Modelos Preditivos Séries Temporais

Fernando Mourão

2020

Modelos Preditivos Séries Temporais

Fernando Mourão

Sumário

Capítulo 1. Introdução à Modelagem Preditiva

Este capítulo tem como objetivo tornar o aluno apto à:

- Definir formalmente Modelagem Preditiva.
- Perceber a relevância prática da Modelagem Preditiva.
- Conhecer os principais termos relacionados à área.
- Identificar os cinco passos principais do processo de criação de Modelos preditivos.
- Assimilar os principais desafios relacionados à Modelagem Preditiva em cenários reais.

"Previsões são muito difíceis, especialmente sobre o futuro."

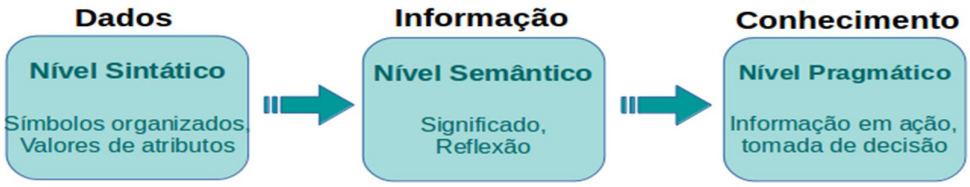
Nils Bohr (Prêmio Nobel de Física - 1922)

Tomada de decisão é uma das ações mais frequentes e relevantes que executamos ao longo de nossas vidas. Diariamente, nos deparamos com momentos em que precisamos tomar desde decisões corriqueiras, como escolher uma roupa para um evento formal, até decisões complexas, como selecionar o melhor candidato para uma vaga de emprego. Decisões que afetam vidas pessoais ou vidas profissionais, tanto de simples indivíduos quanto de grandes corporações. Em qualquer destes cenários há, porém, um requisito importante para que possamos tomar decisões corretas: a necessidade de conhecimento prévio sobre cada cenário. De fato, um conceito chave para a tomada de decisão é conhecimento.

Mas o que é conhecimento? De maneira prática, a Ciência da Informação [9,10] propõe uma diferenciação entre **dados**, **informação** e **conhecimento**. Tal como mostra a figura 1.1, dados é somente um conjunto de símbolos organizados, representando apenas um nível sintático, sem significado ou contexto. Por exemplo, '25.000,00' é um dado representado por uma sequência de dígitos e pontuações. Por

sua vez, informação é dados analisados em um contexto. Informação possui significado e representa o nível semântico. No exemplo anterior, uma possível informação seria que um usuário X possui renda bruta mensal de '25.000,00' reais. Há assim um significado claro no mundo real associado a este dado. Por fim, temos o conhecimento, que representa o nível pragmático, ou seja, a informação em ação. Em nosso exemplo, conhecimento seria identificar que nosso usuário é um potencial cliente da linha premium de um banco e que deveria ser encaixado em campanhas de marketing direcionado a tal perfil.

Figura 1.1 - Diferenciação entre dados, informação e conhecimento.



Definição 1.1: Conhecimento

Processo de uso da informação para a tomada de decisão.

Percebe-se assim a relevância prática da obtenção de conhecimento. A humanidade por muito tempo se mantém obcecada na busca por conhecimento no intuito de, no presente, ser capaz de tomar decisões corretas sobre o futuro. Para tanto, a estratégia geral é bem simples, observamos dados do passado, extraímos informações e derivamos conhecimentos potencialmente úteis para o futuro, ou para eventos passados com resultados desconhecidos. Note que não trata-se de uma estratégia nova, mas sim algo que vem sendo usado e aperfeiçoado ao longo dos séculos. Com o advento da internet e abundância de dados oriundos de variadas fontes do mundo real, essa estratégia ganhou novos contornos, ferramentas e possibilidades, passando a ser conhecida mais recentemente como **Modelagem Preditiva**. Para entender melhor esta estratégia, precisamos defini-la formalmente,

nos familiarizarmos com seus principais conceitos, identificar seus passos principais, bem como os maiores desafios relacionados à tarefa de se realizar previsões sobre o futuro. Neste capítulo, passaremos por cada um destes pontos.

1.1. O que é Modelagem Preditiva?

Recentemente, a Modelagem Preditiva vêm adquirindo uma crescente notoriedade em Ciência da Computação e áreas afins. A principal razão para isto é a diversidade de domínios e problemas para os quais a Modelagem Preditiva é útil. Serviços financeiros, e-commerce, sistemas de saúde, fraudes econômicas, detecção de spams, análise de risco e análise de *trending topics* estão entre os casos de sucesso de aplicação da Modelagem Preditiva. Formalmente, podemos definir Modelagem Preditiva como apresentado na definição 1.2.

Definição 1.2: Modelagem Preditiva

Uso de dados, algoritmos e métodos oriundos da Estatística, Aprendizado de Máquinas e Mineração de Dados para se determinar as chances de resultados futuros, ou desconhecidos, com base em dados passados. O objetivo é ir além de saber o que aconteceu ao fornecer uma melhor estimativa do que acontecerá no futuro, ou aconteceu e ainda não é conhecido.

Ao longo deste material, usaremos um exemplo guia para facilitar o entendimento de alguns dos principais conceitos relacionados à Modelagem Preditiva. Ou seja, construiremos gradativamente ao longo deste material um cenário exemplo no intuito de se apresentar e aprofundar discussões sobre conceitos específicos. Assim, um cenário prático de aplicação da Modelagem Preditiva é apresentado na definição do exemplo guia a seguir:

Exemplo Guia: Definição do Cenário

Suponha que você trabalhe como Cientista de Dados para a Nozama, uma grande companhia de e-commerce concorrente da Amazon. Mais especificamente, você foi alocado para o time de Anúncios On-line cujo objetivo é propor estratégias eficientes para aumentar o lucro que sua empresa obtêm ao exibir anúncios de produtos para seus clientes, bem como auxiliar seus clientes a acharem produtos de seus interesses. O modelo de negócio da Nozama é baseado em clicks dos clientes nos anúncios que sua plataforma apresenta a eles durante a navegação. Ou seja, a cada click a empresa anunciante paga à Nozama R$0,05. Além disso, a Nozama tem métricas de negócio específicas que precisam ser satisfeitas. Por exemplo, a venda de anúncios é feita por número de clicks por semana. Ou seja, cada anunciante pode comprar um pacote de X clicks distintos para seu anúncio em uma semana. O preço do pacote varia em função do número de clicks comprados. Assim, é fundamental garantir que ao fim de uma semana, o anúncio alcance os X clicks vendidos.

Neste cenário, podemos enumerar diversas tarefas para as quais a **Modelagem Preditiva** se mostra crucial para obtenção de sucesso. Por exemplo, uma tarefa potencialmente relevante para seu time seria prever as chances de cada cliente clicar em um anúncio específico na próxima vez que ele logar na plataforma. Ou ainda, prever quais são os anúncios que alcançarão X clicks primeiro, ou quais terão mais dificuldade em alcançar tal número. O objetivo destas tarefas é utilizar dados e informações passadas sobre clientes, produtos, anunciantes e comportamentos na plataforma, no intuito de prever comportamento futuro sobre métricas de negócio, permitindo, assim, com que a empresa alcance seus objetivos de expansão de lucro e satisfação dos clientes.

Um conceito frequentemente relacionado ao de Modelagem Preditiva é a denominada **Análise Preditiva.** Para muitos pesquisadores e profissionais da área, trata-se somente de um termo sinônimo para Modelagem Preditiva. Porém, um crescente número de profissionais vem adotando uma diferenciação entre ambos termos. Análise Preditiva passou a ser vista como um conceito mais amplo que

engloba a Modelagem Preditiva. Formalmente, podemos interpretar Análise Preditiva tal como apresentado na definição 1.3.

Definição 1.3: Análise Preditiva

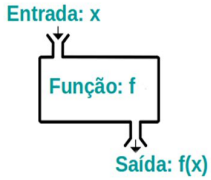
Área de estudo da Estatística que abrange diversas disciplinas intimamente relacionadas à análise de dados para tomada de decisão e predição, tais como Modelagem Preditiva, Modelagem Descritiva, Modelos de Decisão e Otimização.

1.2. Terminologia

De forma a entender melhor a Modelagem Preditiva, é importante diferenciarmos os vários tipos de modelos existentes, bem como nos familiarizarmos com a terminologia utilizada pela área.

O primeiro passo para entendermos esta terminologia consiste em definirmos de maneira concreta o que é **modelo**, uma vez que modelagem refere-se ao processo de criação de modelos. Há diversas definições para modelo, desde mais abstratas e conceituais às mais concentras e matemáticas. Considerando-se o escopo da Modelagem Preditiva, sem perda de generalidade, podemos considerar um modelo simplesmente como uma função matemática. Tal como ilustrado na figura 1.2, o tipo de modelo que estamos interessados recebe dados de entrada, realiza algum tipo de transformação sobre tais dados e gera dados de saída.

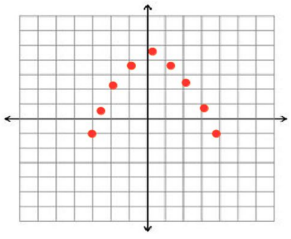
Figura 1.2 - Ilustração de funcionamento de uma função matemática.



Considerando o escopo de Modelagem Preditiva, consideramos modelo como sinônimo de função.

Por essa definição, podemos fazer uma analogia entre o processo de modelagem a partir de dados com o ajuste de funções a pontos no plano cartesiano, tal como aprendemos no ensino médio. Considere o exemplo de ajuste de função aos pontos da figura 1.3. Primeiro, ao observar como os pontos estão distribuídos no plano temos um palpite sobre o tipo de função que melhor se ajustaria a tais pontos. Por conhecimento prévio, sabemos que uma reta (i.e., uma função de 1º grau) não se ajustaria bem aos pontos. Nosso primeiro palpite seria, então, uma função de 2º grau, embora outros tipos de funções também se ajustem aos pontos neste caso (e.g., uma função de 4º grau). Uma vez selecionado o tipo de função que desejamos, nosso próximo passo consiste em achar os valores exatos que definem a função que melhor se ajustar aos pontos. Estes valores são denominados **parâmetros** da função na matemática. Uma função de 2º grau possui a seguinte fórmula $y = ax^2 + bx + c$, apresentando três parâmetros (*a*, *b* e *c),* uma variável de entrada (*x*) e uma variável de saída ou resposta (*y*). Em nosso escopo, a variável *x* é, muitas vezes, denominada **variável independente,** e a variável *y* é conhecida como **variável dependente**. Ao definirmos os valores exatos destes parâmetros, definimos completamente a função que melhor se ajusta aos pontos. Por exemplo, podemos definir a função da figura 1.3 como: $y = -1x^2 + 1x + 4$.

Figura 1.3 - Ajuste de função a pontos no plano cartesiano, tal como aprendemos no ensino médio.

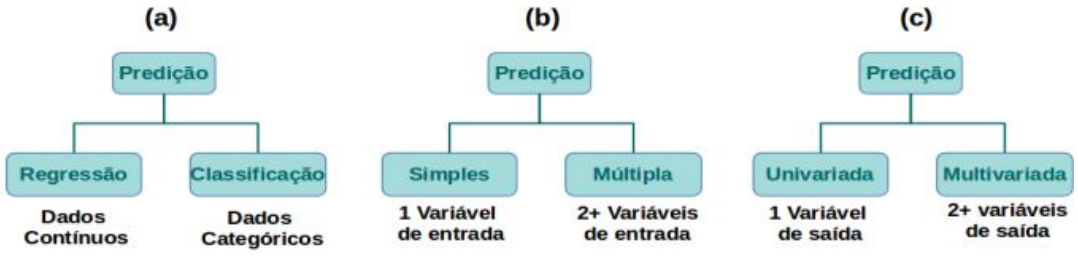


Em Modelagem Preditiva, o processo é exatamente o mesmo. Primeiro, selecionamos o tipo de modelo (i.e., função). Fazemos isso ao selecionar o algoritmo específico a ser utilizado. Cada algoritmo define um tipo específico de função, e a seleção do algoritmo ideal, em geral, baseia-se em nosso conhecimento prévio sobre o comportamento dos dados ou do problema. Por exemplo, ao usar o algoritmo de Árvore de Decisão para realizar uma classificação, estamos definindo uma família de

funções lineares compostas. De maneira análoga ao exemplo de ajuste de função aos pontos, selecionar os parâmetros do algoritmo significa selecionar os valores exatos de parâmetros que definem a função que melhor se ajusta aos dados de entrada.

Nota: de forma concreta, podemos definir o processo de aprendizado como o processo de se definir os valores dos parâmetros para a função que melhor se ajusta aos dados de entrada. Não existe uma estratégia de se selecionar o algoritmo ideal para cada problema, tampouco um único algoritmo capaz de resolver bem todos os problemas. Esta seleção é quase uma heurística e se baseia fundamentalmente em nosso conhecimento sobre dados e domínios. Por isso, é sempre necessário que um Cientista de Dados ou especialista em Modelagem Preditiva seja também um especialista sobre o domínio do problema ou trabalhe em conjunto com um profissional com este perfil.

Figura 1.4 - Classificação dos modelos de predição de acordo com (a) tipos de dados de entrada; (b) número de variáveis independentes; (c) número de variáveis dependentes.



Tal como funções, Modelos de Predição podem ser classificados de acordo com o tipo de dados de entrada e saída. Quando consideramos os dados de entrada, temos duas classes distintas de modelos, como ilustrado na figura 1.4. São considerados como dados categóricos quaisquer tipos de dados contáveis, finitos e, de maneira prática, de baixa cardinalidade. Por exemplo, o alfabeto é um tipo de dados contável, finito e de baixa cardinalidade (apenas 27 letras). Por outro lado, apesar do conjunto de números inteiros ser contável, não é finito e possui alta

cardinalidade. Já o conjunto de números reais sequer é contável. Ambos exemplos se enquadram no conjunto de dados contínuos.

Quando realizamos predições sobre dados categóricos, estamos realizando a tarefa de **Classificação**, ao passo que predição sobre dados contínuos é denominada **Regressão**. O principal exemplo de classificação presente em nosso dia a dia é a Previsão do Tempo. Como entrada, temos variáveis tais como temperatura, velocidade do vento, estação do ano observada tanto em dias recentes quanto em anos anteriores, e almejamos como resposta uma única variável: o clima para amanhã. A variável clima pode assumir poucos valores discretos distintos (e.g., chuvoso, nublado, ensolarado, etc.). Com o exemplo de regressão, podemos citar a previsão de valor do dólar para semana que vem. Como variáveis de entrada temos diversos fatores econômicos internos e externos, bem como fatores políticos; como saída desejamos uma única variável: valor da moeda. Note que o valor da moeda pode assumir valores do conjunto real, que são contínuos.

Quando consideramos o número de variáveis dependentes, podemos ainda classificar os Modelos de Predição em **Simples** ou **Múltipla**. Predição simples ocorre quando temos apenas uma única variável de entrada. Um exemplo de predição simples seria estimar a renda familiar de um indivíduo de acordo apenas com o bairro onde mora. Por sua vez, a predição múltipla usa mais de uma variável como entrada. Este é o caso do nosso exemplo de Previsão do Tempo.

Por fim, ao considerar o número de variáveis independentes, temos predição **Univariada** e **Multivariada**. Na predição univariada, estamos interessados em uma única variável resposta. Todos os exemplos discutidos até aqui referem-se à predição univariada. A predição multivariada visa obter mais de uma variável de saída simultaneamente, por exemplo: o Ministério da Saúde pode realizar um estudo sobre hábitos populacionais dos brasileiros tentando prever simultaneamente os níveis de colesterol, pressão arterial e peso de indivíduos adulto a partir de alguns hábitos alimentares na infância, tais como consumo diário de carne, peixe e laticínios. Note que neste exemplo temos três variáveis independentes (colesterol, pressão arterial e peso).

Exemplo Guia: Tipo de Modelagem

Suponha que sua primeira tarefa no time de Anúncios On-line da Nozama consiste em determinar perfis distintos dos usuários. O interesse da empresa é diferenciar usuários frequentes compradores, frequentes navegadores, infrequentes compradores e esporádicos.

Para tanto, você fará uso de informações de frequência de acesso de cada usuário, tempo de duração de cada sessão, número de páginas visitadas dentro da plataforma, clicks em produtos, número de produtos comprados e preço médio dos produtos comprados. Neste caso, temos uma tarefa de classificação, pois temos como saída dados categóricos (i.e., os perfis distintos que almejamos enquadrar cada usuário). Além disso, nosso modelo de classificação é univariado, pois temos apenas uma variável independente (o perfil) e múltipla, visto que temos diversas variáveis dependentes.

Momento de reflexão:

Suponha que o MEC deseje avaliar o desempenho de alunos nas universidades federais ao longo do tempo. O objetivo é identificar quais critérios melhor preveem o desempenho dos alunos que ingressam nas universidades através do ENEM. Como critérios atuais de avaliação, eles possuem as notas de cada aluno no ENEM segmentada por disciplina (e.g., matemática, Redação, Biologia, etc.). A ideia é, por exemplo, verificar se a nota individual na prova de Biologia seria capaz de prever o desempenho de alunos de Enfermagem no último semestre do curso melhor que a nota geral do ENEM o faria. Com isso, o MEC poderia aperfeiçoar os critérios de seleção de alunos por curso no intuito de se garantir que o desempenho dos alunos seja consistentemente alto. Qual tipo de modelagem temos neste exemplo?

R: Modelo de Regressão Múltipla Multivariada.

1.3. Como criar Modelos Preditivos?

A criação de Modelos Preditivos e sua aplicação em cenários reais é um processo complexo. Tal como em qualquer atividade de pesquisa, a chave para o sucesso neste processo é a aplicação sistemática de uma sequência de passos bem delineados. A definição exata destes passos, bem como a ordem correta de aplicação, pode variar ligeiramente de acordo com os objetivos específicos e cenário de análise. Porém, há um crescente consenso na área que aponta a existência de pelo menos cinco passos principais a serem aplicados na ordem descrita abaixo.

- **Definição do problema:** parece intuitivo e lógico que os objetivos devem estar definidos antes de se começar a resolver um problema, porém, muitas vezes pelo curto tempo de execução de projetos, o nível de definição fica aquém do necessário. Uma definição clara e objetiva do problema e dos objetivos finais é subestimada diversas vezes. Por exemplo, não é raro que profissionais passem para o passo de análise dos dados sem obter uma fórmula clara de cálculo do Principal Indicador de Desempenho (i.e., *KPI*) do projeto. É necessário que se defina pontualmente quais são as variáveis de saída do projeto, as métricas de avaliação, bem como a forma exata de cálculo de cada uma dessas métricas. Além disso, o escopo de execução do projeto, tempo de execução e mão de obra humana disponível são fundamentais para se definir e dimensionar de maneira apropriada os próximos passos. Note, porém, que a definição correta do objetivo requer um amplo conhecimento do domínio de análise, bem como dos dados disponíveis para condução da Modelagem Preditiva.

- **Coleta de dados:** uma vez definidos os objetivos, é importante definirmos quais dados serão explorados bem como a forma de obtê-los. Na indústria, de modo geral, a obtenção de dados não é um problema. Empresas interessadas em aplicar Modelagem Preditiva comumente possuem uma vasta quantidade de dados sobre seus clientes. Na academia, porém, a obtenção de dados recentes, consistentes e não enviesados sobre serviços reais, é muitas vezes um primeiro desafio a ser superado. E mesmo em casos em que têm-se acesso

a todos os dados existentes sobre os clientes, é importante verificar se o escopo de execução do projeto permite utilizar todos estes dados. Por exemplo, na elaboração de um protótipo muitas vezes limitamos a primeira versão a um subconjunto de usuários com perfil específico, ou ainda sobre uma amostra do conjunto total de usuários. Nestes casos, entra em cena os denominados métodos de amostragem. Há uma máxima na área de que seus modelos não conseguem ser melhor que seus dados. Se os dados trazem comportamentos enviesados ou parciais sobre um domínio, os resultados da Modelagem Preditiva podem ser seriamente comprometidos. Definir quais tipos de dados, as fontes destes dados, bem como os métodos de amostragem mais pertinentes, são tarefas relacionadas a este passo. Além disso, é importante definirmos neste passo indicadores de qualidade sobre os dados coletados.

- **Pré-processamento de dados:** o próximo passo a ser executado é o correto tratamento de todos os dados coletados. O tipo de tratamento a ser aplicado a cada dado varia fundamentalmente em função do tipo de dado, dos objetivos do projeto e do método de predição a ser aplicado sobre estes dados. Este é talvez o passo mais demorado e trabalhoso de todo o processo. Não é raro casos em que se gaste mais de 50% de tempo de execução de projetos apenas conduzindo este tratamento. De forma a facilitar e agilizar todo o tratamento necessário, diversas ferramentas e bibliotecas especializadas neste tipo de tarefa vem sendo disponibilizadas na literatura. A definição e execução de todo o pré-processamento, bem como ferramentas a serem utilizadas, é a principal tarefa deste passo. Como resultado temos um conjunto final de dados pós-processados que serão efetivamente utilizados como dados de entrada para os métodos de predição selecionados no próximo passo.

- **Aprendizado de modelos:** neste passo, o especialista em Modelagem Preditiva deve selecionar o método que melhor se adeque aos objetivos traçados, bem como dados disponíveis para análise, e identificar os valores de parâmetros ideais para o método selecionado. Dada a diversidade de métodos já existentes na literatura, via de regra não é necessário que o profissional

proponha novos métodos. Porém, há cenários específicos de análise que demandam novos métodos de predição ou, pelo menos, ajustes em métodos existentes. A escolha do método a ser utilizado depende de um profundo conhecimento do profissional sobre premissas intrínsecas ao problema e dados, e premissas associadas a cada método. O uso de métodos com premissas dissonantes daquelas inerente ao problema e/ou dados é uma das principais razões de insucesso na área. Por exemplo, usar Árvores de Decisão sobre dados altamente correlacionados pode não ser uma boa estratégia. Neste passo, o profissional deve ponderar sobre cada um destes aspectos de forma a construir um modelo próximo ao ótimo para o problema.

- **Avaliação e comparação de modelos:** por fim, é fundamental certificarmos que os modelos aprendidos no passo anterior tenham alta taxa de acerto sobre dados passados, que usamos para o processo de aprendizado, bem como alta taxa de acerto sobre dados futuros não informados previamente, como dados de entrada ao método selecionado. A esta taxa de acerto sobre dados não vistos dá-se o nome de Generalização. É fundamental que modelos aprendidos tenham alta generalização, caso contrário não servirão para prever o futuro. O cerne deste passo consiste no uso de métodos estatísticos de seleção de dados e comparação de modelos para se garantir a capacidade de generalização do modelo resultante.

Note que o passo 1 é completamente dependente do domínio de análise da Modelagem Preditiva e, por esta razão, foge ao escopo deste material. Nos capítulos subsequentes discutiremos em detalhes conceitos e decisões relacionados aos passos de 2 a 5.

Exemplo Guia: Definição do Problema

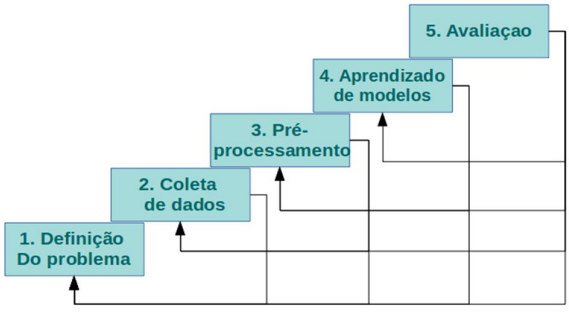
Como primeiro passo do processo de Modelagem Preditiva no time de Anúncios On-line da Nozama, temos a formalização do problema que o time irá analisar. Usaremos este problema ao longo dos capítulos seguintes à medida que formos executando os demais passos do processo. O problema selecionado é o de se recomendar anúncios para os clientes

da empresa, de forma a aumentar em 10% o número de clicks únicos por usuário em anúncios pagos.

O time formalizará com a empresa um projeto relacionado a este problema, e nossa primeira tarefa consiste em detalhar melhor o problema, definindo a KPI relacionada, as premissas assumidas e limitações de escopo. Neste sentido, primeiro definimos que o escopo do projeto limita-se aos usuários ativos. A Nozama considera como usuários ativos usuários que realizaram pelo menos uma compra na plataforma nos últimos três meses. Dessa forma, a KPI do projeto é definida como o total de anúncios distintos clicados por usuários ativos, dividida pelo total de usuários ativos.

A partir da análise de negócios sobre a plataforma, a meta inicial foi definida como aumentar em pelo menos 10% o valor da KPI. Como premissas básicas do problema destacamos: (1) a recomendação deve ser contextualizada com o que o usuário está procurando a cada momento na plataforma; (2) o número de clicks por anúncio deve atingir ao valor mínimo estabelecido no pacote comprado pelo anunciante; (3) o ambiente possui alta rotatividade de clientes e anunciantes; e (4) as recomendações devem ser geradas em tempo real gastando-se no máximo 300 ms.

Figura 1.5 - Interação entre os principais passos do processo de Modelagem Preditiva.



Nota: a figura 1.5 ilustra a interação entre os distintos passos. Na prática, observamos uma alta dependência entre os passos. De fato, a execução destes passos não é estritamente linear. Trata-se, sobretudo, de um processo cíclico no qual a definição de um passo afeta sensivelmente a definição de passos anteriores. Assim, é comum que ao definirmos um passo tenhamos que voltar e redefinir passos anteriores. Por exemplo, uma análise mais aprofundada sobre os dados disponíveis para coleta no passo 2 pode nos revelar que o objetivo traçado no passo 1 não seja factível. Ou ainda, os resultados de comparação de modelos podem apontar a necessidade de se adotar outro método de predição para se resolver o problema.

1.4. Desafios relacionados

Prever comportamentos futuros a partir de observações passadas é certamente uma tarefa desafiante. A diversidade de cenários, tipos de dados e variedade de técnicas existentes, tornam este problema ainda mais complexo. Há diversos desafios práticos que especialistas em Modelagem Preditiva devem ser capazes de identificar e propor soluções. A seguir, discutimos resumidamente alguns destes principais desafios. Maiores detalhes e estratégias conhecidas para solucioná-los serão apresentados nos capítulos seguintes.

- **Identificar as premissas corretas para cada problema:** como já mencionado, a dissonância entre premissas intrínsecas ao problema e assumidas pelos métodos de Modelagem Preditiva é uma grande fonte de insucesso na aplicação deste tipo de método. Além de conhecimento sobre o problema, um domínio sobre os fundamentos e princípios base dos principais métodos existentes é um dos principais diferenciais que um especialista em Modelagem Preditiva deve possuir para contornar este desafio.

- **Limitação dos dados:** Muitas vezes, os dados disponíveis para análise são limitados ou possuem graves deficiências decorrentes do processo de coleta. Por exemplo, informações importantes sobre os usuários de uma plataforma de serviços bancários, tal como ocupação ou escolaridade, podem estar

ausentes nos dados existentes. Este é um cenário denominado de Valores Ausentes (i.e., *Missing Values*). Uma segunda limitação seria o de dados enviesados, em que dados disponíveis involuntariamente priorizam um comportamento específico. Por exemplo, suponha que uma empresa de *Apps* para celular lance um novo aplicativo para auxiliar usuários na escolha de um carro novo. Contas grátis foram feitas para todos os usuários que frequentam um shopping da Zona Sul de São Paulo. Usuários contemplados com contas grátis podem ainda convidar outros usuários. Por construção, o conjunto de usuários que se tornam clientes deste novo *App* tende a ser de clientes com poder aquisitivo elevado e, consequentemente, com interesse em carros de luxo. Análises considerando apenas usuários da plataforma, erroneamente apontariam um grande interesse por acessórios de luxo na compra de carros. Outro cenário tipicamente relacionado à limitação de dados é a existência de fatores que não podem ser mensurados diretamente. Tais fatores são denominados de fatores latentes ou ocultos (i.e., *Hidden Factors*). Por exemplo, em estudos sobre a população não podemos mensurar diretamente qualidade de vida. Este é um fator abstrato que só podem ser mensurado indiretamente a partir de fatores secundários, tais como existência de saneamento básico, renda familiar, escolaridade, etc.

- **Qualidade dos dados:** outro desafio comum em análises associadas à Modelagem Preditiva é a qualidade dos dados de entrada. Fatores como ruídos, escalas diferentes, problemas com encoding, formatos não esperados, dentre outros, afetam significativamente a eficácia da Modelagem Preditiva em cenários reais. Por exemplo, suponha que utilizemos o número de música ouvidas por dia como uma informação importante para modelar usuários de uma plataforma de streaming de música. A existência de usuários que por alguma razão ouçam mais de 2.000 músicas em um único dia, o que pode ocorrer devido à presença de robôs em serviços como este, prejudicaria a análise deste fator em uma escala normal, uma vez que grande parte dos usuários reais ouvem menos de 100 músicas por dia.

- **Complexidade de modelo:** complexidade de modelo refere-se à quão

complexa é a função que o modelo representa. Por exemplo, um modelo representado por uma função linear possui baixa complexidade, ao passo que uma função de quinta potência possui uma complexidade bem maior. Essa noção de complexidade está associada com o tipo de comportamento que conseguimos modelar nos dados. Modelar uma reta é muito mais simples que modelar uma elipse. O grande desafio neste caso é identificar qual é a função que melhor descreve os dados. Usar uma função mais simples que a real pode gerar resultados de baixa qualidade. Por outro lado, usar funções mais complexas que a real podem reduzir a capacidade de generalização do nosso modelo (i.e., capacidade de acertar predições sobre o futuro).

- **Volume de dados e desempenho:** o volume de dados a serem processados por métodos de predição, muitas vezes em tempo real, é gigantesco em diversos cenários reais. Este fator quantitativo impõe restrições severas quanto ao uso de recursos computacionais, tais como memória RAM e tempo de processamento. Tratar de maneira eficiente grandes volumes de dados é um dos requisitos básicos de métodos de predição atualmente. Neste caso, o real desafio surge quando o volume de dados é tamanho que o problema deixa de ser meramente quantitativo e passa a ser qualitativo, fazendo com que soluções clássicas não sejam mais adequadas. Por exemplo, sabemos que, qualitativamente, o problema de se ordenar 10 ou 1 milhão de números é o mesmo, e que o *QuickSort* é o algoritmo com menor custo computacional a ser utilizado. Porém, o problema de ordenar 100 bilhões de números passa a ser outro, uma vez que exige estruturas de dados diferentes e estratégias de interações diferentes. O *QuickSort* tal como conhecemos se torna inadequado para este problema. Casos similares ocorrem quando temos que realizar predições em cenários com grande volume de dados, fazendo-se necessário a elaboração de novos métodos que atendam aos requisitos mínimos de eficiência e uso de recursos computacionais disponíveis.

Momento de reflexão:

Considere um estudo sobre o comportamento de clientes de uma plataforma de *streaming* de música, tal como o Spotify. O objetivo do estudo é diferenciar o comportamento de clientes de não clientes da plataforma. Com isso, almeja-se aprender de que forma os clientes da plataforma são diferentes e aperfeiçoar estratégias de marketing direcionadas a este perfil.

Para conduzir o estudo, os profissionais envolvidos resolveram utilizar todos os dados históricos disponíveis na plataforma. Dentre os desafios mencionados nessa seção, qual seria o principal desafio associado a estratégia adotada pelos profissionais?

R: limitação dos dados, pois não temos dados sobre não clientes.

Capítulo 2. Coleta de Dados

Este capítulo tem como objetivo tornar o aluno apto à:

- Entender o papel da amostragem em Modelagem Preditiva.
- Conhecer os principais termos e conceitos estatísticos relacionados à amostragem.
- Elaborar planejamentos amostrais de maneira robusta.
- Diferenciar as principais técnicas de amostragem existentes.
- Identificar os erros mais comuns em amostragem.

"É um erro capital teorizar antes que se tenha dados. Insensivelmente, alguns tentam ajustar os fatos às teorias, ao invés de ajustarem as teorias aos fatos."

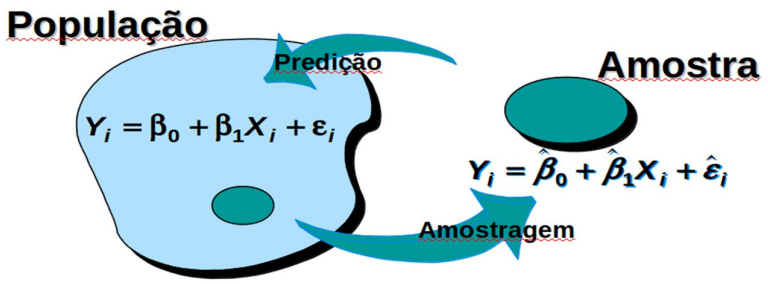
Arthur Conan Doyle (Autor de Sherlock Holmes)

Como discutido no capítulo 1, o processo de Modelagem Preditiva baseia-se na observação de comportamentos passados. Para se ter tais observações, faz-se necessário a existência de dados. Precisamos ser capazes de coletar dados que representam os comportamentos que almejamos prever, caso contrário não poderemos aprender nada. Por este motivo, a coleta de dados é um dos primeiros passos da Modelagem Preditiva.

De forma a entendermos melhor a importância da coleta de dados, precisamos definir os objetivos da Modelagem Preditiva em termos dos conceitos estatísticos de **população** e **amostra** de dados. Na estatística, entendemos como população o conjunto de todos os objetos que queremos abranger em um estudo e para os quais desejamos obter predições. Por sua vez, amostra representa um subconjunto qualquer de uma população. Assim, quando definimos o principal objetivo da Modelagem Preditiva como uma melhor estimativa do que acontecerá no futuro (definição 1.2), intrinsecamente estamos dizendo que almejamos prever o

comportamento esperado para toda uma população. Porém, muitas vezes, não podemos utilizar dados de toda a população para aprender tais comportamentos e, consequentemente, gerar as predições. Nestes casos, utilizamos uma amostra desta população como entrada para a Modelagem Preditiva, gerando predições a partir do comportamento observado nesta amostra apenas. A premissa que assumimos é que todas as predições realizadas para a amostra sejam igualmente válidas para a população, tal como ilustrado na figura 2.1. A principal questão, porém, é: podemos garantir que tal premissa seja sempre válida?

Figura 2.1 - Processo de geração de predição válida para toda uma população de interesse a partir do aprendizado gerado, considerando-se apenas uma amostra da população total.



Quando amostrar?

- Quando não temos acesso a todos os dados disponíveis;
- Quando for caro usar todos os dados disponíveis;
- Quando desejamos avaliar a qualidade das predições.

Uma correta coleta e amostragem dos dados é a estratégia que temos disponível para aumentar nossas certezas sobre a premissa mencionada acima. Neste contexto, as principais ferramentas que garantem a corretude da coleta e amostragem são a Estatística e a Teoria da Probabilidade. Técnicas oriundas dessas áreas são vitais não somente para o passo de coleta de dados, mas também para o passo de avaliação das predições geradas por nossos modelos, como discutiremos

no capítulo 5. A capacidade de um especialista em definir o tamanho correto da amostra e formas de se amostrar os dados é fundamental para o sucesso da Modelagem Preditiva. Neste intuito, revisaremos neste capítulo alguns dos principais conceitos da Estatística e Teoria da Probabilidade associados à amostragem. Além disso, discutiremos as principais decisões a serem tomadas no processo de coleta, definindo um planejamento amostral apropriado para a Modelagem Preditiva. Por fim, precisamos diferenciar as principais técnicas de amostragem existentes, bem como os erros mais comuns observados em amostras de dados reais.

2.1. Noções básicas de amostragem

A Modelagem Preditiva tem auxílio de uma importante ferramenta oriunda da Estatística para obtenção dos dados de entrada, a Teoria da Amostragem. Tal teoria visa estudar as relações existentes entre uma população e as amostras geradas a partir daquela. As técnicas de amostragem, bem como o planejamento amostral, são amplamente utilizadas em pesquisas científicas e de opinião para se conhecer alguma característica da população. O intuito é definir uma metodologia a ser seguida de forma a garantir que os dados coletados representem adequadamente a população alvo do estudo. Neste sentido, uma das primeiras questões que surgem referente à coleta de dados é quando amostrar? Em alguns cenários, é possível utilizar toda a população de objetos no estudo. Em outros, porém, tais dados não estão disponíveis ou sua utilização por completo se torna inviável ou ineficiente.

A amostragem se mostrar particularmente útil quando observamos pelo menos um dos cenários abaixo:

- **População infinita ou arbitrariamente grande:** em diversos domínios o conjunto de objetos alvo de estudo é arbitrariamente grande, tornando a coleta de toda a população inviável. Por exemplo, suponha que desejamos avaliar todas as buscas realizadas por brasileiros no último ano no intuito de se identificar quais são os assuntos mais buscados. A quantidade de buscas realizadas durante o período alvo de estudo torna inviável o armazenamento e

utilização de toda a população.

- **Estudo de comportamentos dinâmicos**: neste cenário, os comportamentos de interesse sobre uma população mudam frequentemente ao longo do tempo. Assim, para grandes populações, o custo de atualização dos dados torna-se caro. Por exemplo, análise de *trending topics* no Twitter é um estudo com alta dinamicidade sobre um grande volume de dados.

- **Restrições de tempo e custo computacional:** não são raros os domínios em que um estudo ou modelagem tem restrições quanto ao tempo máximo de execução e/ou custos computacionais envolvidos. Muitas vezes, coletar e armazenar os dados referentes à toda a população acaba elevando estes custos e tempos a valores acima do permitido.

- **Predições sobre dados desconhecidos:** em Modelagem Preditiva é comum a realização de experimentações que visam aferir a qualidade do modelo de predição sobre dados desconhecidos para o algoritmo (i.e., não utilizados para a etapa de aprendizado e ajuste de modelos). A forma mais barata e fácil de se conduzir este tipo de análise é utilizar amostras distintas para os passos de aprendizado do modelo e análise de qualidade.

- **Identificação de comportamentos verdadeiramente significativos:** um objetivo de pesquisa muito comum é a validação de determinados comportamentos ou características observadas na população. Alguns destes valores ou correlações obtidas entre distintas características podem ser casuais (i.e., acontecem ao acaso) e não exprimem o real comportamento da população. Uma forma de se validar tais comportamentos consiste em verificar sua ocorrência ou não em distintas amostras da mesma população. A premissa neste caso é que amostras distintas, corretamente coletadas, tendem a exprimir comportamentos ou características intrínsecas à população como um todo.

Nos cenários descritos acima, a utilização de amostras mostra-se mais barata, rápida de obtenção e fácil de ser controlada por envolver operações menores. Porém, há cenários em que o uso de toda a população pode ser mais vantajoso:

- Quando a população é pequena e o custo de se amostrar for similar ao de se utilizar toda a população.
- Quando o tamanho da amostra necessária para conduzir o estudo for muito grande.
- Quando for necessária uma precisão completa sobre os resultados.
- Quando a população for muito heterogênea e o erro introduzido pelo uso da amostra se tornar grande.

Nota: apesar de assumirmos que uma amostra de qualidade exprime fidedignamente os comportamentos e características da população como um todo, na prática, sempre há um desvio. A este desvio damos o nome de erro amostral. Um planejamento amostral sempre leva em consideração o erro introduzido pela amostragem. Quando definimos que um estudo necessita de precisão completa, estamos afirmando que o erro amostral tolerado é zero.

A fim de entendermos melhor os passos e definições necessários para uma correta amostragem de dados, precisamos também entender alguns conceitos básicos oriundos das áreas de Estatística e Teoria da Probabilidade:

- **Experimento ou estudo**: processo com objetivos claros e passos bem delineados, cujo resultado não é determinado com certeza.
- **População ou espaço amostral:** conjunto de todos os objetos que queremos abranger em um estudo e para os quais desejamos obter resultados.
- **Amostra:** subconjunto de objetos da população usado para se obter conclusões acerca da população.
- **Ponto do espaço amostral**: um resultado possível (i.e., um membro do

espaço amostral).

- **Variável aleatória**: variável quantitativa cujo resultado (i.e., valor) depende de fatores aleatórios.
- **Parâmetro:** característica ou comportamento desconhecido da população a qual se tem interesse em estudar. Parâmetros representam quantidades numéricas que podem ser interpretadas pelo pesquisador (e.g., média, mediana, variância, etc.).
- **Predição ou estimativa:** valor estimado a partir dos dados obtidos pela amostra para o parâmetro.
- **Unidade amostral:** qualquer elemento individual da população. As unidades amostrais podem ser os próprios objetos da amostra ou podem ser compostas por grupos de elementos denominados conglomerados.

Momento de reflexão:

Suponha que você trabalhe para o setor de Inteligência Artificial de uma empresa que desenvolve Apps de jogos para celular. Sua empresa deseja conduzir um estudo sobre as dificuldades que os usuários têm em cada fase dos cinco principais games já lançado pela empresa. Mais especificamente, eles gostariam de saber quais perfis de usuários possuem mais dificuldade em cada uma das fases. Perfil, neste caso, é definido em função da idade, sexo, frequência com que joga, escolaridade, horário do dia que joga, região de onde joga, dentre outras informações relevantes. O intuito é a partir dos resultados deste estudo propor um recomendador de jogos para novos usuários, garantindo uma maior satisfação dos usuários da empresa. Considerando que a empresa possui atualmente 30 mil clientes e que tal estudo tenha prioridade total do time para ser corretamente executado, você utilizaria toda a população para estudo ou tentaria obter amostras? Apresente um argumento para a sua escolha.

2.2. Planejamento Amostral

O planejamento amostral é um importante procedimento usado para controle de amostras. Nos planejamentos amostrais, a coleta dos dados deve ser realizada observando-se uma metodologia adequada para que os resultados possam ser extrapolados para a população como um todo. A definição do plano amostral envolve a definição das seguintes informações:

- População alvo: subconjunto de objetos pertinentes para o estudo.
- Unidade amostral: indivíduos ou grupos de indivíduos.
- Tamanho da população: número total de unidades amostrais na população alvo.
- Tamanho da amostra: número total de unidades amostrais na amostra.
- Técnica de amostragem: método de seleção das amostras.

O primeiro passo do planejamento amostral consiste em se definir com clareza qual a população alvo do estudo. Em seguida, é importante especificar qual a unidade amostral a ser considerada. Em grande parte das vezes, estamos interessados em indivíduos do espaço amostral. Há casos específicos, porém, que requer considerarmos conglomerados. Por exemplo, análises de filmes assistidos por usuários em uma plataforma de streaming por categoria de filme (e.g., romance, comédia, etc.). É importante estimarmos o tamanho total da população alvo definida para o estudo. Tal estimativa é importante para as decisões a serem tomadas nos próximos passos.

O próximo passo consiste na determinação do tamanho da amostra. Este passo é, talvez, o principal passo do planejamento amostral. Tal escolha deve levar em conta um erro tolerável e a probabilidade de se cometer tal erro. O erro tolerável é uma margem de erro das estimativas em relação ao parâmetro de interesse, para mais ou para menos, o qual o pesquisador está disposto a aceitar. A forma de se calcular o erro tolerável é amplamente discutida na literatura e pode variar de acordo

com o estudo conduzido [1,3]. Por fim, temos a seleção do método de amostragem. Há diversos métodos com características e premissas bem distintas.

Um erro muito comum em Modelagem Preditiva é o de se subestimar a complexidade e/ou custo do processo de amostragem. Por falta de tempo, muitas vezes, projetos acabam abreviando a correta definição deste processo. Todavia, os custos de uma amostragem mal projetada e/ou conduzida acabam superando em muito os custos necessários para a correta elaboração de um plano amostral. Pode-se, por exemplo, chegar ao fim de um projeto e descobrir que hipóteses e modelos estão sendo validados sobre dados inválidos ou que não representam de forma satisfatória a população alvo.

Nota: há de se considerar ainda dois fatores que afetam o planejamento amostral: custo de execução do plano traçado e heterogeneidade da população alvo. Ambos aspectos afetam incisivamente a escolha do plano amostral mais adequado para cada estudo. Algumas vezes, um plano amostral deve ser redefinido de forma a se adequar aos custos de execução disponíveis (em termos humanos, monetários, computacionais e de tempo necessário).

2.3. Técnicas de amostragem

É a parte da estatística que define os procedimentos para os planejamentos amostrais e as técnicas de estimação utilizadas. As técnicas de amostragem, tal como o planejamento amostral, são amplamente utilizadas nas pesquisas científicas e de opinião para se conhecer alguma característica da população. Técnicas de amostragem são atualmente classificadas em dois grandes grupos: amostragem probabilísticas e não-probabilísticas.

2.3.1. Amostragem probabilísticas

São técnicas de amostragem realizadas segundo critérios bem definidos da teoria estatística das probabilidades. Na amostragem probabilística, todas as unidades da população amostral devem ter a mesma probabilidade de serem selecionadas. A principal premissa, neste caso, é que uma amostra é representativa da população da qual foi selecionada se todos os objetos da população tiverem a mesma chance (i.e., probabilidade) de serem selecionados para compor a amostra.

- **Aleatória simples:** para se obter uma amostra de tamanho *n*, realiza-se *n* iterações selecionando um elemento da população aleatoriamente através de uma distribuição uniforme. Tal amostragem podem ser sem reposição, em que o item selecionado a cada iteração é removido do conjunto de itens candidatos das próximas iterações; ou com reposição, em que um mesmo item pode ser sorteado mais de uma vez.

Exemplo: suponha uma população com 1000 objetos, que numeramos de 000 a 999 para selecionar uma amostra aleatória de n=100 objetos. O processo termina quando for sorteado o elemento 100. A probabilidade de cada elemento ser selecionado a cada iteração é p=1/1000.

- **Aleatória sistemática:** uma amostra de tamanho *n* é obtida em quatro etapas. No primeiro passo, gera-se uma lista ordenada dos elementos da população. No segundo, define-se um intervalo de seleção K. Este intervalo consiste nos K primeiros números naturais e é definido como o tamanho da população (N) dividido pelo tamanho da amostra (*n*). O terceiro e último passo consiste em se definir o ponto de partida (a1), que é um número selecionado aleatoriamente (via amostragem aleatória simples). Por fim, determina-se cada elemento ai da amostra, através da progressão aritmética ai = a1 + (i-1) K.

Exemplo: suponha uma população de 5.000 indivíduos e desejamos obter uma amostra com 100 deles. Em primeiro lugar, ordenamos todos os indivíduos da população. Em seguida, dividimos a população em 100 fragmentos de 50 indivíduos. Selecionamos um número aleatório entre 1 e 50 para extrair o primeiro indivíduo de

forma aleatória, por exemplo, o número 24. A partir deste indivíduo, está definida como será extraída a amostra, com intervalos de 50 unidades, conforme a equação: 24, 74, 124, 174, …, 4.974

- **Aleatória estratificada:** Os elementos da população são divididos primeiramente em grupos chamados estratos, de forma que cada elemento da população pertença a um e somente um estrato. A estratificação é feita para compor grupos mais homogêneos. Existem dois tipos de amostragem estratificada: (1) de mesmo tamanho; (2) proporcional. No primeiro tipo, sorteia-se igual número de elementos em cada estrato. Esse processo é utilizado quando o número de elementos por estrato for aproximadamente o mesmo. No segundo caso, utiliza-se de proporção para determinar o número de elementos de cada estrato que a amostra será composta.

Exemplo: numa localidade com 150.000 habitantes, 45.000 têm menos de 20 anos de idade, 75.000 têm entre 30 e 50 anos de idade e 30.000 têm mais de 50 anos de idade. Uma amostra estratificada de 30 habitantes com tamanho proporcional desta população, seria composto por 9 indivíduos com menos de 20 anos (30% da população), 15 indivíduos entre 30 e 50 anos (50% da população) e 6 indivíduos com mais de 50 anos (20% da população).

- **Por conglomerado de um estágio:** é uma amostra aleatória simples, na qual cada unidade de amostragem é um grupo, ou conglomerado, de elementos. O primeiro passo para se usar este processo é especificar conglomerados apropriados. Os elementos em um conglomerado devem ter características similares. Como regra geral, o número de elementos em um conglomerado deve ser pequeno em relação ao tamanho da população, e o número de conglomerados, razoavelmente, grande.

Exemplo: ao estimar a proporção de pessoas idosas em certa cidade, pode-se considerar como conglomerados os bairros deste município, as ruas, os quarteirões ou as residências. Neste caso, sorteia-se alguns conglomerados e os objetos destes constituirão a amostra desejada.

- **Por conglomerado de múltiplos estágios:** caracterizada por unidades populacionais arranjadas em uma hierarquia. Em cada estágio amostra-se unidades amostrais de um nível da hierarquia, usando-se a amostragem por conglomerado de um estágio.

Exemplo: deseja-se avaliar o nível de conhecimento em matemática de crianças do ensino básico no estado de Minas Gerais. Utilizando-se a amostragem por conglomerado de múltiplos estágio, primeiro, seleciona-se aleatoriamente quais escolas participaram do estudo. Posteriormente, em cada escola selecionada, selecionamos aleatoriamente quais turmas serão analisadas. Por fim, dentro de cada turma selecionada, amostramos algumas crianças.

Nota: tanto no caso da amostragem estratificada como no da amostragem por conglomerado, a população deve estar dividida em grupos. Na estratificada, entretanto, seleciona-se uma amostra aleatória simples dentro de cada grupo (estrato). Por outro lado, na amostragem por conglomerado selecionam-se amostras aleatórias simples dos grupos (conglomerados), e todos os itens dentro dos grupos selecionados farão parte da amostra.

Momento de reflexão:

Numa fábrica, produz-se em média 1.000 bobinas de aço por dia. Chega-se à conclusão de que é necessário avaliar no controle de qualidade 40 dessas bobinas. Determine qual técnica de amostragem utilizar e quais bobinas poderiam compor a amostra, de modo que esta seja uma representativa da produção diária.

R: amostragem sistemática com intervalo de seleção entre 1 e 25.

2.3.2. Amostragem não-probabilística

Na amostragem não-probabilística, ou intencionada, há uma escolha deliberada da amostra. Neste caso, objetos distintos possuem chances distintas de entrarem na amostra. A premissa assumida é que, para se analisar com profundidade e de maneira adequada algumas hipóteses, é necessário gerar amostras enviesadas que nos permitem observar a ocorrência ou não da hipótese estudada. Um olhar homogêneo sobre toda a população, pode ocultar determinados comportamentos que estamos interessados em entender melhor. Em geral, são vistos como técnicas pragmáticas, pois possibilitam um estudo mais rápido e com menores custos.

- **Por conveniência:** elementos são incluídos na amostra sem probabilidades previamente especificadas, ou conhecidas, de eles serem selecionados. Não tem valor científico. Tem a vantagem de permitir que a escolha de amostras e a coleta de dados sejam relativamente fáceis de acordo com o que for mais conveniente para quem está realizando a pesquisa. Este tipo de amostragem é bom para fazer um teste piloto de um questionário que será utilizado em uma pesquisa posterior.

Exemplo: os fabricantes e as agências de propaganda costumam fazer entrevistas em shoppings para obter informações sobre os hábitos dos consumidores e a eficiência de anúncios. Uma amostra de clientes de um shopping é rápida e barata. “A entrevista em shoppings resulta principalmente de um problema de custo”, afirmou um perito ao New York Times.

- **Intencional ou por julgamento:** o pesquisador avalia quais pessoas detém maior conhecimento do tema a ser estudado e escolhe os elementos que julga serem os mais representativos da população. Seleciona-se as unidades da amostra segundo um determinado perfil definido de acordo com os objetivos da pesquisa. No estudo comparativo, certas características são comparadas em duas, ou mais, populações, através de amostras escolhidas por julgamento. Nos estudos comparativos, normalmente não se busca a generalidade, mas sim as diferenças entre os grupos em análise. Nesse contexto, as amostras devem ser o mais similares possíveis, diferindo apenas em relação ao fator de

comparação.

Exemplo: estudo sobre a percepção do conceito de morte em crianças de diferentes períodos de desenvolvimento cognitivo (subperíodo pré-operacional, subperíodo das operações concretas, período formal). Estudo comparativo da incidência de câncer de pulmão em grupos de Fumante e Não Fumantes.

- **Snowball:** método de amostragem frequentemente utilizado em pesquisas por entrevista. Primeiramente, deve-se encontrar um conjunto inicial de indivíduos que atenda os objetivos da pesquisa. A cada um que se enquadra, o entrevistador pede que este lhe indique onde é possível encontrar outro para entrevistar, até chegar ao número de entrevistas desejadas. É uma técnica de amostragem bastante útil quando se pretende estudar pequenas populações muito específicas. No entanto, pode originar resultados enviesados, uma vez que as pessoas tendem a indicar amigos ou conhecidos que tendem a ter o mesmo comportamento ou opinião.

Exemplo: opinião dos torcedores do América sobre o time. Primeiro, encontra-se um torcedor do América. Feita a entrevista, o entrevistado poderá indicar onde encontrar outros torcedores do América e assim sucessivamente.

- **Por quotas:** método usualmente trabalhado em levantamento de mercado e em prévias eleitorais. É a amostragem por estratificação, porém não existem sorteios. Para cada entrevistador é atribuída uma cota de entrevistas e este escolherá pessoas que estejam dentro do perfil da pesquisa.

Exemplo: classificação da população em termos de propriedades que se sabe, ou presume, serem relevantes para a característica a ser estudada.

Exemplo Guia: Planejamento amostral

Voltando a nossa tarefa de recomendar anúncios para usuários da plataforma da Nozama, nosso próximo passo consiste em projetar um plano amostral completo. Neste intuito, primeiro definimos que a população alvo da nossa tarefa é o subconjunto de usuários ativos.

Ou seja, de todos os usuários cadastrados na plataforma, estamos interessados apenas em usuários que fizeram pelo menos uma compra na plataforma nos últimos três meses. Além disso, sabemos que a nossa unidade amostral é usuário, uma vez que a KPI do projeto foi definida em função do número de usuários ativos. O tamanho da população alvo é obtido ao analisarmos os dados da plataforma. Sabemos que a Nozama possui 15 milhões de usuários cadastrados. Porém, apenas 1,2 milhão foram considerados ativos. Em seguida, com base em uma taxa de erro amostral, custos de execução, tempo de experimentação dentre outros fatores analisados previamente pelo time, definimos que para a etapa de aprendizado utilizaremos apenas 10% da população alvo como tamanho da amostra. Por fim, uma vez que não há restrições e condições específicas e claras num primeiro momento, e que o objetivo é que haja um aumento da KPI considerando toda a população de usuários ativos, adotaremos uma amostragem aleatória simples.

2.4. Qualidade das amostras

Ao utilizar amostras para a condução de qualquer estudo científico, tal como a Modelagem Preditiva, uma primeira preocupação deve ser quanto a qualidade destas amostras. Quando falamos em qualidade de amostras, consideramos especificamente três características fundamentais que qualquer amostra deve apresentar e que pesquisadores devem ser capazes de identificar e mensurar.

Primeiro, é importante que os pesquisadores compreendam a população caso a caso e testem a amostra para obter **Consistência** antes de prosseguir com o estudo. Isso é especialmente crítico para pesquisas que visam analisar mudanças comportamentais ao longo do tempo, em que precisamos ter confiança de que qualquer comportamento ou mudança observados nos dados reflete um comportamento ou mudança real na população. A segunda característica refere-se à **Diversidade** da amostra. Garantir a diversidade da amostra é também crucial, pois alcançar de maneira homogênea partes da população com comportamentos distintos, pode ser difícil.

Para ser verdadeiramente representativa, uma amostra deve ser tão diversificada como a própria população e sensível às diferenças locais inerentes às diferentes partes da população. Por fim, temos a **Transparência** do processo de amostragem. Existem várias restrições que determinam o tamanho e a estrutura da população. É fundamental que pesquisadores discutam essas limitações e mantenham a transparência sobre os procedimentos seguidos ao selecionar a amostra.

Dessa forma, o uso de amostras requer diversos cuidados e um bom planejamento, a fim de se evitar erros. Na prática, amostras são imperfeitas e incompletas. Cabe ao pesquisador controlar tais imperfeições durante o processo de coleta, de forma a minimizar os erros existentes em cada amostra extraída da população. Basicamente, a literatura classifica estes erros inerentes ao processo de amostragem em dois tipos: erros amostrais e não-amostrais. Discutiremos em detalhes cada um destes tipos de erros no restante desta seção.

2.4.1. Erros amostrais

O erro amostral é definido como sendo a diferença entre a estimativa obtida para um parâmetro e o seu verdadeiro valor. Este tipo de erro é decorrente da variabilidade natural das unidades amostrais, uma vez que as observações são feitas apenas em uma amostra e não em toda a população. Ou seja, é plausível esperar que amostras tenham comportamentos similares ao da população, mas não idêntico. O erro amostral depende de fatores tais como o tamanho da amostra, a variabilidade das características de interesse na população, projeto amostral e o método de seleção das amostras. Por exemplo, para um determinado tamanho de amostra, o erro amostral dependerá do procedimento de estratificação empregado, da escolha das unidades de amostragem e do método de seleção.

A característica mais importante da amostragem é que o erro amostral pode ser medido a partir da própria amostra. Normalmente, o erro amostral é medido pela variabilidade esperada da estimativa quando comparada ao valor verdadeiro,

expressa como uma porcentagem da estimativa. Esta medida é denominada **Coeficiente de Variação** (CV) [6,7].

Como regra de ouro, diversos estudos assumem que amostras com valores de CV abaixo de 10% são boas, enquanto aquelas com CV entre 10% e 20% são aceitáveis. Amostras com valores de CV acima de 20% devem ser descartadas e o projeto experimental deve ser revisado por completo.

2.4.2. Erros não-amostrais

Denominamos como erro não amostral qualquer tipo de erro decorrente da coleta dos dados. Os erros não-amostrais podem ser introduzidos em qualquer etapa da coleta de dados (e.g., falta de resposta, diferenças na interpretação de perguntas ou informações incorretas dos entrevistados) e processamento dos dados (e.g., como codificação, entrada de dados, transformação, ponderação, tabulação, etc.).

É fundamental minimizar erros de não-amostragem através de um controle de qualidade e análise de dados. Caso contrário, as possíveis distorções introduzidas por este tipo de erro podem comprometer seriamente um plano amostral tecnicamente perfeito e invalidar os resultados obtidos. De maneira geral, sobre um grande número de observações, erros ocorridos aleatoriamente terão pouco efeito nas estimativas da pesquisa. No entanto, erros que ocorrem sistematicamente contribuirão para um viés nas estimativas da pesquisa. Os erros não-amostrais são classificados em quatro tipos principais:

- **Erros de valores ausentes:** este erro consiste na incapacidade de se medir em todas as unidades amostrais todas as variáveis de interesse. Os valores ausentes produzem erros nas estimativas da pesquisa de duas maneiras. Em primeiro lugar, os valores ausentes muitas vezes possuem características que se diferem dos valores presentes nas amostras, resultando em amostras enviesadas caso a ausência de valores não seja corrigida adequadamente. Quanto maior a taxa de valores ausentes, maior será o risco de viés presente na amostra. Em segundo lugar, ter um número maior de valores ausentes reduz

o tamanho efetivo da amostra. Como resultado, a precisão das estimativas diminui (i.e., o erro de amostragem nas estimativas aumenta). Este segundo aspecto pode ser contornado selecionando-se um tamanho de amostra maior. No entanto, essa estratégia não reduzirá o viés nas estimativas.

- **Erros de cobertura:** os erros de cobertura ocorrem quando há diferenças entre a população alvo e a população amostrada (i.e., população coberta). A cobertura aumentada, geralmente, ocorre quando a população amostrada é maior que a população alvo e não é um problema, uma vez que as unidades fora do escopo na amostra são tipicamente identificadas durante a coleta de dados e podem ser removidas posteriormente. No entanto, a cobertura reduzida em que a população amostrada torna-se menor que a alvo é problemática. Neste caso, uma primeira tarefa seria identificar as características da população alvo ausente na coleta. Na maioria dos casos, é improvável que os itens ausentes sejam aleatórios.

- **Erros de medição:** os erros de medição ocorrem quando uma resposta fornecida difere do valor real. Por exemplo, quando pedimos a distância de uma viagem, poucas pessoas conhecem essas informações com um alto nível de precisão. Neste caso, os erros podem ser atribuídos ao entrevistado, ao entrevistador, ao questionário, ao método de coleta ou ao sistema de registro de dados, dentre outros. Existem várias fontes de erro de medição: viés social sobre o comportamento (e.g., tipo de viagens que as pessoas realizam), efeito entrevistador (e.g., pessoa versus web); influência política; memória; imprecisão (e.g., arredondamentos). De maneira similar ao tipo de erro de valores ausentes, os métodos para se reduzir erros de medição baseiam-se, em grande parte, em melhores design de pesquisa.

- **Erros de processamento:** podem ocorrer erros em várias etapas do processamento e armazenamento das amostras, incluindo captura, codificação e transformação dos dados. Neste caso, é necessário se definir um conjunto de procedimentos completos e sistematizados para avaliar a qualidade de cada variável de interesse e corrigir os erros identificados. Alguns exemplos das

verificações de processamento de dados são: revisão de todos os fluxos de perguntas em questionários, verificação do intervalo de valores de cada variável e análise de distribuição de valores ausentes antes e após cada etapa de transformação, dentre outros.

Capítulo 3. Pré-processamento de Dados

Este capítulo tem como objetivo tornar o aluno apto à:

- Realizar os principais passos de pré-processamento sobre dados numéricos.
- Realizar os principais passos de pré-processamento sobre dados categóricos.
- Realizar os principais passos de pré-processamento sobre dados textuais.
- Identificar os principais desafios no tratamento de dados.
- Identificar os erros mais comuns que afetam a qualidade dos dados.

"Torture os dados e eles confessarão alguma coisa."

Ronald Coase (Prêmio Nobel de Economia - 1991)

Consistência e confiabilidade são pré-requisitos fundamentais para que possamos explorar qualquer tipo de dados em Modelagem Preditiva. Porém, dados do mundo real, muitas vezes, não possuem tais características. Não são raros os cenários em que observamos dados incompletos, ruidosos ou mesmo duplicados. Por este motivo, o correto pré-processamento de tais dados é uma das etapas mais importantes do processo de Modelagem Preditiva. Como dito anteriormente, dado o cuidado necessário, bem como os possíveis casos de inconsistência existentes, o pré-processamento é usualmente a mais trabalhosa e demorada etapa de todo processo de Modelagem Preditiva. Note que, além da consistência e confiabilidade, o pré-processamento adequado dos dados afeta outras dimensões de qualidade, tais como Acurácia, Completeza, Interpretabilidade, dentre outras.

Nota: diferentemente do capítulo 2 ao qual nos referimos à qualidade das amostras, neste capítulo estamos assumindo que o processo de amostragem está adequado. As dimensões de qualidade aqui mencionadas referem-se às características intrínsecas aos dados coletados. Assim, por exemplo, uma amostra pode estar consistente com sua população de origem, mas os dados presentes nessa

população podem não ter consistência entre si, já que podem apresentar duplicações ou erros referentes aos dados em si.

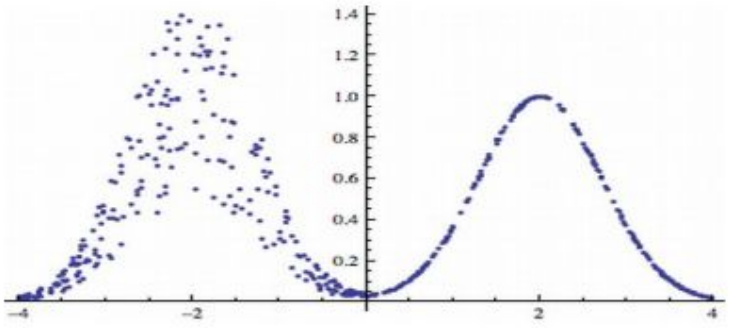
Para um correto tratamento sobre os dados coletados é importante, primeiramente, um conhecimento sobre os tipos de problemas que podem ocorrer nos dados. Existem quatro grandes problemas usualmente tratados de maneira distinta na literatura: **Incompletude**, **Inconsistência, Duplicação** e **Ruídos**.

Afirmamos que um conjunto de dados está incompleto quando há ausência de valores de alguns atributos, ausência de determinados atributos de interesse ou quando a coleção contém apenas dados agregados (e.g., médias). Por exemplo, ao coletar informações demográficas sobre os usuários de um sistema, a informação de renda per capita pode estar ausente para diversos usuários. Dados Incompletos decorrem de valores de dados não aplicáveis a algumas observações durante a coleta, ou mesmo devido a diferentes considerações e metodologias entre o tempo de coleta e tempo de análise.

Por sua vez, dados inconsistentes são aqueles que contém discrepâncias entre valores esperados e observados. Este tipo de erro decorre da consolidação de diferentes fontes de dados ou violação de dependências. Por exemplo, um mesmo usuário pode conter em fontes distintas valores diferentes de idade.

Como terceiro grande problema temos a duplicação de registros, que ocorre em virtude de uma metodologia de coleta equivocada ou erros de verificação. Por fim, ruídos são vistos como erros ou *outliers* presentes na coleção de dados. No exemplo acima, um *outlier* seria observar um usuário com renda de R$ 100.000,00 por mês. Para determinadas análises e objetivos de aprendizado poderia ser relevante retirar este registro. Tais erros podem ocorrer devido a falha nos instrumentos de coleta; erros humanos; erros na entrada de dados; erros na transmissão de dados; ou variações naturais dos dados. A figura 3.1 exemplifica as distorções que os ruídos podem introduzir ao analisarmos uma população.

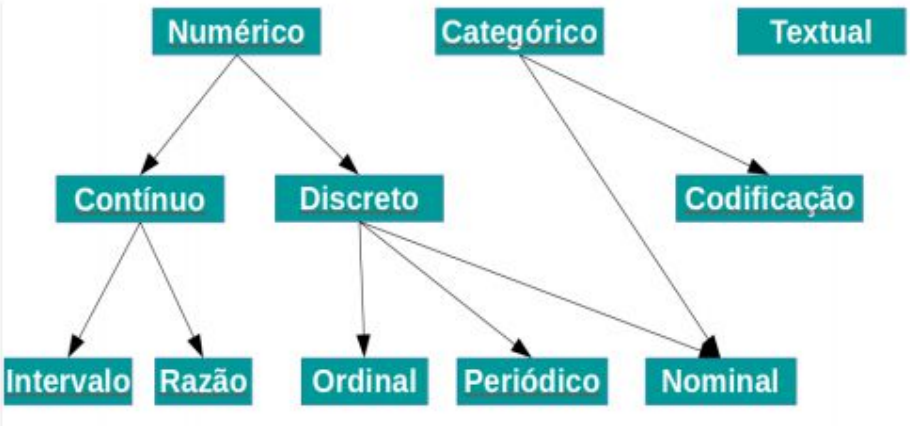
Figura 3.1 - Distribuição de ocorrência de dados ruídos versus dados reais. Note que o excesso de ruídos pode alterar fundamentalmente a modelagem que teremos sobre a população.



3.1. Tipos de dados

De modo geral, podemos explorar diversos tipos de dados na Modelagem Preditiva. Cada tipo define um conjunto de operações pertinentes, significados e transformações que podem ser executadas. A figura 3.2 sumariza os principais tipos de dados comumente utilizados na literatura. Na tabela abaixo descrevemos cada um destes tipos, bem como apresentamos as operações e transformações relevantes, além de alguns exemplos práticos de dados provenientes de cada tipo.

Figura 3.2 - Representação esquemática dos principais tipos de dados existentes.



Tipo de dados	Descrição	Exemplo	Operações	Transformações
Nominal	Os valores são apenas atributos nominais que proveem informação suficiente para distinguir um objeto do outro (=,).	CEP, número de identidade, cor dos olhos e gênero.	Entropia, correlação, contingencia, etc.	Qualquer permutação dos valores.
Ordinal	Os valores proveem informação suficiente para se definir uma ordem relativa entre os objetos (<,>).	Número de ruas, escala de rigidez de minerais e tempo de vida no sistema.	Média, percentis, correlação de rank, etc.	Operações que alteram os valores mas preservam a ordem relativa.
Intervalo	Para este tipo de dados, a diferença entre valores possui significado, existindo uma unidade de medida (+,-).	Datas, temperatura em Celsius ou Fahrenheit.	Média, desvio padrão, Pearson's correlation, t-test, etc.	Operações lineares (e.g., novo_valor = a* valor_antigo + b).
Razão	Para este tipo de dados, ambos, diferença e razão, possuem significado (*, /)	Temperatura em Kelvin, quantidades monetárias, idade, tamanho e peso.	Média geométrica, média harmônica, porcentagem de variação, etc.	Operações lineares (e.g., novo_valor = a* valor_antigo + b).

Periódico	Este tipo de dados armazena informações sobre unidades de repetição de determinados eventos.	Semana, meses e horas.	Frequência de ocorrência, tamanho do período e concatenação de períodos.	Operações que permitem dilatar ou encolher os períodos analisados, bem como alterar a unidade de mensuração.
Codificação	Esquema que converte campos categóricos em um subcobjunto especial de numerais com significado categórico.	*Dummy coding* ou *Effects coding.*	Operações devem ser feitas sobre o método de codificação e não sobre dados de saída.	Mapeamento entre esquemas de codificação distintos.
Textual	Dados não estruturados representados por uma sequência de símbolos alfanuméricos usados na comunicação verbal humana.	Posts, avaliações textuais de apps e páginas web.	Análise Léxica, remoção de stopwords, aplicação de Stemming, etc.	*Raw textual data* pode ser convertido em sequência de palavras processadas.

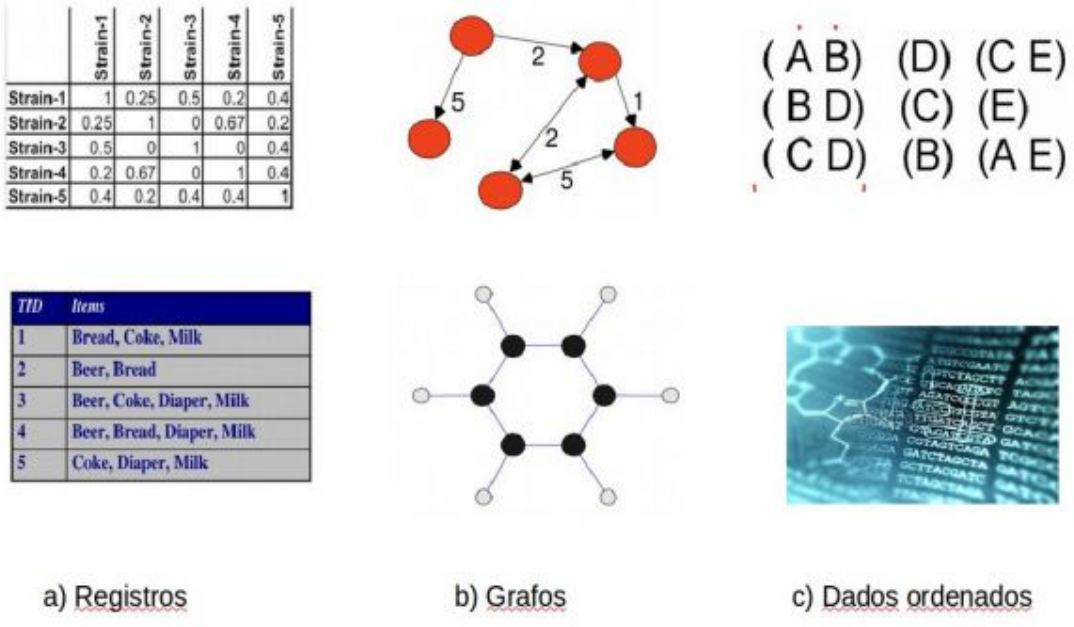
Além de tipos diferentes, é importante mencionarmos que existem representações, ou modelos, distintos para os dados. Essa modelagem determina como os dados serão armazenados e processados pelos algoritmos de Modelagem Preditiva, afetando inclusive as estruturas de dados necessárias para uma manipulação eficiente de grande volume de dados. De modo geral, podemos distinguir três modelos de dados principais:

1. **Registros**: dados que consistem em uma coleção de registros, cada um dos quais possui um conjunto fixo de atributos. *Exemplos: matrix de dados, documentos e transações.*

2. **Grafos**: uma estrutura de dados de grafo consiste em um conjunto finito (e possivelmente mutável) de vértices ou nós, juntamente com um conjunto de pares não ordenados desses vértices para um grafo não direcionado ou um conjunto de pares ordenados para um grafo direcionado. Esses pares são conhecidos como arestas ou linhas para um grafo não direcionado e como setas, ou arestas, direcionadas para um grafo direcionado. Os vértices podem ser parte da estrutura do grafo, ou podem ser entidades externas, representadas por índices ou referências inteiras. *Exemplos: web, estruturas moleculares e redes sociais.*

3. **Tipos ordenados**: sequência de eventos ou itens em que a ordem de ocorrência carrega uma informação relevante e útil para a análise e compreensão dos dados. *Exemplos: dados espaciais, dados temporais e DNA.*

Figura 3.3 - Principais modelos de representação de dados.



3.2. Tratamento de dados numéricos e categóricos

O pré-processamento de dados é um passo frequentemente negligenciado, mas importante, no processo de Modelagem Preditiva. Os métodos de coleta de dados geralmente são fracamente supervisionados, resultando em dados com valores fora do intervalo aceitável (e.g., renda familiar: -100), combinações de dados impossíveis (e.g., sexo: masculino, grávida: sim), valores faltantes, etc. Utilizar dados que não foram cuidadosamente examinados podem produzir resultados enganosos. Assim, a representação e a qualidade dos dados vem antes de tudo. Como dito, este tratamento varia de acordo com o tipo de dados de entrada. Nesta seção, discutiremos o tratamento de dados numéricos e categóricos, que possui cinco passos principais:

1. Limpeza dos dados;
2. Integração;
3. Transformação;
4. Redução;
5. Discretização.

Limpeza dos dados:

Consiste no processo de se preencher valores ausentes, atenuar distorções em dados ruidosos, identificar e remover outliers e duplicações, bem como resolver inconsistências observadas nos dados. Dados sujos podem causar confusão para o procedimento de aprendizado. Embora a maioria das rotinas de Modelagem Preditiva tenham alguns procedimentos para lidar com dados incompletos ou ruidosos, eles nem sempre são robusto. Portanto, um passo de pré-processamento útil consiste em executar seus dados através de algumas rotinas de limpeza de dados. Como principais estratégias para o tratamento de valores ausentes, temos:

1. Ignorar observações;

2. Preencher manualmente;
3. Usar uma constante global ou valores médios;
4. Preencher automaticamente via Inferência Bayesiana, método de Expectation Maximization (EM) ou Árvores de Decisão.

No caso de remoção de ruídos, as estratégias mais comuns são análise de variabilidade das variáveis, detecção manual de valores suspeitos ou métodos de bonificação e atenuação de valores. Por fim, para o correto tratamento de outliers, destacamos o uso de algoritmos de agrupamento: *fitting* de curvas ou teste de hipótese sobre algum modelo assumido como verdadeiro para os dados.

Integração:

Objetiva combinar dados de múltiplas fontes de dados e metadados em uma única fonte coerente. Essas fontes podem incluir vários bancos de dados, cubos de dados ou arquivos planos. Há uma série de questões a serem consideradas durante a integração de dados. Por exemplo, a integração do esquema pode ser complicada. Outro desafio refere-se ao problema de resolução de entidades. Como ter certeza que entidades do mundo real, oriundas de múltiplas fontes de dados, são correspondentes? Por exemplo, uma dada ocorrência da palavra Lula se refere ao ex-presidente Luis Inácio Lula da Silva ou a um animal marinho.

Redundância de dados é outra questão importante. Um atributo pode ser redundante se puder ser derivado de outra tabela. Por fim, esta etapa também envolve a detecção e resolução de conflitos de valores. Para uma mesma entidade do mundo real, valores oriundos de diferentes fontes podem divergir. Uma integração cuidadosa dos dados de várias fontes pode ajudar a reduzir/evitar redundâncias e inconsistências, bem como melhorar a qualidade dos dados.

Transformação:

Na transformação de dados, os dados são transformados ou consolidados em formas apropriadas de acordo com os métodos de Modelagem Preditivas a serem

aplicadas posteriormente. A transformação de dados pode envolver as seguinte tarefas:

1. **Agregação ou sumarização de valores:** aplicação de alguma função de agregação de distintos valores em um único. *Exemplo: uso de média, mediana.*

2. **Generalização**: uso de ontologia ou hierarquia de conceitos para substituir conceitos específicos por outros mais genéricos. *Exemplo: Ontologia de cargos para representar empregos.*

3. **Normalização**: alteração dos valores dos atributos, para que caiam sobre um determinado intervalo de interesse. *Exemplo: uso do z-score e normalização min-max.*

4. **Construção de features (*Feature Engineering*)**: criação de novos atributos/variáveis a partir dos dados. *Exemplo: geração de tuplas a partir dos dados.*

Discretização:

Consiste em transformar valores numéricos contínuos em valores discretos. Diversos dos algoritmos de Modelagem Preditiva existentes são capazes de extrair conhecimento de bancos de dados que armazenam atributos discretos apenas. Se o atributo for contínuo, os algoritmos podem ser integrados com algoritmos de discretização que os transformam em características discretas. Os métodos de discriminação são usados para reduzir o número de valores para determinados atributos contínuos, dividindo o intervalo do atributo em distintos valores discretos. Dessa forma, a discretização pode tornar a etapa de aprendizado mais precisa e mais rápida. Os resultados do processo, geralmente, são mais compactos, mais curtos e mais precisos quando recursos discretos são usados em comparação com recursos contínuos. Além disso, o critério de desempenho mais importante do método de discretização é a taxa de precisão. Existem um grande número de métodos de discretização. Alguns dos principais métodos estão listados abaixo:

- **Discretização não-supervisionada:**
  - Intervalos iguais: divide-se os números em intervalos de mesmo tamanho.
  - Intervalos uniformes: usa-se intervalos contendo o mesmo número de valores.
- **Discretização supervisionada:**
  - Limites de classe: identifica-se intervalos de valores associados a distintas classes.
  - Entropia: gera-se intervalos de valores que gere a menor entropia sobre algum atributo usado como classe.

Redução:

O aprendizado de modelos preditivos sobre enormes quantidades de dados pode demorar muito tempo, tornando este tipo de análise impraticável ou inviável. Técnicas de redução de dados são úteis para se obter representação reduzida do conjunto de dados, sem comprometer a integridade dos dados originais e ainda produzindo um conhecimento de qualidade. Reduzir consiste em representar de uma forma mais compacta, sem perda de informações importantes. A redução é comumente entendida como redução do volume ou redução das dimensões (número de atributos).

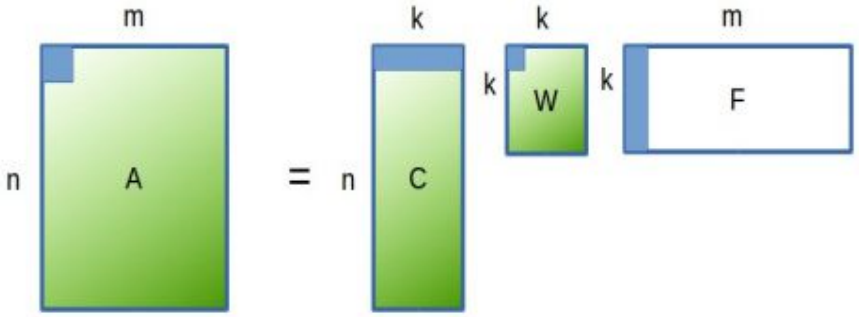
Dessa forma, técnicas de redução da dimensionalidade se tornaram fundamentais para diversas aplicações e métodos de Modelagem Preditiva. Podemos dividir técnicas de redução em três grandes grupos:

1. **Seleção de características (*Feature selection*):** neste conjunto de técnicas o objetivo é atribuir a cada atributo um valor de relevância, frente as demais informações que já existem na coleção. Um exemplo deste tipo de métrica é o *Information Gain.* Uma vez assinalado a um valor de relevância, apenas métricas que possuem uma relevância superior a um *threshold* são mantidas na coleção. Todas as demais são descartadas.

2. **Redução de características (*Feature Reduction*):** este conjunto de técnicas não descarta nenhuma informação original. A premissa é que todas as informações originalmente presentes na coleção podem ser expressas de maneira mais compacta, usando-se outros 'atributos' derivados. Estes atributos derivados podem ser vistos como combinações lineares dos atributos originais. O principal exemplo deste tipo de abordagem são as técnicas de Fatoração de Matrizes.
3. **Compressão de dados:** conjunto de técnicas usadas, principalmente, para se reduzir o volume de dados de entrada, selecionando um subconjunto de dados representativos. A premissa é que em coleções grandes, uma amostra representativa da população é mais que suficiente para se aprender sobre as características desejadas de toda a população. Métodos baseados em amostragem compõem esta categoria de métodos.

Dada a popularidade de uso e eficiência em cenários práticos, discutiremos em mais detalhes as técnicas baseadas em Fatoração de Matrizes (Redução de Características). Fatorar uma dada matriz de entrada, que representa os nossos dados, consiste em representar tal matriz através do produto de outras matrizes, tal como mostra a figura 3.5. A grande vantagem de se utilizar estas matrizes resultantes é que além de terem um número menor de dimensões, elas, em geral, possuem características úteis para o métodos de Modelagem Preditiva.

Figura 3.4 - Fatoração de matrizes.



Nesta nova representação, cada entrada da matriz A torna-se um tipo de combinação ponderada de pedaços de informação contido nas matrizes C e F. Esses

pedaços de informação são denominados **fatores latentes**. O papel de k é forçar uma representação mais compacta, em geral, muito menor que o número de dimensões originais m. A matriz **W** é sempre diagonal e define a relevância de cada dimensão de informação contida em C e F. Algumas técnicas de decomposição não geram a matriz **W.** Essas técnicas diferem-se sobre as premissas feitas sobre os fatores latentes.

Dentre os vários métodos aplicados em Modelagem Preditiva para se realizar a Fatoração de Matrizes, destacamos a Decomposição por Valor Singular (SVD - *Singular Valor Decomposition*) e a Análise de Componentes Principais (PCA - *Principal Component Analysis*), dada a popularidade e generalidade de aplicação de ambos. SVD tem por objetivo realizar a redução de posto e a aproximação de redução de baixo-posto de uma matriz, que representa dados expressos em um espaço N dimensional. A importância deste método decorre, sobretudo, das propriedades algébricas úteis e atrativas apresentadas, derivadas de observações teóricas e geométricas sobre transformações lineares. Em virtude dessas propriedades, SVD é comumente aplicada visando um de três objetivos distintos:

1. Eficiência de manipulação de dados matriciais;
2. Redução da dimensionalidade dos dados para consequente diminuição da complexidade de um problema;
3. Remoção de ruídos dos dados originais.

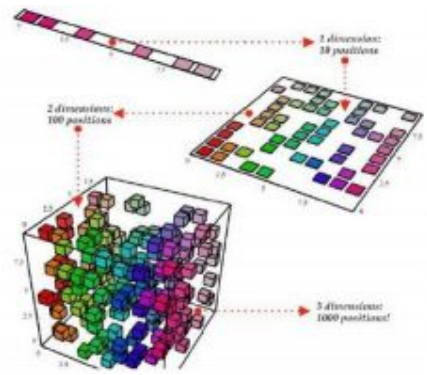
O PCA, por outro lado, é uma técnica estatística multivariada que possui como objetivo explorar a estrutura de variabilidade dos dados. Seus objetivos são, de maneira geral, (1) a redução de dados e (2) a interpretação dos dados. A partir da decomposição dos dados em um espaço ortonormal, o PCA representa grande parte da variabilidade dos dados originais através de um número usualmente menor de dimensões. A aplicação de técnicas como o SVD e PCA durante a fase de pré-processamento dos dados possui três funções importantes na análise de dados:

1. **Limpeza dos dados:** esse conjunto de técnicas nos permite separar automaticamente os diferentes processos subjacentes aos dados.

2. **Reduzir redundâncias**: representa uma maneira elegante de identificar similaridades e redundâncias entre objetos ou atributos em coleções de dados.
3. Endereçar um problema conhecido como o **Mal da Dimensionalidade.**

O Mal da Dimensionalidade refere-se a um problema resultante do grande número de variáveis (i.e., atributos) envolvidas. Este problema resulta do fato que um número fixo de pontos se tornam crescentemente “esparsos” a medida que o número de dimensões aumenta. Um exemplo prático que nos permite entender este fato no mundo real é o seguinte: suponha que você seja um caçador de animais (qualquer tipo). Seria muito fácil caçar uma barata que consegue caminhar apenas em uma tubulação de esgoto linear (uma aproximação para um cenário unidimensional). Também é fácil caçar um cachorro e talvez pegá-lo se estivesse correndo por uma planície (duas dimensões). Porém, é muito mais difícil caçar pássaros, que agora têm uma dimensão extra, uma vez que podem voar pela planície. Se fingimos que os fantasmas são seres de dimensão superior, esses são ainda mais difíceis de se caçar.

Figura 3.5 - Ilustração do problema conhecido como Mal da Dimensionalidade.



A principal consequência do Mal da Dimensionalidade é que muitas técnicas da Modelagem Preditiva dependem criticamente de medidas baseadas em distância ou similaridade de objetos no espaço. Estudos demonstraram que tais medidas são significativamente afetadas em espaços com altas dimensões:

$$\lim_{d\to\infty} \frac{MaxDist - MinDist}{MinDist} \to 0$$

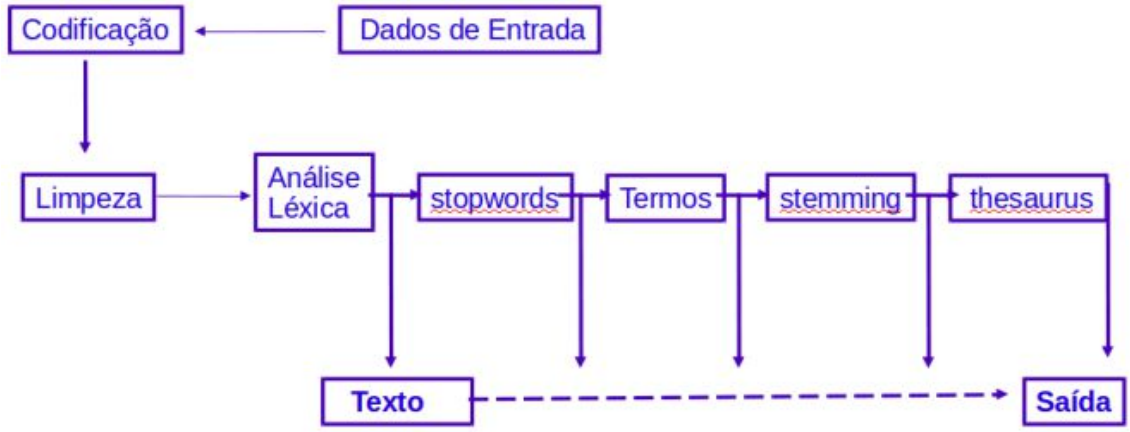
Em espaços com alta dimensão, distâncias entre pontos se tornam relativamente uniformes. Ou seja, algoritmos de Modelagem Preditiva baseados em distância não são capazes de dizer quais objetos estão próximos uns dos outros em altas dimensões. O Mal da Dimensionalidade ainda é frequentemente considerada, em termos de viabilidade computacional de tentar se estimar funções em altas dimensões.

3.3. Tratamento de dados textuais

Nem sempre documentos estão representados de maneira adequada e precisamos transformá-los a fim de melhorar a precisão na posterior manipulação dos dados. Existem vários passos neste tratamento. Os principais passos aplicados em Modelagem Preditiva são listados abaixo. A figura 3.6 apresenta a sequência em que comumente estas etapas são aplicadas sobre os dados.

1. Identificação da codificação;
2. Limpeza do texto;
3. Análise léxica/transformações;
4. Eliminação de Stopwords;
5. Stemming;
6. Indexação;
7. Aplicação de Thesaurus.

Figura 3.6 - Fluxo de tratamento dados textuais



Identificação da codificação:

Dados textuais são representados computacionalmente através de códigos definidos por um sistema de codificação, tal como ASCII ou Unicode. Dessa forma, o primeiro passo na manipulação deste tipo de dados é identificar corretamente a codificação que o texto se encontra. Ao falar sobre codificação de dados, temos dois processos:

- **Encoding**: transforma um conjunto de caracteres de uma codificação específica para uma sequência de bytes.
- **Decoding**: transforma uma sequência de bytes em um conjunto de caracteres especificados por uma codificação qualquer.

Como nem sempre sabemos a codificação que um texto possui, precisamos primeiro tentar identificar esta codificação e, em seguida, realizar o processo de **encoding** sobre os dados, transformando os dados textuais em sequências de bytes. Por fim, realizamos o processo de **decoding** sobre os dados, transformando as sequências de bytes na codificação textual desejada. Este processo pode ser trabalhoso, uma vez que dados da web podem possuir mais de uma codificação.

Limpeza do texto:

Tal como realizado sobre dados numéricos e categóricos, nesta etapa de limpeza visamos remover ou corrigir dados incompletos, inconsistentes ou mesmo errados. Preenchimento de valores inexistentes, atenuação de dados ruidosos, resolução de inconsistências, identificação e remoção de desvios são as principais tarefas realizadas nesta etapa. Abaixo ilustramos alguns problemas comumente observados em dados textuais, que são resolvidos nesta etapa de limpeza de texto.

- **Valor ausentes:** falta de preenchimento de atributos obrigatórios.
- **Erro ortográfico:** anomalia encontra-se associada aos atributos textuais.
  - Exemplo: cidade = 'Brga'
- **Utilização de sinônimos:** Informações sintaticamente distintas, mas semanticamente iguais.
  - Exemplo profiss = 'professor' profiss = 'docente'.
- **Inexistência de representação padrão:** o valor do atributo aparece sob variados formatos.
  - Exemplo: data = 04/05/2004, data = 4 de Julho de 2004.

Note que é importante sistematizar a limpeza de dados, adaptando-se a diferentes domínios e a variedade de tipos de atributos presentes em cada coleção. Para facilitar este processo, há na literatura diversas ferramentas comerciais e acadêmicas disponíveis para este fim. Dada a complexidade e custo de uma limpeza de textos ampla e robusta, muitas vezes estratégias simples são adotadas.

Análise léxica/transformações:

Processo de conversão de uma sequência de caracteres em uma sequência de palavras. Envolve, dentre outras coisas, reconhecimento de dígitos, remoção de hífens, pontuações e caixa alta. Além disso, transformações nessa fase são aplicadas aos dados numéricos, uma vez que este tipo de dados não é útil para análise de

dados textuais, dado a variedade de valores que possam assumir. O primeiro passo da análise léxica consiste, basicamente, na remoção de caracteres especiais e pontuações dos documentos de entrada. Para documentos HTML, há ainda a remoção das tags. Por fim, desabilita-se a caixa alta de todas as letras, remove-se caracteres não pertencentes ao vocabulário desejado, bem como acentuações. Note que o objetivo é reduzir as variações de escrita relacionadas a cada termo, aumentando assim a cobertura de documentos que possuem cada termo distinto presente na coleção de entrada. Por sua vez, no tratamento de hifenação, separamos palavras constituídas por hífen.

Apesar de ser uma transformação útil que visa normalizar a ocorrência de alguns termos, ela pode distorcer a semântica das palavras. É necessário definir uma regra geral bem como exceções. Por exemplo, como representar os termos guarda-chuva, guarda chuva e guardachuva?

Eliminação de Stopwords:

Palavras que são muito frequentes entre os documentos de uma coleção não são boas como discriminantes. Por exemplo, uma palavra que ocorre em 80% dos documentos de uma coleção é inútil para os propósitos de classificação. Tais palavras são frequentemente chamadas de stopwords e são normalmente removidas dos documentos antes de se aplicar os métodos de aprendizado da Modelagem Preditiva. Assim, esta etapa é vista como uma filtragem de termos não-úteis para diversos propósitos. Como termos removidos nesta etapa, temos: artigos, preposições, conjunções (e.g., portanto, logo, pois, como, dentre outras) e alguns verbos.

Como principais vantagens do processo de stopwords temos a redução do tamanho da estrutura de indexação e o foco do aprendizado a partir de dados textuais, em termos semanticamente mais relevantes. Apesar dos benefícios, a eliminação de stopwords pode atrapalhar o aprendizado em alguns casos, reduzindo a revocação dos modelos. Por exemplo, se o objetivo é modelar a semântica por trás de frases tais como 'Ser ou não ser, eis a questão', a remoção de stopwords comprometeria totalmente o aprendizado. Além disso, o processo de definição das stopwords é usualmente manual, sendo dependente da linguagem dos dados e trabalhos.

Stemming:

Stemming é o processo de extrair o prefixo de uma palavra, baseada em seu radical ou forma raiz. O objetivo é tratar a variação sintática de termos. Stemming é útil, por exemplo, para máquinas de busca, facilitando a expansão de consultas. Os algoritmos de criação funcionam cortando o fim da palavra e, em alguns casos, também o começo, enquanto procura a raiz. Para realizar o stemming, usualmente criamos uma tabela que mapeia cada inflexão para um radical. Com isso, basta procurar cada palavra do texto na tabela e substituí-la pelo seu radical.

A desvantagem desta estratégia consiste no tempo dispendido para construir essa tabela. Além disso, remover o sufixo de uma palavra pode acarretar em perda semântica. Por exemplo, ao aplicar stemming nos termos “presidente” e “presidiário”, ambos remeteriam ao mesmo radical: “presid”. Este corte indiscriminado pode ser bem sucedido em algumas ocasiões, mas nem sempre, e por isso esta é uma abordagem que oferece algumas limitações.

Com o intuito de se contornar as limitações do processo de stemming, alguns estudos propõem o uso da lematização. O objetivo de ambos os processos é o mesmo: reduzir as formas e derivações inflexivas de cada palavra para uma base ou raiz comum. Quando estamos modelando o significado de um termo, queremos encontrar tantos documentos quanto possível, e isso inclui não apenas a palavra exata, mas também documentos que possuem outros termos com a mesma raiz. Por exemplo, quando analisamos a palavra estudante, irá enriquecer nossas modelagens se tivermos resultados que contenham palavras como estudar ou estudioso. Para obter essas palavras relacionadas, a lematização leva em consideração a análise morfológica das palavras. Para tanto, o processo faz uso de dicionários capazes de mapear as palavras ao seu lema. A principal diferença é que um lema é a forma básica de todas as inflexões de um termo.

Indexação:

Um índice é uma estrutura de dados construída sobre dados textuais e usada para acelerar o acesso a este tipo de informação. O objetivo do armazenamento de

um índice é otimizar velocidade e desempenho no acesso de documentos relevantes. Seu uso mais comum é em máquinas de busca, porém alguns algoritmos de Modelagem Preditiva podem fazer uso de índices, no intuito de se manipular grandes volumes de textos mais rapidamente. Sem um índice, o acesso às palavras presentes nos documentos exigiria a varredura de cada documento na coleção, o que exigiria um tempo de execução e poder de computação consideráveis. Por exemplo, enquanto um índice de 100.000 documentos pode ser consultado em milissegundos, uma varredura sequencial de cada palavra em 100.000 documentos grandes pode levar minutos. O armazenamento de dados adicionais necessários para armazenar o índice, bem como o aumento considerável no tempo necessário para uma atualização deste, são trocados pelo tempo economizado durante o acesso às informações.

Aplicação de Thesaurus:

O thesaurus é um instrumento que reúne termos escolhidos a partir de uma estrutura conceitual previamente estabelecida e destinados à indexação e à recuperação de documentos e informações num determinado campo do saber. Não é simplesmente um dicionário, mas um instrumento que garante aos pesquisadores o processamento e a busca destas informações. O intuito é agrupar as palavras com mesmo significado. Dessa forma, um thesaurus é visto como um vocabulário controlado, utilizado para as tarefas de indexação e processamento de textos. O objetivo é prover um vocabulário padronizado, realizando uma “classificação” semântica dos termos. As principais características que diferenciam um thesaurus de um simples vocabulário controlado são: no thesaurus cada termo corresponde a um conceito. Uma vez aceito, esse termo torna-se um "descritor" ou um "indexador". Caso o termo não seja aceito como "descritor", ele pode ser aceito como "remissivo", isto é, remete a um termo descritor; e todos os termos estão relacionados entre si. Nenhum termo pode figurar no thesaurus sem estar relacionado a algum outro, sendo essa relação determinada pelo seu significado.

Capítulo 4. Aprendizado de Modelos Preditivos

Este capítulo tem como objetivo tornar o aluno apto à:

- Conhecer os principais métodos de regressão.
- Conhecer os principais métodos de classificação.
- Identificar as premissas de utilização dos principais métodos da Modelagem Preditiva.
- Ajustar corretamente modelos aprendidos para a tarefa de predição.
- Identificar os principais desafios relacionados à aprendizagem de modelos preditivos.

"Todo método de aprendizado deve incorporar algum conhecimento ou premissas, além dos dados que lhe são apresentados, para ser capaz de generalizar."

Pedro Domingos

Poucas áreas em Ciência da Computação apresentaram um crescimento tão notório quanto o experimentado pelo Aprendizado de Máquinas (AM) recentemente. Embora sua definição não seja consensual, muitos estudiosos concordam que o AM está intimamente relacionado com a tradicional área de Inteligência Artificial, dedicando-se, sobretudo, ao desenvolvimento de técnicas e algoritmos que permitem ao computador "aprender" com experiências passadas.

Essa capacidade de aprendizado é justamente o que muitos pesquisados buscam para resolver problemas relacionados à Modelagem Preditiva.

Mais formalmente, o processo de Aprendizado de Máquinas refere-se ao projeto de sistemas capazes de "aprender" regras a partir de dados, adaptar-se a mudanças, e melhorar sua eficácia com base em observações analíticas e experiência adquiridas sobre um domínio. Ou seja, dado um conjunto de dados D, uma tarefa T a ser desempenhada e uma medida de desempenho M, um sistema é

dito aprender a partir de D para desempenhar uma tarefa T, se após o sistema aprender, seu desempenho sobre T, mensurado por M, melhora.

De maneira geral, há diversas formas de se aprender a partir dos dados contidos em D. Trabalhos diversos na literatura classificam os métodos de aprendizado de acordo com o tipo de dados de entrada exigidos. Um primeiro grupo é denominado de **Aprendizado Supervisionado**, por necessitar que todos os exemplos contidos em D sejam rotulados.

Este primeiro grupo gera como saída uma função de mapeamento das entradas para as saídas desejadas (i.e., rótulos). Um segundo grupo é composto por técnicas chamadas **Não Supervisionado**, uma vez que os dados de entrada não são rotulados. Neste capítulo vamos revisar alguns dos principais métodos oriundos da AM, tanto supervisionados quanto não supervisionados, usados para aprender modelos usados em Modelagem Preditiva.

Cabe ainda salientar que, mesmo dentro de cada grupo, há uma grande diversidade de técnicas que assumem um conjunto de premissas e conceitos distintos para o processo de aprendizado. Técnicas baseadas em Árvores de Decisão, modelos Probabilístico Bayesianos, Redes Neurais e *Support Vector Machines* estão entre as técnicas mais populares e eficazes na literatura. Cada uma destas técnicas difere desde o modelo de descrição dos dados à forma como proveem as informações de saída.

Dada a grande variedade de técnicas existentes, um dos problemas primordiais que surge consiste em selecionar a melhor técnica para o problema tratado. De fato, não há uma técnica igualmente eficaz para todos os problemas. A qualidade do aprendizado depende de características intrínsecas a cada domínio, bem como de características que definem o conjunto de dados D. Tais características podem ser relevantes o suficiente para invalidar o conjunto de premissas assumidas por uma técnica específica. Por exemplo, utilizar um algoritmo probabilístico Bayesiano para um conjunto de dados cujos atributos sejam altamente dependentes pode não ser uma boa estratégia. Isto porque tais técnicas assumem, por simplificação, a independência entre atributos. Dessa forma, entender a fundo cada

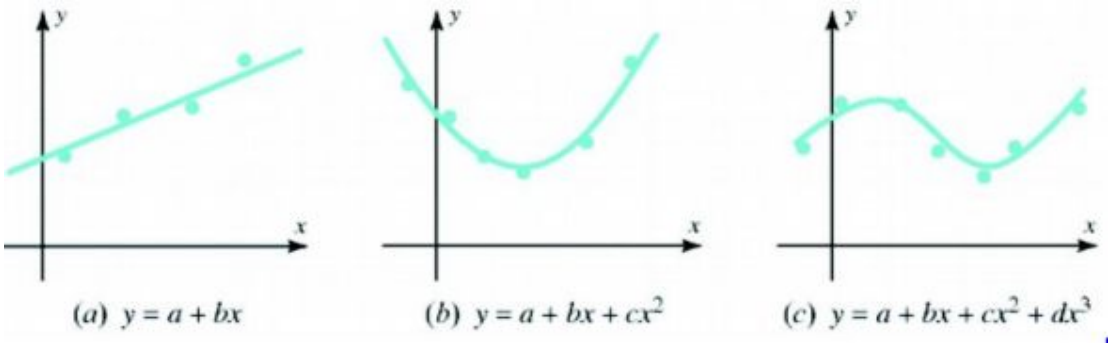
domínio de análise, bem como os compromissos presentes na tarefa de aprendizado em questão é uma tarefa extremamente relevante.

4.1. Modelos de regressão

Regressão linear:

Análise de regressão é um processo estatístico que visa estimar a relação entre variáveis. Regressão nos ajuda a entender como o valor de uma variável dependente muda quando uma das variáveis independentes se altera. Consequentemente, somos capazes de fazer predições mais precisas uma vez que entendemos tais variações e conseguimos mensurar a variação de algumas dessas variáveis (i.e., as variáveis independentes). O objetivo de predição, neste caso, é uma função das variáveis independentes, denominada de **função de regressão**. Em regressão linear, dados são modelados usando-se funções de regressão lineares e parâmetros desconhecidos do modelo são estimados a partir dos dados.

Figura 4.1 - Exemplos de funções de regressão linear.



Para se determinar uma função de regressão linear, as variáveis dependentes e independentes são, em geral, organizadas de forma a compor um sistema de equação linear a ser resolvido. Um sistema linear é uma coleção de duas ou mais equações lineares envolvendo o mesmo conjunto de variáveis. Por sua vez, uma equação linear é uma equação algébrica na qual cada termo é uma constante, ou o

produto de uma constante e uma variável simples (operadores lineares), tal como mostra a equação abaixo.

$$yi = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in} + \varepsilon_i$$

Dessa forma, o termo "linear" refere-se à ocorrência dos coeficientes de regressão β_j. É importante frisar que "linear" é uma restrição muito mais fraca que aparenta ser. Por exemplo, uma equação do tipo $y = \beta_0 + \beta_1 x + \beta_2 x^2$, tal como na Figura 4.1 (b), pode ser considerada uma equação linear se pensarmos que as variáveis x e x^2 são variáveis distintas (i.e, dimensões distintas).

Para se determinar a solução de um problema de regressão linear, precisamos determinar os coeficientes de regressão β_j? Tais coeficientes são usualmente determinados resolvendo-se o denominado **problema dos Quadrados Mínimos**. Dado um sistema linear de *m* equações sobre *n* variáveis, o problema dos Quadrados Mínimos visa encontre o vetor *x* que minimiza a equação abaixo com respeito ao produto euclidiano no espaço *R*.

$$y = Ax + \boldsymbol{\varepsilon}$$

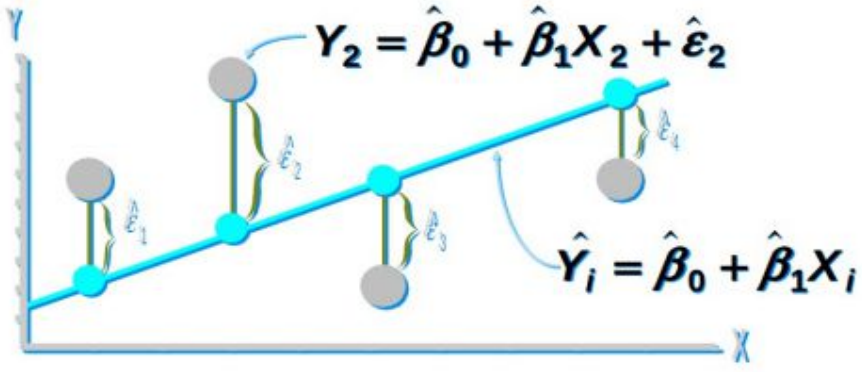
Denominados o vetor *x* como a solução quadrados mínimos do sistema; A como uma matriz representando os dados de entrada; e **ε** como o vetor de erros quadrados. Tal como ilustra a figura 4.2, o termo "solução quadrados mínimos" resulta do fato de que minimizar:

$$||y\text{-}Ax - \boldsymbol{\varepsilon}||$$

Representa minimizar:

$$||y\text{-}Ax||^2 = \boldsymbol{\varepsilon 1^2 + \varepsilon 2^2 + \ldots + \varepsilon m^2}$$

Figura 4.2 - Representação do problema dos quadrados mínimos.



A minimização dos quadrados ao invés da função linear é amplamente utilizada devido a quatro razões principais:

1. Ao somar quadrados, evitamos de misturar valores positivos e negativos;

2. Funções quadráticas enfatizam magnitudes maiores e, consequentemente, penalizam erros maiores;

3. A forma quadrática é a mais simples função que possui exatamente um único ponto de mínimo (ou de máximo);

4. Lembramos que achar o mínimo da função quadrático é extremamente simples e barato, bastando derivar a função e igualá-la a zero.

Regressão logística:

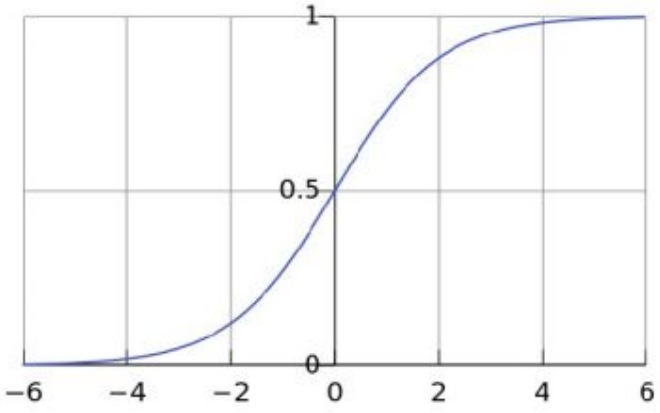
A regressão logística é o principal método recomendado para situações em que desejamos obter uma função de regressão e a variável dependente é de natureza dicotômica ou binária. Neste caso, a função de regressão é definida através de uma função logística, ao invés de uma função linear utilizada na regressão linear. A figura 4.3 apresenta um exemplo de função logística. Este método produz uma estimativa de **probabilidade** para cada objeto pertencer a uma de duas classes. Os resultados

da análise ficam contidos no intervalo de zero a um. Trata-se de um método rápido de se aplicar que funciona para conjuntos de dados relativamente grandes.

Como algumas principais aplicações da regressão logística podemos destacar:

- **Previsão de risco na área tributária:** calcular a probabilidade do contribuinte ser inadimplente após o parcelamento de tributos.
- **Classificação de empresas:** classificar se a empresa encontra-se no grupo de empresas solventes ou insolventes.
- **Identificação de pacientes doentes:** determinar os fatores que caracterizam um grupo de indivíduos doentes em relação aos indivíduos sãos.

Figura 4.3 - Exemplo de função logística.



Tal como na regressão linear, o aprendizado da função logit consiste em se estimar os coeficientes β_0, β_1, ... , β_p a partir do conjunto dados, pelo método estatístico da Máxima Verossimilhança (MLE - *Maximum likelihood Estimation*). Tais coeficientes maximizam a probabilidade de a amostra ter sido observada.

Como restrições para aplicação deste método, é importante mencionar que não deve haver outliers nos dados, os quais podem ser removidos normalizando-se

as variáveis independentes via *z-score* e removendo valores maiores que 3.29 e menores que -3.29. Além disso, não deve haver intercorrelações elevadas entre as variáveis independentes. Outra prática comumente adotada no aprendizado da regressão logística é o uso da função *probit,* ao invés da função logística, em alguns cenários. Na maioria dos casos, um modelo é construído com ambas as funções e a função com melhor ajuste é escolhida. A função *probit* assume a distribuição normal da probabilidade do evento, enquanto a função *logit* assume a distribuição do log. Na função *logit* a distribuição condicional P($y|x$) é uma distribuição de Bernoulli, em vez de uma distribuição Gaussiana.

Como principais vantagens da regressão logística destacamos:

- Facilidade para lidar com variáveis independentes categóricas.
- Fornece resultados em termos de probabilidade.
- Facilidade de classificação de indivíduos em categorias.
- Requer pequeno número de suposições.
- Alto grau de confiabilidade.

4.2. Modelos de classificação

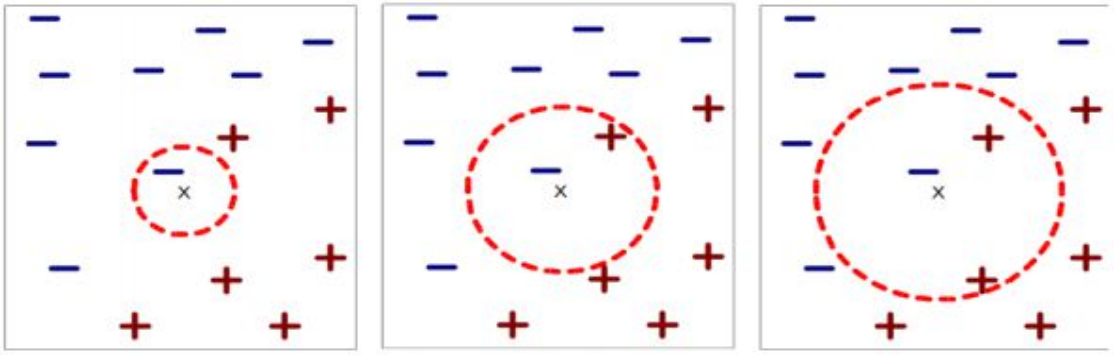
KNN:

O KNN (*k-Nearest Neighbor*) é um algoritmo supervisionado e postergado de classificação, uma vez que a construção de um modelo de classificação é adiada até um exemplo de teste ser apresentado para a classificação. O KNN amostra o conjunto de treinamento (i.e., conjuntos de dados em que as classes de todos os exemplos são conhecidos) no momento da classificação com base nos atributos presentes no exemplo de teste. Dessa forma, este algoritmo decide a qual classe c_i um exemplo

de teste d_t deve ser atribuído, verificando as classes dos k exemplos mais similares à d_t presentes em um conjunto de treinamento.

Cada um destes exemplos similares é tido como um voto para a classe à qual ele pertence, e cada voto é ponderado pela similaridade entre o exemplo de teste d_t e o exemplo considerado. Os dados são representados em um espaço vetorial e os vizinhos são definidos em termos de distância. No cálculo do vizinho mais próximo, a função alvo pode retornar valores discretos ou reais. Para valores discretos, o kNN retorna a moda entre os k exemplos de treinamento mais próximos. Normalmente, utiliza-se valores reais e essa distância é definida em termos da distância cosseno ou distância euclidiana. A figura ilustra esse processo de classificação realizado pelo KNN.

Figura 4.4 - Exemplo de funcionamento do KNN. Uso de vizinhanças de tamanho 1,2 e 3 para classificar um exemplo x com classe desconhecida.



Está técnica é tipicamente uma técnica de pesquisa empregada principalmente na análise de prognósticos. Por exemplo, podemos estimar a renda de um indivíduo de uma população, pesquisando k=20 vizinhos mais próximos do mesmo, pelos valores dos atributos bairro de moradia, profissão, escolaridade e idade. Um dos problemas da aplicação dessa técnica é a necessidade de existir nos registros um número de atributos suficientes para determinação da vizinhança. Outra questão importante relacionada ao uso prático do KNN refere-se ao valor de K. Não

existe um único K ótimo para todos os cenários, tampouco uma forma de se selecionar o melhor valor de K em cada caso. Em geral, este valor é determinado experimentalmente através de uma heurística simples:

- Comece com k = 3 e use um conjunto de teste para validar a taxa de erro do classificador.
- Repita com k = k + 2.
- Escolha o valor de k para o qual a taxa de erro é mínima.
- Tente sempre utilizar um número ímpar para K, de forma a reduzir as chances de empates durante a classificação

CART:

As árvores de classificação e regressão (CART) são árvores binárias não-paramétricas, que produzem árvores de classificação ou regressão dependendo se a variável dependente é categórica ou numérica, respectivamente. Desenvolvido por Breiman, Friedman, Olshen e Stone no início dos anos 80, este método introduziu modelagem estatística baseada em árvores na tarefa de classificação. Trata-se de uma abordagem rigorosa envolvendo validação cruzada para selecionar a árvore ideal. O funcionamento consiste na construção iterativa de árvores binárias através do seguinte algoritmo:

1. Comece pelo nodo raiz;
2. Analise todos os dados de entrada;
3. Inspecione todos os possíveis valores de todas as variáveis (força-bruta);
4. Selecione o variável/valor (X=vi) que produz a melhor separação dos dados de acordo com a função objetivo;
5. Para cada observação, se X < vi, então coloque a observação no nodo esquerdo;
    1. Caso contrário, assinale-a para o nodo direito.

6. Repita o processo para cada nodo.

O principal ponto do algoritmo é, a cada iteração, selecionar o valor de variável que produza a melhor separação dos dados de entrada, de acordo com a função objetivo. Essa separação pode ser definida de diversas formas:

- Árvores de regressão: possui uma função objetivo contínua e a medida de separação usada é a soma dos erros quadrados (SSE).
- Árvores de classificação: possui uma função objetivo categórica e as principais medidas de separação são: Gini measure e Entropia, dentre outras que visam estimar a pureza de cada nodo na árvore.

Árvores de Decisão e Ensemble Trees:

Amplamente utilizadas em algoritmos de classificação, as árvores de decisão são representações simples do conhecimento, e um meio eficiente de se construir classificadores que predizem ou revelam classes ou informações úteis, baseadas nos valores de atributos de um conjunto de dados. Uma árvore de decisão é, essencialmente, uma série de declarações *if-then-else*, que quando aplicados a um registro de uma base de dados, resulta na classificação daquele registro. Trata-se de uma ferramenta de suporte à decisão que usa estrutura hierárquica para representar visualmente decisões. A árvore de decisão chega a sua decisão pela execução de uma sequência de testes. Cada nó interno da árvore corresponde a um teste do valor de uma das propriedades, e os ramos deste nó são identificados como os possíveis valores do teste. Cada nó folha da árvore especifica o valor de retorno se a folha for atingida.

O mais interessante sobre árvores de decisão não é a sua construção a partir de classificação de um conjunto de treinamento, e sim a sua habilidade de aprendizado. Quando o treinamento é finalizado, é possível alimentar sua árvore de decisão construída a partir de exemplos com novos casos, a fim de classificá-los.

Além disso, uma das principais vantagens dessa técnica é que ela gera uma das estruturas de dados mais fáceis de se entender com uma boa representação gráfica. A figura 4.5 apresenta uma árvore de decisão construída para decidir em que

circunstâncias climáticas podemos jogar tênis. Note que a representação em si é altamente intuitiva, bastando percorrer a árvore a partir da raiz até uma folha para se chegar a uma decisão.

Figura 4.5 - Exemplo de árvore de decisão para se avaliar em quais condições climáticas devemos jogar tênis.

Na prática, um grande problema no aprendizado de árvores de decisão é sobre a árvore ótima. Em geral, quanto maior a árvore, menor é a sua capacidade de generalização. Recentemente, a solução mais adotada para resolver este problema consiste na utilização de um *ensemble* (i.e., conjunto) de árvores de decisões distintas para se tomar uma decisão. A premissa básica é que muitos classificadores fracos (i.e., que não são capazes de generalizar as soluções), conseguem ser bons em diferentes partes do espaço de entrada, tornando seu uso conjunto melhor que o uso de um único classificador fraco. Neste material, discutiremos mais afundo dois métodos específicos baseados em ensembles de árvores de decisão: *Random Forests* e *Boosted trees*.

Random Forest:

O algoritmo de Random Forest (RF) foi introduzido por Breiman em 2001 e é um método de ensemble que utiliza classificadores do tipo árvore. O RF constrói uma grande quantidade de árvores de decisão a partir de um único conjunto de treinamento. Basicamente, o método gera diferentes amostras aleatórias a partir do

conjunto de treino original, e cada amostra é usada para construir cada uma das árvores de decisão que compõe a RF. Cada árvore de decisão é apontada como um componente preditor.

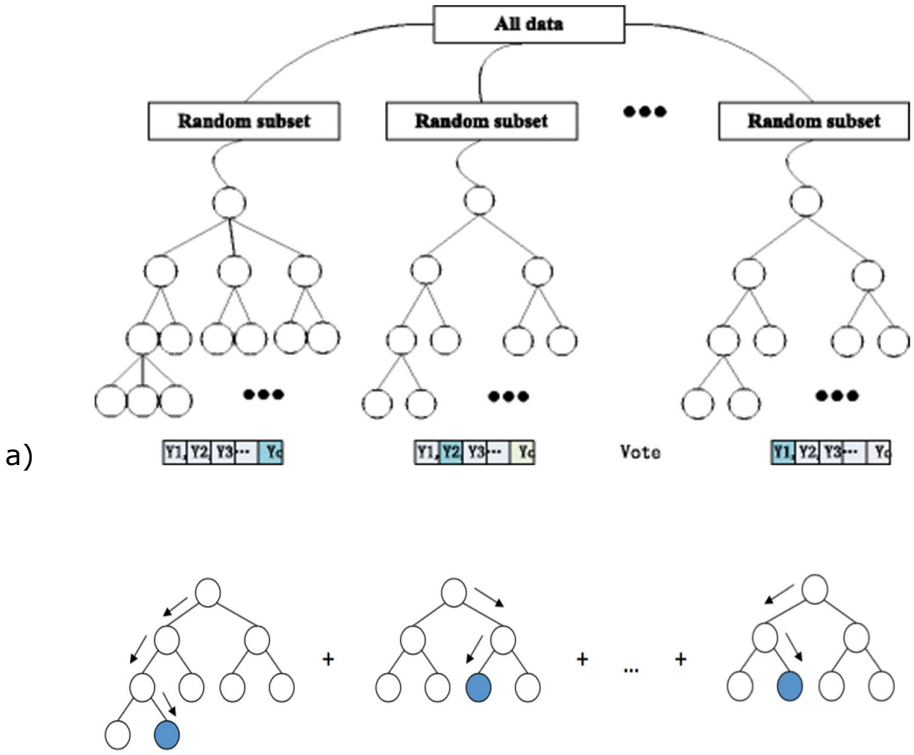
O algoritmo inclui duas fases importantes: o período de criação de cada árvore e o período de votação. A primeira fase consiste em treinar cada árvore de decisão a partir de uma amostra gerada. Na segunda, os dados de teste são classificados pela votação majoritária em que cada árvore define um voto para uma única classe. O RF é um dos classificadores mais populares para dados densos. Tal popularidade deve-se a facilidade de implementação e paralelização.

Boosted Trees:

Diferentemente do RF, neste método uma única árvore de decisão é gerada como classificador final. Porém, para se obter esse classificador final parte-se de uma árvore inicial que é iterativamente modificada através do aprendizado de distintas árvores de decisão sobre o conjunto de treinamento modificado. É necessário que os classificadores base do algoritmo sejam treinados sequencialmente, visando definir os padrões que irão constituir os próximos conjuntos de treinamento. A cada passo, os objetos “difíceis” (i.e., objetos cujo algoritmo apresentou predições erradas) são reponderados, de forma a ganhar mais representatividade na coleção, e o algoritmo é reexecutado nesse novo cenário. Assim, o mesmo algoritmo é executado de forma recursiva até convergir.

A figura 4.6 ilustra a diferença entre o método *Random Forest* e *Boosted Trees*.

Figura 4.6 - Ilustração do processo de aprendizado realizado pelos métodos Random Forest (a) e Boosted Trees (b).



a)

b)

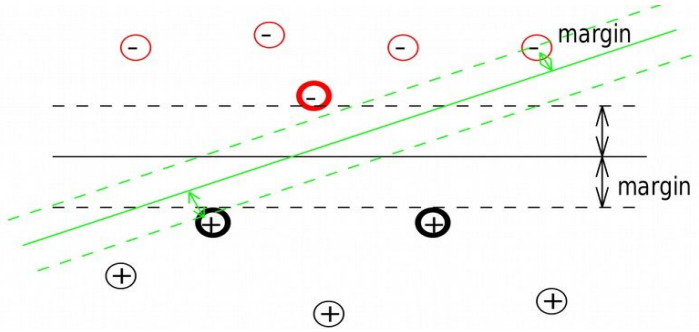
SVM:

Support Vector Machines é um dos principais métodos de classificação já propostos na literatura. Este método é baseado no princípio de Minimização de Riscos Estruturais da teoria da aprendizagem computacional. A ideia principal de minimização de risco estrutural consiste em encontrar um hiperplano separador, para que possamos garantir o menor erro esperado, tal como ilustra a figura 4.7. O erro esperado de h e a probabilidade de h não classificar corretamente um exemplo de teste não conhecido e selecionado aleatoriamente.

O SVM utiliza um método de treinamento muito eficiente que permite encontrar o hiperplano ótimo, ou próximo ao ótimo, para um grande número de cenários reais. Por este motivo, é um método amplamente utilizado para classificação de documentos, manuscritos, imagens dentre outros. Note que, por definição, o SVM

executa uma classificação binária, uma vez que usa um único hiperplano que divide o espaço de features em duas áreas. Para se realizar classificação multiclasse (i.e., quando se tem mais de duas possíveis classes de saída), geralmente, modela-se o problema como N problemas de classificação binária (i.e., para cada uma das N classes possíveis verificamos se um exemplo teste pertence ou não àquela classe) e, ao fim, seleciona-se o hiperplano que retorna o menor erro.

Figura 4.7 - Hiperplano separador usado pelo SVM para realizar uma classificação binária.



Como o SVM sempre gera um separador linear, outro questionamento comum quanto ao seu uso é sobre essa restrição de linearidade do problema. É possível usar SVM para resolver problemas não lineares? A resposta é sim. Sabe-se que, se os dados forem mapeados em um espaço de dimensão suficientemente alta, então eles sempre serão linearmente separáveis. O problema é que nem sempre é possível identificar um hiperplano separador em dimensões elevadas devido ao mal da dimensionalidade, tal como discutido no capítulo 3.

A solução adotada neste caso é utilizar as denominadas ***funções kernel.*** Funções kernel são funções que retornam o produto escalar das imagens de seus argumentos. Em outras palavras, podemos dizer que funções kernel nos permitem calcular o produto escalar de pontos no espaço euclidiano original, como se estivessem em um espaço maior (ou menor). Como o SVM usa apenas o produto escalar entre pares de pontos para definir os hiperplanos, o uso de kernels se tornou predominante. Existem diversas funções kernel que podem ser usadas, tais como

classificadores polinomiais, funções de base radial, a abordagem de margem máxima, dentre outros.

Redes neurais:

As redes neurais compreendem procedimentos computacionais que envolvem o desenvolvimento de estruturas matemáticas com habilidade de aprendizado. Representam o esforço de investigações acadêmicas para implementar, computacionalmente, a maneira com a qual o cérebro humano funciona. São programas que implementam detecções sofisticadas de padrões e algoritmos de aprendizado de máquina para construir modelos, principalmente, de prognóstico de grandes bancos de dados históricos. Está baseada nos conceitos de como um cérebro humano está organizado e como ele aprende. Existem duas estruturas principais: (1) o nó, que corresponde ao neurônio; (2) o link, ou vínculo orientado, que corresponde as conexões entre neurônios.

Um vínculo da unidade *j* para a unidade *i,* serve para propagar a ativação *aj* desde *j* até *i*. Cada vínculo tem um peso numérico *Wji* associado a ele, o qual determina a intensidade e o sinal da conexão. Cada unidade *i* calcula primeiro uma soma ponderada de suas entradas:

$$in_i = \sum_{j=0}^{n} W_{j,i} a_j$$

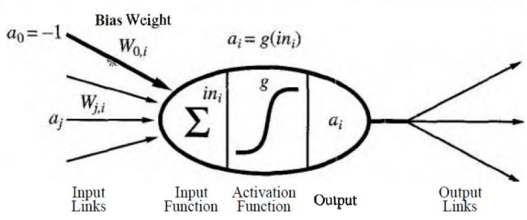
Então, ela aplica uma função de ativação *g* a essa soma para derivar a saída:

$$a_i = g(in_i) = g\left(\sum_{j=0}^{n} W_{j,i} a_j\right)$$

A função de ativação *g* é projetada para atender a duas aspirações. Primeiro, que a unidade seja “ativa” (1) quando as entradas “corretas” forem recebidas e que a unidade seja “inativa” (0) quando as entradas “erradas” forem recebidas. Segundo, a ativação precisa ser não-linear, caso contrário a rede neural inteira entrará em colapso, tornando-se uma função linear simples. A figura 4.8 ilustra como todos estes

conceitos são mesclados de forma a representar o funcionamento dos neurônios em redes neurais.

Figura 4.8 - Representação de neurônios em redes neurais.



Existem duas categorias de estruturas de redes neurais. Primeiro, temos as denominadas redes acíclicas ou **redes de alimentação direta.** Este tipo de rede representa uma função de sua entrada atual e não tem nenhum estado interno além dos pesos propriamente ditos. A segunda categoria compreende as redes cíclicas ou **redes recorrentes**. Neste caso, utiliza-se as saídas da rede para alimentar de volta suas próprias entradas. Os níveis de ativação da rede formam um sistema dinâmico que pode atingir um estado estável ou exibir oscilações, ou mesmo apresentar um comportamento caótico. A resposta da rede a uma determinada entrada depende de seu estado inicial, que pode depender de entradas anteriores.

4.3. Meta-Aprendizado

O desempenho dos diferentes algoritmos de aprendizado varia de acordo com o domínio e o conjunto de dados sob análise. Cada método apresenta pontos fortes e pontos fracos, não existindo um único método ótimo para todos os cenários. Além disso, dada uma mesma quantidade de informação para treinamento, é amplamente aceito que o desempenho e a confiabilidade de muitos classificadores é geralmente melhor do que um único classificador. Por este motivo, um tipo de aprendizado que vem ganhando notoriedade em Modelagem Preditiva é denominado meta-aprendizado. Meta-aprendizado significa aprender por meio da experiência quando diferentes vieses podem ser utilizados para tratar um problema. Meta-

aprendizado difere do aprendizado convencional ou base, no escopo do nível de adaptação. O aprendizado base ocorre no nível dos exemplos e o viés é fixado a priori, enquanto em meta-aprendizado é escolhido dinamicamente com base no acúmulo da experiência para diferentes conjuntos de dados. Assim, a ideia básica é combinar essas experiências distintas de forma a compor uma experiência final mais robusta. Há três métodos básicos amplamente utilizados na literatura para se combinar essas experiências: Bagging, Boosting e Stacking.

Bagging:

O Bagging (Bootstrap Aggregating), um método proposto por Breiman em 1996, gera um conjunto de dados por amostragem bootstrap dos dados originais. O conjunto de dados gera um conjunto de modelos utilizando um algoritmo de aprendizagem simples, por meio da combinação por votos para classificação. O seu uso é particularmente atraente quando a informação disponível é de tamanho limitado. No Bagging, os classificadores são treinados de forma independente por diferentes conjuntos de treinamento através do método de inicialização. Para construí-los é necessário montar k conjuntos de treinamento idênticos e replicar esses dados de treinamento de forma aleatória para construir k modelos independentes por reamostragem com reposição. Em seguida, deve-se agregar os k modelos através de um método de combinação apropriada, tal como a maioria de votos.

Boosting:

No Boosting, de forma semelhante ao Bagging, cada classificador é treinado usando um conjunto de treinamento diferente. A abordagem por Boosting original foi proposta por Schapire em 1990. A principal diferença em relação ao Bagging é que os conjuntos de dados reamostrados são construídos especificamente para gerar aprendizados complementares, e a importância do voto é ponderado com base no desempenho de cada modelo, em vez da atribuição de mesmo peso para todos os votos. Essencialmente, esse procedimento permite aumentar o desempenho de um limiar arbitrário simplesmente adicionando modelos mais fracos. Dada a utilidade desse achado, Boosting é considerado uma das descobertas mais significativas em aprendizado de máquina.

Stacking:

Stacking é considerado métodos de combinação de classificadores heterogêneos. Ou seja, seu objetivo é tentar combinar classificadores distintos para explorar os benefícios de todos e atenuar as fraquezas de cada um. O método proposto é uma estrutura em duas camadas: no nível-0, vários algoritmos de aprendizado recebem o conjunto de treinamento, gerando os classificadores de nível-0. A camada seguinte (nível-1) tem como entrada as predições da camada anterior (nível-0), na qual um meta-algoritmo de nível-1 as combina para fornecer o meta-classificador final.

4.4. Ajuste de modelos

Um modelo é dito ser um bom modelo de aprendizado de máquina, se generalizar qualquer novo dado de entrada do domínio do problema de forma adequada. Isso nos ajuda a fazer previsões sobre dados futuros ou desconhecidos, que o modelo de dados nunca viu. Agora, suponha que desejemos verificar o quão bem o nosso modelo de aprendizagem de máquinas aprende e generaliza para os novos dados. Neste intuito, definimos o conceito de função objetivo e dizemos que o objetivo de aprendizado é minimizar os erros da função objetivo. Minimizar estes erros tanto sobre os dados de conhecimentos, quanto sobre dados desconhecidos, é um desafio e requer que sejamos capazes de abordar dois grandes problemas: *underfitting* e *overfitting*. Estes são os principais problemas responsáveis pelo desempenho ruim dos algoritmos de aprendizado da máquina.

Overfitting & Underfitting:

Underfitting: um modelo estatístico ou um algoritmo de aprendizado de máquina é considerado como inadequado quando não pode capturar a tendência subjacente dos dados. É como tentar calçar sapatos menores que o seu pé. O *underfitting* destrói a precisão do nosso modelo de aprendizagem de máquinas. Sua ocorrência simplesmente significa que nosso modelo ou o algoritmo não se encaixam bem nos dados. Geralmente acontece quando temos poucos dados para construir um

modelo preciso e também quando tentamos construir um modelo linear com dados não-lineares. Nesses casos, as regras do modelo de aprendizado da máquina são muito amplas e flexíveis para serem aplicadas em coleções pequenas e, portanto, o modelo provavelmente fará muitas previsões erradas. O *underfitting* pode ser evitado usando-se mais dados ou reduzindo-se o número de features no espaço de entrada.

Overfitting: um modelo estatístico apresenta *overfitting* quando captura todas as tendências e comportamentos dos dados. Quando um modelo é treinado com muitos dados, ele começa a aprender com o ruído e as entradas de dados imprecisas em nosso conjunto de dados. Consequentemente, o modelo não categoriza os dados corretamente, por causa de muitos detalhes e ruídos. As causas do *overfitting,* geralmente, estão relacionadas com escolhas equivocadas de parâmetros para métodos não-paramétricos e não-lineares. Esses tipos de algoritmos de aprendizado de máquina têm mais liberdade na construção do modelo com base no conjunto de dados e, portanto, eles realmente podem criar modelos irrealistas. Uma solução para evitar overfitting é utilizar modelos menos complexos.

Mais especificamente, há diversas propostas de se corrigir e selecionar o modelo mais adequado para cada problema, evitando assim o overfitting e o underfitting. As metodologias comumente utilizadas para se evitar overfitting são:

- Reamostragem (Resampling): aprende-se sobre distintas subamostras, visando assim estimar os erros inerentes ao processo de aprendizado.
- Early Stopping: tenta estimar quantas iterações de aprendizado podem ser executadas antes que o modelo gerado comece a apresentar *overfitting*.
- Poda: a poda é amplamente utilizada ao construir modelos relacionados. Ele simplesmente remove os nós que adicionam pouca força de previsão para o problema em questão.
- Regularização (Regularization): introduz um termo de custo para penalizar modelos mais completos.

Regularization:

Uma técnica que é muitas vezes usada para controlar o fenômeno de *overfitting* é a regularização, que envolve a adição de um termo de penalidade (regularização) à função de erro, para desencorajar os coeficientes de alcançarem grandes valores. O termo de regularização mais simples leva a forma de uma soma de quadrados de todos os coeficientes, levando a uma função de erro modificada da forma:

$$\widetilde{E}(\mathbf{w}) = \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Onde:

$$\|\mathbf{w}\|^2 \equiv \mathbf{w}^{\mathrm{T}}\mathbf{w} = w_0^2 + w_1^2 + \ldots + w_M^2$$

O coeficiente λ governa a importância do termo de regularização em comparação com a soma do quadrado dos erros. Técnicas como esta são conhecidas na literatura estatística como métodos de encolhimento porque reduzem o valor da coeficientes. O caso particular de um regulador quadrático é chamado de ridge regression (Hoerl e Kennard, 1970). No contexto das redes neurais, esta abordagem é conhecida como *descending weight*.

O termo de regularização impacta diretamente o erro de generalização do modelo. De fato, λ controla a complexidade efetiva do modelo e, portanto, determina o grau de *overfitting*. Assim, a regularização é uma das formas mais simples de se controlar a complexidade de modelos. Um cuidado a ser tomado é que este processo de escolha da complexidade do modelo não deve ser feito considerando-se todo o conjunto de treinamento. É importante separamos o conjunto de treinamento em treinamento e validação e utilizamos o conjunto de validação para ver o desempenho dos vários modelos que obtemos ao variar o valor de λ. Em muitos casos, no entanto, essa abordagem acaba gerando um grande desperdício de valiosos dados de treinamento, e temos que buscar abordagens mais sofisticadas.

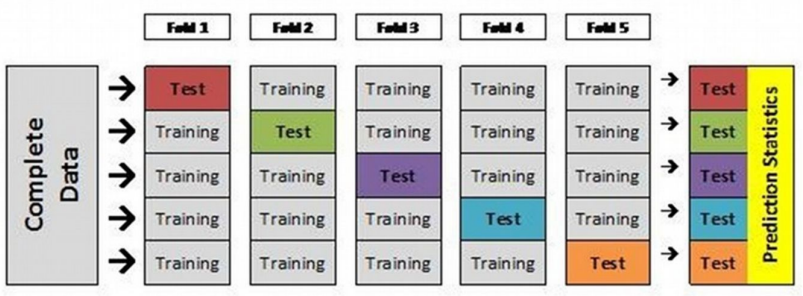
Resampling:

Uma estratégia mais robusta e aplicável a um número maior de cenários é aprender usando-se distintas partes da coleção de treino. Conseguimos fazer isso se conseguirmos amostrar corretamente o conjunto de treinamento sem prejudicar o aprendizado, processo este denominado reamostragem. Ao longo dos anos, estatísticos desenvolveram diversos procedimentos para criar as múltiplas amostras necessárias para a reamostragem a partir da amostra original. Agora uma amostra pode gerar um grande número de outras amostras que podem ser empregadas para gerar a distribuição amostral empírica de uma estatística de interesse.

Uma diferença chave entre os vários métodos de reamostragem é se as amostras são extraídas com ou sem reposição. A amostragem com reposição obtém uma observação a partir da amostra e então a coloca de volta na amostra para possivelmente ser usada novamente. A amostragem sem reposição obtém observações da amostra, mas uma vez obtidas eles não estão mais disponíveis. O verdadeiro poder da reamostragem vem de amostragem com reposição. Abaixo descrevemos os quatro principais métodos de reamostragem utilizados em Modelagem Preditiva atualmente.

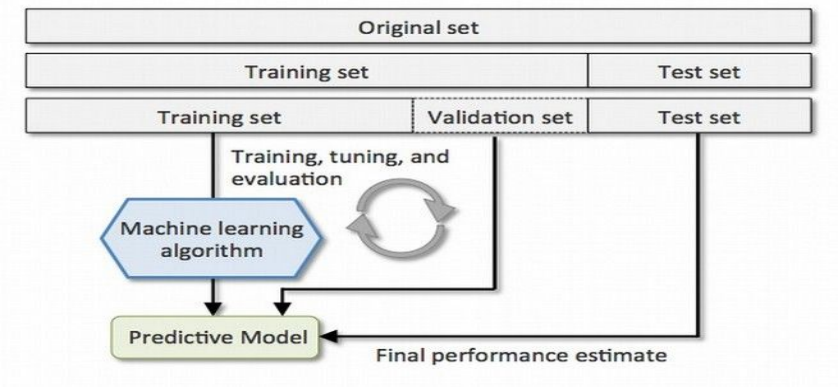
- **K-Fold Cross-Validation:** em k-fold cross-validation, a amostra original é dividida aleatoriamente em k subamostra de tamanho igual. Das subamostrações k, uma única subamostra é mantida como dados de teste para mensurar a qualidade do modelo gerado a partir das k-1 subamostras restantes, que são usadas como dados de treinamento. O processo de validação cruzada é então repetido k vezes, com cada uma das subamostras usada exatamente uma vez como dados de teste. A Figura 4.9 ilustra este procedimento.

Figura 4.9 - Representação do método de reamostragem K-Fold cross-validation.



- **Método Holdout:** este método pode ser considerado a variação mais simples da validação cruzada do k-fold, embora não seja validação cruzada. Nós atribuímos aleatoriamente pontos de dados para dois conjuntos d0 e d1, geralmente chamados de conjunto de treinamento e o conjunto de teste, respectivamente. O tamanho de cada um dos conjuntos é arbitrário, embora, normalmente, o conjunto de teste seja menor do que o conjunto de treinamento. Em seguida, treinamos em d0 e teste em d1. A figura 4.10 ilustra este procedimento.

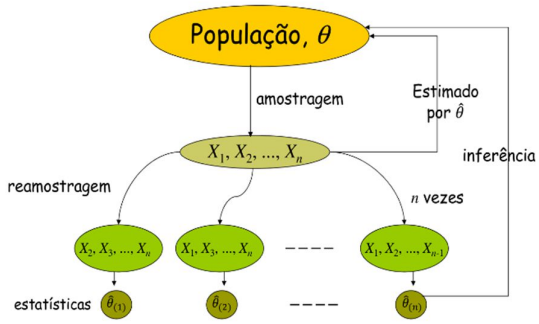
Figura 4.10 - Representação do método de reamostragem holdout.



- **Jackknife:** o jackknife é um método não paramétrico destinado a estimar o enviesamento, e portanto reduzi-lo, e a variância de estimadores em condições teoricamente complexas, ou em que não se tem confiança no modelo

especificado. Foi introduzido por Quenouille em 1949, retomado por Tukey em 1958 e desenvolvido na última década. Tal como ilustrado na Figura 4.11, o método jackknife computa n subconjuntos (n=tamanho da amostra) pela eliminação sequencial de um caso de cada amostra. Assim, cada amostra tem um tamanho de n-1 e difere apenas pelo caso omitido em cada amostra. Por esta razão, este método é também conhecido como *"leave-one-out"*.

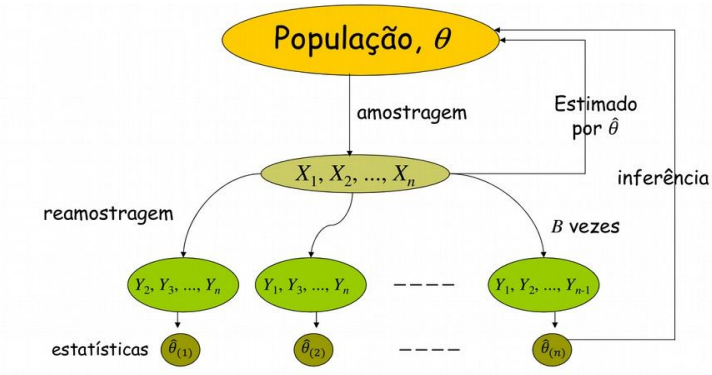
Figura 4.11 - Representação do método de reamostragem jackknife.



- **Bootstrap:** o bootstrap, introduzido por Efron no final dos anos 70 vem historicamente na linha do jackknife, e pode-se dizer que é uma técnica não paramétrica que procura substituir a análise estatística teórica (insuficiente em muitos casos, como se exemplifica na utilização da análise fatorial) pela força bruta da computação, cada vez mais acessível e menos dispendiosa. A terminologia, introduzida por Efron (1979), é basicamente uma técnica de reamostragem que permite aproximar a distribuição de uma função das observações pela distribuição empírica dos dados baseada em uma amostra de tamanho finita. A amostragem é feita com reposição da distribuição da qual os dados são obtidos, se esta é conhecida (bootstrap Paramétrico) ou da amostra original (bootstrap não-paramétrico). Esta técnica foi extrapolada para resolução de muitos problemas de difícil resolução através de técnicas de análise estatística tradicionais, baseadas na hipótese de um elevado número de amostras. A técnica bootstrap tenta realizar o que seria desejável na prática se tal fosse possível: repetir a experiência. As observações são escolhidas de

forma aleatória e as estimativas recalculadas. A figura 4.12 ilustra o funcionamento do bootstrap.

Figura 4.12 - Representação do método de reamostragem bootstrap.



Capítulo 5. Avaliação e Comparação de Modelos Preditivos

Este capítulo tem como objetivo tornar o aluno apto à:

- Diferenciar e selecionar adequadamente as principais métricas de qualidade em Modelagem Preditiva.
- Projetar adequadamente experimentos para comparação de distintos modelos de predição.
- Identificar as principais diferenças entre avaliação off-line e avaliação on-line.
- Utilizar corretamente estratégias de avaliação capazes de garantir generalização dos resultados obtidos.
- Identificar os principais desafios relacionados à avaliação de modelos preditivos.

"Qualidade significa fazer o certo quando ninguém mais está observando."

Henry Ford

As diferenças de desempenho entre os distintos métodos de aprendizado existentes são suportadas pelo denominado teorema ***No Free Lunch***, o qual afirma a inexistência de um algoritmo universal capaz de obter desempenho preditivo superior aos demais, para todos os problemas. Portanto, idealmente, um algoritmo adequado (ou uma combinação deles) deve ser escolhido de acordo com as características do problema sob análise. Para isso, especialistas em Aprendizado de Máquina ou, no nosso caso, em Modelagem Preditiva, devem ser capazes de, além de entender e selecionar adequadamente modelos de aprendizado, avaliar adequadamente os modelos selecionados.

Por este motivo, a etapa de avaliação torna-se tão importante quanto as anteriores no processo de Modelagem Preditiva. Além disso, somente através de um processo de avaliação correto somos capazes de estimar a capacidade de generalização dos modelos gerados. Ou, em outras palavras, somente através de um

processo de avaliação adequado somos capazes de avaliar a utilidade prática do que se está propondo. Tal como nas etapas anteriores, temos diversos desafios relacionados a este processo.

Um primeiro desafio consiste em saber o que mensurar. Há uma grande diversidade de métricas na literatura que abrangem diferentes perspectivas de análise. Por exemplo, enquanto métricas orientadas a acurácia focam na quantidade de acertos feitos pelos modelos, métricas centradas no usuário focam na satisfação dos usuários com a solução proposta, o que não necessariamente está relacionado apenas com os acertos. Ou ainda podemos considerar métricas mais próximas à visão de negócios, tal como conversão de usuários visitantes em pagantes.

Um segundo desafio refere-se a metodologia de mensuração. Como mensurar? Qual tipo de experimento projetar? Podemos nos basear apenas em dados históricos para conduzir algumas análises. Porém, essa abordagem pode ser insuficiente para alguns objetivos de análise. Por exemplo, dificilmente conseguimos mensurar satisfação do usuário utilizando dados históricos. Isso porque satisfação é um conceito dinâmico e muito relacionado ao contexto ao qual apresentamos resultados aos usuários.

Outro tipo de decisão comum nesta etapa refere-se ao uso de análises qualitativas ou quantitativas para se obter conclusões. Algumas vezes precisamos combinar ambos tipos de análises, principalmente quando desejamos, além de verificar que nossa proposta é melhor do que um método base (denominado baseline), entender porque somos melhores (ou piores).

Por fim, é importante revisarmos alguns conceitos, estatísticas e estratégias de comparação de sistemas. Garantir que as comparações realizadas sejam justas e replicáveis por outros pesquisadores é fundamental. Neste sentido, precisamos revisar conceitos básicos de experimentação, tais como intervalos de confiança, observações pareadas e não pareadas, além de testes de hipótese.

5.1. Métricas de qualidade

Encontramos na literatura diversas métricas para avaliar modelos de predição. Tais métricas podem ser agrupadas de acordo com os objetivos de análise. Dentre os principais objetivos comumente avaliados destacamos:

- **Medidas de erro**: MAE, MSE e RMSE.
- **Médias centradas em acertos**: acurácia, ROC AUC, Breese score e precision/recall.
- **Médidas centradas em usuários**: cobertura, user retention, satisfação e reversals.
- **Medidas de correlação**: Spearman e Pearson.

No meio acadêmico, grande parte dos trabalhos focam em medidas centradas em acertos e erros, por proverem uma forma simples de se mensurar a qualidade das predições. Porém, métricas como acurácia acabam não sendo muito úteis em diversos cenários reais. Por exemplo, na indústria usualmente um bom preditor é aquele que traz mais dinheiro. Assim, métricas centradas em negócios, tal como lift, cross-salve e up-sales tornam-se muito mais relevantes. Mais recentemente, pesquisas na área de modelagem preditiva vêm, também, reconhecendo a importância de métricas focadas na experiência do usuário.

Além do objetivo da avaliação é importante ressaltarmos que a seleção de métricas de avaliação depende também do tipo de tarefa de predição sendo executada. Há basicamente dois tipos de tarefas a serem consideradas neste caso. Quando nosso interesse de análise são os valores preditos, tal como em classificação de documentos, estamos realizando a **tarefa de predição** pura. Por outro lado, se nosso objetivo de análise é a ordenação que podemos gerar a partir dos valores preditos, nossa tarefa é denominada de **tarefa de *ranking***. Como exemplo de tarefa de ranking temos a recomendação de filmes. Abaixo discutiremos algumas das principais métricas existentes em três grupos distintos de métricas: Medidas de erro, Medidas centradas em acertos e Medidas de Correlação.

Medidas de erro:

Este conjunto de métricas visa estimar o quanto os valores preditos se distanciam dos valores reais. Dessa forma, são métricas focadas na tarefa de predição. Como principais métricas deste grupo, temos:

- **MAE (Mean Absolute Error)**: calcula a norma da diferença entre o valor predito e o valor real para cada predição gerada. Em seguida, calcula-se a média sobre todas as diferenças.
- **MSE (Mean Squared Error)**: calcula o quadrado da diferença entre o valor predito e o valor real para cada predição gerada. Em seguida, calcula-se a média sobre todas as diferenças ao quadrado.
- **RMSE (Root Mean Squared Error)**: aplica raiz quadrada à métrica MSE.

É importante observar que grande parte das medidas de erro são equivalentes. Diversos estudos usam MAE, então utilizar esta métrica facilita a comparação com o que já foi feito. Porém, erros quadrados podem ser mais apropriados para domínios com escalas de avaliação maiores, permitindo capturar erros de maior magnitude. Por outro lado, MSE não é definido na mesma escala que os dados, dificultando sua intepretação. Neste caso, a RMSE torna-se mais apropriada. De modo geral, uma desvantagem de métricas centradas em erros é que erros podem ser dominados por itens irrelevantes.

Medidas centradas em acertos:

Este conjunto de métricas de avaliação tem suas origens em uma área de pesquisa denominada Recuperação da Informação, que nada mais é que uma teoria que estuda a forma em que os documentos são recuperados de forma automática.

Tais métricas surgiram dentro dessa teoria como uma forma de avaliar se o sistema de Recuperação de Informação estava se comportando da maneira na qual ele foi proposto. Isso é, o objetivo é avaliar não somente se o sistema estava

recuperando as informações corretas, mas também verificando o quanto o sistema estava abstraindo as informações erradas.

Ao longo do tempo, muitos pesquisadores viram que essas métricas de Recuperação da Informação poderiam ser aplicadas em classificadores e modelos de predição para determinar as características de cada um dos modelos. Como principais métricas deste grupo, destacamos:

- **Acurácia:** a proporção de predições corretas, sem levar em consideração o que é positivo e o que é negativo. Esta medida é altamente suscetível a desbalanceamentos do conjunto de dados, e pode facilmente induzir a uma conclusão errada sobre o desempenho do modelo. Mais especificamente, acurácia é calculada como a razão entre o total de acertos do modelo, pelo total de dados no conjunto de teste.

- **Precisão:** esta métrica visa estimar o quanto um modelo é capaz de acertar, considerando-se diversas classes de saída. É calculada como a porcentagem de acertos sobre o total de predições realizadas para cada classe.

- **Revocação:** através dessa métrica, mensuramos o quanto do total de objetos pertencentes a cada classe um modelo é capaz de apontar corretamente. A revocação é definida como a porcentagem de acertos sobre o número total de instâncias existentes em cada classe.

- **Curvas de precisão e revocação:** as definições de precisão e revocação assumem que todos os objetos de A foram examinados previamente. Na prática, usuários não estão interessados em todos os objetos retornados pelo modelo. Eles avaliam apenas os top-N objetos, em que N é um número usualmente baixo. Assim, precisão e revocação variam a medida que o número de objetos avaliados aumenta. Com isso, torna-se mais apropriado apresentar as chamadas curvas de precisão e revocação. Este tipo de curva define o valor de precisão do modelo para cada item que ele é capaz de acertar. Ou seja, calculamos o número de tentativas necessárias até o próximo acerto do modelo.

- **P@5 e P@10**: em muitos cenários, alta revocação não é um requisito forte. Na verdade, quanto maior o número de documentos relevantes nos primeiros ranks, melhor a impressão do usuário. *Precision at 5* (P@5) e *Precision at 10* (P@10) medem, respectivamente, a precisão quando temos 5 e 10 objetos retornados. O objetivo dessas métricas é verificar se os usuários estão tendo acesso aos documentos relevantes no topo do rank.

- **MAP:** calcula a média de precisão após recuperar cada objeto relevante.

- **F-Measure:** média harmônica entre precisão e revocação.

- **DCG**: precisão e revocação permitem apenas medições binárias de relevância. Ou seja, não existe distinção entre objetos altamente relevantes e pouco relevantes. Essas limitações podem ser superadas adotando-se definições de relevâncias graduais. Uma forma de se definir relevâncias graduais é através da métrica DCG. Essa métrica considera dois aspectos relevantes ao avaliar o resultado de uma tarefa de ranking: (1) um resultado bom é o que apresenta objetos altamente relevantes nas primeiras posições do rank; (2) objetos relevantes presentes nas últimas posições do rank são menos úteis.

Medidas de correlação:

Há situações em que não podemos diretamente mensurar a relevância das predições, ou estamos mais interessados em saber quanto um modelo varia em relação a um outro. Neste caso, dizemos que estamos interessados em realizar uma **análise de correlação**. Análise de correlação é definida como um tipo de análise bivariada que mede a relação entre duas variáveis. Quando desejamos observar a correlação entre os valores preditos, em geral, a principal métrica utilizada é a Pearson, a qual mede a relação entre variáveis linearmente relacionadas. Uma premissa dessa métrica é que ambas as variáveis devem ser normalmente distribuídas. Em outros casos, estamos interessados em comparar a ordem relativa produzida por duas funções de ranking. Isso é feito através de métricas estatísticas denominadas *rank correlation metrics*. O coeficiente Spearman é a métrica de

correlação mais utilizada neste cenário. Ela se baseia na diferença entre as posições de um mesmo documento em dois rankings.

5.2. Comparação de modelos

Muitas vezes, quando propomos um novo modelo preditivo, nosso intuito é que este novo modelo seja melhor que o modelo atualmente empregado para resolver o problema abordado. Dessa forma, nosso objetivo é substituir um modelo "ultrapassado" ou simplista por outro mais robusto. Este objetivo é muito recorrente, principalmente na indústria. Dessa forma, uma pergunta chave que devemos nos fazer é: como nos certificarmos que o método proposto é, de fato, melhor que o anterior?

Responder essa pergunta requer um amplo conhecimento na área da estatística denominada **Métodos Quantitativos**. De fato, a estatística nos fornece diversas ferramentas para realizar este tipo de comparação, de maneira confiável, considerando-se todas as peculiaridades de cada cenário de análise. Basicamente, quando almejamos comparar dois modelos quaisquer, usamos um dos três métodos abaixo:

- Comparação não pareada.
- Comparação pareada.
- Teste de hipótese.

A fim de entender um pouco mais sobre cada uma, primeiramente, precisamos revisar alguns conceitos estatísticos básicos utilizados por elas. Dessa forma, nas seções subsequentes, primeiro apresentaremos tais conceitos, e em seguida discutiremos em detalhes cada método de comparação.

Conceitos básicos:

- **Hipótese alternativa**: é a hipótese estatística sobre o resultado ou conclusão desejado. Aquilo que gostaríamos de provar que é verdadeiro para um dado domínio e modelo. Por exemplo, em modelagem preditiva, uma hipótese alternativa muito comum é que o modelo proposto é melhor que o já existente.
- **Hipótese nula**: é a conjectura de que sua hipótese alternativa é falsa. O método científico de investigação estabelece que pesquisadores devem sempre provar que a hipótese nula é falsa, ao invés de provar a hipótese alternativa diretamente. Essa estratégia reduz as chances de construirmos experimentos enviesados que beneficiam a hipótese alternativa.
- **Teorema Central do Limite:** este teorema descreve a distribuição da média de uma amostra aleatória de uma população com variância finita. Quando o tamanho amostral é suficientemente grande, a distribuição da média é uma distribuição aproximadamente normal. O teorema aplica-se independentemente da forma da distribuição da população. Muitos procedimentos estatísticos comuns requerem que os dados sejam aproximadamente normais. O teorema central do limite permite a aplicação destes procedimentos úteis para populações que são fortemente não-normais.
- **Intervalo de confiança:** um intervalo de confiança (IC) é um intervalo estimado de um parâmetro de interesse de uma população. Em vez de estimar o parâmetro por um único valor, é dado um intervalo de estimativas prováveis. O quanto estas estimativas são prováveis será determinado pelo coeficiente de confiança (1-α), para $\alpha \in (0, 1)$. Intervalos de confiança são usados para indicar a confiabilidade de uma estimativa. Por exemplo, um IC pode ser usado para descrever o quanto os resultados de uma pesquisa são confiáveis. Sendo todas as estimativas iguais, uma pesquisa que resulte num IC pequeno é mais confiável do que uma que resulte num IC maior. Para calcular o intervalo de confiança, utilizamos uma de duas fórmulas. Quando nossa amostra possui mais de 30 observações e os dados possuem qualquer distribuição de probabilidade, utilizamos a fórmula baseada na distribuição-z. Caso a amostra possua menos de 30 observações e os dados são normalmente distribuídos, devemos utilizar a distribuição-t. Um erro muito comum é usar a distribuição-t

para populações não normalmente distribuídas.

- **Teste estatístico:** o teste estatístico é uma fórmula matemática que permite aos pesquisadores avaliar se as diferenças observadas entre os dois grupos podem ser meramente justificadas por fatores casuais ou se tais diferenças são reais. Basicamente, este tipo de teste determina a probabilidade de se obter os resultados observados, quando assumimos que a hipótese nula é verdadeira.

Comparação não pareada:

Algumas vezes, ao tentar comparamos dois modelos, precisamos avaliá-los considerando amostras distintas. Isso pode ocorrer, por exemplo, quando temos apenas resultados históricos de um dos modelos e não conseguimos aplicá-lo sobre uma população recente. Ou ainda, o número de predições geradas por um dos modelos é maior que o outro. Há também cenários em que os objetos pertencentes a cada amostra não são diretamente comparáveis. É o caso por exemplo, de um algoritmo que tenta prever o peso de homens e mulheres à medida que envelhecem. O membro de cada grupo fornece apenas uma observação do seu peso para compor a análise. Neste tipo de cenário, consideramos que as amostras são independentes e, por conseguinte, podem apresentar comportamentos diferentes. Neste caso, devemos conduzir o método de comparação não pareada entre os modelos.

A comparação não pareada é baseada na comparação de médias entre dois grupos. Assim, primeiro calculamos a média das amostras para cada uma das alternativas. Em seguida, calculamos o intervalo de confiança associado à média de cada grupo. Por fim, verificamos a interseção entre os dois intervalos de confiança. Caso não haja sobreposição, os modelos avaliados são diferentes e aquele que apresentar maior média é considerado o melhor. Em caso de sobreposição, primeiramente verificamos se o intervalo de confiança de um dos grupos sobrepõe à média do outro. Neste caso, os modelos não são considerados como estatisticamente diferentes. Caso não haja esta sobreposição, a única forma de verificarmos se ambos são estatisticamente diferentes ou não é através da aplicação de um teste estatístico para amostras independentes.

Comparação pareada:

Estudos com amostras pareadas são muito comuns em Modelagem Preditiva. Esses estudos consistem em realizar mais de uma medida em uma mesma unidade amostral e verificar se houve diferença entre essas medidas, onde a primeira informação será pareada com a segunda informação, com a terceira e assim por diante. Por exemplo, é comum contrastarmos o impacto de um novo algoritmo de predição e um algoritmo correntemente adotado, verificando qual seria a efetividade de ambos para um mesmo indivíduo. Sendo assim, é de se esperar que as medidas de um mesmo indivíduo sejam similares, enquanto que as medidas de indivíduos distintos sejam diferentes. Dessa forma, as observações de um mesmo indivíduo são dependentes, o que faz com que o uso dos testes usuais de comparação de duas ou mais amostras não sejam adequados, uma vez que existe violação da suposição de independência das observações.

Um dos métodos mais utilizados para realizarmos a comparação pareada entre dois modelos, consiste em tratar o problema como uma única amostra de n pares de mensurações distintas. Para cada par, calculamos a diferença dos resultados obtidos por cada modelo. Em seguida, calculamos o intervalo de confiança para a diferença média derivada a partir de todos os pares. Se este intervalo incluir o valor 0, os modelos não são estatisticamente diferentes com a confiança utilizada. Por outro lado, se o intervalo não inclui o valor 0, o sinal da diferença indica qual modelo é o melhor.

Teste de hipótese:

Teste de hipótese ou teste de significância é um método para testar uma reivindicação ou hipótese sobre um parâmetro em uma população, usando dados medidos em uma amostra. Baseia-se em estatísticas inferenciais porque nos permite medir comportamentos em amostras para aprender mais sobre o comportamento em populações, que geralmente são muito grandes ou inacessíveis. A confiabilidade deste método é suportada pelo Teorema Central do Limite, o qual garante que a probabilidade de se observar o valor médio de um parâmetro dessa população é normalmente distribuída.

No teste de hipótese, começamos assumindo que a hipótese nula é verdadeira. Em seguida, realizamos algumas análises para verificar se a hipótese nula, de fato, é verdadeira para a amostra de entrada. Note que a única razão pela qual estamos testando a hipótese nula é porque achamos que ela está errada. O objetivo do teste de hipótese é determinar a probabilidade de que um parâmetro de população, como a média, seja verdade. O método possui quatro passos principais:

1. Indique as hipóteses;
2. Defina os critérios para uma decisão;
3. Calcule a estatística do teste;
4. Tome uma decisão.

Para tomar uma decisão, precisamos avaliar quão provável é o resultado observado na amostra, se a média da população declarada pela hipótese nula é verdadeira. Utilizamos um teste estatístico para determinar essa probabilidade. Especificamente, um teste estatístico nos diz até onde, ou quantos desvios de padrão, a média da amostra está da média da população. Quanto maior o valor do teste estatístico, maior a distância entre a média da população e a média estabelecida pela hipótese nula.

Quando estamos conduzindo uma comparação não pareada, o teste estatístico a ser utilizado é o t-test (ou student-t). Já para comparação pareada, devemos utilizar o t-test pareado. Por exemplo, suponha um estudo onde os indivíduos foram submetidos a um algoritmo de predição de consumo e deseja-se verificar se houve diferença entre o consumo do usuário no sistema antes e depois dessa alteração. Nesse caso, a variável de interesse é numérica e o objetivo é verificar se existe diferença significativa dessa variável entre dois grupos de interesse. Note que o objetivo é o mesmo que o do t-test utilizado para comparar duas amostras, porém, a diferença é que no t-test pareado, as amostras são dependentes.

É importante também lembrarmos que, tal como o t-test, o t-test pareado é paramétrico, ou seja, possui a suposição de que a variável de interesse seja

normalmente distribuída. Quando essa suposição de normalidade da variável de interesse é violada, o ideal seria usar o teste de Wilcoxon.

5.3. Avaliação off-line

Avaliação off-line, ou retrospectiva, é a principal estratégia de avaliação utilizada em Modelagem Preditiva para se avaliar os modelos construídos. Nesta abordagem, olhamos apenas como os modelos geram predições para itens já consumidos ou avaliados no passado. Para tanto, exploramos dados históricos dos sistemas e/ou domínios estudados. A popularidade de uso dessa estratégia deve-se, sobretudo, a abundância de dados históricos disponíveis para análise na maioria dos cenários, além da relativa facilidade em se configurar avaliações off-line. Note, porém, que a avaliação off-line é capaz de estimar apenas quão bem um modelo é capaz de replicar comportamentos passados. A principal pergunta relacionada a utilidade de modelos preditivos, que se refere a estimar quão bem um modelo é capaz de prever comportamentos futuros, não pode ser corretamente respondida por esta abordagem. Neste caso, a abordagem mais apropriada seria utilizar avaliação on-line, a qual discutiremos na próxima seção.

Metodologia de mensuração:

Como discutido na seção 5.1, existem diversas métricas na literatura que visam avaliar distintas dimensões de qualidade em modelos preditivos. A avaliação off-line reside, basicamente, em gerar predições para dados passados e calcular diversas dessas métricas. Este tipo de cálculo só é possível porque em dados históricos conhecemos a priori os comportamentos ou saídas observadas. Por exemplo, em sistemas de recomendação é comum avaliarmos através de avaliações off-line o número de itens já consumidos por cada usuário que um Sistema de Recomendação é capaz de acertar. Uma preocupação, neste caso, é como calcular essas métricas de forma correta, sem enviesar nosso modelo com informações que ele não deveria conhecer a priori?

O cálculo dessas métricas de qualidade (e.g., acurácia e erro), idealmente, devem seguir uma metologia *leave-one-out*. Ou seja, retiramos do conjunto de treinamento apenas o elemento para o qual desejamos gerar uma predição e contrastamos a predição gerada com o valor observado para este elemento. Entretanto, este processo é muito caro e complicado para a maioria dos cenários reais. Nestes casos, adota-se uma simplificação, utilizando o processo de amostragem de *hold-out* ou de *cross-validation*. Ou seja, retiramos X% da coleção de treino e avaliações predições geradas para estes elementos removidos. Via de regra, o valor de X% varia entre 10% a 30% de toda a coleção de treino.

É importante reforçar que o processo de aprendizado deve utilizar apenas os dados de treinamento. Qualquer tipo de informação vinda do teste, ou de um tempo futuro ao que define o conjunto de treinamento, representa um viés indesejável para as análises. Identificar este tipo de viés, algumas vezes, pode ser complicado. Por exemplo, em determinados domínios, algumas informações de *id* dos objetos são geradas através de uma função *auto increment*, na qual os *ids* são valores monoatômicos e crescentes. Ao selecionar aleatoriamente este tipo de dados para o conjunto de treino, podemos incluir *ids* não ordenados. Com isso, estamos implicitamente dando indícios de tamanho da coleção de dados no teste para alguns algoritmos de aprendizado. Por esta razão, realizar um corte temporal nas coleções de dados é sempre que possível uma boa estratégia de separação entre treino e teste.

Análise de fatores:

Muitas vezes estamos interessados em entender quais fatores são responsáveis por um modelo ser melhor que outro. É possível realizar este tipo de estudo através da avaliação off-line também. De fato, há uma grande diversidade de estudos e propostas para se conduzir este tipo de análise a partir de dados históricos. Abaixo listamos algumas das principais estratégias existentes:

- Educated guesses.
- Um fator por vez (projeto simples).
- Fatorial completo.

- Fatorial fracionado.
- Fatorial 2K (e variantes).
- Quasi-Experimentos.
- Field Experiment.

Projeto Simples:

A estratégia para análise de fatores mais utilizada na indústria é o Projeto Simples. Em prol da simplicidade de projeto e agilidade de execução, esta estratégia assume diversas simplificações sobre a interação entre os fatores. A principal delas é que o impacto de cada fator independe dos demais. Dessa forma, podemos analisar o impacto que cada fator traz sobre o resultado final individualmente. Claramente, isso não é verdade para todos os cenários, mas representa um bom ponto de partida para se identificar os fatores mais relevantes. Caso necessário, pode-se investigar o impacto das interações deste subconjunto de fatores relevantes apenas, limitando a complexidade da análise.

Para realizarmos a análise de fatores via o Projeto Simples, primeiramente, precisamos definir um ponto de partida ou ponto de referência. Este ponto de partida é um conjunto de valores que todos os fatores possuem. Por exemplo, os valores *default* de parâmetros de entrada de um algoritmo. Posteriormente, alteramos o valor de um único fator, mantendo os demais constantes. O intervalo de variação do valor dependerá essencialmente do próprio fator, bem como características da coleção e do algoritmo. Após avaliar todos os possíveis valores para este fator, repetimos o processo para outro fator. Este processo termina quando avaliamos todos os fatores individualmente. Apesar de simples, esta estratégia provê bons resultados de análise quando não há interação entre os fatores. Outro ponto importante a se ressaltar sobre o Projeto Simples é que, em geral, ele requer mais esforço de análise que se pensa, principalmente se o número de fatores for alto. Por essa razão, recomenda-se em alguns cenários evitar esse enfoque de experimentação. Abordagens via avaliação on-line ou mesmo um Projeto Fatorial, apesar de mais trabalhosos, podem gerar conclusões importantes mais rapidamente.

5.4. Avaliação on-line

Muitas vezes, novo alvo de experimentação são os usuários ou sistemas que interagem diretamente com usuários. Nestes casos, os produtos, tanto quanto possível, devem ser projetados para satisfazer os usuários. Assim, para algumas organizações, experimentos randomizados centrados nos usuários desempenham um importante papel no projeto destes produtos e na tomada de decisão. Tais experimentos podem ser usados para, por exemplo, explorar diversas opções de projeto ou nos auxiliar a entender como usuários reagem a mudanças. Empresas estão cientes da importância deste tipo de experimentação. Especialmente, empresas que oferecem produtos e serviços na web, que conduzem um grande número de experimentos on-line centrados em usuários.

A condução de avaliações on-line apresenta grandes diferenças quando comparada às avaliações off-line. Uma das principais diferentes é a dificuldade e custo elevado de se projetar e conduzir avaliações on-line. Em geral, este tipo de avaliação não pode ser repetido de forma trivial. Além disso, deve-se ter um cuidado maior para que o usuário compreenda o que está sendo avaliado. O foco deste tipo de avaliação também costuma ser diferente da avaliação off-line. Muitas vezes, ao realizarmos experimentos on-line, estamos mais interessados em questões qualitativas ou em métricas de negócio. Dessa forma, as perguntas a serem feitas são completamente diferentes das perguntas de experimentação centrada em dados. Cabe ainda ressaltar que avaliações on-line não são tão escaláveis quanto experimentos centrados em dados.

Quando projetamos experimentos centrados nos usuários, os dados vêm, essencialmente, de observação e entrevistas. Observações podem ser obtidas em três momentos distintos:

1. Antes do projeto do produto/sistema;
2. Durante a projeto e prototipação do produto/sistema;
3. Quando o produto/sistema já está pronto.

Além disso, há diversos tipos distintos de experimentos que podemos conduzir para obter essas observações. Neste material, focaremos em três tipos principais:

1. Surveys/Questionários.
2. Teste A/B.
3. Interleaving.

Surveys/Questionários:

Surveys ou Questionários são estudos retrospectivos de uma situação para documentar relacionamentos e resultados. Em geral, são feitos pessoalmente ou de maneira on-line com usuários através de entrevistas ou questionários. São sempre feitos após se realizar alterações no sistema. Em geral, não há controle sobre a situação ou manipulação das variáveis de interesse. Os dados vêm da memória dos entrevistados. Este método de avaliação é definido como uma avaliação qualitativa sobre o comportamento, percepção ou interesse dos usuários, e visa, sobretudo, identificar causas ou explicações para tais comportamentos, percepções e interesses. A partir deste entendimento, pode-se, por exemplo, refinar soluções propostas e realizar nova etapa de avaliações qualitativas, ou mesmo partir para avaliações quantitativas. Como principais etapas deste tipo de avaliação, destacamos:

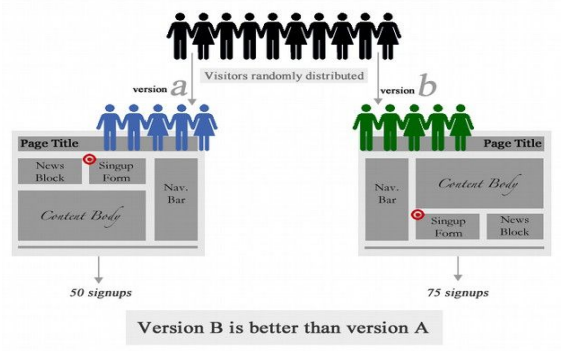
- Escolha de objetivos específicos e mensuráveis.
- Planejamento e escalonamento do survey.
- Obtenção dos recursos necessários.
- Design do survey.
- Preparação do instrumento de coleta de dados.
- Validação do instrumento.
- Seleção de participantes.

- Execução.
- Análise de dados.
- Relatar os resultados.

Teste A/B:

O teste A/B é um experimento controlado com duas variantes, A e B, bem como uma métrica clara que define sucesso. É uma das formas mais populares de teste de hipóteses estatísticas, ou "teste de hipóteses de duas amostras". O teste A/B é uma maneira de comparar duas versões de uma única variável. Tipicamente, este teste avalia a resposta de uma população para a variável A contra a variável B, determinando qual das duas variáveis é mais efetiva. Em testes A/B, ambas versões são idênticas, exceto pela alteração que introduzimos e cujo impacto almejamos mensurar. Em geral, a versão A é a versão atualmente usada (controle), enquanto a versão B é modificada em algum aspecto. Idealmente, os usuários que terão acesso a cada lado do teste são selecionados aleatoriamente, de tal forma que se garanta que ambas amostras de usuários sejam comparáveis (i.e., possuem conjuntos de variáveis de interesse similares). Além disso, uma vez sorteado para um dos lados, por consistência de navegação e comportamento, usuários sempre acessam apenas o lado para o qual foi sorteado inicialmente. A figura 5.1 ilustra o processo de avaliação do teste A/B.

Figura 5.1 - Representação da aplicação do teste A/B.



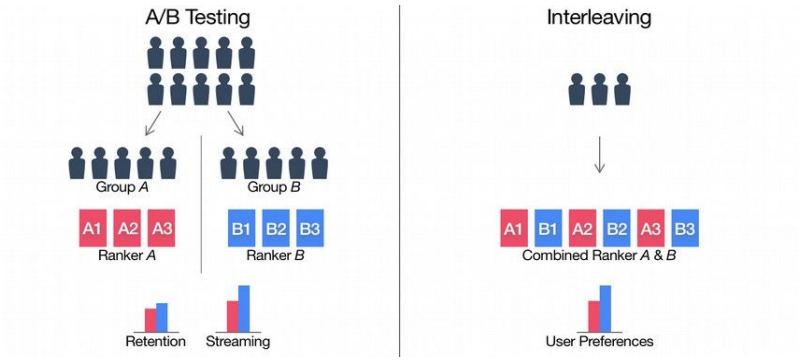
Ao projetar um teste A/B alguns cuidados devem ser tomados. Primeiro, saiba quanto tempo executar um teste antes de desistir. Neste sentido, é importante lembrar que pode ser necessário considerar e excluir um tempo de familiarização do usuário com as alterações introduzidas. Neste período, comportamentos observados podem estar relacionados à dificuldade ou resistência do usuário quanto a alteração na forma de interação e não com as alterações propriamente. Outro cuidado importante é tornar o teste A/B consistente em todo o site ou sistema. Por fim, faça vários testes A/B, a fim de garantir que todo o arcabouço de configuração e implantação do teste esteja correto.

Outra prática frequentemente adotada, que nos auxiliar a verificar a confiabilidade de testes A/B em cenários reais, é o uso de uma variante de teste denominada A/B/C. Nesta variação, cria-se um terceiro grupo que utiliza o mesmo produto/sistema de A. Todos os três grupos possuem o mesmo número de usuários e são comparáveis. O objetivo, neste caso, é garantir que o teste A/B esteja configurado e executando de maneira adequada. O comportamento esperado é que A e C tenham o mesmo resultado, já que usam o mesmo produto/sistema. Além disso, a diferença entre A e C nós dá o intervalo de erro bem como o tempo necessário para convergência do teste A/B.

Interleaving:

Os experimentos de intercalação formulam avaliação de interação dos usuários como um teste de comparação pareada entre dois rankings. Os testes de comparação pareada são um dos projetos de experimentos centrais usados na análise sensorial. Ao testar uma qualidade perceptual de um item (por exemplo, gosto ou som), é amplamente reconhecido que avaliações absolutas (tal como a escala de Likert) são pouco informativas. Em vez disso, os sujeitos são apresentados com duas ou mais alternativas e são convidados a identificar uma diferença relativa ou a indicar uma preferência. A figura 5.2 ilustra o método de avaliação pareada de interleaving.

Figura 5.2 - Representação do método de avaliação pareada interleaving.



A ideia básica por trás de todas as variantes da abordagem de intercalação é realizar comparações em linha pareadas de dois rankings. Isso envolve a fusão dos dois rankings em um único ranking intercalado e, em seguida, apresentando o ranking intercalado ao usuário. O algoritmo usado para produzir o ranking intercalado é projetado para ser "justo", de modo que os cliques dos usuários podem ser interpretados como julgamentos imparciais sobre a qualidade relativa dos dois rankings. Desta forma, a intercalação modifica interativamente, em uma experiência controlada, os resultados de pesquisa apresentados ao usuário, para que o comportamento observado do usuário (ou seja, clique) seja mais interpretável. Isso evita o problema da interpretação pós-hoc de dados observacionais comuns à maioria das outras abordagens para interpretar um feedback implícito.

Capítulo 6. Aplicação de Modelos Preditivos

Este capítulo tem como objetivo tornar o aluno apto à:

- Entender o uso da Modelagem Preditiva em variados cenários reais.
- Identificar as principais informações e decisões a se documentar para fins de gerenciamento de projetos relacionados a Modelagem Preditiva.
- Assimilar as principais práticas para se aumentar as chances de sucesso de aplicação da Modelagem Preditiva.

"Guerra é noventa por cento informação."

Napoleão Bonaparte

Ao longo deste material, discutimos a relevância de Modelagem Preditiva, bem como a execução do processo de modelagem tanto na indústria quanto na academia. Como mencionado anteriormente, diversas organizações estão investindo cada vez mais em análises preditivas para prever com precisão os resultados de seus negócios, melhorar o desempenho do negócio e aumentar a lucratividade. O desafio, porém, é transformar todo este potencial e capacidade de auxílio à tomada de decisão que discutimos na teoria, em prática. Fazer todo o processo discutido acima funcionar no mundo real requer diversos cuidados.

Primeiramente, quando estamos falando de aplicação de Modelagem Preditiva para fins de negócio (i.e., na indústria) é fundamental lembrar que, neste caso, Modelagem Preditiva é antes de tudo um problema de engenharia! Diferentemente da academia em que o enfoque é o avanço da área através de pesquisas sobre algoritmos, estruturas de dados e metodologias, na indústria o enfoque é consolidar um produto com apelo prático. Uma boa analogia neste caso é imaginar que Modelagem Preditiva é uma peça nova e potente capaz de turbinar um motor que já está em funcionamento. O desafio é encaixar esta peça no motor sem fazê-lo parar.

Neste contexto, um segundo ponto a se considerar é que para gerar excelentes produtos, precisamos conduzir a Modelagem Preditiva como o grande Engenheiro de Software que somos, e não como o excelente especialista em Modelagem Preditiva que ainda não somos. A maioria dos problemas que você enfrentará são, de fato, problemas de engenharia. Além disso, relatos de experiência em diversos cenários apontam que grande parte dos ganhos que a Modelagem Preditiva traz para o seu produto provêm de excelentes *features*, e não de excelentes algoritmos de *Machine Learning*. De fato, poucos cenários práticos requerem, em um primeiro momento, o projeto de novos algoritmos.

Para usar modelos preditivos com sucesso, sobretudo, é necessário gerenciar adequadamente modelos desde a fase de desenvolvimento até o seu ambiente de produção. Este gerenciamento de modelos não deve ser uma atividade única, mas um processo comercial contínuo. Dessa forma, este capítulo foca em discutir algumas decisões e melhores práticas relacionadas ao processo de gerenciamento de projetos. Iniciamos o capítulo discutindo alguns cenários de aplicação de Modelagem Preditiva no mundo real. Em seguida, discutimos o processo de documentação do ciclo de vida de projetos relacionados a Modelagem Preditiva. Por fim, apresentamos algumas recomendações e dicas de melhores práticas a serem seguidas, de forma a se economizar tempo e maximizar as chances de sucesso de aplicação dos modelos.

6.1. Aplicação em cenários reais

A abrangência de uso de Modelagem Preditiva é enorme, e sua efetividade vem sendo comprovada em um número crescente de cenários distintos. Aplicações da área médica, previsão do tempo, análise de mercado financeiro, previsão de cenários políticos e sociais, análises comportamentais, aplicações *e-commerce*, dentre outras, estão entre os principais cenários de aplicação. Nesta seção, discutiremos resumidamente alguns cenários e tarefas que podem ser resolvidas com sucesso, utilizando-se Modelagem Preditiva:

- **Detecção de fraude:** a combinação de vários métodos de análise pode melhorar a detecção de padrões e prevenir comportamentos criminosos. À medida em que a segurança cibernética se torna uma preocupação crescente, a análise comportamental de alto desempenho examina todas as ações em uma rede em tempo real para detectar anormalidades que podem indicar fraude, vulnerabilidades e ameaças persistentes. Por exemplo, bancos comumente modelam o comportamento recente de seus clientes, considerando valores de transições, horários, destinatários e tipos de operações mais frequentes, bem como interesses de consumo. Dessa forma, transações individuais podem ser classificadas automaticamente como potenciais fraudes, baseado em Modelagem Preditiva, verificando-se as chances do cliente realizar aquele tipo de transação, nos valores, horário e tipos de compras observados.

- **Otimização de campanhas de marketing:** Modelagem Preditiva é também usada para determinar as respostas ou compras dos clientes, bem como promover oportunidades de venda cruzada. Os modelos preditivos ajudam as empresas a atrair, reter e desenvolver seus clientes mais lucrativos. Por exemplo, o Youtube atualmente seleciona quais propagandas colocar nos vídeos de cada usuário, de acordo com o tipo de vídeo, comentários, frequência de uso e outras informações que o Google consegue extrair de cada usuário. O intuito é maximizar as chances do usuário assistir à propaganda, aumentando assim o lucro do Google com este tipo de negócio.

- **Otimização de operações:** muitas empresas usam modelos preditivos para prever inventário e gerenciar recursos. Por exemplo, as companhias aéreas usam análises preditivas para definir os preços das passagens. Os hotéis procuram prever o número de hóspedes para uma noite determinada para maximizar a ocupação e aumentar a receita. A análise preditiva permite, assim, que as organizações funcionem de forma mais eficiente.

- **Redução de risco:** as pontuações de crédito são usadas para avaliar a probabilidade de inadimplência de um comprador e são um exemplo bem

conhecido de Modelagem preditiva. Uma pontuação de crédito é um número gerado por um modelo preditivo que incorpora todos os dados relevantes sobre a credibilidade de uma pessoa, tais como valores de suas últimas compras, salário, tempo de trabalho, se paga aluguel ou não, se possui filhos, dentre outras. Outros usos relacionados ao risco incluem reivindicações e cobranças de seguros.

- **Sistemas de recomendação:** uma das aplicações mais populares de Modelagem Preditiva, atualmente, são sistemas capazes de gerar recomendações de itens, serviços e produtos de maneira personalizada para cada usuário. O intuito é aprender com o comportamento passado do usuário, ou de usuários com interesses similares, de forma a prever quais itens desconhecidos para o usuário ele teria mais interesse a cada momento. A Netflix é um dos maiores exemplos deste tipo de aplicação. Ao analisar os filmes e seriados que cada usuário assistiu, ela é capaz de modelar os interesses de cada usuário e sugerir novos itens.

- **Segurança de pacientes em UTI**: na University of California Davi, pesquisadores estão usando dados de rotina coletados de pacientes em UTI, como entrada para um algoritmo que gera aos médicos um alerta precoce sobre sepsis (um tipo de complicação que apresenta uma taxa de mortalidade de 40% e é difícil de ser detectar com antecedência). Os pesquisadores afirmam que através de Modelagem Preditiva encontraram uma maneira precisa e rápida de se determinar quais pacientes estão em alto risco de desenvolver a doença.

- **Gestão da saúde populacional:** segmentar os pacientes com base em seus comportamentos passados pode ajudar a prever eventos futuros, como um diabético que acabou na sala de emergência porque ele não ministrou sua medicação ou uma criança com asma que exige uma internação devido a fatores ambientais relacionados a sua doença (e.g., a umidade relativa do ar ficar por alguns dias abaixo de um valor mínimo). Modelagem preditiva atua diretamente sobre este problema, correlacionando fatores frequentes em

eventos passados, de forma a prever novas ocorrências em momentos futuros.

Os cenários acima apresentam de maneira resumida a aplicação de Modelagem Preditiva na prática. Em geral, conseguir cenários de aplicação na indústria, que sejam mais detalhados, é difícil devido a questões tais como segurança dos dados e privacidade dos usuários. Uma estratégia mais efetiva de se obter cenários bem detalhados e com explicações claras sobre decisões, algoritmos, avaliações e informações utilizadas para a Modelagem, é voltarmos nossa atenção para a área acadêmica. Conferências científicas focadas em Modelagem Preditiva possuem bons exemplos para se obter um entendimento prático de aplicação da Modelagem Preditiva. Abaixo listamos algumas dessas conferências na área de Ciência da Computação:

- *ACM KDD.*
- *ACM SIGIR.*
- *ACM RecSys.*
- *The Web Conference.*
- *SIAM DM.*
- *ECML PKDD.*
- *AAAI Conference.*

6.2. Documentação do ciclo de vida

Uma importante parte do processo de gerenciamento é a geração de documentações. Tais documentações darão suporte à manutenção, bem como correções e melhorias futuras. Vale lembrar que definimos o processo de Modelagem Preditiva como um ciclo. Como tal, a cada etapa podemos nos deparar com informações ou situações que nos obrigam a voltarmos em etapas anteriores e reexaminar decisões tomadas. Isso pode acontecer, inclusive após a publicação do

produto. Neste cenário, é vital que tenhamos documentadas todas as decisões relevantes que tomamos ao longo da execução do projeto.

Considerando que o resultado final do processo de Modelagem Preditiva será um produto, precisamos também incluir duas etapas no ciclo de vida do processo de Modelagem Preditiva descrita no capítulo 1. A primeira dessas etapas é a etapa de **implantação**, que se preocupa com a implantação do modelo em ambiente de produção. Ou seja, visa definir passos e operações capazes de transformar um protótipo em um produto final. Idealmente, esta etapa deve ser toda automatizada e possuir um processo de aprovação de cada passo rigorosamente definido.

A segunda etapa refere-se à etapa de **monitoramento**. Devemos suportar o monitoramento contínuo do produto gerado. Além disso, precisamos de ferramentas capazes de avaliar o impacto no mercado que este novo produto gera. Análises de retenção de usuários, taxas de custo e livro de qualidade empresarial, estão entre as medidas mais comuns monitoradas nesta etapa. Além de contínuo, este monitoramento deve ser automático, dinâmico e intuitivo. Deve ser possível, por exemplo, agendar tarefas de monitoramento do modelo para executar em uma variedade de períodos de tempo. Além disso, os índices de desempenho devem ser portáteis, no sentido de que podem ser aplicados a diversos modelos. Uma boa ferramenta de monitoramento de modelo deve também permitir notificações de casos espúrios, e fornecer monitoramento através de gráficos ou relatórios de tendências. De fato, gráficos que mostram como a degradação do modelo evoluem ao longo do tempo são críticos para o processo de monitoramento do modelo.

Uma vez que temos o ciclo de vida para a consolidação de produtos finais, baseados em Modelagem Preditiva, precisamos entender como documentar cada etapa deste ciclo. Abaixo listamos as principais informações que usualmente constam em documentações deste tipo:

- A fase de **definição do problema**, também conhecida como fase de Entendimento de Negócio, deve consolidar um documento que descreva de forma objetiva e clara métricas de avaliação, metas específicas e objetivo geral.

- A fase de **coleta de dados** conclui com a consolidação de uma documentação que descreve:
  - Dados coletados;
  - Conclusões sobre como os dados devem ser explorados;
  - Discussão sobre quais questões relacionadas a qualidade dos dados devem ser abordadas.
- A fase de **pré-processamento dos dados** conclui com a construção de documentação, que contém:
  - A localização dos arquivos de entrada usados na modelagem dos bancos de dados;
  - O código usado para limpar os dados e criar o banco de dados de modelagem;
  - A localização do banco de dados de modelagem final;
  - Uma descrição conceitual de como os aspectos técnicos relacionados ao pré-processamento e transformação foram abordados.
- A fase de **aprendizado de modelo** conclui com um relatório documentando:
  - As variáveis preditoras e os coeficientes no modelo final;
  - Os principais testes de significância estatística e visualizações empregados para as variáveis de preditores individuais;
  - Os testes do desempenho geral do modelo.
- Na conclusão da fase de **avaliação do modelo,** é elaborado um relatório que:
  - Demonstra que os objetivos de modelagem foram cumpridos.
  - Documentos de assinatura dos principais interessados.

  - Resume o protótipo do impacto.

- A fase de **implantação** deve terminar com um documento contendo:
  - Etapas do processo de implantação;
  - Especificações do ambiente de produção;
  - Script de implantação a ser seguido;
  - Script de publicação de novas versões.
- A fase de **monitoramento** deve terminar com um documento contendo:
  - Descrição das ferramentas de monitoramento;
  - Endereço e descrição de forma de acesso e interpretação do monitoramento;
  - Descrição das metodologias e métricas empregadas.

6.3. Melhores práticas

O grande uso de Modelagem Preditiva em variados cenários, por profissionais oriundos de diversas formações, *backgrounds* de conhecimento e domínios de estudo, permitiram-nos obter uma grande base de conhecimento sobre o que, em geral, funciona ou não quando tentamos aplicar Modelagem Preditiva na prática. Em 2017, *Martin Zinkevich* publicou um relatório que sumariza diversos destes conhecimentos, transformando-os em práticas a serem adotadas por especialistas na área. Em nosso material, discutiremos brevemente algumas das principais práticas presentes neste documento.

De maneira mais ampla, a principal e primeira prática a se adotar é: faça funcionar primeiro. Antes termos um motor muito simples que funciona a ter uma 'geringonça' super potente e inovadora que não funciona. Neste sentido, alguns

cuidados básicos devem ser tomados ao longo do processo. Primeiro, certifique-se de que o seu pipeline é sólido de ponta a ponta. Além disso, comece com um objetivo razoável e, quando necessário, torne-o gradativamente mais complexo. Ou seja, trabalhe de maneira gradual. Assim, iterativamente adicione recursos de senso comum de forma simples e, posteriormente, certifique-se de que seu pipeline permaneça sólido. Essa abordagem gerará muito dinheiro e/ou fará muita gente feliz por um longo período de tempo. Divirja apenas quando não há mais soluções simples. A adição de complexidade retarda as versões futuras. Além dessa perspectiva simplificada e gradual, listamos algumas práticas muito úteis:

1. **Primeiro, projete e implemente métricas**: acompanhe o máximo possível o seu sistema atual. É mais fácil obter permissão do usuário do sistema para obter dados históricos no início. Projete seu modelo com a instrumentação de métricas em mente. Você notará o que muda ou permanece igual ao longo do tempo.

2. **Prefira Machine Learning sobre uma heurística complexa**: uma heurística simples pode criar um novo produto. Porém, uma heurística complexa não é sustentável. Com dados e uma ideia básica do que você está tentando realizar, use ML. Como na maioria das tarefas de Modelagem Preditiva, você estará constantemente atualizando sua abordagem, seja através de uma heurística ou de ML, e ML é mais fácil de atualizar e manter.

3. **Mantenha o primeiro modelo simples e obtenha a infraestrutura correta**: Um primeiro corte sobre o que "bom" e "mau" significa para o seu sistema. Escolher recursos simples torna mais fácil garantir que as features sejam utilizadas pelo algoritmo de aprendizagem corretamente.

4. **Teste a infraestrutura independentemente do modelo**: teste a obtenção de dados no algoritmo. O modelo em seu ambiente de treinamento dá o mesmo resultado que o modelo no seu ambiente de produção?

5. **Comece com um modelo interpretável**: facilite a depuração. Sempre que possível tente primeiro um modelo linear. Obtenha informações sobre a

qualidade de cada recurso. Tente entender porque funciona ou não.

6. **Planeje a publicação de várias versões**: espere lançar mais de um modelo. Analise o impacto da complexidade em futuros lançamentos. Verifique o quão fácil é adicionar, remover ou recombinar features. Crie uma nova cópia do pipeline. Verifique a correção do pipeline. Tenha duas ou três cópias em paralelo.

7. **Explore features simples**: com toneladas de dados é mais fácil aprender milhões de features simples do que poucas complexas. Mantenha grupos de features em que cada feature se aplica a uma fração muito pequena de seus dados, mas a cobertura geral é superior a 90%. Use a regularização para eliminar as features que se aplicam a poucos exemplos.

8. **Não avalie o produto, você não é um típico usuário final**: porque você não deve avaliar os resultados do modelo: você está muito perto do código e pode procurar um aspecto particular dos dados. Além disso, seu tempo é muito valioso. De modo geral, a melhor estratégia para avaliar o modelo é pagando a leigos para responder perguntas em uma plataforma de *crowdsourcing*. Ou ainda, se possível, tente executar um experimento em tempo real com usuários reais (avaliações qualitativas).

9. **Não perca tempo com novas features, se objetivos conflitantes existem**: à medida que suas medidas se estendem, sua equipe começará a analisar questões que estão fora do escopo dos objetivos de sua ML atual. Você precisa alterar o objetivo ou os objetivos do seu produto.

10. **Decisões de publicação são um proxy para objetivos de longo prazo**: as únicas decisões de publicação fáceis são quando todas as métricas melhoram (ou pelo menos não pioram). As decisões de publicação dependem de múltiplos critérios, apenas alguns dos quais podem ser otimizados diretamente. As métricas mensuráveis nos testes A/B em si são apenas um proxy para objetivos a mais longo prazo: satisfazer os usuários, aumentar os usuários, satisfazer os parceiros e lucrar. Nenhuma métrica cobre a preocupação final

da equipe: "Como este produto estará em cinco anos?"

11. **Quando nada muda, procure fontes de informação qualitativamente novas**: quando é difícil ver melhoria significativa nas métricas, embora você esteja adicionando novas features, significa que é hora de começar a construir a infraestrutura para recursos radicalmente diferentes. Talvez seja a hora de usar *deep learning*, por exemplo. Além disso, comece a ajustar suas expectativas quanto ao retorno esperado no investimento. Por fim, sempre pese o benefício de se adicionar novas features contra o custo de uma complexidade maior.

Referências

ABBOTT, Dean. *Applied predictive analytics:* Principles and techniques for the professional data analyst. John Wiley & Sons, 2014.

BISHOP, Christopher M. *Pattern Recognition and Machine Learning*. