Machine learning

**Durante os últimos anos devido aos avanços da tecnologia, o armazenamento de dados tem sido uma prática recorrente. O que levou a um crescente interesse na prospecção de dados, ou na utilização de dados históricos para descobrir regularidades e melhorar decisões futuras. (Mitchell, 1999)**

Inteligência Artificial (IA) é um vasto conceito, que tradicionalmente se refere a criações artificiais que podem imitar o funcionamento da inteligência humana para resolver problemas. Machine Learning (ML) é um campo da Inteligência Artificial, baseado na ideia de que os sistemas podem aprender com dados, identificar padrões e tomar decisões com um mínimo de intervenção humana. A aprendizagem automática é frequentemente utilizada para fazer previsões utilizando computadores. **Apesar de não ser nova, tem vindo a ganhar importância nos últimos anos e é agora utilizada numa grande variedade de aplicações.**

Este aglomerado de dados requer métodos automatizados de análise de dados, que é o que a aprendizagem automática proporciona. Em particular, definimos a aprendizagem mecânica como um conjunto de métodos que podem detectar automaticamente padrões em dados, **e depois utilizar os padrões descobertos para prever dados futuros, ou para realizar outros tipos de tomada de decisão sob incerteza. (Murphy, 2012)**

Os princípios básicos da aprendizagem mecânica consistem em dar dados de formação a um algoritmo de aprendizagem. O algoritmo de aprendizagem gera então um novo conjunto de regras, produz decisões e resultados fiáveis, incluindo previsões, com base em inferências dos dados e generalizando a partir de exemplos. Este processo é revolucionário porque as instruções não têm de ser programadas passo a passo. Em vez disso, o algoritmo aprende e ajusta, num processo iterativo, o seu modelo de conhecimento representativo, a fim de melhorar o seu desempenho. Os computadores podem aprender e construir modelos matemáticos sem dependerem de serem explicitamente programados para realizar tarefas específicas e podem adaptar-se quando expostos a novos dados. Quanto mais dados estiverem disponíveis para formar o algoritmo, mais aprenderá e os resultados serão mais precisos. Os modelos de aprendizagem mecânica podem aplicar uma mistura de diferentes técnicas, mas os métodos são tipicamente categorizados em três tipos gerais:

- Aprendizagem supervisionada: os dados de entrada etiquetados e a saída desejada são fornecidos previamente ao algoritmo de aprendizagem. Utilizamos este tipo de algoritmo quando temos dados que queremos prever ou explicar.

- Aprendizagem não supervisionada: os dados não etiquetados são fornecidos ao algoritmo de aprendizagem. O algoritmo é pedida a padrões reconhecidos nos dados fornecidos. Este tipo é utilizado quando se pretende relacionar e agrupar dados sem ter um alvo de saída.

- Reforço da aprendizagem: É definido um ambiente dinâmico a fim de interagir com o algoritmo de aprendizagem com o objetivo de fornecer uma avaliação sobre a resposta do sistema.

As duas maiores limitações da aprendizagem de máquinas envolvem sobreajustamento e dimensionalidade. O primeiro caso acontece quando os dados são tendenciosos e o modelo não generaliza a novos dados, enquanto o segundo acontece quando temos dificuldade em compreender os dados como resultado de múltiplas dimensões. Outro problema frequente é ter acesso a um conjunto de dados suficientemente grande.

Machine Learning (ML) está relacionado com a concepção e desenvolvimento de algoritmos e técnicas que permitem aos computadores "aprender". O principal objetivo da investigação do ML é extrair informação dos dados automaticamente, através de métodos computacionais e estatísticos. Está assim intimamente relacionado com a prospecção de dados e estatísticas". [Svensson e Söderberg, 2008] Com o rápido desenvolvimento da inteligência artificial (IA), o ML e o reconhecimento inteligente têm sido aplicados à vida humana. (Xia, Wang, Yan, Dong, & Wang, 2019) **Além de que a IA está a provar ser cada vez mais aplicável aos cuidados de saúde e existe uma lista crescente de tarefas em que os algoritmos têm desempenho médico igual ou superior ao dos médicos.(Stewart, Sprivulis, & Dwivedi, 2018)**

