Inteligência Artificial aplicada à Bioinformática

By Joicy Xavier

4 de julho de 2021

Inteligência Artificial aplicada à Bioinformática

Danilo Castro , Paulo Ferreira , Quêzia Bernardes , Kaíssa Barbosa , Wandré Veloso ,
Pâmela Marinho , Joicymara S. Xavier

Revisão: Diego Mariano

BIOINFO – Revista Brasileira de Bioinformática. Edição #01. Julho, 2021.

DOI: [10.51780/978-6-599-275326-14](https://doi.org/10.51780/978-6-599-275326-14)

Nos últimos anos, a Inteligência Artificial tem sido uma ferramenta essencial em diversas áreas de estudo e também de mercado. Suas aplicações vão desde automatização de tarefas humanas, passando por recomendação de conteúdos até a predição de eventos futuros. Muitas dessas aplicações têm contribuído com avanços em pesquisas biológicas e também na medicina. Na Bioinformática, a Inteligência Artificial vem sendo cada vez mais utilizada, principalmente no desenvolvimento de aplicações que auxiliam e guiam experimentos diversos. Este capítulo tem como objetivo apresentar, de forma geral, a história da Inteligência Artificial focada no Aprendizado de Máquina (*Machine Learning*), uma das áreas mais utilizadas por bioinformatas, e também apresentar alguns dos principais tipos de aprendizado.

Inteligência Artificial: Um breve histórico

AS MÁQUINAS PODEM PENSAR?

-Alan Turing

Em 1950, Alan Mathison Turing, matemático e cientista da computação, natural do Reino Unido fez tal questionamento, que viria a impulsionar o estudo e o desenvolvimento de uma nova área da Computação. Essa área, que depois de mais de meio século ainda está longe de seu ápice, a Inteligência Artificial (IA), foi nomeada somente alguns anos após o pontapé inicial dado por Turing [1]. A IA é a área da Computação que tem como objetivo o emprego de máquinas em atividades que exigem uma capacidade de raciocínio. Uma IA é então, uma máquina capaz de executar funções que, caso fossem executadas por um ser humano, seriam consideradas inteligentes.

Nas últimas décadas, mesmo sem perceber, diariamente um grande número de pessoas estão interagindo constantemente com uma IA. Por exemplo, imagine que uma pessoa em um dia estava interessada em um celular lançado recentemente no mercado, e usou algum mecanismo de busca na web para saber um pouco mais sobre o aparelho. Em outro dia, essa mesma pessoa, navegando por suas redes sociais se depara com anúncios apresentando ofertas sobre o tal aparelho celular. Essa situação é um acontecimento bem corriqueiro. Por trás desse acontecimento está uma IA que tem como objetivo personalizar anúncios na web de forma a melhorar as vendas de uma empresa, usando dados de navegação, para recomendar o que interessa ao cliente, e consequentemente, a ela.

Da mesma forma, na Medicina, sistemas de diagnóstico têm utilizado IA para otimizar filas de espera em leitos de UTIs, monitorar, em tempo real, o quadro de pacientes graves e em muitas outras aplicações. Nesse sentido, o uso de Inteligência Artificial tem sido cada vez mais frequente nas mais diversas áreas de estudo e uma importante ferramenta dentro da sociedade.

Embora estejamos vivendo uma grande expansão atualmente, o surgimento do estudo de IA não é recente. Os primeiros passos no desenvolvimento de IAs se deram em 1950 através de Alan Turing. No artigo *Computing Machinery and Intelligence* [2], Turing propôs o **Jogo da Imitação**. Esse jogo, seria disputado por três indivíduos, entre eles um interrogador, que tem a função de determinar qual dos outros dois seria uma máquina e qual seria humano. Caso, durante a interação, a máquina consiga ser confundida com um humano pelo interrogador, diz-se que essa máquina passou no *Teste de Turing*. Até hoje, passar no Teste de Turing é um grande desafio no desenvolvimento de IAs.

Após Turing, a pesquisa na área de IA foi impulsionada com a realização da **Conferência de Dartmouth** [1], em 1956, conforme visto na linha do tempo na Figura 1, que apresenta os principais eventos no campo de estudo de IA. A conferência foi idealizada por Marvin Minsky, também cientista da computação e matemático. Ao fim da conferência, foi definido como objetivo criar uma máquina capaz de raciocínio, o que chamou a atenção de investidores.

