

Regressão Linear

Henrique Ferreira e Felipe Resende

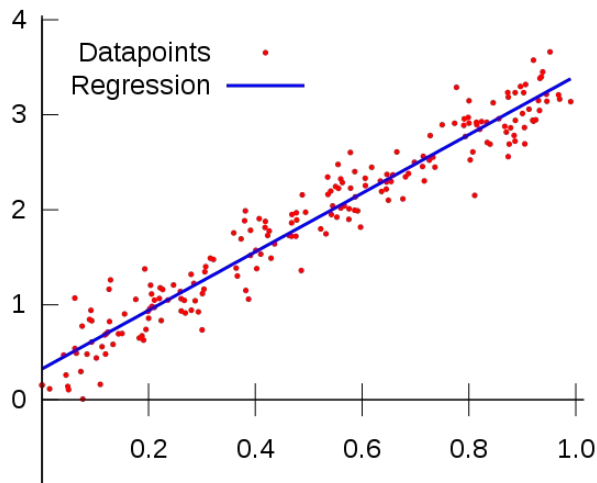
Justificativa

A regressão linear é um dos métodos mais antigos para análise de dados devido a sua facilidade de modelagem de dados correlacionados e é uma técnica muito utilizada para estudo de política ou análise de sistemas lineares.

Um bom uso do sistema linear é quando os dados de um sistema não possuem classificadores e os resultados são valores de acurácia.

Teoria - O que é regressão linear?

- Em estatística ou econometria, regressão linear é uma equação para se estimar as condições de uma variável Y , dado as condições de uma variável X que não podem ser estimadas inicialmente.



Teoria - Fundamentos

Podemos estabelecer uma regressão linear simples, cujo modelo estatístico é

$$Y = \alpha + \beta X + u$$

- α é o coeficiente linear da reta e β é o coeficiente angular
- X é a variável explanatória
- Y é a variável dependente.

Teoria - Características do modelo

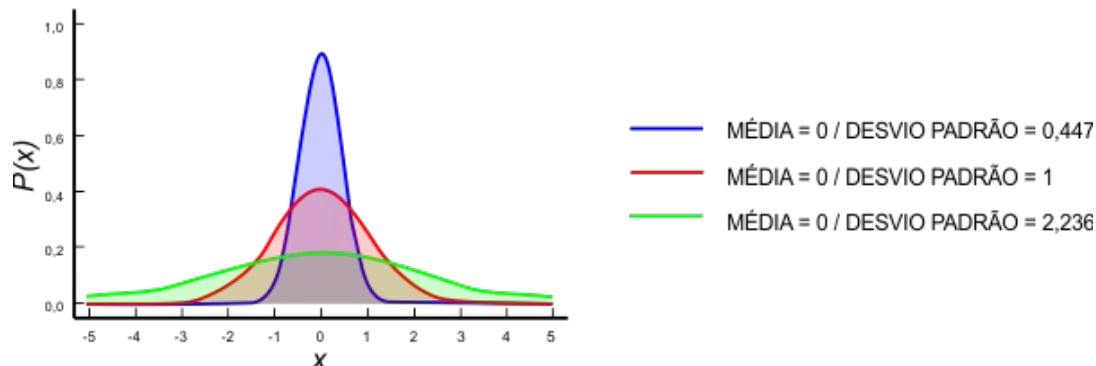
Ao estabelecer o modelo de regressão linear simples, pressupomos que:

- A relação entre X e Y é linear.
- Os valores de X são fixos, isto é, X não é uma variável aleatória.
- A média do erro é nula, isto é, 0.
- Os erros tem distribuição normal.

Teoria - Desvio padrão

Um desvio padrão grande significa que os valores amostrais estão bem distribuídos em torno da média, enquanto que um desvio padrão pequeno indica que eles estão condensados próximos da média. Em poucas palavras, quanto menor o desvio padrão, mais homogênea é a amostra.

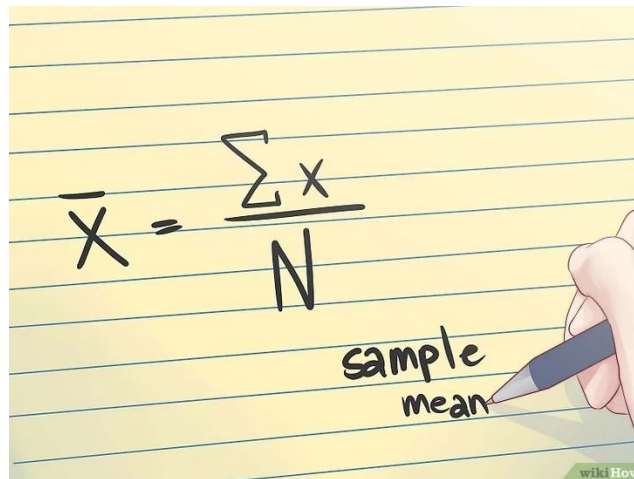
DIFERENÇA ENTRE DISTRIBUIÇÕES COM MESMA MÉDIA E DESVIOS PADRÃO DIFERENTES



Teoria - Calculando o erro padrão

O erro padrão é uma medida de variação de uma média amostral em relação à média da população. Sendo assim, é uma medida que ajuda a verificar a confiabilidade da média amostral calculada.

Para se chegar a uma estimativa do erro padrão, basta dividir o desvio padrão pela raiz quadrada do tamanho amostral. O resultado obtido também estará na mesma unidade de medida do valor amostral.



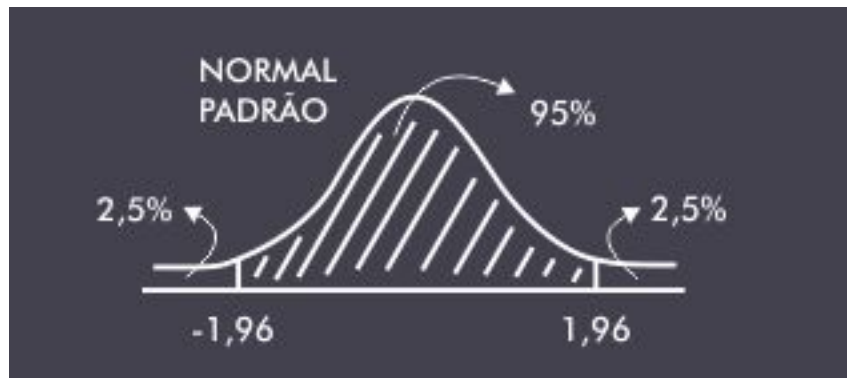
A hand-drawn illustration on yellow lined paper showing the formula for the sample mean. The formula is $\bar{X} = \frac{\sum x}{N}$. Below the formula, the words "sample mean" are written in a cursive-like font. A hand holding a blue pen is shown writing the word "mean". In the bottom right corner, there is a small green logo that says "wikiHow".

$$\bar{X} = \frac{\sum x}{N}$$

sample mean

Teoria - Grau de confiabilidade

Para calcular o intervalo de confiança basta multiplicar o erro padrão pelo percentil associado ao nível de significância observado em uma distribuição normal padrão, ou seja, que apresenta média 0 e desvio-padrão igual a 1.



Exemplo: Estima-se que o intervalo de confiança é de 95% onde o valor mínimo e máximo de uma normal padrão é $\pm 1,96$.

- Valor mínimo: $\text{Media} - 1,96 \cdot \text{erro padrão}$
- Valor máximo: $\text{Media} + 1,96 \cdot \text{erro padrão}$

Aplicações

É uma das primeiras formas de análise regressiva a ser estudada rigorosamente, e usada extensamente em aplicações práticas. Modelos que dependem de forma linear dos seus parâmetros desconhecidos, são mais fáceis de ajustar que os modelos não-lineares aos seus parâmetros.

No Brasil a utilização pode ser considerada tímida, sobretudo se comparada a norte-americana. Existe uma hostilidade em relação aos métodos quantitativos e à estatística na ciência social brasileira

Este tipo de análise pode ser utilizada para modelar sistemas lineares onde as variáveis são relacionadas de forma diretamente proporcional.

Principais APIs

- **Weka** - Coleção de algoritmos *open source* de machine learn em Java. Contém ferramentas para pré-processamento, classificação, regressão, agrupamento, associação de regras e visualização.
- **Scipy** - É um conjunto de algoritmos científicos em python. Ele possui o método `linregress(x[, y])` que os dados de uma regressão linear entre os vetores `x` e `y`.
- **Pysal** - Biblioteca de análise espacial em python. Utiliza o método `spreg.ols(x,y)` e possui muitos outros parâmetros de configuração além do Scipy.
- **Skit-Learn** - Biblioteca em python específica para aprendizado de máquinas que inclui algoritmos de regressão classificação e agrupamentos.

Artigo Relacionado 1

Título: Análise Preditiva do Desempenho Acadêmico de Alunos de Graduação da UnB Utilizando Mineração de Dados.

Objetivo: Concepção de um sistema previsor capaz de indicar quais alunos estão com maior risco de não conseguirem formar.

Dados: Dados descaracterizados de alunos de graduação de cursos da área de computação que ingressaram de 2000 até 2016 e já saíram da universidade foram utilizados

Algoritmos: Naive Bayes, ANN, SVR, Regressor Linear e Random Forests

Artigo Relacionado 1 - Informações

- Considerando a idade do aluno, os dados foram divididos em alunos jovens (ingressaram com até 30 anos) e alunos seniores (ingressaram com mais de 30 anos).
- Atributos considerados são: sexo, idade, cotista, curso, forma de ingresso, taxa de aprovação, taxa de trancamento, índice de rendimento acadêmico (IRA), taxa de melhora acadêmica, créditos integralizados, taxa de aprovação em disciplinas difíceis, estar em condição e posição no ranque.
- Os melhores resultados foram obtidos com regressão linear multivariada: alunos jovens da FT (F-measure de 0,8), alunos jovens da licenciatura (0,86), alunos jovens da computação (0,77) e alunos seniores (0,75). Nesse último caso, a SVR obteve F-measure de 0,79.

Artigo Relacionado 1 - Resultados

- Os resultados obtidos apontam a viabilidade da utilização de mineração de dados para análise preditiva de alunos em risco de evasão na UnB nos cursos da área de computação. Como a metodologia utilizada não empregou nenhum conceito específico dessa área do conhecimento, pode-se usá-la para outros cursos de graduação da UnB.

Artigo Relacionado 2

Título: Aplicação da ferramenta estatística de análise de regressão numa fazenda de cultivo de camarão marinho no estado do Rio Grande do Norte

Objetivo: observar nos ajustes dos modelos de regressão alguma influência dos parâmetros físico-químicos da água dos viveiros com relação ao peso do camarão.

Dados: Coletados em uma fazenda em Piau, no Rio Grande do Norte. A fazenda possui uma área de 43 hectares, com 8 viveiros de engorda de camarões e os dados foram coletados durante 9 meses.

Algoritmos: Regressão linear.

Artigo Relacionado 2 - Informações

- As variáveis analisadas são: Temperatura manhã, Temperatura tarde, Salinidade, Turbidez, Lâmina d'água, Ração consumida, Pluviosidade.
- As variáveis que melhor se adequam ao modelo são: Oxigênio tarde, temperatura manhã, temperatura tarde, lâmina d'água, ração e pluviosidade. Ou seja, a cada unidade acrescida na variável temperatura tarde considerando as demais variáveis no modelo fixas, estima-se em média um aumento de 4,02g no peso do camarão.
- Conclui-se que 99% da variação total do peso do camarão é explicada pela inclusão das variáveis acima mencionadas no modelo e apenas 1% por uma parte aleatória ou fatores desconhecidos.

Artigo Relacionado 2 - Conclusão

- Os ajustes dos modelos de regressão evidenciaram que as variáveis oxigênio, temperatura, ração consumida pelo camarão e a pluviosidade da região foram as que mais influenciaram na variável dependente, ou seja, o peso do camarão, sendo portanto recomendado aos pesquisadores uma maior atenção com o controle das mesmas.
- O controle dessas variáveis se faz necessário pelo fato do camarão ser um animal muito sensível a fatores externos, qualquer variação súbita que haja em algum desses fatores pode causar a perda de toda a produção de um viveiro, acarretando assim um prejuízo enorme para o produtor.

Estudo de caso 1 - Energy efficiency Data Set

- **Descrição:** Através de um software de simulação foram construídos 12 formatos diferentes de edifícios. Essas edificações diferem com relação à área de cobertura de vidro, área de distribuição do vidro, orientação e outros parâmetros. Para poder se usar algum classificador neste dataset, deve-se converter todas os dados para valores inteiros.
- **Instâncias:** 768
- **Tipos de Dados:** Inteiros e reais
- **Objetivo:** Tentar prever a taxa de temperatura ou de resfriamento no prédio com base na sua forma de construção.
- **Observações:** Não foram encontrados valores nulos ou incoerentes no dataset.

Energy efficiency Data Set

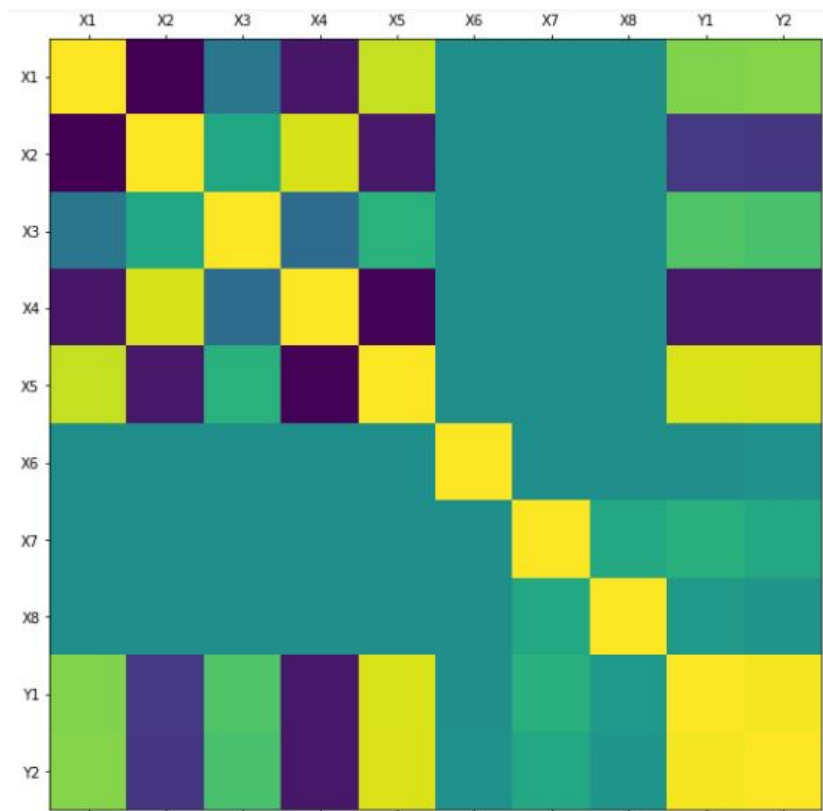
Dados:

- X1 - Compactação Relativa
- X2 - Área da Superfície
- X3 - Área da Parede
- X4 - Área do Telhado
- X5 - Altura Geral
- X6 - Orientação
- X7 - Área de Vidro
- X8 - Distribuição da área de vidro

Dados:

- Y1 - Carga de Aquecimento
- Y2 - Carga de Resfriamento

Matriz de correlação de dados



A área de superfície e a área do telhado pouco possuem relação com o resultado final do dataset enquanto que a altura afeta bem a distribuição do calor no edifício.

Resultados

Algoritmo:Regressão Linear Sklearn

Porcentagem de Treino/Teste: 80/20 %

Resultado para Y1: 91% de acerto

Resultado para Y2: 88% de acerto

Com esses resultados, é possível prever como o formato de construção afeta a dispersão de calor em um prédio com uma boa taxa precisão.

Resultados

Algoritmo:K-folding e Árvore de Decisão

Porcentagem de Treino/Teste: 80/20 %

Resultado para Y1: 100% de acerto

Resultado para Y2: 100% de acerto

Estudo de caso 2 - Bike Sharing Dataset Data Set

- **Descrição:** Este dataset apresenta informações sobre sistemas de compartilhamento de bicicletas e a quando elas são usadas além das condições climáticas como temperatura, visibilidade e velocidade do vento.
- **Instâncias:** 17389
- **Tipos de Dados:** Inteiros e reais
- **Objetivo:** Detectar os eventos mais importantes da cidade através do uso das bicicletas
- **Observações:** O dataset apresenta dados já normalizados e outros dados de classificação sobre forma de valores inteiros.

Bike Sharing Dataset Data Set

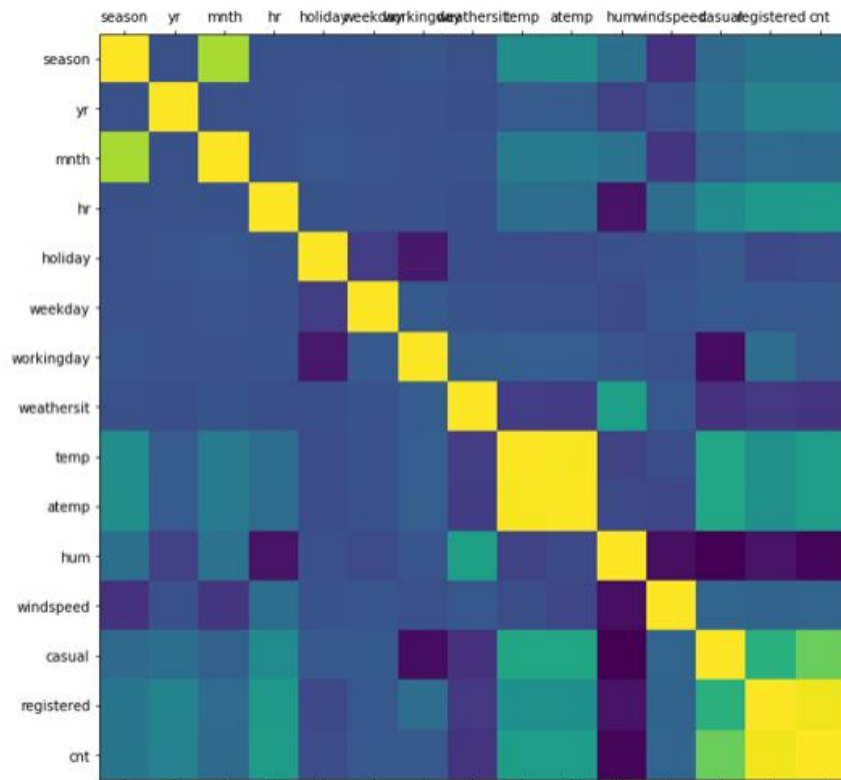
Dados:

- dteday (data)
- season
- yr (ano)
- mnth (mes)
- hr (hora)
- holiday (feriado)
- weekday
- workingday
- weathersit (clima)

Dados:

- temp (temperatura)
- atemp(sensação termica)
- hum (umidade)
- windspeed
- casual (usuários casuais)
- registered
- cnt (total de bicicletas alugadas)
- Instant

Matriz de correlação dos dados



Na matriz, é possível observar que o aluguel de bicicletas é mais influenciado pelo horário, temperatura e sensação térmica enquanto que é pouco influenciado pela umidade do ar, o estado do clima ou até mesmo se é um feriado.

Avaliando os dados

- **Instant** - É apenas o número do registro no banco de dados. Essa coluna pode ser removida.
- **Year** - O ano que aconteceu pode não interferir na predição de resultados futuros. Também pode ser removida.
- **Dteday** - O dia específico não interfere na relação dos dados pois o modelo pois ele é um valor contínuo e ordenado que não apresenta informação. Outras informações como feriados ou estação do ano.
- Existem 3 possíveis resultados para Y que são a quantidade de aluguel casuais, aluguel por usuários registrados e aluguel totais.

Resultados

Algoritmo:Regressão Linear Sklearn

Porcentagem de Treino/Teste: 80/20 %

Resultado para Aluguel totais: 39% de acerto

Resultado para Usuários Registrados: 34% de acerto

Resultado para Usuários Casuais: 45% de acerto

Resultados

Algoritmo: KNeighborsRegressor

Porcentagem de Treino/Teste: 80/20 %

Resultado para Aluguel totais: 49% de acerto

Resultado para Usuários Registrados: 51% de acerto

Resultado para Usuários Casuais: 38% de acerto

Conclusão: Com esta avaliação, é possível prever as épocas e momentos onde mais se alugam bicicletas e é possível prever com um pouco mais de precisão quando pessoas que estão registradas no sistemas além dos casuais.

Referências

- HOFFMANN, Rodolfo et al. Análise de regressão: uma introdução à econometria. O autor, 2016. Página 44.
- O que é desvio padrão. ABG CONSULTORIA ESTATÍSTICA. Disponível em <<http://www.abgconsultoria.com.br/blog/desvio-padrao-e-erro-padrao/>>, acessado em 17/03/2018.
- Estimação de parâmetros do modelo. Portal Action. disponível em <<http://www.portalaction.com.br/analise-de-regressao/12-estimacao-dos-parametros-do-modelo>> acessado em 17/03/2018.
- Calculadora da distribuição normal. Site Prof. Bertolo. Disponível em <<http://www.bertolo.pro.br/FinEst/Estatistica/DistribuicaoProbabilidades2/normal/index.html>> acessado em 17/03/2018.

Referências

- SILVA, Gabriel Ferreira. Análise preditiva do desempenho acadêmico de alunos de graduação da UnB utilizando mineração de dados. 2017.
- Mineração de dados com WEKA, Parte 1: Introdução e regressão. Autor: Michael Abernethy. Disponível em <https://www.ibm.com/developerworks/br/opensource/library/os-weka1/index.html>> acessado em 17/03/2018.
- Statistical functions. Scipy.org. Disponível em <https://docs.scipy.org/doc/scipy/reference/stats.html>> acessado em 17/03/2018.
- spreg.ols — Ordinary Least Squares. Pysal. Disponível em <http://pysal.readthedocs.io/en/v1.7/library/spreg/ols.html>> 17/03/2018.

Referências

- Linear Regression Example. Scikit learn. Disponível em <http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html> acessado em 17/03/2018.