**Com a emergência de novas tecnologias, uma grande quantidade de dados é registada,** oferecendo perspectivas interessantes com a aprendizagem de máquinas para a análise preditiva de dados. A aprendizagem mecânica é um conjunto de métodos que processam dados para modelar um problema de aprendizagem. Algoritmos de aprendizagem supervisionada por máquinas consistem em utilizar dados anotados para construir o modelo. Esta categoria permite resolver problemas de análise de dados de previs**ão. (Jamin, Abraham, & Humeau-Heurtier, 2021)**

O cuidado dos pacientes do serviço de urgência (SU) depende de decisões clínicas rápidas e precisas baseadas em informação limitada e está a tornar-se cada vez mais difícil.(Kareemi, Vaillancourt, Rosenberg, Fournier, & Yadav, 2021)

**Kareemi et al. destacam o potencial do ML implementado nos cuidados de saúde proporcionando cuidados de saúde de maior qualidade e mais eficientes. As ferramentas de ML aproveitam muitos dos mesmos princípios centrais das abordagens estatísticas tradicionais, ao mesmo tempo que relaxam as limitações sobre o número de variáveis em estudo, variedades de dados de entrada, e os tipos de relações entre as variáveis. (Kareemi et al., 2021)**

**A revisão sistemática sugere que que os modelos de aprendizagem de máquinas parecem ter melhor desempenho de diagnóstico e prognóstico em comparação com os cuidados habituais para pacientes admitidos em SU (Serviço de Urgência) com uma variedade de atuações. (Kareemi et al., 2021) Esta revisão conclui que o ML tem um desempenho superior em quase todas as tarefas, mas também chama a atenção para várias deficiências generalizadas, incluindo a adesão limitada às diretrizes de comunicação e a falta de avaliação através de ensaios intervencionais.** Estas conclusões sublinham a necessidade de uma nova fase no apoio à decisão clínica (CDS) para cuidados de emergência, com investigação e prática centradas em sistemas CDS integrados, dirigidos pelo ML, que sejam utilizáveis, interpretáveis e eficazes.

A IA por si só ou em parceria com o ML parece ser uma solução eficaz para melhorar a qualidade da medicina personalizada e para acelerar o ritmo de evolução de técnicas complexas de diagnóstico e terapêuticas, tais como no campo da genética, pequenas moléculas, e terapias super-alvo. A transformação digital ao serviço da medicina deve basear-se tanto em conhecimentos clínicos - para garantir a máxima eficácia - como em orientações informáticas precisas, a fim de ultrapassar limitações.(Alsuliman, Humaidan, & Sliman, 2020)

A Aprendizagem Automática é uma área de investigação que utiliza conceitos de inteligência artificial e

estatística. É uma disciplina extensa usando diversos métodos de aprendizagem, como por exemplo as

redes neuronais, e, tendo como pontos de aplicação a robótica, entre outras.

Aprendizagem Automática considera métodos e os respetivos programas de *software* para extrair conhecimento útil (padrões, modelos, relações) de grandes bases de dados que frequentemente contem informação caótica e redundante. A maior utilidade e impacto do conhecimento extraído a partir de dados e eventos históricos é na previsão de eventos e alterações similares no futuro.

por um lado sabe-se que os algoritmos de aprendizagem produzem resultados fiáveis, incluindo previsões, com base em inferências dos dados e generalizando a partir de exemplos. Por outro lado tem-se a necessidade dos pacientes do serviço de urgência (SU) que dependem de decisões clínicas rápidas e precisas baseadas em informação limitada. (Kareemi et al., 2021)

torna-se importante encontrar soluções que auxiliem a tomada de decisão nesta área.

### História e Evolução

História da aprendizagem mecânica

O primeiro caso de redes neurais foi em 1943, quando o neurofisiologista Warren McCulloch e o matemático Walter Pitts escreveram um artigo sobre os neurónios, e como eles funcionam. Decidiram criar um modelo disto usando um circuito eléctrico, e assim nasceu a rede neural.

Em 1950, Alan Turing criou o mundialmente famoso Teste de Turing. Este teste é bastante simples - para que um computador passe, tem de ser capaz de convencer um humano de que é um humano e não um computador.

1952 viu o primeiro programa de computador que podia aprender à medida que corria. Era um jogo que jogava damas, criado por Arthur Samuel.

Frank Rosenblatt concebeu a primeira rede neural artificial em 1958, chamada Perceptron. O principal objectivo deste era o reconhecimento de padrões e formas.