Figura 1 – Linha do Tempo das origens da IA (Fonte: Autoria própria)

Um ano depois, o psicólogo americano Frank Rosenblatt apresentou ao mundo o **Perceptron**, um algoritmo que teria capacidade de simular um neurônio humano. Entretanto, levando em conta a tecnologia da época, o modelo desenvolvido apresentava muitas limitações, como foi apontado pelo próprio Minsky. Somado aos problemas apontados por Minsky, o matemático James Lighthill em 1973 publicou o documento *Artificial Intelligence: A General Survey* [3]. A publicação de Lighthill apontava que pesquisas em IA já estavam muito caras e não apresentavam retorno satisfatório. Com isso, os investimentos na área foram cortados, levando ao **Primeiro Inverno da IA**, época em que houve um baixo interesse no estudo da área.

Em 1980, surgiu um novo conceito, os **Sistemas Especialistas**, que realizavam com maestria uma função em uma área específica. Esses sistemas impulsionaram novamente as pesquisas. Entretanto, logo apresentaram problemas relacionados ao custo das tecnologias e foram abandonados, dando início ao **Segundo Inverno da IA**.

A era do esquecimento da Inteligência Artificial perdurou até a década de 1990, onde o retorno do interesse na pesquisa foi impulsionado pelo **Deep Blue** [4], a primeira máquina a ganhar uma partida de xadrez de um campeão mundial. Em 1996, Garry Kasparov, enxadrista campeão disputou uma série de seis partidas contra o *Deep Blue*, no qual perdeu uma e empatou duas com a máquina. Já no ano seguinte, a máquina foi capaz de ganhar duas vezes e empatar três com o campeão, novamente em uma série de seis partidas. Iniciava-se a **Primavera da IA**.

Saiba mais...

Para saber um pouco mais dessa história, ouça o **Podcast do Escriba Café** que descreve toda essa evolução.

Desde então, diversos avanços no campo da Inteligência Artificial aconteceram, e é claro, a evolução da tecnologia nos últimos anos contribuiu para tal. Hoje há um grande interesse na pesquisa e aplicação de IAs em diferentes áreas, fazendo o uso de diversos modelos desenvolvidos para realização de diferentes tarefas. Como por exemplo, a predição de acontecimentos futuros a partir de dados passados e até mesmo o mapeamento de regiões de interesse a partir de imagens de diferentes fontes.

Anualmente, a Universidade de Stanford realiza publicações do *AI Index Report*, relatando dados relevantes sobre a pesquisa em Inteligência Artificial. O gráfico apresentado na Figura 2 foi retirado da edição de 2021 [5] e mostra o crescimento do número de publicações na área a partir do ano 2000. O crescimento das publicações é notável e, segundo os autores, sinaliza que

podemos estar vivendo o **Verão da IA**. Os autores observam que entre 2000 e 2019 os artigos publicados na área passaram de 0.8% para 3.8% de todos os artigos revisados por pares publicados. Vale lembrar que o ano 2000 foi pouco depois de uma máquina vencer um campeão de xadrez. Ainda, entre 2019 e 2020 houve um aumento de 34,5% no número de publicações, um crescimento bastante significativo.

Figura 2 – Número de publicações em periódicos científicos sobre IA entre 2000 e 2020. (Fonte: AI Index Report [5])

As tendências apresentadas no documento mostram que em 2020 houve um grande investimento na aplicação de IAs para desenvolvimento de fármacos. O que pode justificar o crescimento na pesquisa, visto o cenário mundial de uma pandemia. Esse investimento é um bom argumento para se provar como, depois de tantos contratempos, a Inteligência Artificial se mostra em ascensão.

Diante do grande crescimento e importância da IA para a Bioinformática e pesquisas relacionadas, vamos abordar aqui os principais elementos que podem contribuir para o início dos estudos de quem está ingressando na área. A IA é uma área com diversas subáreas, então o foco aqui será no Aprendizado de Máquina (*Machine Learning*). Nas seções seguintes, veremos os principais conceitos e algoritmos de forma generalista.

