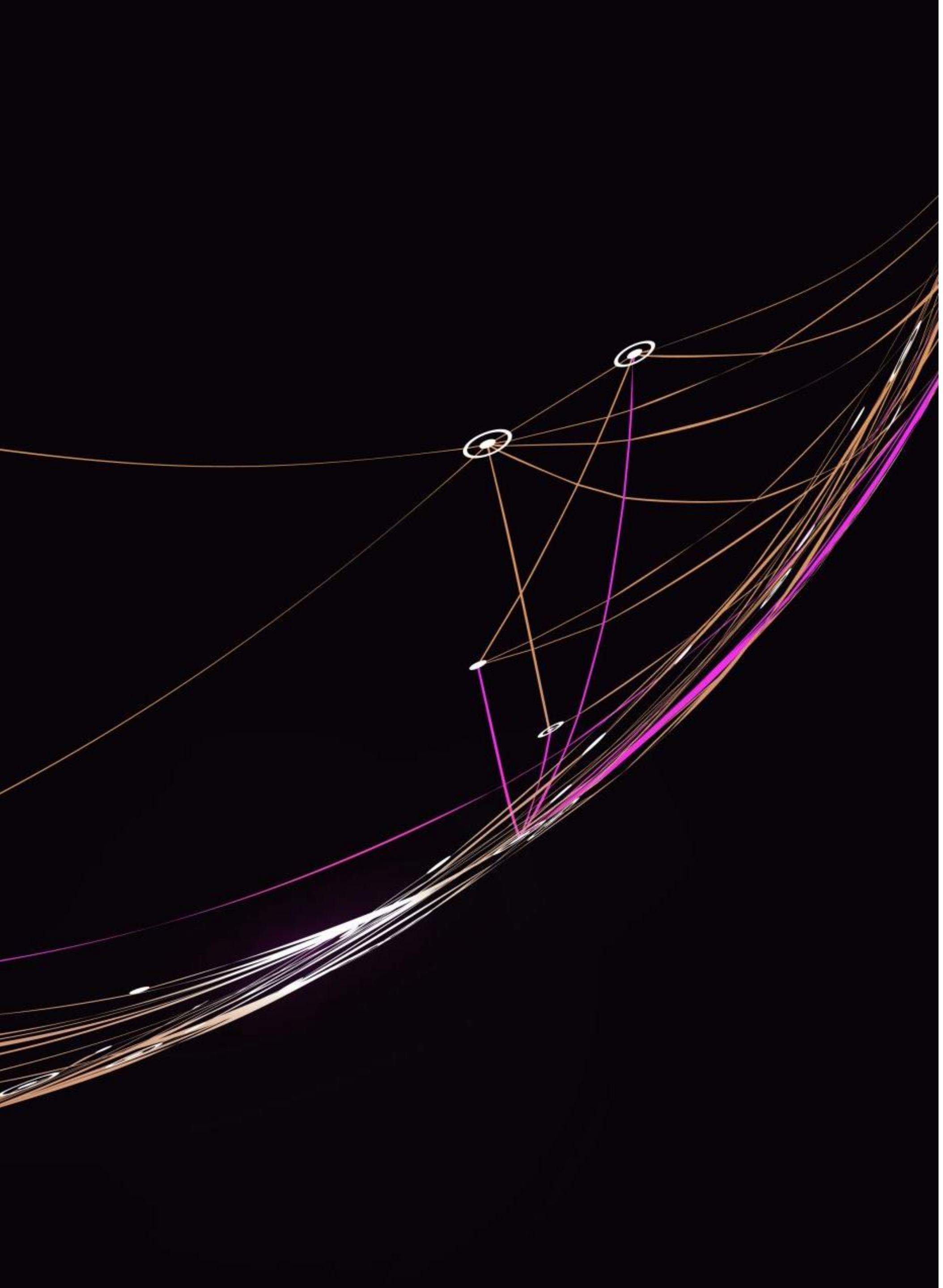


ADMINISTRAÇÃO

IBM0112 DATA MINING

Regressão Linear e Logística

Cassius Figueiredo

Regressão Linear

Função Objetivo: Regressão

- Desejamos que a variável dependente possa ser escrita como uma relação das variáveis independentes:

$$Y \approx f(X)$$

- Chamamos a função f de função objetivo do nosso problema de regressão.
- Nosso modelo pode ser escrito como:

$$Y \approx f(X) + \epsilon$$

- Onde ϵ é o ruído, e captura discrepâncias entre Y e $f(x)$.

Regressão Linear

- É uma técnica de *aprendizagem supervisionada* que supõe uma dependência linear entre a saída Y e a entrad $X_1, \ldots, X_p$.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- A função objetivo é raramente é linear.
- Apesar da simplicidade o modelo linear é extremamente útil de um ponto de vista prático e teórico.
- Diversos modelos mais complexos se baseiam no modelo linear.