Outro exemplo extremamente precoce de uma rede neural surgiu em 1959, quando Bernard Widrow e Marcian Hoff criaram dois modelos deles na Universidade de Stanford. O primeiro chamava-se ADELINE, e conseguia detectar padrões binários. Por exemplo, num fluxo de bits, podia prever qual seria o próximo. A geração seguinte chamava-se MADELINE, e podia eliminar o eco nas linhas telefónicas, pelo que tinha uma aplicação útil no mundo real. Ainda hoje está em uso.

Apesar do sucesso de MADELINE, não houve muito progresso até finais dos anos 70 por muitas razões, principalmente a popularidade da arquitectura Von Neumann. Esta é uma arquitectura onde instruções e dados são armazenados na mesma memória, o que é indiscutivelmente mais simples de compreender do que uma rede neural, e muitas pessoas construíram programas com base nisto.

Anos 80 e 90

1982 foi o ano em que o interesse em redes neurais começou a despertar novamente, quando John Hopfield sugeriu a criação de uma rede com linhas bidireccionais, semelhante à forma como os neurónios realmente funcionam. Além disso, em 1982, o Japão anunciou que se concentrava em redes neuronais mais avançadas, que incentivavam o financiamento americano na área, e assim criavam mais investigação na área.

As redes neurais utilizam a propagação posterior (explicada em detalhe na Introdução às Redes Neurais), e este importante passo veio em 1986, quando três investigadores do departamento de psicologia de Stanford decidiram estender um algoritmo criado por Widrow e Hoff em 1962. Isto permitiu, portanto, a utilização de múltiplas camadas numa rede neural, criando o que é conhecido como "aprendiz lento", que aprenderá durante um longo período de tempo.

Os finais dos anos 80 e 90 não trouxeram muito para o campo. Contudo, em 1997, o computador IBM Deep Blue, que era um computador de xadrez, venceu o campeão mundial de xadrez. Desde então, houve muitos mais avanços no campo, como em 1998, quando a investigação nos Laboratórios AT&T Bell sobre o reconhecimento de dígitos resultou numa boa precisão na detecção de códigos postais manuscritos do Serviço Postal dos EUA. Isto utilizou a retropropagação, que, como acima referido, é explicada em pormenor na Introdução às Redes Neurais.

Século XXI

Desde o início do século XXI, muitas empresas aperceberam-se de que a aprendizagem de máquinas aumentará o potencial de cálculo. É por esta razão que estão a pesquisar mais sobre o assunto, a fim de se manterem à frente da concorrência.

Machine Learning é um subcampo da informática que evoluiu do estudo do reconhecimento de padrões e da teoria da aprendizagem computacional em Inteligência Artificial (IA).

Em 1959, Arthur Samuel, um pioneiro americano no campo dos jogos de computador, *machine learning*, e inteligência artificial estudou procedimentos de *machine learning* e verificou que um computador poderia ser programado para que aprendesse a jogar um jogo de damas num curto espaço de tempo, tal como as pessoas. Para tal acontecer, apenas seria necessário programar todas as diretrizes do jogo. E conclui que este mecanismo de aprendizagem poderia ser aplicado a problemas da vida real.(Samuel, 1959)

*Machine learning* é um campo da ciência da computação que envolve a utilização de métodos estatísticos para criar programas que ou melhoram o desempenho ao longo do tempo, ou detetam padrões em enormes quantidades de dados que os humanos dificilmente encontrariam.

A aprendizagem automática explora o estudo e construção de algoritmos que podem aprender e fazer previsões sobre os dados. Tais algoritmos funcionam através da construção de um modelo a partir de exemplos de entradas, a fim de fazer previsões ou decisões orientadas por dados, em vez de seguir instruções de programas estritamente estáticos.

Todas as definições acima são corretas; em suma, "Aprendizagem por Máquina é uma coleção de algoritmos e técnicas utilizadas para criar sistemas computacionais que aprendem com os dados a fim de fazer previsões e inferências".

A área de aplicação da aprendizagem de máquinas é abundante. Vejamos algumas das aplicações quotidianas mais comuns da Aprendizagem Automática que acontece à nossa volta.

Sistema de Recomendação: O YouTube traz vídeos para cada um dos seus utilizadores, com base num sistema de recomendação que acredita que o utilizador individual estará interessado. Da mesma forma, a Amazon e outros retalhistas electrónicos sugerem produtos nos quais o cliente estará interessado e com probabilidade de comprar, analisando o histórico de compras de um cliente e um grande inventário de produtos. Um outro exemplo são os fornecedores de serviços de e-mail que utilizam um modelo de *machine learning* que pode detectar e mover automaticamente as mensagens não solicitadas para a pasta de spam.