Aprendizado de Máquina: Principais conceitos e algoritmos

A Inteligência Artificial é uma grande área de estudos, pesquisas, e aplicações que se subdividem em diversas outras subáreas. O Aprendizado de Máquina é uma subárea da Inteligência Artificial que se baseia na premissa de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana.

O Aprendizado de Máquina, ou comumente referido na terminologia em inglês, *Machine Learning*, explora o estudo e construção de algoritmos que podem aprender com seus erros e fazer previsão sobre dados [6]. Essa é uma

área de muita importância e destaque dentro da IA em decorrência do grande número de benefícios e do impacto que o aprendizado pode gerar. Quando uma máquina consegue “aprender” e resolver problemas, os processos de produção e a execução de tarefas tornam-se muito mais eficientes, rápidas e lucrativas.

O *Machine Learning* (ML) utiliza técnicas de diversas áreas como estatística, álgebra, neurociência e teoria da informação para embasar e executar os processos de treinamento e aprendizagem em si. Esses processos consistem em, a partir de um conjunto de dados, gerar modelos capazes de representar as instâncias dos dados de entrada, de forma que esses modelos possam ser utilizados para a predição e resolução de problemas genéricos.

Os algoritmos de ML podem ser utilizados para diversas tarefas diferentes, essas tarefas podem ser divididas em tarefas preditivas ou descritivas. As tarefas preditivas, ou **supervisionadas**, têm como objetivo processar os dados de entrada e encontrar um modelo (ou hipótese) que possa ser utilizado para prever um **rótulo (ou classe)**. Já as tarefas descritivas, ou **não-supervisionadas**, consistem em descrever ou classificar um conjunto de dados de interesse, com base nos dados de entrada [7]. Sendo assim, podemos dividir as tarefas de ML, basicamente, entre não supervisionadas e supervisionadas.

Você sabia?

Além dessas duas categorias de aprendizado, há também o aprendizado por reforço, que visa ensinar a máquina de forma experimental, como um ser humano aprendendo a executar uma determinada tarefa. Se um bebê aprende algo novo, os adultos em volta não reforçam o aprendizado aplaudindo e o incentivando? Então, é essa a ideia! Mas vamos deixar pra falar dele em outro momento.

Independente do tipo de aprendizado empregado, vamos definir alguns conceitos-chave presentes em uma atividade de ML padrão. A Figura 3 mostra um exemplo de classificação de uma cesta de frutas: esses são os nossos **dados**. Precisamos passar para o modelo de ML alguns **atributos** (ou **características**) que são responsáveis por descrever esse conjunto de dados. Para esse exemplo, os atributos podem ser: cor, tamanho, peso, acidez, etc. Ou seja, precisamos extrair dos dados informações que os diferenciem.

Suponhamos que esse modelo de ML seja capaz de diferenciar as frutas limão e laranja, essas duas classificações serão os nossos **rótulos** ou **classes**. Em algumas terminologias, são utilizadas por padrão as letras X e Y para nos referirmos aos atributos e às classes, respectivamente.

Figura 3: Exemplo de uma tarefa padrão de classificação. Neste exemplo, os dados correspondem à cesta de frutas. O objetivo é classificar as frutas em laranjas ou limões. Um modelo genérico de Machine Learning (ML) recebe os atributos (ou características) de cada item de dados (representado pela variável X) e também os rótulos associados a eles (variável Y). Esse algoritmo é treinado até que seja capaz de classificar novos dados (diferentes daqueles utilizados no treinamento). Ou seja, ao passar os atributos de uma nova fruta para o modelo de ML treinado, ele deverá ser capaz de dizer se ela é uma laranja ou um limão. (Fonte: Autoria Própria)

A seguir, faremos um breve resumo de cada um dos principais tipos de aprendizado: supervisionado e não supervisionado.

Aprendizado Supervisionado

O Aprendizado Supervisionado é utilizado quando o conjunto de dados de entrada já se encontra rotulado. Ele se subdivide em duas classes diferentes de métodos: os de **classificação** e os de **regressão**. Os métodos de classificação são utilizados quando os dados de entrada são categóricos e descritivos (grupos de palavras que caracterizam um dado). Já os de regressão, são utilizados quando os dados de entrada assumem valores numéricos (contínuos ou discretos). Números contínuos podem assumir qualquer valor em um intervalo infinito de valores, como por exemplo, um valor decimal. Valores discretos assumem valores de um conjunto específico e limitado de valores, a exemplo de um conjunto binário, que pode assumir valores iguais a 0 ou 1.

