است که prediction نگاه black box به موضوع دارد وبه مکانیزم کاری ندارد و مهم این است که برای این ورودی این خروجی را می‌گیریم ولی در inference نگاه white box وجود دارد وفرایند و رابطه ی میان داده های موجود را مورد بررسی قرار میدهد.

حال فرض کنید D که نمونه های ماست از یک توزیع احتمالاتی مانند $p(x, y)$ که در جمعیت وجود داشته است آمده باشند . یک مدل مانند $y = f_\beta(x) + \epsilon$ برای predict کردن در نظر می‌گیریم و باید یک تابع loss برای ارزیابی مدل ارائه شده تعریف کنیم و نمونه ها را به دو قسمت train و test تقسیم می‌کنیم و برای ارزیابی، امید ریاضی تابع loss را بر روی نمونه های train و test محاسبه می‌کنیم اگر این امید ریاضی برای نمونه های train و test نزدیک به هم باشد، یعنی مدل ما توانایی تعمیم خوبی دارد.

به عنوان مثال اگر مدل ما $y = ۵x$ باشد، $l(x, y) = (y - ۵x)^۲$ تابع loss می‌باشد و برای ارزیابی مدل برای تعمیم یافته بودن از $\mathbb{E}((y - ۵x)^۲)$ استفاده می‌کنیم.

همچنین طبق قانون اعداد بزرگ برای محاسبه امید ریاضی بر روی داده ها می‌توانیم از $\mathbb{E}(l) = \frac{۱}{m}\Sigma(l_i)$ استفاده کنیم.

رگرسیون خطی Linear Regression

در جمعیت:

$$Y = \beta_۱^* X + \beta_۰^* + \varepsilon$$

ε از توزیع نرمال $\varepsilon \sim \mathrm{N}(0,\sigma^2)$ پیروی می کند و X از توزیع $P[X]$ می آید.

ε و X مستقل از همند.

$$P[X, \varepsilon] = P[X].P[\varepsilon]$$

$$P[X, Y] = P[Y|X].P[X] \rightarrow P[(\varepsilon = Y - \beta_۱^* X + \beta_۰^*) \mid X]\ P[X]$$

در حد جمعیت می دانیم که چه اتفاقی افتاده است ولی آن را در نظر نمی گیریم و فکر می کنیم که فقط دیتا از جمعیت داریم.

Data: $(X_1, Y_1), (X_2, Y_2)$ ... (X_n, Y_n)

برای مثال:

$$f_\beta(X) = \beta_1 X + \beta_0$$

$$(\hat{\beta}_0, \hat{\beta}_1) \longrightarrow \hat{y} = \hat{\beta}_1 X + \hat{\beta}_0$$

$$l(y, \hat{y}) = |y - \hat{y}|^2$$

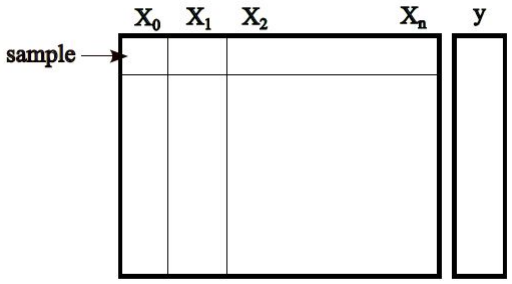
در داده های train:

$$(\hat{\beta}_0, \hat{\beta}_1) = arg\ min_{(\beta_0, \beta_1)} \frac{1}{n} \sum_{i=1}^{n} \left| y - (\beta_1 X + \beta_0) \right|^2$$

پارامترهای $\hat{\beta}_0, \hat{\beta}_1$ بدست آمده را در فرمول زیر قرار می دهیم:
برای test:

$$\longrightarrow \frac{1}{m} \sum_{i=1}^{m} \left| y - (\hat{\beta}_1 X + \hat{\beta}_0) \right|^2$$