Prospeção da identificação do cliente: Bancos, companhias de seguros e organizações financeiras têm modelos de *machine learning* que desencadeiam alertas para que as organizações comecem a intervir na altura certa para se envolverem com as ofertas certas para o cliente e persuadi-los a converterem-se mais cedo. Estes modelos observam o padrão de comportamento de um utilizador durante o período inicial e mapeiam-no para os comportamentos passados de todos os utilizadores para identificar aqueles que irão comprar o produto e aqueles que não irão.

Em 1950, Alan Turing, um conhecido cientista informático, propôs um teste conhecido como Turing no seu famoso artigo "Computing Machinery and Intelligence". O teste foi concebido para fornecer uma definição operacional satisfatória de inteligência, o que exigia que um ser humano não fosse capaz de distinguir a máquina de outro ser humano, utilizando as respostas às perguntas colocadas a ambos.

Para poder passar no teste de Turing, o computador deve possuir as seguintes capacidades:

- Processamento de linguagem natural, para ser capaz de comunicar com sucesso numa língua escolhida

- Representação de conhecimentos, para armazenar a informação fornecida antes ou durante o interrogatório que pode ajudar a encontrar informação, a tomar decisões e a planear. Isto é também conhecido como 'Expert System'.

- Raciocínio automatizado (discurso), para utilizar a informação do mapa de conhecimento armazenada para responder a perguntas e para tirar novas conclusões, quando necessário

- Aprendizagem da máquina, para analisar dados para detetar e extrapolar padrões que ajudarão a adaptar-se a novas circunstâncias

- Visão por computador para perceber objetos ou a análise de imagens para encontrar características das imagens

- Dispositivos robóticos que podem manipular e interagir com o seu ambiente. Isso significa movimentar os objetos com base nas circunstâncias

- Planeamento, programação e otimização, o que significa descobrir formas de tomar planos de decisão ou alcançar objetivos especificados, bem como analisar o desempenho dos planos e desenhos.

As sete áreas de capacidade de IA acima mencionadas têm assistido a uma grande quantidade de investigação e crescimento ao longo dos anos. Embora muitos dos termos nestas áreas sejam utilizados de forma intercambiável, podemos ver pela descrição que os seus objetivos são diferentes. Em particular, a aprendizagem mecânica tem visto um âmbito de atuação transversal a todas as sete áreas da IA.

![Diagram

Description automatically generated]()

Figura 2 - Areas do machine learning

O estudioso alemão Gottfried Achenwall introduziu a palavra "Statistics" em meados do século XVIII (1749). A utilização desta palavra durante este período significou que ela estava relacionada com o funcionamento administrativo de um Estado, fornecendo os números que refliam a atualidade periódica relativamente às suas várias áreas de administração. A origem da palavra estatística pode ser atribuída à palavra latina "Estado" ("conselho de estado") ou à palavra italiana "Statista" ("estadista" ou "político"); ou seja, o significado destas palavras é "Estado político" ou um Governo. Shakespeare usou uma palavra Statist no seu drama Hamlet (1602). No passado, as estatísticas foram utilizadas por governantes que designaram a análise de dados sobre o estado, significando a "ciência de estado".

No início do século XIX, as estatísticas atingiram o significado da recolha e classificação dos dados. O político escocês, Sir John Sinclair, apresentou-a aos ingleses em 1791 no seu livro Statistical Account of Scotland. Por conseguinte, o objetivo fundamental da o nascimento das estatísticas foi em torno de dados a serem utilizados pelo governo e pela administração centralizada organizações para recolher dados de recenseamento da população para estados e localidades.