Os algoritmos de classificação compreendem técnicas focadas no processo de tomada de decisão ou de cálculos probabilísticos. Esses algoritmos são divididos em tipos de classificadores, que são: Lineares, Vetores, Árvores de decisão e Coleções. Dentre os principais e mais tradicionais algoritmos de classificação, se encontram os seguintes: Naïve Bayes, Árvores de decisão e *Support Vector Machine* (SVM) [8]. De forma genérica, uma tarefa de aprendizado supervisionado consiste em: extração de características, seleção de um modelo (de classificação ou regressão), treinamento, avaliação do resultado e finalmente a realização da predição e geração da saída (Figura 4).

Figura 4 – Etapas de geração e aplicação de um modelo. (Fonte: Autoria Própria)

A **extração de características** consiste na preparação do conjunto de dados a ser utilizado, extraindo as características de interesse de cada item. Após a extração, segue-se para a **escolha do modelo**. Nessa etapa, vários experimentos podem ser conduzidos a fim de descobrir qual algoritmo melhor se adapta aos dados. A etapa de **treinamento** tem como objetivo encontrar uma função ou um padrão que melhor descreva o conjunto de dados de forma genérica.

Com o modelo ajustado, segue-se então para a **avaliação do resultado**. Nessa etapa é verificada qual a performance do modelo treinado para dados que ele ainda não conhece. Para essa etapa, utilizamos conjuntos diferentes dos dados para a fase de treino e para a fase de teste do modelo. Por fim, com o modelo validado, ele já estará apto a ser utilizado para realizar novas previsões.

Algo importante de se observar quando treinamos modelos de aprendizado supervisionado, usando técnicas de avaliação e amostragem, é a tendência no super ajuste dos dados. Um modelo com performance boa não é aquele em que sua métrica atingiu próximo dos 100% na avaliação, e sim o que consegue se comportar como esperado no mundo real. Existe uma diferença entre aprender e memorizar!

Existem três situações que podem ocorrer quando se está treinando um modelo de ML. A primeira, do inglês **Underfitting**, é quando temos um sub-ajuste dos dados ao nosso modelo, ou seja, o modelo provavelmente não foi treinado o suficiente.

A segunda situação é quando há um super ajuste dos dados, do inglês, **Overfitting**. Os casos de *overfitting* são muito prováveis de acontecer quando ajustamos tanto o nosso modelo que ele **decora** o conjunto de dados que estamos usando para treinar. Dessa forma, quando ele precisar lidar com dados que ainda não conhece, provavelmente vai errar as previsões.

O modelo desejado (terceira situação) é o que consegue generalizar a solução ao ponto de produzir saídas corretas sem memorizar. Dessa forma, o modelo será extremamente útil para generalizar situações e problemas reais. Por isso, a avaliação e validação dos modelos é muito importante.

Aprendizado Não Supervisionado

O Aprendizado Não Supervisionado (do inglês *Unsupervised Learning*), também conhecido como aprendizado descritivo, é utilizado quando o conjunto de dados de entrada não está rotulado. O objetivo do Aprendizado Não Supervisionado é encontrar padrões que descrevam os dados. Para isso, o algoritmo de aprendizado irá agrupar os dados de entrada conforme padrões encontrados nos atributos que podem ser medidos através de estruturas, distância entre as variáveis, etc. No exemplo da cesta de frutas, mencionado na seção anterior, se ao invés de rótulos para cada um do conjunto de características, tivéssemos apenas as imagens das frutas, por exemplo, seria útil utilizar um algoritmo não supervisionado. Nesse caso, a identificação de laranjas e limões seria feita através da diferenciação das estruturas, considerando o formato, tamanho e cor da fruta (Figura 5).