پارامترهای $\hat{\beta}_0$ و $\hat{\beta}_1$ از داده تولید شده اند و تابعی از داده هستند. اگر یک دیتاست دیگری داشته باشیم، این پارامترها نیز متفاوت خواهند بود.



$$\hat{y} = f_\beta(X) = \sum_{i=1}^{p} \beta_i X_i + \beta_0 = \beta^T X = X^T \beta$$

$$X = \begin{bmatrix} X_0 \\ \vdots \\ X_p \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$l(y, \hat{y}) = |y - \hat{y}|^2$$

$$L = E[|y - \hat{y}|^۲]$$

آیا می توان loss function را صفر کرد؟

اگر دیتای کمی در نظر بگیریم مثلا p+1 تا feature یعنی p+1 ستون داده داشته باشیم، می توانیم β را طوری بیابیم که تابع loss دقیقا صفر شود اما این خوب نیست چون قابلیت تعمیم پذیری ندارد.
یکی از مسایل مهم شناختن ابعاد است و در این شناخت اشتباهات زیادی رخ می دهد. باید در فرایند training حواسمان به دیتا و پارامترها باشد.
این که کدام در سطح population است و کدام در سطح sample باید دقت کرد.
دیتا را به n, m تقسیم می کنیم و یکی بر روی population است و دیگری بر روی sample.

$$E[|y - \hat{y}|^۲] \Longrightarrow \frac{۱}{n}\sum_{i=۱}^{n}\left|y - \hat{y}\right|^۲$$

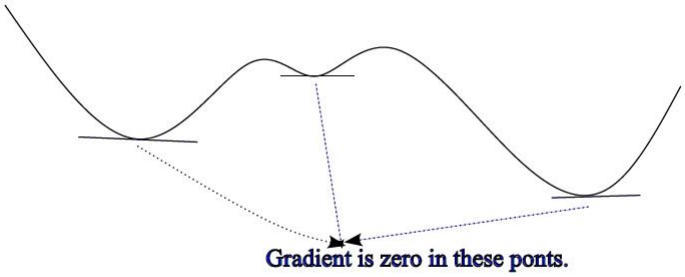
با تخمین از عبارت بالایی β را پیدا می کنیم و از روی آن مدل را ارزیابی می کنیم.

$$L_{test} = \frac{۱}{m}\sum_{i=۱}^{m}\left|y - \hat{y}\right|^۲$$

که m تعداد جمعیت اولیه است.

$$arg\ min_{\hat{\beta}}\ \frac{۱}{n}\sum_{i=۱}^{n}(y_j - X_j^T\beta)^۲ = arg\ min\ g(\beta) \Longrightarrow \nabla g(\beta) = ۰ \Longrightarrow \hat{\beta}$$

با β عبارت بالایی را مینیمم می کنیم. آیا $\hat{\beta}$ که بدست آورده ایم، یک Global Minimum است؟
اگر تابع ما convex باشد، یک مینیمم جهانی دارد. اما اگر non-convex باشد، β لزوما مینیمم جهانی نیست.
مساله رگرسیون خطی، قطعا convex بوده و β قطعا جهانی است.



عوض کردن پی در پی مدل برای به جواب رسیدن اشتباه است. تفکر درست این است که ابتدا فضای مدل را مشخص کنیم. مثلا مشخص کنیم که می خواهیم از مدل Neural Net استفاده کنیم یا Random Forest، بعد β ای را پیدا کنیم که loss کمینه شود. بحث fine tune کردن هم مطرح می شود که باید ببینیم برای مساله چه

مناسب است.

$$\langle \nabla g \cdot \mathbf{d} \rangle = \lim_{\varepsilon \to ۰} \frac{g(\beta + \varepsilon d) - g(\beta))}{\varepsilon} \Longrightarrow$$