### Aprendizagem Supervisionada

Regressão Logística

As técnicas de regressão são versáteis na sua aplicação à investigação médica porque podem medir as associações, prever resultados, e controlar para efeitos variáveis confusos. Como uma dessas técnicas, a regressão logística é uma forma eficiente e poderosa de analisar o efeito de um grupo de variáveis independentes sobre um resultado binário, quantificando a contribuição única de cada variável independente. Utilizando componentes da regressão linear refletidos na escala logítica, a regressão logística identifica iterativamente a combinação linear mais forte de variáveis com a maior probabilidade de detectar o resultado observado. (Stoltzfus, 2011)

Considerações importantes na condução da regressão logística incluem a selecção de variáveis independentes, a garantia de que os pressupostos relevantes são cumpridos, e a escolha de uma estratégia adequada de construção de modelos. Para a selecção de variáveis independentes, deve-se orientar por factores tais como teoria aceite, investigações empíricas anteriores, considerações clínicas, e análises estatísticas univariadas, com reconhecimento de potenciais variáveis confusas que devem ser contabilizadas. Os pressupostos básicos que devem ser cumpridos para a regressão logística incluem independência de erros, linearidade no logit para variáveis contínuas, ausência de multicolinearidade, e ausência de outliers fortemente influentes. Além disso, deve haver um número adequado de eventos por variável independente para evitar um modelo de sobreajustamento, com "regras de polegar" mínimas geralmente recomendadas, variando de 10 a 20 eventos por covariada. Relativamente às estratégias de construção de modelos, os três tipos gerais são direct⁄ standard, sequencial ⁄ hierárquico, e passo a passo ⁄ estatístico, tendo cada um deles uma ênfase e um objetivo diferentes. Antes de se chegar a conclusões definitivas a partir dos resultados de qualquer destes métodos, deve-se quantificar formalmente a validade interna do modelo (ou seja, a replicabilidade dentro do mesmo conjunto de dados) e a validade externa (ou seja, a generalizabilidade para além da amostra atual). A adequação global do modelo de regressão logística resultante aos dados da amostra é avaliada utilizando várias medidas de adequação, com uma melhor adequação caracterizada por uma menor diferença entre os valores observados e os valores previstos no modelo. Recomenda-se também a utilização de estatísticas de diagnóstico para avaliar melhor a adequação do modelo. Finalmente, os resultados para variáveis independentes são normalmente reportados como odds ratios (ORs) com intervalos de confiança de 95% (CIs). (Stoltzfus, 2011)

Conclusão:

A regressão logística é uma forma eficiente e poderosa de avaliar contribuições variáveis independentes para um resultado binário, mas a sua exatidão depende em grande parte de uma cuidadosa seleção das variáveis com satisfação dos pressupostos básicos, bem como da escolha apropriada da estratégia de construção do modelo e validação dos resultados. Além disso, é evidente que um modelo de regressão logística bem construído não é o único determinante de uma investigação de alta qualidade - desenvolver uma hipótese clinicamente relevante e objetivamente mensurável, implementar uma concepção de estudo e um plano de análise estatística apropriados, e relatar com precisão tanto os resultados como as conclusões são todas considerações importantes. Portanto, os leitores que prestarem muita atenção aos parâmetros da sua análise de regressão logística no contexto de um estudo bem concebido e bem executado, darão o contributo mais significativo para a medicina de emergência baseada em provas.(Stoltzfus, 2011)

O resultado a prever é um número contínuo em relevância com um dado conjunto de dados de entrada. Exemplos de casos de utilização são previsões de vendas a retalho, previsão do número de funcionários necessários para cada turno, número de lugares de estacionamento necessários para uma loja de retalho, pontuação de crédito, para um cliente, etc.

Classificadores

O resultado a prever é o real ou a probabilidade de um evento/classe e o número de classes a serem previstas podem ser duas ou mais. O algoritmo deve aprender os padrões em a entrada relevante de cada classe a partir de dados históricos e ser capaz de prever a classe invisível ou evento no futuro, considerando a sua contribuição. Um exemplo de caso de utilização é a filtragem de correio eletrónico não desejado em que o resultado esperado é classificar um e-mail em "spam" ou "não spam".

A construção de modelos de aprendizagem supervisionada por máquinas tem três fases:

1. Formação: O algoritmo será fornecido com dados históricos

dados com a saída mapeada. O algoritmo aprenderá o padrões dentro dos dados de entrada para cada saída e representam que como equação estatística, que também é vulgarmente conhecida como modelo.

2. Teste ou validação: Nesta fase é avaliado o desempenho do modelo formado, geralmente aplicando-o num conjunto de dados (que não foi utilizado como parte da formação) para prever a classe ou evento.

3. Predição: Aqui aplicamos o modelo treinado a um conjunto de dados que não fazia parte nem da formação nem dos testes. A previsão será utilizada para orientar as decisões empresariais.

Árvores de decisão

Em 1986, J. R. Quinlan publicou Indução de Árvores de Decisão resumindo uma abordagem para sintetizar árvores de decisão usando a aprendizagem de máquinas com um conjunto de dados ilustrativos de exemplo, onde o objetivo é tomar uma decisão sobre se se deve jogar ao ar livre numa manhã de sábado.

Como o nome sugere, uma árvore de decisão é uma estrutura semelhante a uma árvore onde os nós internos representam um teste sobre um atributo, cada ramo representa o resultado de um teste, e cada nó de folha representa uma etiqueta de classe, e a decisão é tomada após o cálculo de todos os atributos. Um caminho da raiz à folha representa regras de classificação. Assim, uma árvore de decisão consiste em três tipos de nós.

- Nó radicular

- Nó de ramificação

- Nó de folha (etiqueta de classe)

Máquina Vectorial de Apoio (SVM)

Vladimir N. Vapnik e Alexey Ya. Chervonenkis em 1963 propuseram a SVM. O principal objetivo da SVM é desenhar um hiperplano que separe as duas classes de forma ótima, de modo a que a margem seja máxima entre o hiperplano e as observações. A figura 3-15 ilustra que existe a possibilidade de diferentes hiperplanos. Contudo, o objetivo da SVM é encontrar aquele que nos dá uma margem elevada.

k Vizinhos mais próximos (kNN)

A classificação K do vizinho mais próximo foi desenvolvida a partir da necessidade de efetuar análises discriminantes quando estimativas paramétricas fiáveis de densidades de probabilidade são desconhecidas ou difíceis de determinar. Fix e Hodges em 1951 introduziu um método não paramétrico de classificação de padrões que desde então se tornou conhecido o k vizinho mais próximo governar. Como o nome sugere, o algoritmo funciona com base numa votação maioritária da sua classe k vizinha mais próxima. Na Figura 3-16, k = 5 vizinhos mais próximos para o ponto de dados desconhecido são identificados com base na medida de distância escolhida, e o ponto desconhecido será classificado com base na classe maioritária entre as classes de pontos de dados mais próximos identificadas. A principal desvantagem da kNN é a complexidade na pesquisa dos vizinhos mais próximos para cada amostra.

Coisas a lembrar:

- Escolher um valor de k ímpar para um problema de duas classes

- k não deve ser um múltiplo do número de classes.

### Aprendizagem Não-Supervisionada

Há situações em que a classe/evento de saída desejada é desconhecida para os dados históricos. O objetivo em tais casos seria estudar os padrões no conjunto de dados de entrada para obter uma melhor compreensão e identificar padrões semelhantes que possam ser agrupados em classes ou eventos específicos. Como estes tipos de algoritmos não requerem qualquer intervenção prévia dos peritos na matéria, são chamados aprendizagem não supervisionada. Vejamos alguns exemplos de aprendizagem não supervisionada.

Clustering

O agrupamento é um problema de aprendizagem não supervisionado. O principal objetivo é identificar grupos (chamados clusters) com base em alguma noção de semelhança dentro de um dado conjunto de dados.

As origens da análise de agrupamentos podem ser rastreadas até à área da Antropologia e Psicologia na década de 193. As técnicas de agrupamento mais utilizadas são k-means (divisivo) e hierárquico (aglomerante).

Alguns exemplos são o agrupamento de artigos de notícias semelhantes, o agrupamento de clientes semelhantes com base no seu perfil, etc.

K-means

O objectivo-chave de um algoritmo k-means é organizar os dados em clusters de modo a que haja elevada semelhança intra-cluster e baixa semelhança inter-cluster. Um item só pertencerá a um agrupamento, não vários, ou seja, gera um número específico de desarticulado, não hierárquico clusters.

K-means utiliza a estratégia de dividir e conquistar, e é um exemplo clássico de um algoritmo de maximização de expectativas (EM). Os algoritmos EM são constituídos por duas etapas: a primeira etapa é conhecida como expectativa (E) e é utilizada para encontrar o ponto esperado associado a um agrupamento; e a segunda etapa é conhecida como maximização (M) e é utilizada para melhorar a estimativa do agrupamento utilizando o conhecimento da primeira etapa. As duas etapas são processadas repetidamente até se alcançar a convergência.

K-means é concebido apenas para a distância Euclidiana.

|  |  |  |
| --- | --- | --- |
|  |  | (2.3) |

Análise de Componentes Principais (PCA)

A existência de um grande número de características ou dimensões torna a análise computacionalmente intensiva e difícil para a realização de tarefas de aprendizagem de máquinas para identificação de padrões. A PCA é a técnica de transformação linear não supervisionada mais popular para a redução da dimensionalidade. A PCA encontra as direções de variância máxima em dados de alta dimensão, de tal forma que a maior parte da informação é retida e projeta-a para um subespaço dimensional mais pequeno.

### Frameworks para construção de modelos de *Machine Learning*

### Metodologia

Semma vs crisp-dm

A metodologia foi criada há pouco mais de 20 anos, pela necessidade dos profissionais de Data Mining. Apesar de existir uma série de ferramentas capazes de nortear esses profissionais, quando o assunto é Big Data e o seu grande volume de dados, elas deixam a desejar. O CRISP DM surgiu justamente para atender aos projetos que estão diretamente envolvidos com o processamento e a análise de um grande volume de dados. O DM (Data Mining) faz parte de Data Science, que utiliza estatística e matemática como base para cruzamento de dados, por meio de técnicas de indução para propor hipóteses e solucionar questões empresariais. Simplificando, é a mineração de dados que vai conseguir transformar todo o volume de dados em informações úteis para o gerenciamento e a tomada de decisões.

A metodologia CRISP DM define o ciclo de vida do projeto, dividindo-o em seis etapas:

1. **Entendimento do problema:** A primeira coisa a ser feita é entender de fato qual o problema a ser resolvido, buscando todos os detalhes sobre o impacto dele na empresa e quais os objetivos em relação ao trabalho.

2. **Compreensão dos dados:** Essa etapa consiste em organizar e documentar todos os dados que se encontram disponíveis. É aqui que começa de facto o trabalho de mineração de dados, pois o profissional deve ser capaz de identificar quais são os dados importantes para a resolução do problema.

3. **Preparação dos dados:**

4. Modelação

5. Avaliação

6.Implementação

### ML Python Packages

Existe um rico número de bibliotecas de código-fonte aberto disponíveis para facilitar a máquina prática aprendizagem. Estas são principalmente conhecidas como bibliotecas científicas Python e são geralmente postas em uso na execução de tarefas elementares de ML. A um nível elevado, podemos dividir estas bibliotecas em análise de dados e bibliotecas centrais de aprendizagem de máquinas, com base na sua utilização/fim.

Os pacotes de análise de dados são conjuntos de pacotes que fornecem pacotes matemáticos e funcionalidades científicas que são essenciais para realizar o pré-processamento e transformação de dados.

Pacotes de aprendizagem de máquinas nucleares são o conjunto de pacotes que fornecem todos os algoritmos e funcionalidades de aprendizagem de máquina necessários que podem ser aplicados em um dado conjunto de dados para extrair os padrões.

Existem quatro pacotes chave que são mais amplamente utilizados para análise de dados.

- NumPy

- SciPy

- Matplotlib

- Pandas

Pandas, NumPy, e Matplotlib desempenham um papel importante e têm o âmbito de utilização em quase todas as tarefas de análise de dados.

NumPy é a biblioteca central para a computação científica em Python. Fornece um objeto de matriz multidimensional de alto desempenho, e ferramentas para trabalhar com estas matrizes. É um sucessor do pacote Numérico. Em 2005, Travis Oliphant criou NumPy ao incorporar características do Numarray concorrente no Numeric, com extensas modificações. Penso que os conceitos e os exemplos de código em grande medida foram explicados na forma mais simples no seu livro Guide to NumPy. Aqui só vamos olhar para algumas das chaves NumPy conceitos que são obrigatórios ou bons de conhecer em relevância para a aprendizagem mecânica.

Python tem sido sempre ótimo para a mistura de dados; no entanto, não foi ótimo para análise em comparação com bases de dados que utilizam frames de dados SQL ou Excel ou R. Os Pandas são uma fonte aberta Pacote Python que fornece estruturas de dados rápidas, flexíveis e expressivas, concebidas para fazer trabalhar com dados "relacionais" ou "etiquetados", tanto fáceis como intuitivos. Os Pandas foram desenvolvidos por Wes McKinney em 2008 quando estava na AQR Capital Management por necessidade de um alto desempenho, ferramenta flexível para realizar análises quantitativas sobre dados financeiros. Antes de deixando a AQR, conseguiu convencer a gerência a permitir-lhe abrir a biblioteca. Os Pandas são bem adequados para dados tabulares com colunas datilografadas de forma heterogénea, como em um Tabela SQL ou folha de cálculo Excel.