Figura 5 – Exemplo de classificação utilizando Aprendizado Não Supervisionado. Nesse exemplo, os dados correspondem às frutas, porém, não são fornecidos os rótulos para as entradas de dados. Um algoritmo de Machine Learning (ML) genérico irá receber esses dados e, através de uma heurística, identificar padrões que os diferencie. Nesse exemplo, a estrutura que diferencia uma fruta da outra pode ser a cor da casca. (Fonte: Autoria Própria)

De forma genérica, o Aprendizado Não Supervisionado pode ser de três tipos: (i) o agrupamento de dados (ou clusterização), onde os dados são divididos de acordo com a similaridade; (ii) a associação, que busca conectar dados relacionados; ou (iii) a sumarização de dados, que tem como objetivo encontrar uma descrição simples e sucinta a partir de um conjunto de dados maior. Essas técnicas utilizam medidas estatísticas simples, como valor mínimo, média, desvio padrão, mas também técnicas mais complexas de visualização e determinação de relações funcionais entre atributos [7]. A seguir, daremos uma breve introdução à clusterização e à associação por serem as duas técnicas mais utilizadas.

Agrupamento

No agrupamento ou clusterização (do inglês *clustering*), o objetivo é **encontrar grupos** com base nas características conhecidas (Figura 6). A máquina escolhe a melhor maneira e divide automaticamente um conjunto de dados em grupos de acordo com medidas de similaridade ou de distância. As estruturas de agrupamento podem se associar por localização (agrupamento de genes de acordo com sua localização nas bibliotecas), por formato (agrupamento de formatos mecânicos em imagens 3D) ou por densidade (agrupamento por densidade populacional).

A clusterização normalmente é utilizada na segmentação de mercado para definir, por exemplo, tipos de consumidores ou a fidelidade. Também é usada para mesclar pontos em um mapa, na compressão de imagens, para analisar e rotular novos dados, para detectar comportamento anormal, dentre outros.

Figura 6 – Exemplo de um algoritmo de agrupamento com uma divisão em três clusters.
(Fonte: Autoria Própria)

Existem várias formas de agrupar objetos, sendo que, cada abordagem utiliza estruturas ou modelos para descrever os dados [7]. Os algoritmos de agrupamento mais comuns são: *K-means clustering*, *Mean-shift* e DBSCAN.

No *K-means*, primeiramente, é definido o K que é o número de *clusters* que queremos encontrar nos dados. Em seguida é definido o centro desses K *clusters* que serão os centróides. De forma resumida, o algoritmo faz essa atribuição e segue otimizando, diminuindo a distância quadrática total do centro do *cluster* até os pontos. Essa etapa se repete até que os centróides parem de se mover ou, de maneira equivalente, até que os pontos parem de alternar entre os *clusters*. No exemplo da Figura 7, foram escolhidos 4 *clusters* e, após 11 interações, os centróides foram posicionados definindo 4 associações.

Dica:

Algumas visualizações podem ser executadas em **Naftali Harris**.

Figura 7 – Funcionamento do k-means com $k = 4$. Nesse exemplo é possível perceber que os centróides de cada grupo vão se movimentando e à medida que isso acontece, objetos dos grupos também são atualizados. O algoritmo k-means inicializa k centroids, sendo k o número de agrupamentos desejado. A cada iteração, o algoritmo calcula a semelhança entre cada centroid e os objetos que pertencem àquele agrupamento e cria grupos relacionando cada objeto com o centroid mais semelhante. Esse processo é repetido até que não exista mudanças significativas dos elementos em cada grupo. (Fonte: Jadson Lucio. *Introdução ao Aprendizado Não-supervisionado*. LICA-UFAL. 2019)

Associação

Na técnica de associação do Aprendizado Não Supervisionado, a procura é por **padrões no fluxo de dados**. O objetivo é identificar regras e correlações de um conjunto de itens que frequentemente ocorrem juntos. Por exemplo, pessoas que comprem carne e pão podem ter interesse em comprar outros itens para fazer hambúrguer (Figura 8).

Figura 8 – Exemplos de mineração de dados por associação. (Fonte: Autoria Própria)

Os algoritmos baseados em regras de associação normalmente são utilizados para prever vendas e descontos, analisar mercadorias compradas em conjunto, colocar produtos nas prateleiras e analisar padrões de navegação na web. Os algoritmos mais comuns são: Apriori, Eclat e FP-growth. O Apriori, além de encontrar os termos que são adquiridos de forma conjunta, também é muito útil para saber quais combinações de termos podem ser descobertas e qual o nível de interesse nessas combinações [7].

