linear regression

مدرس :الهام حیدری

linear regression

رگرسیون خطی یکی از ساده ترین و پرکاربرد ترین الگوریتم‌های یادگیری ماشین است که برای پیش‌بینی یک متغیر پیوسته (مانند قیمت خانه، دمای هوا) بر اساس یک یا چند متغیر مستقل (مانند مساحت خانه، زمان) استفاده می شود. معمولاً روی داده‌های پیوسته کار می‌کندیک رابطه خطی بین متغیر مستقل و متغیر وابسته را فرض می کند و هدف آن یافتن بهترین خطی است که این رابطه را توصیف می کند

رگرسیون خطی

در یک رگرسیون خطی ساده، یک متغیر مستقل و یک متغیر وابسته وجود دارد. مدل شیب و خط برازش را تخمین می زند که نشان دهنده رابطه بین متغیرها است. شیب نشان دهنده تغییر در متغیر وابسته برای هر تغییر در متغیر مستقل است

رگرسیون خطی برای تحلیل پیش‌بینی در یادگیری ماشین استفاده می‌شود. رگرسیون خطی رابطه خطی بین متغیر مستقل (پیش بینی کننده) یعنی محور Xو متغیر وابسته (خروجی) یعنی محور Yکه رگرسیون خطی نامیده می شود را نشان می دهد

رگرسیون خطی

چرا رگرسیون خطی اهمیت دارد؟

- **سادگی و تفسیرپذیری:** مدل رگرسیون خطی بسیار ساده است و به راحتی می‌توان آن را تفسیر کرد. هر ضریب در معادله رگرسیون نشان می‌دهد که با تغییر یک واحد در متغیر مستقل، چقدر متغیر وابسته تغییر می‌کند.
- **کاربرد گسترده:** رگرسیون خطی در بسیاری از زمینه‌ها مانند اقتصاد، مالی، مهندسی و علوم اجتماعی کاربرد دارد.
- **مبنای الگوریتم‌های پیچیده‌تر:** بسیاری از الگوریتم‌های پیچیده‌تر یادگیری ماشین بر اساس مفاهیم رگرسیون خطی ساخته شده‌اند.

رگرسیون خطی

هدف از رگرسیون خطی رسیدن به یک معادله‌ی خطی مناسب با کمترین خطاست.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n + \varepsilon$$

$\beta_1, \beta_2 x_2, \beta_3$: شیب

$x_1, x_2, x_3 + \ldots + x_n$: متغیر مستقل

β_0 : ضریب ثابت (Intercept)

y : متغیر وابسته است.

ε :عبارت خطا

رگرسیون خطی

بتای صفر β_0 در رگرسیون خطی، به عنوان عرض از مبدأ شناخته می‌شود و نشان‌دهنده مقدار متغیر وابسته Y زمانی است که تمام متغیرهای مستقل برابر با صفر باشند. به عبارت ساده‌تر، بتای صفر نقطه تلاقی خط رگرسیون با محور عمودی (محور Y) را مشخص می‌کند.

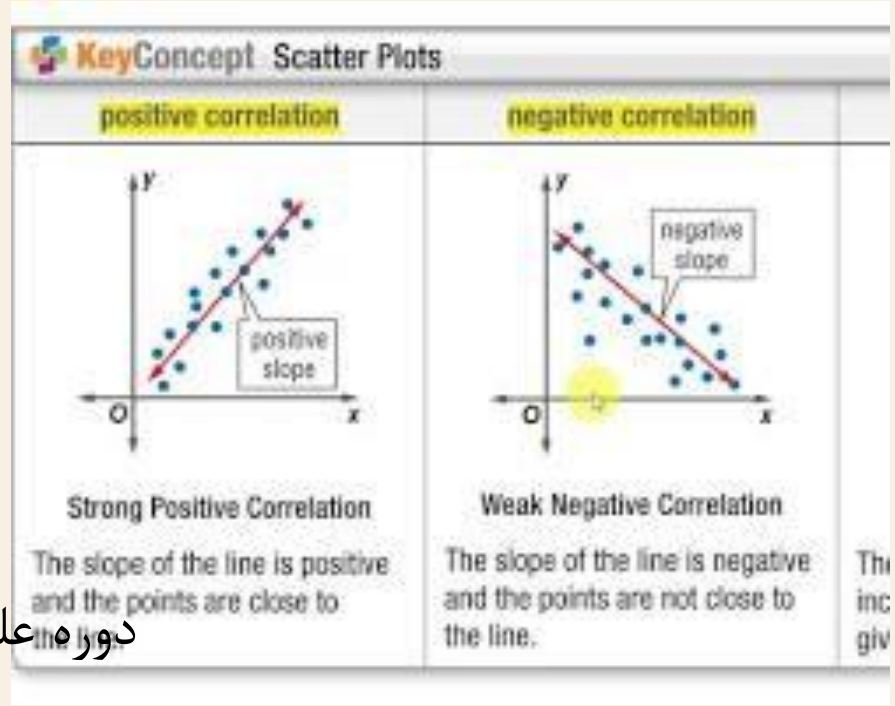
اهمیت بتای صفر

- تفسیر مدل: بتای صفر به ما کمک می‌کند تا مدل رگرسیون را بهتر درک کنیم. برای مثال، اگر بتای صفر مثبت باشد، به این معنی است که حتی اگر تمام متغیرهای مستقل صفر باشند، متغیر وابسته همچنان مقداری مثبت خواهد داشت.
- پیش‌بینی: با داشتن مقدار بتای صفر، می‌توانیم مقدار متغیر وابسته را برای مقادیر صفر متغیرهای مستقل پیش‌بینی کنیم.

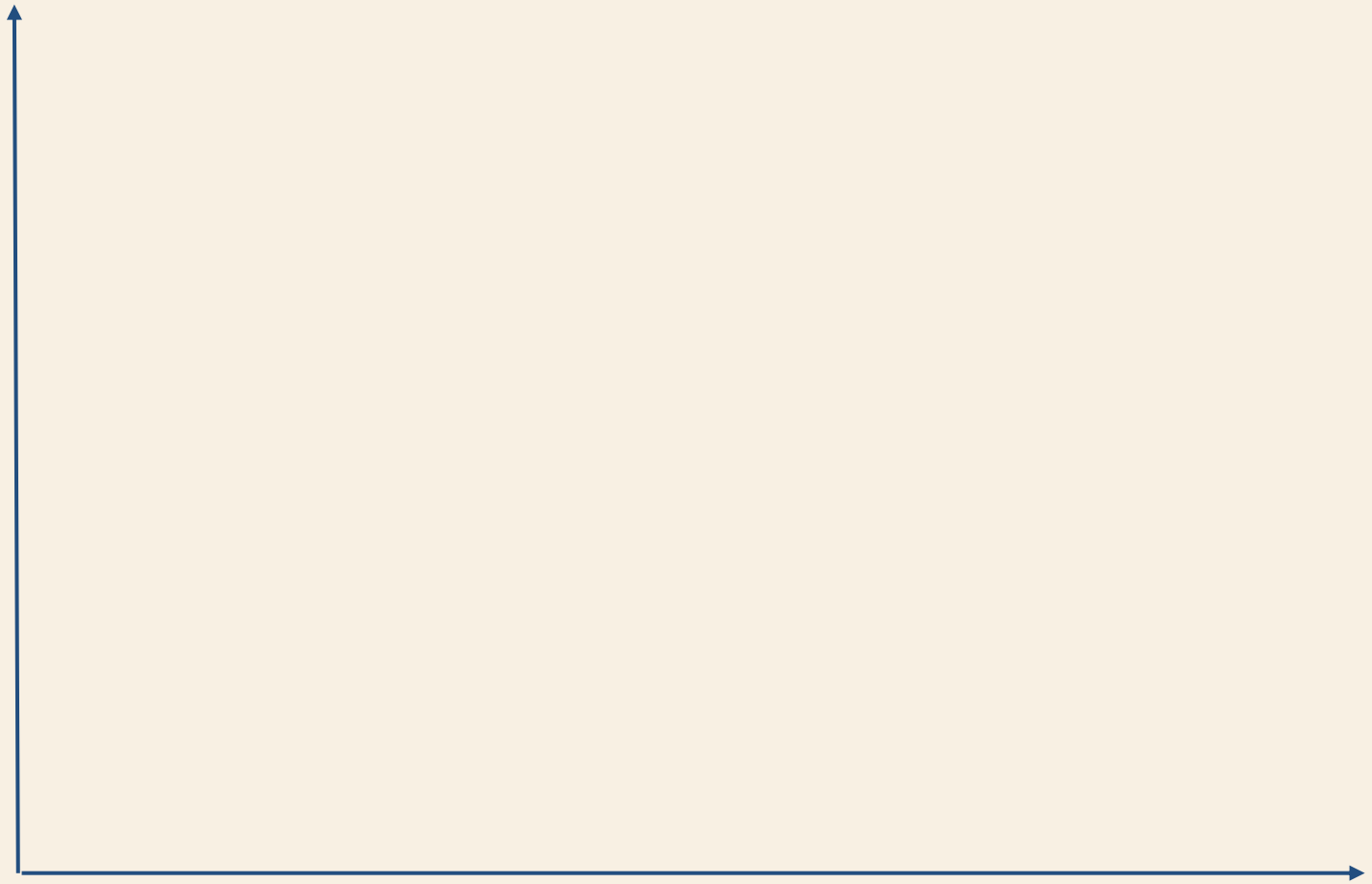
رگرسیون خطی

بتای ۱ در رگرسیون خطی: شیب خط و میزان تاثیر متغیر مستقل

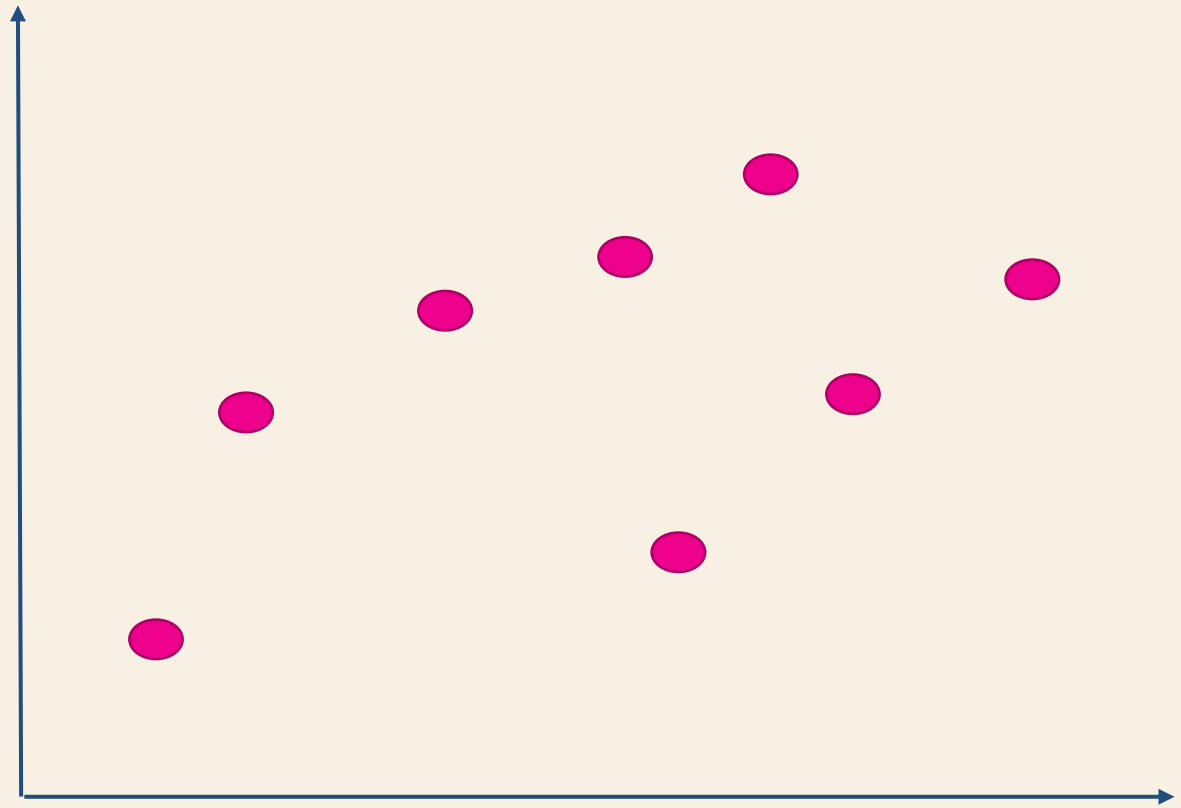
β_1 در مدل رگرسیون خطی، ضریبی است که به متغیر مستقل Xمرتبط است. این ضریب نشان‌دهنده شیب خط رگرسیون و بیانگر میزان تغییری است که در متغیر وابسته Yبه ازای هر واحد تغییر در متغیر مستقل رخ می‌دهد.

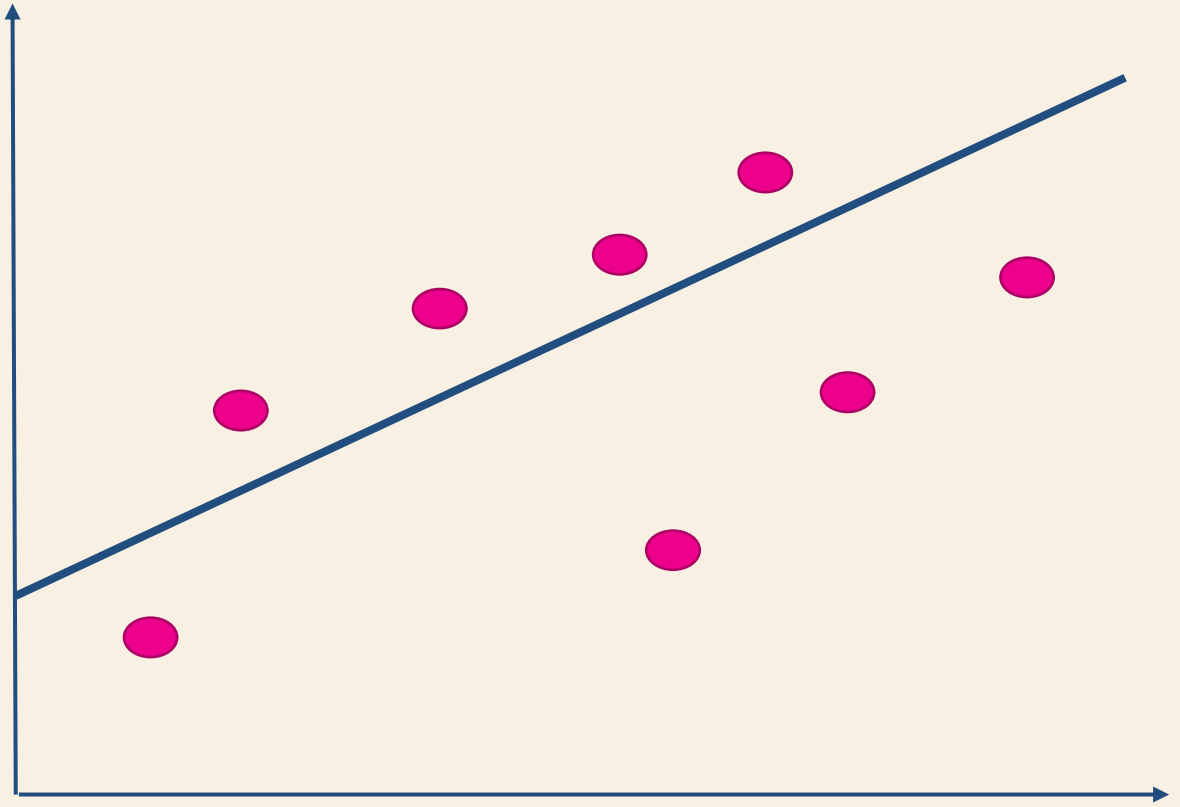


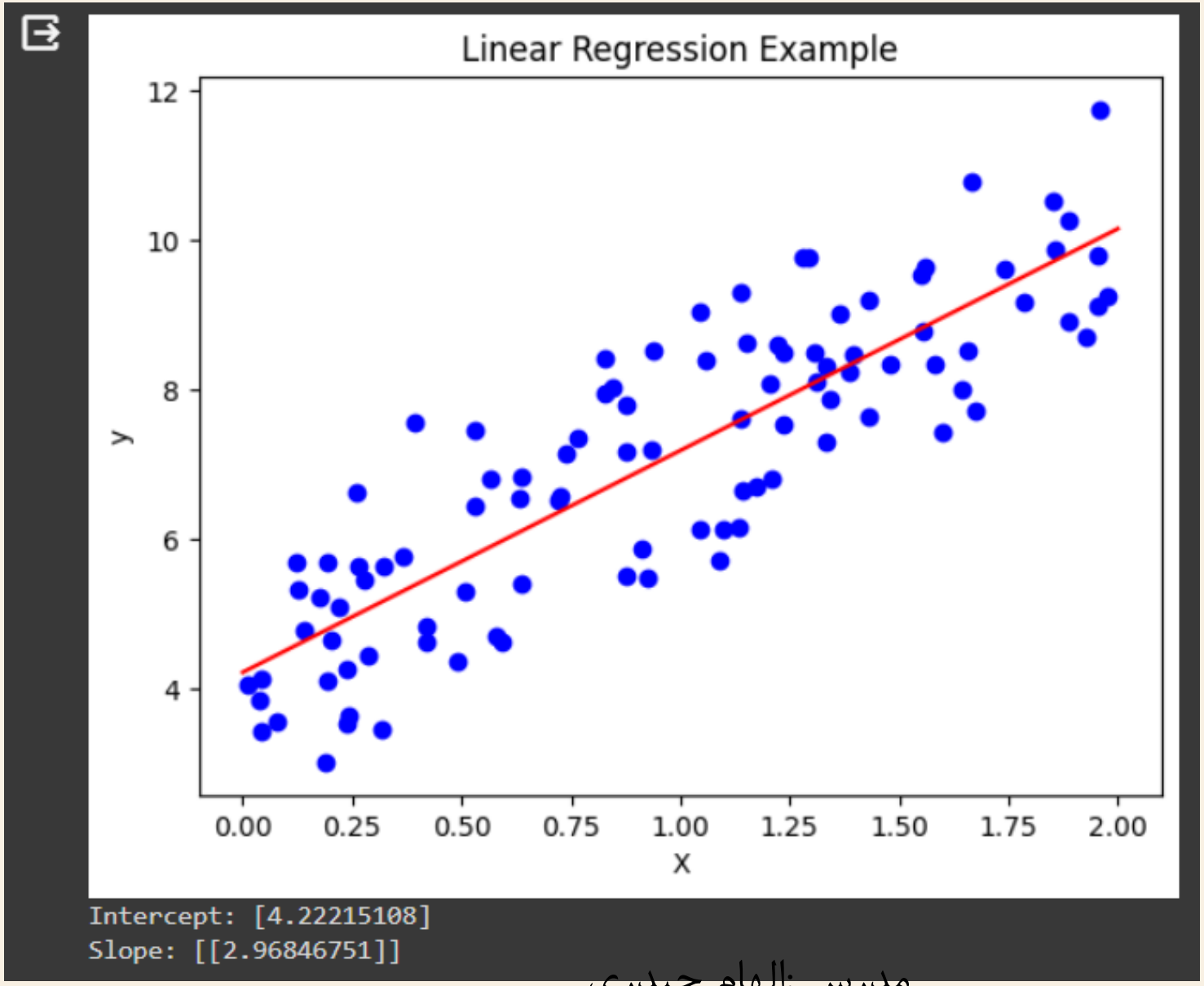
رگرسیون خطی



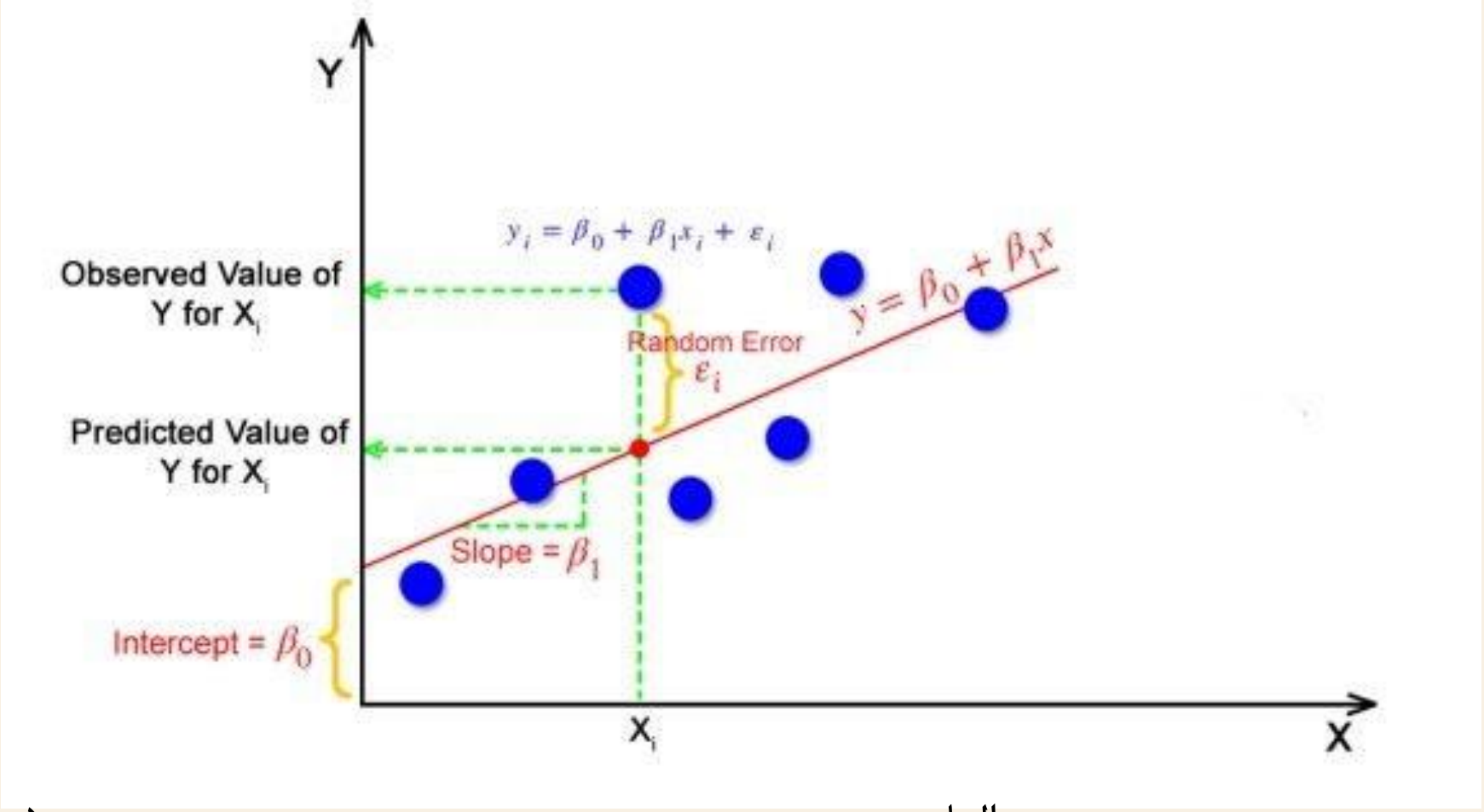
رگرسیون خطی






Linear Regression Example
12
10
8
6
4
y
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
X
Intercept: [4.22215108]
Slope: [[2.96846751]]

رگرسیون خطی


Y
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Observed Value of
Y for X_i
Random Error
ε_i
$y = \beta_0 + \beta_1 x$
Predicted Value of
Y for X_i
Slope = β_1
Intercept = β_0
X_i
X

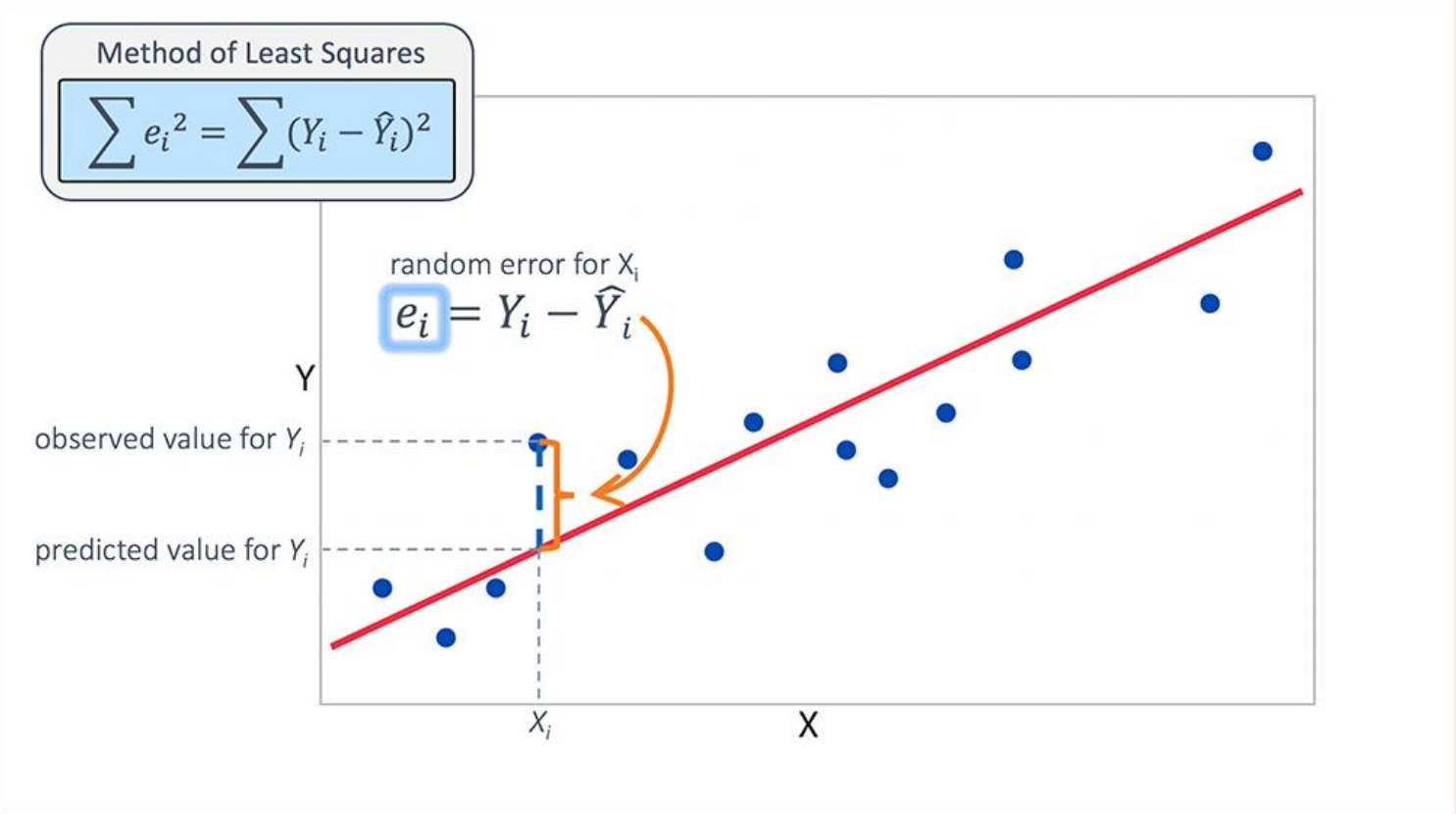
بهترین خط

هدف از الگوریتم رگرسیون خطی بدست آوردن بهترین مقادیر برای B0و B1 برای یافتن بهترین خط مناسب است.
بهترین خط مناسب خطی است که کمترین خطا را داشته باشد به این معنی که خطا بین مقادیر پیش بینی شده و مقادیر واقعی باید حداقل باشد.

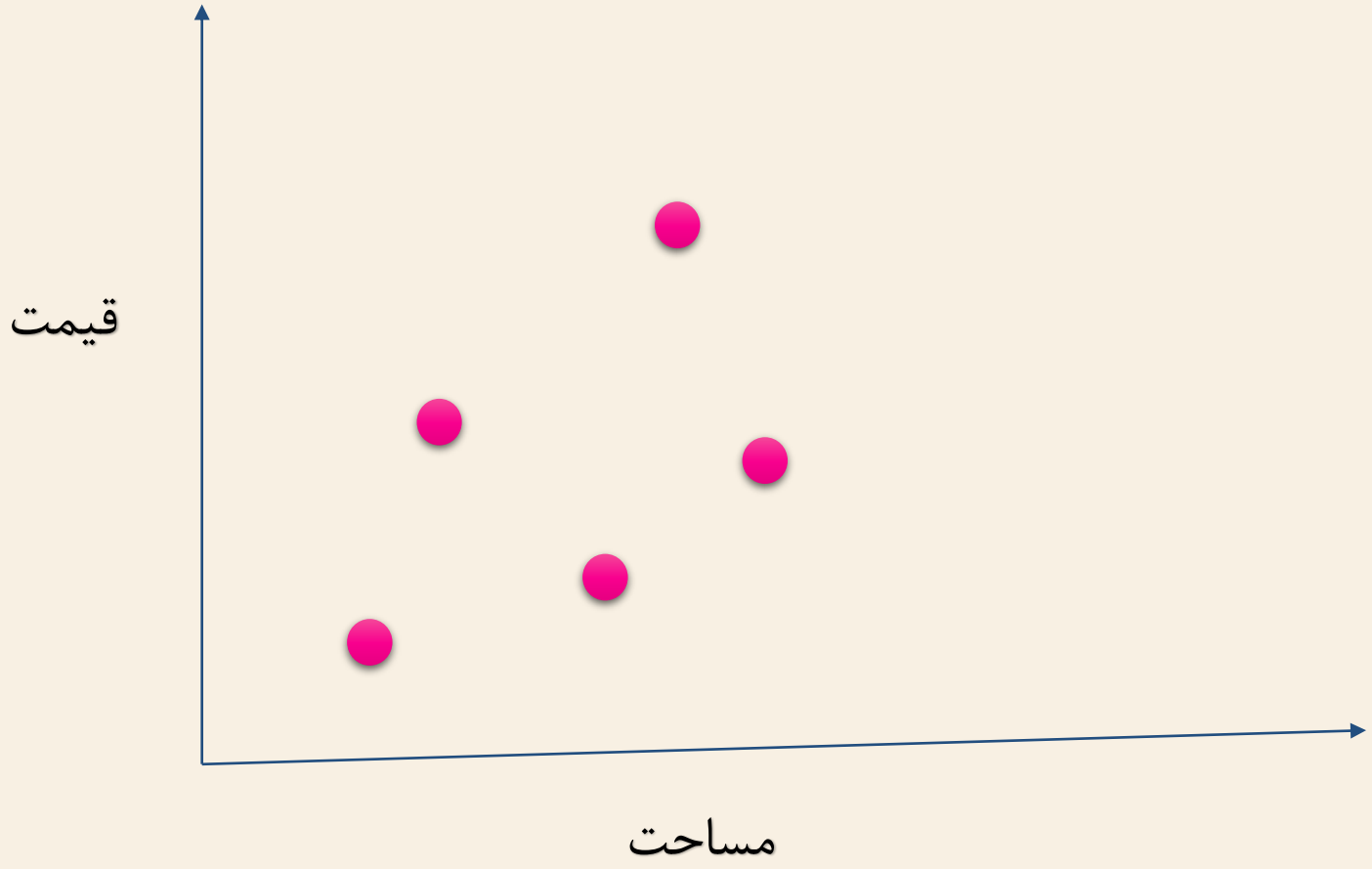
در رگرسیون، تفاوت بین مقدار مشاهده شده متغیر وابسته yiو مقدار پیش بینی شده ($y_{predicted}$) خطا نامیده می شود.

$$\varepsilon_i = y_{predicted} - y_i$$

least squares error



پیش بینی قیمت خانه



رگرسیون خطی روشی است برای پیدا کردن بهترین خط مستقیم که از میان نقاط داده‌ای ما عبور کند. این خط، رابطه بین متغیر وابسته (آنچه می‌خواهیم پیش‌بینی کنیم) و متغیر مستقل (آنچه برای پیش‌بینی استفاده می‌کنیم) را نشان می‌دهد.

تصور کنید:

- **محور افقی (:(X**مقدار متغیر مستقل (مثلاً مساحت خانه)
- **محور عمودی (:(Y**مقدار متغیر وابسته (مثلاً قیمت خانه)
- **هر نقطه روی نمودار:** یک خانه با مساحت و قیمت مشخص

رگرسیون خطی چه می‌کند؟

- **خط بهترین برازش:** یک خط مستقیم رسم می‌کند که به طور متوسط، کمترین فاصله را با تمام نقاط داده داشته باشد.
- **شیب خط:** نشان می‌دهد که با افزایش یک واحد در متغیر مستقل، چقدر متغیر وابسته تغییر می‌کند.
- **قطع محورهای مختصات:** نقطه‌ای که خط مستقیم محور عمودی را قطع می‌کند، مقدار متغیر وابسته را زمانی نشان می‌دهد که متغیر مستقل صفر باشد.

چرا این مهم است؟

پیش‌بینی: با داشتن این خط، می‌توانیم برای خانه‌هایی با مساحت مشخص، قیمت تقریبی را پیش‌بینی کنیم.

درک رابطه: شیب خط به ما می‌گوید که آیا رابطه بین مساحت و قیمت مستقیم است (با افزایش مساحت، قیمت هم افزایش می‌یابد) یا معکوس (با افزایش مساحت، قیمت کاهش می‌یابد).

سادگی: رگرسیون خطی یکی از ساده‌ترین روش‌های مدل‌سازی است و تفسیر نتایج آن آسان است.

رگرسیون

پراکندگی نقاط: اگر نقاط داده‌ها به طور کامل روی خط رگرسیون قرار بگیرند، به این معنی است که رابطه بین دو متغیر کاملاً خطی و دقیق است. اما در واقعیت، همیشه مقداری خطا وجود دارد و نقاط داده‌ها به طور کامل روی خط قرار نمی‌گیرند.