Enquanto que SciPy suplementa a biblioteca NumPy e tem uma variedade de módulos-chave de ciência e engenharia de alto nível, a utilização destas funções, no entanto, depende em grande medida do caso de utilização a caso.

![Timeline

Description automatically generated]()

### Exploração de dados

Aprendizagem supervisionada

(Libbrecht & Noble, 2015) Os métodos supervisionados são treinados em exemplos com etiquetas (por exemplo, 'gene' ou 'não gene') e são depois utilizados para prever estas etiquetas noutros exemplos, enquanto que os métodos não supervisionados encontram padrões em conjuntos de dados sem a utilização de etiquetas. Os métodos semi-supervisionados combinam estas duas abordagens, aproveitando padrões em dados não rotulados para melhorar o poder na previsão de rótulos. Podem ser necessários diferentes métodos de aprendizagem mecânica para uma aplicação, dependendo se se está interessado em interpretar o modelo de saída ou se está simplesmente preocupado com o poder de previsão. Os modelos gerativos, que apresentam uma distribuição probabilística sobre os dados de entrada, são geralmente melhores para a interpretabilidade, enquanto que os modelos discriminatórios, que procuram apenas modelar etiquetas, são geralmente melhores para o poder de previsão. A informação prévia pode ser acrescentada a um modelo a fim de treinar o modelo mais eficazmente quando lhe são fornecidos dados limitados, para limitar a complexidade do modelo ou para incorporar dados que não são utilizados directamente pelo modelo. As informações prévias podem ser incorporadas explicitamente num modelo probabilístico ou implicitamente através da escolha de características ou medidas de semelhança. A escolha de uma medida de desempenho adequada depende fortemente da tarefa de aplicação. Os métodos de aprendizagem da máquina são mais eficazes quando optimizam uma medida de desempenho apropriada. Os métodos de estimativa de rede são apropriados quando os dados contêm dependências complexas entre os exemplos. Estes métodos funcionam melhor quando têm em conta os efeitos de confusão das relações indirectas.

Resumo

O campo da aprendizagem de máquinas, que visa desenvolver algoritmos informáticos que melhorem com a experiência, tem a promessa de permitir que os computadores ajudem os seres humanos na análise de conjuntos de dados grandes e complexos. Aqui, fornecemos uma visão geral das aplicações da aprendizagem de máquinas para a análise de conjuntos de dados de sequenciação do genoma, incluindo a anotação de elementos de sequência e dados epigenéticos, proteómicos ou metabolómicos. Apresentamos considerações e desafios recorrentes na aplicação de métodos de aprendizagem mecânica supervisionada, semi-supervisionada e não supervisionada, bem como de abordagens de modelação generativa e discriminatória. Fornecemos orientações gerais para ajudar na selecção destes métodos de aprendizagem mecânica e na sua aplicação prática para a análise de conjuntos de dados genéticos e genómicos.(Libbrecht & Noble, 2015)

Na aprendizagem supervisionada os dados utilizados pelo modelo do aprendiz contêm informações sobre o resultado. Este tipo de aprendizagem usa exemplos rotulados, usando entradas com saídas conhecidas e correspondentes para treinar os modelos. É usado quando os dados históricos estão disponíveis e o objetivo é prever resultados futuros. Por isso, é possível comparar os valores gerados pelo sistema com os esperados. Existem muitos exemplos de modelos que usam aprendizagem supervisionada: Árvores de Decisão, Floresta Aleatória, Máquina de Vetor de Suporte, Classificador Naive Bayes, Redes Bayesianas, Redes Neurais, entre outros (Qiu, Wu, Ding, Xu e Feng, 2016).

Em tarefas de aprendizagem supervisionadas, um estimador é treinado em primeiro lugar numa conjunto de dados de formação que contém um conjunto de variáveis de entrada (características) e o valores de resposta (saída) correspondentes. O objectivo é obter uma função que mapeia as variáveis de entrada para a variável de saída, minimizando ao mesmo tempo uma particular função de perda (Friedman 2002). Uma vez construído, o modelo pode ser aplicado a dados anteriormente não vistos para prever os valores de saída correspondentes