A Bioinformática e a Inteligência Artificial

Como a proposta deste capítulo era fornecer uma breve visão sobre Inteligência Artificial, apresentamos alguns dos principais conceitos da área, assim como de Aprendizado de Máquina. A Bioinformática por ser uma área interdisciplinar e trabalhar com uma grande quantidade de dados, utiliza-se dos diversos tipos de ferramentas aqui mencionados para manipulação desses dados.

Não buscamos conduzir você, leitor, direcionando a uma ou outra técnica, e nem desejamos apresentar todas as possibilidades que você deverá optar ao se deparar com um problema com dados biológicos. Aqui apresentamos uma visão geral. Acreditamos que, sem ela, se tornará muito difícil ao bioinformata conseguir evoluir com sua pesquisa, quando se deparar com uma grande quantidade de dados, realidade muito comum entre os pesquisadores da área.

Em termos práticos, aplicações de bioinformática tem evoluído bastante a partir da utilização de técnicas de Inteligência Artificial. Tarefas como o alinhamento ou busca de estruturas em um sistema como o PDB (*Protein Data Bank*; disponível em <https://www.rcsb.org/>), que possui quase 180 mil estruturas biológicas macromoleculares, tendem a utilizar melhor os recursos computacionais quando há a aplicação de algumas das técnicas aqui citadas. O que um sistema como esse pode ganhar com a IA? Melhor tempo de resposta às requisições do usuário, maior organização dos dados além da realização de novas inferências, possíveis a partir de técnicas de Aprendizado de Máquina.

A descoberta de fármacos é um exemplo de como a Bioinformática pode se beneficiar do Aprendizado de Máquina, e vem sendo muito discutido atualmente. Para lidar com bibliotecas de compostos, como o Zinc (<https://zinc.docking.org/>) que contém mais de 750 milhões de compostos, técnicas para trabalhar com uma quantidade massiva de dados como essa sempre muito bem-vindas. Sabendo-se que determinados fármacos têm sido utilizados para tratamento de doenças específicas, a busca por substâncias semelhantes às conhecidas (descobertas a partir de agrupamento, por exemplo) permite uma análise inicial que leve a uma maior eficácia do tratamento ou diminuição dos efeitos colaterais.

Então, o que falta para que o seu próximo problema biológico seja resolvido em menos tempo com uma solução mais assertiva? Se deseja imergir nesse mundo, há muita coisa a ser explorada. Busque por conteúdos introdutórios e comunidades na internet, com certeza isso poderá te ajudar muito no processo de integração dessas duas áreas tão promissoras.

Referências

- [1] MCCARTHY, J.; MINSKY, M. L.; ROCHESTER, N.; SHANNON, C. E.. **A Proposal for the Dartmouth Summer Research Project On Artificial Intelligence**. AI Magazine, Agosto, 1995.
- [2] TURING, A. M.. **Computing Machinery and Intelligence**. Mind, vol. 59, n. 236, p. 433-460, Outubro, 1950.
- [3] LIDTHILL, James. Artificial Intelligence: A General Survey. *In: Artificial Intelligence: a Paper Symposium*. Science Research Council, 1973.
- [4] DEEP Blue. **IBM 100: Icons of Progress**. Disponível em: <<https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/>>. Acesso em: 23 de Junho de 2021.
- [5] ZHANG, Daniel *et al.* **The AI Index 2021 Annual Report**. AI Index Steering Committee, Human-Centered AI Institute. Universidade de Stanford, Stanford, CA. Março, 2021. Disponível em: <<https://aiindex.stanford.edu/report/>>. Acesso em: 23 de Junho de 2021.
- [6] KOHAVI, Ron; PROVOST, Foster . **Glossary of terms**. Machine Learning 30: 271-274, 1998.
- [7] FACELI, Katti [et al.]. Inteligência Artificial: **Uma Abordagem de Aprendizagem de Máquina**. Rio de Janeiro: LTC, 2011.
- [8] FURQUIM, Cristiano. **10 Algoritmos de Aprendizagem de Máquinas (Machine Learning) que você precisa saber**. Towards data Science. Medium, 2017. Disponível em: <<https://towardsdatascience.com/10-machine-learning-algorithms-you-need-to-know-77fb0055fe0>>. Acesso em: 15 de Junho de 2021.