FACULDADE DE INFORMÁTICA E ADMINISTRAÇÃO PAULISTA
TECNOLÓGO EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

Alex Yuji Vieira Isomoura - RM 84432
Daniel de Oliveira Sobanski - RM 84281
Denys Lonkovski Maiolli - RM 86487
Edson de Oliveira - RM 84361
Ruan Vieira da Silva - RM 85631

AI & CHATBOT

Prof. Marcelo Grave

SÃO PAULO
2020

Técnicas de machine learning

Machine learning ou “aprendizado de máquina” é uma parte da inteligência artificial que fornece aos sistemas a habilidade de aprender e melhorar automaticamente por meio da experiência de uso do sistema. Machine learning foca no desenvolvimento contínuo dos programas por meio da utilização e estudo dos dados inseridos pelos usuários.

Existem dois tipos de técnicas de machine learning: as supervisionadas e as não supervisionadas.

As técnicas supervisionadas, também chamada de preditivas, se utilizam de dados categorizados do passado para prever eventos futuros. O algoritmo pode comparar aquilo que foi predito com os valores que realmente foram realizados para corrigir erros e constantemente melhorar o modelo. As predições podem ser realizadas por valor ou por classe. Quando ocorrem predições por valor uma das técnicas utilizada é a regressão linear. Quando as predições se dão com base na classe, podemos utilizar técnicas de regressão logística ou árvore de decisão.

Por outro lado, as técnicas não supervisionadas são utilizadas quando as informações não são categorizadas e nem classificadas. O sistema não prediz o output correto, somente explora o dado, podendo desenhar inferências e descrevendo estruturas escondidas em dados sem categorias.

Regressão linear

O modelo de regressão linear funciona como uma equação matemática que demonstra através de uma linha reta, a relação linear entre duas variáveis x (variável independente) e y (variável dependente). Basicamente, existem duas formas diferentes de se usar a regressão linear: a técnica de regressão univariada e a técnica de regressão multivariada. A regressão univariada considera apenas uma variável independente, enquanto a regressão multivariada leva em conta mais que uma variável independente.

Para se estimar os valores da variável y temos a seguinte fórmula:

$$Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i$$

Onde:

Y_i: Variável dependente – será previsto pelo modelo.

β_0: Constante que representa em que ponto está a interceptação da reta com o eixo vertical.

β_1: Valor que representa a inclinação da reta (coeficiente angular) com relação à variável independente.

X_i: Variável independente.

ε_i: Representa a distância de um ponto observado (x, y) com relação a reta, ou seja, a distância daquilo que foi observado com aquilo que foi previsto.

Antes de se iniciar a predição dos valores, é necessário separar a variável independente da variável dependente que se quer prever. Após isso, é necessário dividir os dados da variável x e y em 30% para teste e 70% para treinamento. É recomendado utilizar essas porcentagens, ou algo próximo, para a divisão das variáveis. Com as divisões realizadas é possível treinar o modelo com os dados de treinamento e, por fim, avaliar o modelo com os dados que foram separados para teste.

Os resultados que podem ser encontrados com a formação da reta linear são:

- O valor da interceptação da reta com o eixo y (β_0) e
- O coeficiente angular da reta (β_1)

Já as métricas de avalição que podem ser calculadas utilizando o modelo são:

- Erro quadrático médio: é a média da distância daquilo que foi previsto daquilo que foi realmente observado elevado ao quadrado. Quanto maior o valor, mais distante os pontos estão da reta ou maior a quantidade de outliers no modelo.
- Erro absoluto médio: é a média da distância absoluta daquilo que foi previsto daquilo que foi realmente observado. Assim como no erro quadrático médio, quanto maior o valor, mais distante os pontos estão da reta ou maior a quantidade de outliers no modelo.

- R^2 ou coeficiente de determinação: valor que varia entre 0 e 1. Números mais próximos de 1 indicam que o modelo criado é melhor que um modelo predito com base na média do target. Números mais distantes de 1 indicam que o modelo é pior.

Um dos cuidados a se ter com esse modelo são os problemas com multicolinearidade. Esse problema ocorre quando se possui variáveis com uma correlação muito elevada entre si, tornando algumas dessas características não tão importantes no modelo como deveriam ser. O modelo pode acabar colocando um coeficiente de correlação muito baixo em algumas delas quando não deveria. Uma das recomendações para lidar com esse problema é remover do modelo variáveis que são altamente correlacionadas, evitando, dessa forma, a multicolinearidade.

Por fim, com a utilização desse modelo é possível obter informações e reconhecer alguns comportamentos com base na relação entre as variáveis que seriam difíceis de visualizar a olho nu. O modelo, com a adição de mais dados nas variáveis, é capaz de buscar resultados cada vez mais confiáveis e melhores para os usuários do sistema.