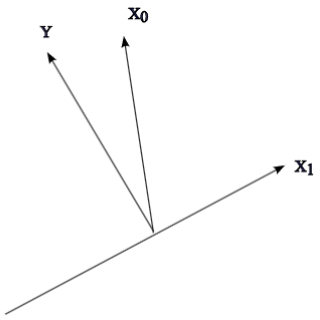
$$\sum_j \Big[(y_j - X_j^T \beta - X_j \varepsilon d)^۲ - (y_j - X_j^T \beta)^۲ \Big] \Longrightarrow$$

$$-۲ \Big[\sum_j (y_j - X_j^T \beta) X_j^T \Big] d \Longrightarrow$$

$$\nabla g(\beta) = \frac{-۲}{n} \Big[\sum_j (y_j - X_j^T \beta) X_j^T \Big] = \frac{-۲}{n} \Big[\sum_j X_j (y_j - \beta^T X_j) \Big] \Longrightarrow \sum_j X_j Y_j = \sum_j X_j X_j^T \beta$$

$X_j X_j^T$ ، ضرب داخلی دو بردار است و بیانگر شباهت بین دو بردار است. و نیز رابطه هر feature با feature های دیگر را نشان می دهد.

اگر حالتی پیش بیاید که ببینیم یک feature خیلی با y شباهت داشته باشد و بقیه ستون ها این شباهت را نداشته باشند، حذف کردن بقیه ستون ها کار اشتباهی است چون ممکن است ستون ها با y, correlation نداشته باشند ولی شاید با همدیگر correlation داشته باشند و این اطلاعات مهمی باشد. مثلا در شکل زیر X_0 با y، correlation دارد اما X_1 ندارد. اگر X_1 را دور بیاندازیم کار اشتباهی است چون X_1 در مورد X_0 به ما اطلاعات میدهد.



مساله Linear Regression بردار ضرایب β با ترکیب خطی ستون‌های ماتریس X، y را می‌سازد(تخمین می‌زند).

$$\Big[X_۰ \quad X_۱ \quad \cdots \quad X_p \Big], \qquad \beta = \Big[\beta_۰ \quad \beta_۱ \quad \cdots \quad \beta_p \Big] \Longrightarrow \hat{y} = X\beta$$

$$\hat{\beta} = \arg\min_\beta \|y - \hat{y}\|^۲ = \|y - X\beta\|^۲ \xrightarrow{\frac{\partial}{\partial \beta}} ۲X^T \Big(y - X\hat{\beta} \Big) = ۰ \longrightarrow \hat{\beta} = \big(X^T X \big)^{-۱} X^T y$$

در نتیجه $\hat{\beta}$ یک تابعی از داده است.

حال فرض کنید یک نمونه جدید به صورت $y^{(\cdot)}, X^{(\cdot)} = \begin{bmatrix} X_{\cdot}^{(\cdot)} \\ \vdots \\ X_p^{(\cdot)} \end{bmatrix}$ داریم. در این صورت:

$$\hat{y}^{(\cdot)} = {X^{(\cdot)}}^T \beta \longrightarrow L = \mathbb{E}\left[(y - \hat{y})^{\text{۲}}\right] = \mathbb{E}\left[\left(y^{(\cdot)} - {X^{(\cdot)}}^T \beta\right)^{\text{۲}}\right]$$

در prediction مقادیر β اهمیتی ندارد و مینیمم بودن امید ریاضی برای ما مهم است.

$$y = \langle \beta^*,\ X \rangle + \epsilon : \qquad \text{Var}(\epsilon) = \sigma^{\text{۲}}$$

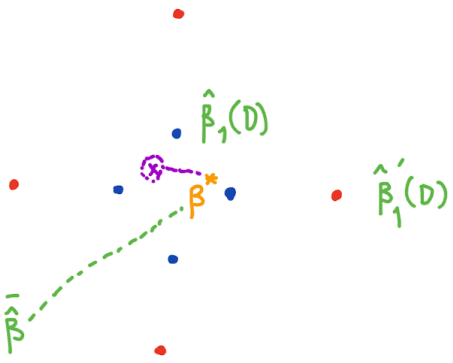
$$\mathbf{L} = \mathbb{E}\left[\left(X^T\beta^* + \epsilon - X^T\hat{\beta}(D)\right)^{\text{۲}}\right] = \mathbb{E}\left[\left\{\left(X^T\left(\beta^* - \mathbb{E}\left(\hat{\beta}\right)\right) - X^T\left(\hat{\beta}(D) - \mathbb{E}\left(\hat{\beta}\right)\right)\right) + \epsilon\right\}^{\text{۲}}\right]$$

در عبارت بالا با توجه به اینکه $\epsilon,\ X$ از یکدیگر مستقل هستند می‌توان مقدار Loss را به صورت زیر نوشته

$$\begin{aligned}\mathbf{L} &= \mathbb{E}\left[\left\{X^T\left(\beta^* - \mathbb{E}\left[\hat{\beta}\right]\right)\right\}^{\text{۲}}\right] + \mathbb{E}\left[\left\{X^T\left(\hat{\beta} - \mathbb{E}\left[\hat{\beta}\right]\right)\right\}^{\text{۲}}\right] + \mathbb{E}\left[\epsilon^{\text{۲}}\right] \\ &= \underbrace{\mathbb{E}_X\left[X^T\left(\beta^* - \mathbb{E}\left[\hat{\beta}\right]\right)\right]^{\text{۲}}}_{\text{bias}^{\text{۲}}} + \underbrace{\mathbb{E}\left[\left\{X^T\left(\hat{\beta} - \mathbb{E}\left[\hat{\beta}\right]\right)\right\}^{\text{۲}}\right]}_{\text{Variance}} + \sigma^{\text{۲}}\end{aligned}$$

Bias-Variance trade off

هدف اصلی Linear Regression این است که بهترین $\hat{\beta}$ را پیدا کنیم. یک روش قدیمی به این صورت است که ابتدا bias و سپس Variance را کم کنیم. اما در روش‌های جدیدتر جمع این دو را کم می‌کنیم.



$$\mathbb{E}\left[\hat{\beta}\right] = \mathbb{E}\left[\left(X^TX\right)^{-1}X^Ty\right] = \mathbb{E}\left[\left(X^TX\right)^{-1}X^T\left(X\beta^* + \epsilon\right)\right] = \mathbb{E}\left[\beta^* + \left(X^TX\right)^{-1}X^T\epsilon\right] = \beta^*$$

با توجه به عبارت بالا داریم که Linear Regression، unbiased است.

- **مدل پیچیده:** در این شرایط فضای جست و جو بزرگتری داریم، شانس رسیدن به β^* بیشتر می‌شود.در نتیجه bias کمتر و variance بیشتر می‌شود.
- **مدل ساده:** در این شرایط فضای جست و جو کوچکتری داریم، شانس رسیدن به β^* کمتر می‌شود.در نتیجه bias بیشتر و variance کمتر می‌شود.
- اگر دیتا به اندازه کافی زیاد نباشد، مدل پیچیدگی کافی را ندارد و شانس کمتری برای رسیدن به جواب وجود دارد و در نهایت مدل ما overfit می‌شود.

فرض کنید که داده‌های ما متعامد باشند

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y = \begin{bmatrix} \frac{1}{\|X_1\|^2} & & \\ & \ddots & \\ & & \frac{1}{\|X_p\|^2} \end{bmatrix} \begin{bmatrix} \langle X_1,\ y\rangle \\ \vdots \\ \langle X_p,\ y\rangle \end{bmatrix} \Longrightarrow \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \frac{1}{\|X_1\|^2}\langle X_1,\ y\rangle \\ \vdots \\ \frac{1}{\|X_p\|^2}\langle X_p,\ y\rangle \end{bmatrix}$$

در نتیجه β_p تنها مربوط به دیتای ستون p است.
اگر دیتاها متعامد به یکدیگر نباشند نیز می‌توانیم با استفاده از الگوریتم Gram-Schmidt داده‌ها را متعامد کنیم و مقادیر β را بدست آوریم.

استنتاج

در بخش های قبلی دیدیم که هدف برای یک prediction خوب کم تر شدن امید ریاضی loss function است.

$$y = < \beta, x > + \epsilon$$

$$l(y, \hat{y})$$

$$\beta(D)$$

$$\mathbb{E}(l(y, \hat{y})) \downarrow ?$$

اما در استنتاج، سوال هایی وجود دارند که قصد داریم به آن ها پاسخ بدهیم:

۱. آیا مقادیر زیر صفر هستند؟ (مثال: دومین سطر به ما نشان می دهد که آیا ستون اول داده با y رابطه ای داشته و informative بوده است یا خیر.)

$$\beta_1^* = 0\ ?$$

$$\beta_2^* = 0\ ?$$

$$(\beta_1^*,\ \beta_2^*,\ \beta_3^*) = 0\ ?$$

۲. آیا مدل خطی مناسب و درست می باشد؟

۳. آیا x_i ها از یک توزیع آمده اند؟

۴. آیا ϵ_i ها از هم مستقل هستند؟

۵. آیا واقعا x_i ها و ϵ_i ها با یکدیگر رابطه ای ندارند؟

در ادامه در ۵ بخش به سوالات مذکور پاسخ می دهیم.

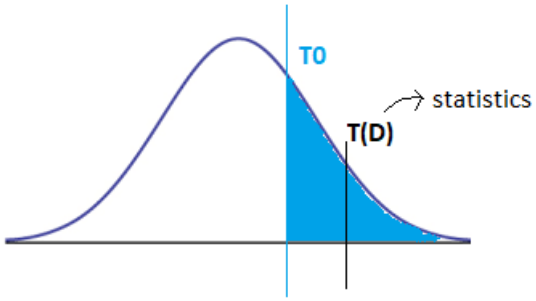
۱

می خواهیم پاسخ این سوال را پیدا کنیم که آیا θ^* برابر صفر هست یا خیر.
می دانیم که بین θ^* و داده رابطه وجود دارد:

$$\theta^* \leftrightarrow D$$

T تابعی از داده ها با توزیعی مستقل از θ^* است که با کمک آن قصد داریم به سوال مورد نظر پاسخ بدهیم. بر اساس این که T(D) از یک threshold مشخص بیشتر یا کمتر باشد می توانیم تصمیم گیری خود را انجام دهیم.پس پاسخ سوال ما به نتیجه مقایسه زیر بررسی دارد:

$$T(D) \lessgtr T.$$



شکل ۲: در مثال مورد نظر خروجی تابع در بازه مورد نظر قرار گرفته و بنابراین فرض اولیه ($\theta^* = ۰$) برقرار است.

برای فهم بهتر به مثال زیر توجه کنید:

مثال

فرض کنید توزیع و داده های زیر را در دست داریم:

$$X \sim N(\mu, \delta^۲)$$

$$D(x_۱, x_۲, ..., x_n)$$

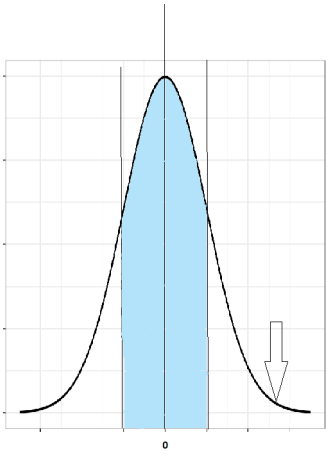
می خواهیم بررسی کنیم که اگر δ^2 برابر ۱ باشد، آیا می توان نتیجه گرفت μ برابر ۰ است:

$$\delta^2 = 2 \Rightarrow \mu = 0 \ ?$$

داریم:

$$T(D) = \frac{\sigma x_i}{n} \Rightarrow T(D) \sim N(\mu.\frac{1}{n}) \Rightarrow N(0.\frac{1}{n})$$

اگر تست را تعداد بسیار بالا تکرار کنیم مشاهده می کنیم که μ حول و اطراف ۰ قرار نمی گیرد. (فرض می کنیم بازه اطمینان ما ۹۹ درصد است.) بنابراین نتیجه گرفته می شود که فرض $\mu = 0$ نادرست بوده است.



شکل ۳: داده در بازه مد نظر قرار نمی گیرد و بنابراین فرض ما صحیح نیست.

اما فرض کنید که ما مقدار δ^2 را در اختیار نداشتیم. در این صورت آماره ما به δ^2 وابسته می شد:

$$T(D) = \frac{\sigma x_i}{n\delta(\hat{D})} \rightarrow \frac{\sigma(x_i - \bar{x})^2}{n-1}$$

در این صورت توزیع به دست آمده نه از توزیع نرمال، بلکه از توزیع T-Student پیروی می کند.

پس به طور خلاصه، اگر توزیعی داشته باشیم:

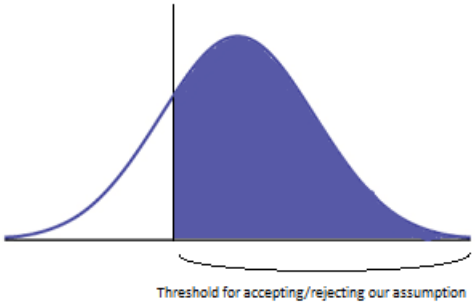
$$p_{\theta_1,\theta_2}(x)$$

$$D = (x_1, x_2, ..., x_n)$$

برای به دست آوردن پاسخ پرسشی مشابه:

$$\theta_1 = 0 \ ?$$

آماره T(D) را روی داده تعریف می کنیم و با استفاده از توزیع آن به پاسخ سوال می رسیم. با توجه به این نکته می توان دریافت که برای ما بسیار راحت تر خواهد بود اگر آماره دارای توزیع مشخصی باشد.
توجه: T(D) زمانی توزیع خوبی خواهد بود که از θ_2 مستقل باشد.



شکل ۴: توزیع T(D)

در ادامه بحث، فرض کنید قصد داریم به سوال اول پاسخ بدهیم و بررسی کنیم آیا مثلا مقدار β_1^* برای با y رابطه ای داشته یا خیر.

جدول داده زیر را درنظر بگیرید. فرض کنید همه مقادیر y برابر میانگین مقدار آن یعنی $\bar{y}$ بود. (تمامی مقدار ها ثابت.) درآن صورت loss برابر ۰ می شد و دیگر برای محاسبه y نیازی به β ها و بررسی داده های x نبود.

اما در شرایطی که چنین شرطی برقرار نباشد، ما نیازمند بررسی داده های x می باشیم. در این صورت دو حالت وجود دارد. اگر با استفاده از داده های x امکان توصیف یا describe کردن y به صورت کامل وجود داشته باشد، مجددا مقدار loss برابر ۰ خواهد بود. اما اگر چنین نباشد، یعنی y به مقدار دیگری به جز x وابسته است که آن را noise می نامیم. اصل کار ما این است که بتوانیم variability در y را با داده describe کنیم.

به مثال زیر توجه کنید.

قصد داریم با استفاده از ستون های مذکور رابطه بین x و y را به دست بیاوریم. اگر مقدار ستونی را برابر ۰ قرار دهیم و مشاهده کنیم که تغییری در تخمین ما از y توسط x ها به وجود نمی آید، متوجه می شویم که این ستون تاثیر چندانی ندارد. informative) نمی باشد.)

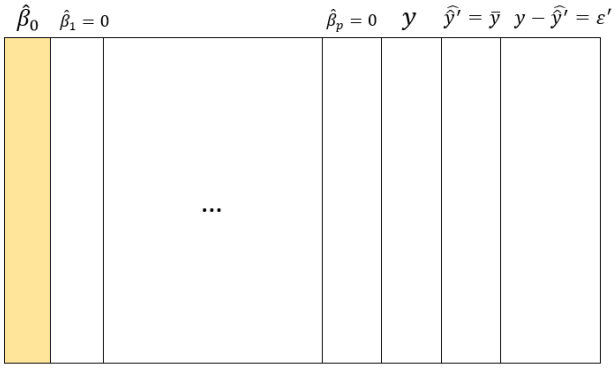
مدل زیر را در نظر بگیرید:

$\hat{\beta}_0$	$\hat{\beta}_1$		$\hat{\beta}_p$	y	$\hat{y}$	$y-\hat{y}=\hat{\varepsilon}$
		...				

شکل ۵: داده ها

پس ابتدا با فرض این که هیچ کدام از β ها صفر نیستند محاسبات را انجام می دهیم و $\hat{y}$ و $\hat{\epsilon}$ به دست می آید.

سپس برای بررسی هر β مقدار آن را نگه داشته و باقی را صفر قرار می دهیم، و به این ترتیب مقادیر $\hat{y}_1$ و $\hat{\epsilon}_1$ به دست می آیند.



شکل ۶:

داریم:

$$RSS = \sum_{n}^{i=۱} (y - \hat{y})^۲$$

$$RSS_۰ = \sum_{n}^{i=۱} (y - \hat{y'})^۲ = \sum_{n}^{i=۱} (y - \bar{y})^۲ = \sum_{n}^{i=۱} \hat{\epsilon'}^۲$$

برای یکی کردن scale ها RSS را به Dof یا degree-of-freedom تقسیم میکنیم.
اگر همه داده ها از هم مستقل بودند، DoF برابر n می شد، اما چون $\hat{\beta}$ ها را از روی تمام داده ها ساخته ایم، dependency داریم. پس درنهایت:

$$RSS = \sum_{n}^{i=۱} \frac{(y - \hat{y})^۲}{n - (p + ۱)}$$

$$RSS_۰ = \sum_{n}^{i=۱} \frac{(y - \hat{y'})^۲}{n - ۱}$$

با مقایسه نسبت RSS و RSS۰ میتوان به سوال های مورد نظر و به طور کلی سری اول سوال ها پاسخ داد.

۲

اکنون به پاسخ سوال دوم می پردازیم.
ما فرض کرده ایم که مدل خطی باشد. یعنی ما مقدار $\hat{y}$ را تنها با استفاده از x به دست آورده ایم. این گزاره زمانی درست است که مدل تنها به x وابستگی داشته باشد. اگر چنین نباشد مدل واقعی دیگر خطی نیست.
فرض کنیم مدل واقعی به صورت زیر باشد:

$$y = f(x) + \epsilon$$

و ما تخمین خطی زیر را روی تابع f انجام داده باشیم:

$$f(x) =< x,\ \beta >$$

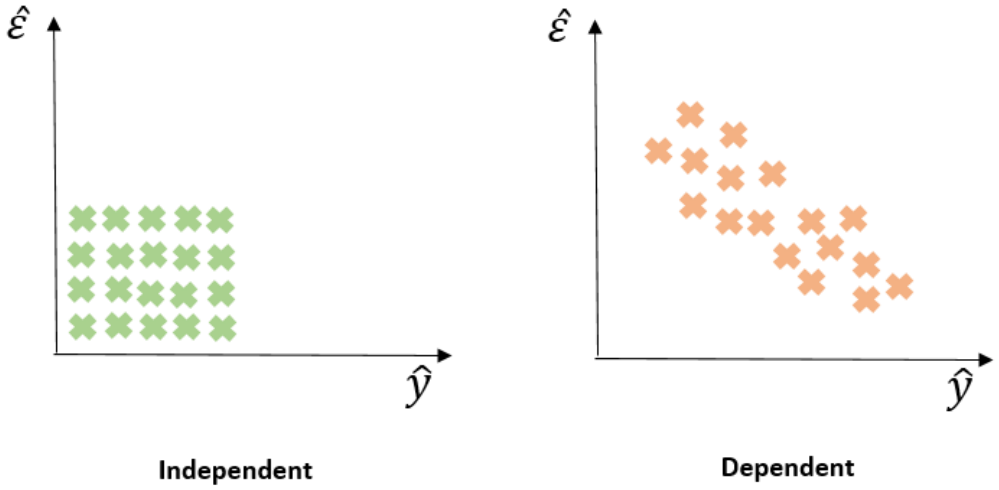
می توان نوشت:

$$y = f(x) - < x,\ \beta > + < x,\ \beta > +\epsilon$$

پس:

$$y - < x,\ \beta >= g(x) + \epsilon \qquad (g(x) = f(x) - < x,\ \beta >)$$

$$\Rightarrow \quad \hat{\epsilon} = g(x) + \epsilon$$

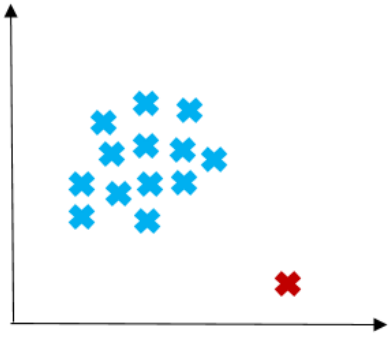


شکل ۷: رابطه $\hat{\epsilon}$ و $\hat{y}$

اگر مقدار g(x) برابر صفر باشد یعنی مدل خطی بوده است. در غیر این صورت مدل خطی نبوده و $\hat{y}$ و $\hat{\epsilon}$ از یکدیگر مستقل نبوده اند.

۳

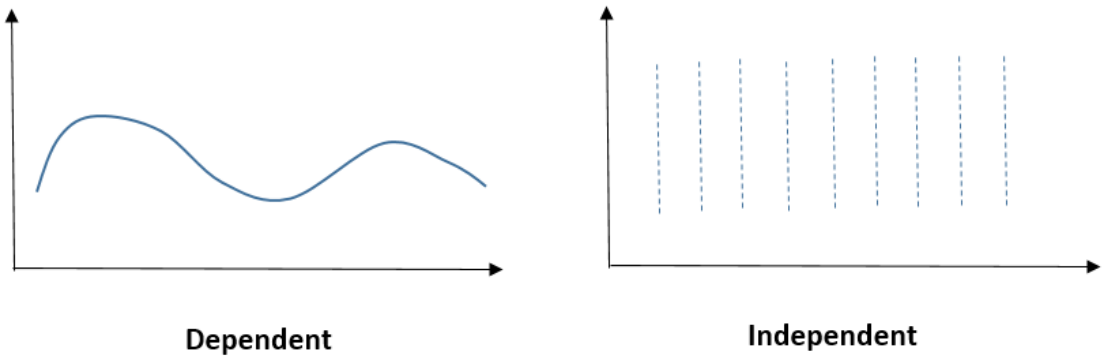
سوال بعدی این است که آیا x_i ها از یک توزیع آمده اند یا خیر.
مثلا فرض کنید تستی انجام شده است و یکی از داده ها بسیار پرت از از سایر مقادیر است. یا این دیتای پرت valid هست و از یک توزیع دیگر امده و یا valid نیست و مثلا اندازه گیری غلط بوده است. یک راه این است که برای سایر داد ها distribution ای فرض کنیم و سپس احتمال این که این داده عضوی از این توزیع باشد را محاسبه کرده و اگر این احتمال خیلی پایین بود از این داده صرف نظر کنیم.
در چنین شرایطی median بسیار کاربری است. median مقدار وسطی بعد از sort داده ها می باشد. (از مزایای median این است که robust می باشد.)



شکل ۸: داده پرت

۴

برای بستگی مستقل بودن یا نبودن ϵ_i ها از یکدیگر باید رابطه آن ها را با زمان بررسی کنیم.اگر پراکندگی آن ها در طول زمان تغییر نکند یعنی ϵ_i ها از یکدیگر مستقل هستند.(اگر مستقل نباشند یعنی مدل خوبی نداشته ایم.)



شکل ۹: ارتباط ϵ_iها