کاربردهای عملی:

پیش‌بینی قیمت: با داشتن مساحت یک خانه جدید، می‌توانیم با استفاده از معادله خط رگرسیون، قیمت تقریبی آن را پیش‌بینی کنیم.

تحلیل بازار: با بررسی شیب خط رگرسیون، می‌توانیم به تغییرات قیمت مسکن در بازار پی ببریم.

تصمیم‌گیری: شرکت‌های املاک می‌توانند از مدل رگرسیون برای تعیین قیمت مناسب برای فروش خانه‌ها استفاده کنند.

محدودیت‌ها:

رابطه خطی: رگرسیون خطی فقط برای روابط خطی بین متغیرها مناسب است. اگر رابطه بین متغیرها غیرخطی باشد، باید از روش‌های دیگری استفاده کرد.

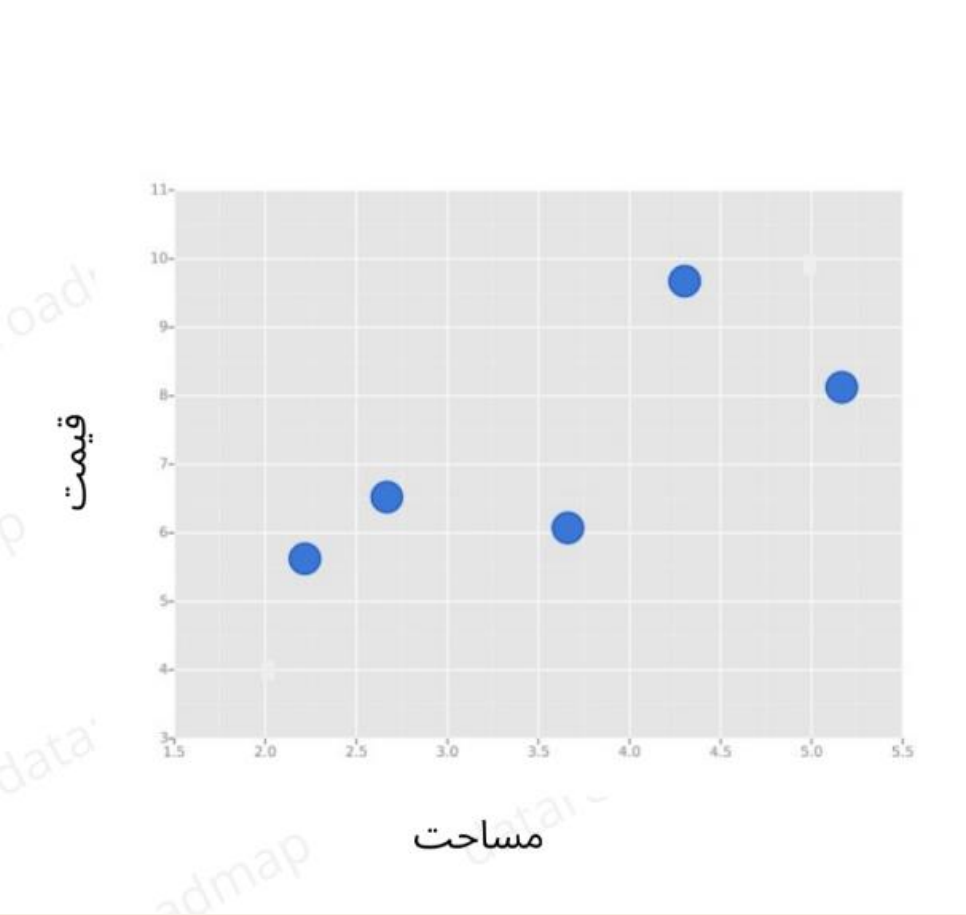
داده‌های پرت: داده‌های پرت می‌توانند تاثیر زیادی بر نتایج رگرسیون خطی بگذارند.

فرضیات: رگرسیون خطی تعدادی فرض دارد که باید رعایت شوند.

دوره علم داده

مدرس :الهام حیدری

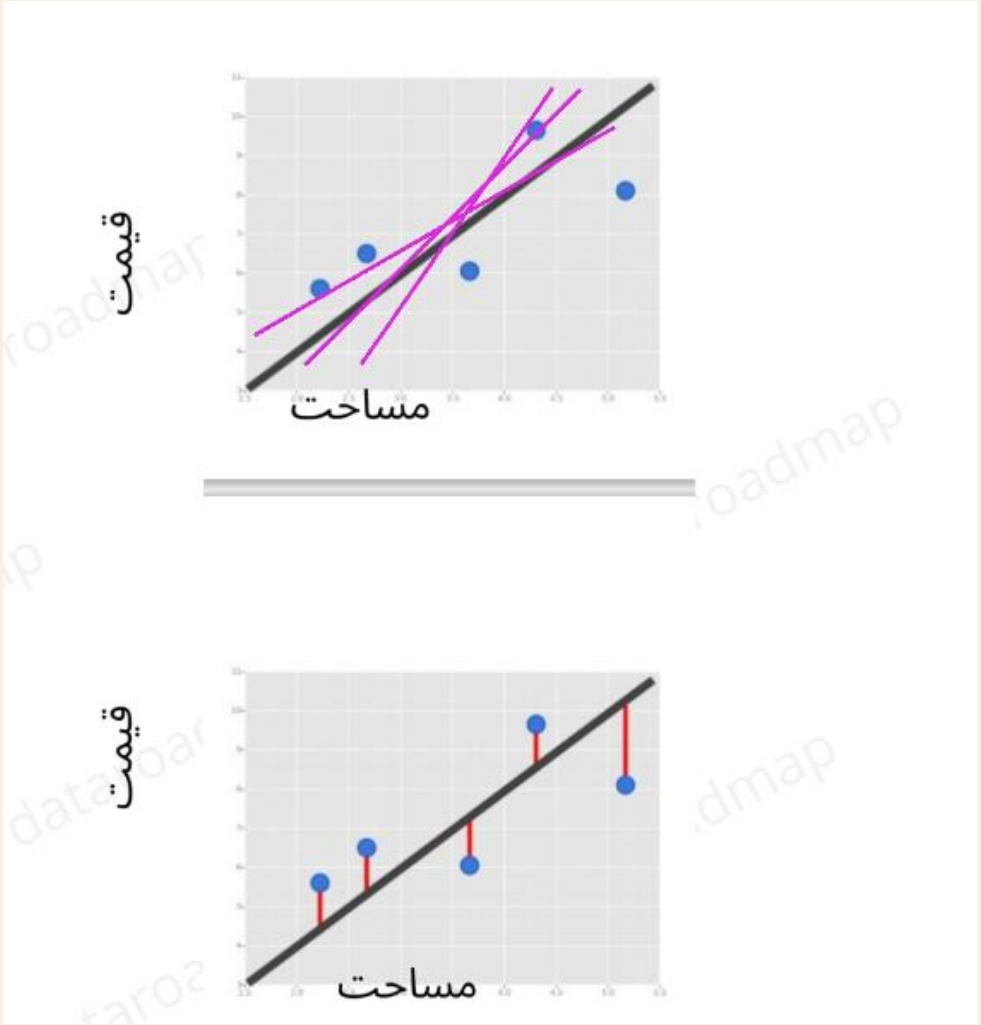
پیش بینی قیمت خانه



مدرس :الهام حیدری

دوره علم داده

پیش بینی قیمت خانه



مدرس :الهام حیدری

دوره علم داده

Regression Evaluation Metrics

Here are three common evaluation metrics for regression problems:

Mean Absolute Error (MAE) is the mean of the absolute value of the errors:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error (MSE) is the mean of the squared errors:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Comparing these metrics:

- **MAE** is the easiest to understand, because it's the average error.
- **MSE** is more popular than MAE, because MSE "punishes" larger errors, which tends to be useful in the real world.
- **RMSE** is even more popular than MSE, because RMSE is interpretable in the "y" units.

All of these are **loss functions**, because we want to minimize them.

پس از رسم خط رگرسیون، سوال مهمی که پیش می‌آید این است که چقدر می‌توان به این خط اعتماد کرد؟ به عبارت دیگر، چقدر این خط می‌تواند داده‌های ما را به خوبی توضیح دهد؟ برای پاسخ به این سوال، از **معیارهای ارزیابی** استفاده می‌کنیم. این معیارها به ما کمک می‌کنند تا کیفیت مدل رگرسیون خود را ارزیابی کنیم.

معیارهای اصلی ارزیابی

1.ضریب تعیین R-squared:

1. این ضریب نشان می‌دهد که چه مقدار از تغییرات متغیر وابسته Yتوسط مدل رگرسیون توضیح داده شده است.
2. مقدار R-squaredبین ۰ تا ۱ متغیر است:
   1. اگر R-squaredبرابر با ۱ باشد، مدل تمام تغییرات Yرا توضیح می‌دهد و یک برازش کامل است.
   2. اگر R-squaredبرابر با ۰ باشد، مدل هیچ تغییری در Yرا توضیح نمی‌دهد.
3. هرچه مقدار R-squaredبه ۱ نزدیک‌تر باشد، مدل بهتر است.

میانگین مربعات خطا MSE:

- MSEمیانگین مربع تفاوت بین مقادیر واقعی Yو مقادیر پیش‌بینی شده توسط مدل را نشان می‌دهد.
- مقدار کمتر MSEنشان‌دهنده برازش بهتر مدل است.

جذر میانگین مربعات خطا RMSE:

- RMSEجذر MSEاست و واحد آن همان واحد متغیر وابسته است.
- RMSEبه ما می‌گوید که به طور متوسط، پیش‌بینی‌های مدل چقدر از مقادیر واقعی فاصله دارند

تفسیر معیارها

میانگین قدر مطلق خطا:MAE

- **MAE**میانگین قدر مطلق تفاوت بین مقادیر واقعی و پیش بینی شده است.
- **MAE**به ما یک تخمین ساده از خطای متوسط مدل می دهد.

R-squared:به ما می‌گوید که مدل چقدر از تغییرات داده را توضیح می‌دهد.

MSEو **RMSE:**به ما می‌گویند که به طور متوسط، پیش‌بینی‌های مدل چقدر دقیق هستند.

MAE:به ما یک تخمین ساده از خطای متوسط مدل می‌دهد.

انتخاب بهترین معیار

انتخاب بهترین معیار به هدف تحلیل شما بستگی دارد. برای مثال:

اگر می خواهید بدانید که مدل شما چقدر از تغییرات داده را توضیح می دهد، R-squaredمناسب تر است.

اگر می خواهید خطای متوسط پیش‌بینی های مدل را بدانید، MSEیا RMSEمناسب تر هستند.

اگر می خواهید یک تخمین ساده از خطای متوسط داشته باشید، MAEمناسب تر است.

مثال

فرض کنید می‌خواهیم رابطه بین مساحت یک خانه و قیمت آن را مدل‌سازی کنیم. پس از انجام رگرسیون خطی، نرم‌افزار به ما خروجی زیر را می‌دهد:

مقدار	معیار
0.85	R-squared
100	MSE
10	RMSE
8	MAE

این نتایج نشان می‌دهند که:

۸۵٪ از تغییرات قیمت خانه توسط مدل توضیح داده شده است.

به طور متوسط، پیش‌بینی‌های مدل با مقدار واقعی قیمت خانه ۱۰ واحد پولی اختلاف دارند.

به طور متوسط، قدر مطلق خطای پیش‌بینی‌ها ۸ واحد پولی است.

الهام حیدری

Linkdin:heidari-ai
Instagaram:heidari_ai