Regressão Linear Simples

Regressão Linear Simples

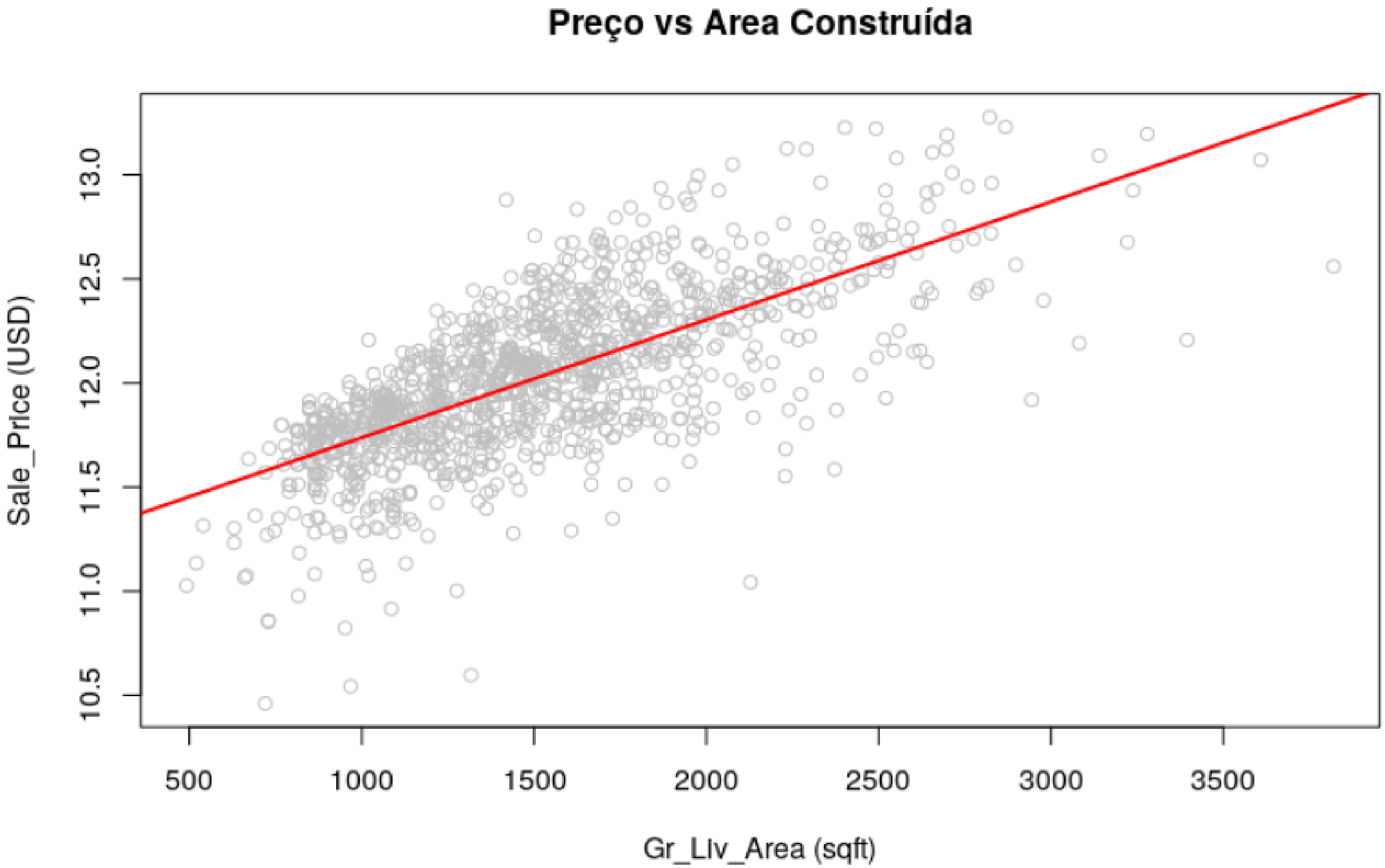
$$Y = \beta_0 + \beta_1 x + \epsilon$$

Parâmetros

- β_0 = intercepto é onde a reta corta o eixo y
- β_1 = inclinação da reta

O meu modelo $\mathcal{F}$ é a coleção de todas as funções lineares $h(x) = \beta_0 + \beta_1 X$. Como escolho os melhores valores para β_0 e β_1?

Regressão Linear Simples



$$\texttt{Sale\_Price} = \beta_0 + \beta_1 \texttt{Gr\_Liv\_Area} + \epsilon$$

Treinamento

O processo de treinamento busca encontrar a função $\widehat{f}$ dentro do modelo $\mathcal{F}$, que minimize $E_{in}(h)$ para todo h em $\mathcal{F}$.

No caso da regressão linear simples $(\widehat{\theta}_0, \widehat{\theta}_1)$ é escolhido de tal forma a minimizar o $E_{in}(\widehat{f})$:

$$\min_{h \in \mathcal{F}} E_{in}(h) = E_{in}(\widehat{f}) \equiv \frac{1}{N} \sum_{i=1}^{N} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.$$

Para encontrar $(\widehat{\beta}_0, \widehat{\beta}_1)$ utilizamos o algoritmo dos mínimos quadrados.

Previsão

O processo de previsão busca prever o valor de novos valores de $X = x$ utilizando o modelo $\widehat{f}$ treinado.

No problema de regressão fazermos a previsão de um novo ponto $\widehat{y}$ fazendo $\widehat{y} = f(x)$

No caso da regressão linear simples,

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

Regressão Linear Múltipla

Regressão Linear Múltipla

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

Comentários

- Não estamos restritos a uma variável, podemos adicionar outras características.
- O modelo é estimado da mesma forma que antes e obtemos $\widehat{f}(\mathbf{x}) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p$.
- Quanto mais regressores mais complexo o nosso modelo.
- Medimos a dimensão do nosso modelo através do número de regressores.

Variáveis categóricas

- Alguns preditores podem ser qualitativos ao invés de quantitativos.
- Dizemos que os preditores são categóricos (ou qualitativos) se eles podem assumir apenas um número finito de valores.
- Regressores categóricos precisam ser recodificados para poderem ser utilizados em um modelo de regressão.

Variáveis Dummy

- Uma variável dummy pode tomar apenas valores 1 ou 0.
- Variáveis categóricas são codificadas como variáveis dummy para serem usadas em modelos de regressão.
- Se a variável categórica possui C categorias, utilizamos C-1 variáveis dummy.
  Ex.: $z \in \{\texttt{casado}, \texttt{solteiro}, \texttt{outros}\}$. Assim definimos *duas* novas variáveis

$$x_{\texttt{cas}} = \begin{cases} 1 & z = \texttt{casado} \\ 0 & z \neq \texttt{casado} \end{cases} \quad \text{e} \quad x_{\texttt{sol}} = \begin{cases} 1 & z = \texttt{solteiro} \\ 0 & z \neq \texttt{solteiro} \end{cases}.$$

Root-Mean-Square Error (RMSE)

- Considerada a mais popular função de perda e métrica de erro para regressões.
- Diferenciável, o que permite utilizá-la como função de perda em processos que utilizem o gradiente descendente como técnica de otimização.
- A perda é simétrica, porém erros maiores tendem a influenciar mais no resultado.
- Por conta da aplicação da raiz quadrada, o erro é apresentado na mesma unidade de medida da variável-alvo.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$\boldsymbol{n}$ is the total number of observations in the (public/private) data set,
$\hat{\boldsymbol{y}}_{\boldsymbol{j}}$ is your prediction of target, and
$\boldsymbol{y}_{\boldsymbol{j}}$ is the actual target for j.

Mean of the Absolute Errors (MAE)

- Simétrica.
- Não leva mais peso para erros maiores.

- Pode ser substituída pela **Median of the Absolute Errors (MedAE)**, que é mais robusta em relação aos outliers (basta chegar à mediana, no lugar de calcular a media).

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Explained Variance Score (R^2 Score)

- Também conhecido por Coeficiente de Determinação.
- Varia entre 0 e 1. Pode ser negativo quando calculado em dados novos (Out-of-Sample) ou no caso de regressões sem intercepto.
- Ao contrário das outras métricas, compara a performance do modelo contra um benchmark.
- O modelo benchmark é um modelo simples que sempre prevê o valor médio da variável-alvo como resultado.

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

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

Regressão Logística

Considerações

- O modelo de regressão logístico é utilizado quando a variável resposta é qualitativa, com dois resultados possíveis.
- Seja a probabilidade de sucesso p.
- A probabilidade de fracasso será $1 - p = q$.
- Chamamos de ‘Chance’ a razão entre a probabilidade de sucesso e a probabilidade de fracasso.
- Ex.: se a probabilidade de sucesso é 0,75, a chance é igual a:

$$\frac{p}{(1-p)} = \frac{p}{q} = \frac{0{,}75}{0{,}25} = 3$$

Logit

- O *logit* equivale ao logaritmo natural (base *e*) da chance:

$$logit(p) = log\left(\frac{p}{1-p}\right) = log(p) - log(1-p)$$

- A função logística será dada pelo logit-inverso, que nos permite transformar o *logit* em probabilidade:

$$p = \frac{exp(x)}{1 + exp(x)}$$

Valor esperado