Regressão logística

O modelo de regressão logística, diferente do de regressão linear, não é utilizado para prever valores em uma determinada situação e não é representada por uma linha reta, mas sim, é utilizado para prever a classe que determinado conjunto de dados faz parte e é representado por uma curva em formato de S. Esse modelo permite a predição de valores binários utilizados para separar uma classe de dados. Apesar de funcionar para prever mais que dois valores, não é recomendado, pois as previsões tendem a obter muitos erros.

Como na regressão linear, também é possível realizar a regressão logística com base em dados univariados ou multivariados. A parte de treinamento e teste também é necessária para dividir o modelo. Definida a variável dependente e as variáveis independentes e realizado o treinamento do modelo, é possível prever a que classes os dados da variável x fazem parte.

Com os dados previstos calculados é possível comparar com os dados realmente observados e mensurar com medidas e métricas o quanto o classificador está correto ou errado. Uma dessas métricas que podem ser utilizadas é a acurácia. Ela indica um valor entre 0 e 1, onde, quanto mais perto de um, maior foi a precisão do modelo e, quanto mais perto de zero, menor foi a precisão do cálculo.

Outra medida que pode ser usada é a matriz de confusão. Ela auxilia na verificação das classes que foram corretamente e erroneamente preditas com uma forma visual simples de entendimento. Além disso, pode-se obter diferentes métricas que ajudam a compreender melhor os resultados. Basicamente, a matriz apresenta a quantidade de dados preditos que foram verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN). A partir dessas quantidades é possível obter a acurácia do modelo com a seguinte fórmula:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

Onde, em resumo, se verifica os acertos do modelo dividido pela soma dos acertos mais a soma dos erros. Essa medida indica a porcentagem de acertos daquilo que foi previsto pelo modelo.

Outra métrica de avaliação da matriz de confusão é a precisão. Ela indica a porcentagem daquilo que foi previsto corretamente por meio das seguintes fórmulas:

$$Precisão = \frac{VP}{VP+FP} \quad \text{e} \quad Precisão = \frac{VN}{VN+FN}$$

Onde a primeira fórmula indica a porcentagem daquilo que, reconhecido como positivo, é realmente positivo e a segunda fórmula indica aquilo que, reconhecido como negativo, é de fato negativo.

Uma outra medida de avaliação é a revocação. Ela indica a porcentagem daquilo que deveria ter sido previsto corretamente o quanto foi certo. Segue as fórmulas:

$$Revocação = \frac{VP}{VP+FN} \quad \text{e} \quad Revocação = \frac{VN}{VN+FP}$$

Onde a primeira fórmula indica a porcentagem daquilo que sendo positivo foi realmente previsto como positivo e a segunda fórmula indica aquilo que sendo negativo foi realmente previsto como negativo.

Por fim, com todas as métricas de precisão e revocação calculadas é possível realizar o cálculo do F1-Score para as classes positivas e falsas. Essa medida realiza uma média harmônica dos valores de precisão e revocação calculados anteriormente através da seguinte fórmula:

$$F = 2 * \frac{Precisão * Revocação}{Precisão + Revocação}$$

Quanto mais próximo de um for o resultado, maior foi a quantidade de acertos do modelo. Quanto mais próximo de zero, maior foi a quantidade de erros do modelo. Serve para avaliar o quanto modelo consegue prever corretamente os dados.

Árvore de decisão

Diferente dos outros modelos que preveem os dados por meio de reta ou curva, a árvore de decisão realiza a divisão dos dados contidos no eixo x, y por meio de retas paralelas aos eixos. Como tem a possibilidade de traçar diversas retas, ela é capaz de aumentar a separação entre as instâncias de diferentes classes. No entanto, ela pode se sair pior que o modelo de regressão logística quando o limite de decisão entre uma classe e outra se encontra em uma diagonal no gráfico (x,y). Sendo assim, ela teria que se ajustar excessivamente para separar as classes, o que uma simples curva traçada na diagonal – como no modelo de regressão logística – já o faria.

Alguns conceitos importantes em uma árvore de decisão são:

- Os nós ou folhas da árvore: indicam como as classes estão sendo separadas com relação ao conjunto de dados.
- O nó raiz: é o início da separação das classes no gráfico (x,y).
- Os nós puros: representam a divisão final da classe em determinado conjunto de dados, ou seja, onde a classe contida dentro dele é única, sem a mistura com outra instância de classe.

Algumas métricas podem ser utilizadas para melhorar a divisão das classes na árvore de decisão: a entropia e a métrica de Gini. Elas ajudam a identificar como um certo dado deve pertencer a um certo subconjunto de dados e não a outro. O cálculo dessas métricas indica o quão homogêneo está um nó da árvore de decisão. Quanto mais próximo de zero, mais puro é o nó. Do contrário, mais impuro é o nó.

Seguindo os passos dos outros modelos, a decisão da variável dependente e das variáveis independentes é necessária. A divisão em dados para testes e para treinamento também deve ser realizada. Após isso, é possível gerar a árvore de decisão e verificar os valores de Gini e a entropia em cada folha da árvore, as amostras contidas em cada nó e sua classificação.

Para tentar aproximar as previsões criadas daquilo que foi realmente observado, existem diversos parâmetros que podem ser utilizados para estruturar a árvore de decisão. Um deles é a profundidade da árvore de decisão, ou seja, a quantidade de divisões que o algoritmo deve realizar para se chegar aos nós finais da árvore. Além disso, existem os hiperparâmetros como o GridSearch e o RandomSearch. Eles são utilizados para treinar o algoritmo em diversas combinações retornando aquele que melhor se comportou com os dados do modelo.

Por fim, com a árvore de decisão já estruturada, é possível realizar as previsões para verificar os acertos e erros do modelo. Como no modelo de regressão logística, é possível utilizar a matriz de confusão para realizar todos os cálculos de precisão, revocação etc. já explicados acima no modelo de regressão logística.

Método do sistema

O aplicativo SkillTest contará com métodos de avaliação de funcionários que irá indicar, ao final da fase de avaliação do novo funcionário, em qual cargo ele demonstrou mais competência. Como um trainee, ele irá passar por diversos cargos nos primeiros meses de sua contratação, onde terá suas habilidades avaliadas por meio da atribuição de notas que os gestores irão definir. Todas as informações inseridas no aplicativo ficarão armazenadas em um banco de dados, onde se realizará a consulta das avaliações. O banco de

dados do sistema contará com informações como: nome, idade, e-mail, cpf, cargos trabalhados, habilidades avaliadas, notas dessas avaliações etc.

Ao final da avaliação, com o cargo já selecionado para o funcionário e após alguns meses de experiência no trabalho, o aplicativo irá buscar do gestor a informação de como aquele funcionário está lidando com o emprego. O gestor irá, dessa forma, classificar o funcionário como ótimo, bom, regular, ruim ou péssimo. Assim, além das notas das avaliações, o banco de dados do aplicativo contará com uma classificação de como cada funcionário está se saindo no seu cargo. Um exemplo com 40 funcionários pode ser verificado na tabela abaixo: (Obs.: exemplo fictício com números gerados aleatoriamente).

A partir desses dados, podemos separar os dados preditores que, no caso, são os cargos, as idades e as notas do dado classificador que é representado pela competência do funcionário. A variável dependente y é o classificador e as variáveis independentes x os preditores. Vale observar que como buscamos a classificação dos dados e não a predição de valores, o método da regressão linear não seria usado para prever as classes, já que não é um método de classificação, como visto anteriormente.

Com as variáveis separadas é necessário dividir os dados para teste e para treinamento. Realizada a divisão dos dados, pode-se testar em qual modelo as previsões seriam mais assertivas – regressão logística ou árvore de decisão – por meio do cálculo da acurácia e pela matriz de confusão. Como a tabela abaixo foi usada para demonstrar apenas as variáveis, seus dados foram formados de forma aleatória, portanto usá-la nas previsões resultaria em muitos erros por não existir nenhuma correlação entre os dados.

Como o aplicativo está em processo de construção, não possuímos dados verídicos do sistema, tornando possível a construção dos modelos apenas com valores aleatórios. Para o modelo de árvore de decisão, testamos com esses dados e a acurácia foi de 10%, um resultado que indica a aleatoriedade dos dados. No entanto, devido a parte teórica da matéria, acreditamos ainda que o modelo ideal para a tarefa de prever os dados seria o da árvore de decisão. Como visto anteriormente, a regressão logística é um método preditivo que se sai melhor quando busca por classes de valores binários. No caso do sistema SkillTest, existem cinco valores possíveis para a

classe, portanto a árvore de decisão tem uma vantagem sobre a regressão logística o que a torna possivelmente melhor.

O objetivo que se pretende buscar é, a partir das notas, saber se aquele é um candidato que irá se adaptar ao ritmo do trabalho ou não. Com a inserção cada vez maior de dados no banco, as previsões ficarão cada vez melhores, facilitando na decisão de empregar o funcionário em determinado cargo ou não.

Cargo selecionado	Idade	Organização	Liderança	Proatividade	Empatia	Respeito	Facilidade com o sist.	Raciocicínio lógico	Competência do func.
Desenvolvedor Java	18	5,00	7,00	1,00	1,00	4,00	5,00	8,00	Bom
Analista de projetos jr.	20	1,00	0,00	6,00	8,00	1,00	7,00	6,00	Ótimo
Desenvolvedor Mobile	25	5,00	8,00	4,00	2,00	4,00	2,00	1,00	Bom
Modelador de dados	23	0,00	4,00	5,00	3,00	9,00	6,00	0,00	Regular
Analista financeiro	18	0,00	9,00	3,00	7,00	8,00	6,00	10,00	Ruim
Auxiliar depto. pessoal	19	9,00	0,00	7,00	10,00	6,00	9,00	5,00	Bom
Analista de contas	21	7,00	4,00	3,00	9,00	10,00	6,00	1,00	Ótimo
Represent. comercial	21	0,00	2,00	0,00	3,00	4,00	8,00	9,00	Bom
Auxiliar de marketing	22	3,00	4,00	5,00	7,00	6,00	7,00	9,00	Péssimo
Analista banco de dados	24	5,00	0,00	9,00	4,00	2,00	1,00	9,00	Ótimo
Front-end	26	7,00	9,00	2,00	3,00	2,00	6,00	0,00	Ruim
Projetista de software	18	4,00	0,00	5,00	8,00	3,00	9,00	9,00	Bom
Auxiliar de contas a pagar	20	6,00	10,00	5,00	3,00	0,00	9,00	7,00	Bom
Desenvolvedor Java	21	3,00	6,00	3,00	5,00	5,00	4,00	0,00	Ótimo
Analista de projetos jr.	20	5,00	1,00	4,00	3,00	2,00	5,00	5,00	Regular
Desenvolvedor Mobile	19	3,00	8,00	10,00	7,00	6,00	5,00	8,00	Regular
Modelador de dados	23	4,00	8,00	4,00	2,00	4,00	10,00	1,00	Bom
Analista financeiro	24	2,00	10,00	8,00	3,00	7,00	4,00	9,00	Ótimo
Auxiliar depto. pessoal	21	4,00	10,00	9,00	1,00	8,00	3,00	2,00	Ruim
Analista de contas	18	1,00	1,00	4,00	6,00	10,00	2,00	2,00	Péssimo
Represent. comercial	19	10,00	6,00	6,00	7,00	5,00	2,00	6,00	Bom
Auxiliar de marketing	25	9,00	8,00	6,00	6,00	4,00	1,00	1,00	Ótimo
Analista banco de dados	23	4,00	5,00	9,00	9,00	8,00	3,00	3,00	Bom
Front-end	26	6,00	3,00	4,00	6,00	7,00	2,00	6,00	Bom
Projetista de software	22	5,00	4,00	6,00	5,00	2,00	2,00	6,00	Regular
Auxiliar de contas a pagar	21	6,00	0,00	5,00	6,00	8,00	10,00	1,00	Bom
Desenvolvedor Java	20	6,00	5,00	3,00	6,00	9,00	2,00	9,00	Péssimo
Analista de projetos jr.	18	0,00	1,00	0,00	5,00	8,00	10,00	5,00	Ótimo
Desenvolvedor Mobile	20	2,00	5,00	4,00	3,00	7,00	5,00	5,00	Ruim
Modelador de dados	22	2,00	9,00	7,00	2,00	3,00	4,00	2,00	Bom
Analista financeiro	23	9,00	2,00	8,00	10,00	8,00	5,00	10,00	Bom

Auxiliar depto. pessoal	21	0,00	3,00	1,00	5,00	7,00	5,00	10,00	Ótimo
Analista de contas	22	5,00	4,00	6,00	7,00	0,00	6,00	7,00	Regular
Represent. comercial	18	4,00	5,00	6,00	7,00	0,00	7,00	5,00	Regular
Auxiliar de marketing	20	3,00	2,00	7,00	10,00	8,00	0,00	0,00	Bom
Analista banco de dados	26	8,00	3,00	4,00	5,00	10,00	2,00	10,00	Ótimo
Front-end	27	8,00	8,00	4,00	9,00	3,00	9,00	7,00	Ruim
Projetista de software	24	0,00	10,00	8,00	5,00	9,00	0,00	2,00	Péssimo
Auxiliar de contas a pagar	25	10,00	10,00	5,00	2,00	2,00	0,00	6,00	Bom
Desenvolvedor Java	24	2,00	10,00	10,00	4,00	2,00	10,00	2,00	Ótimo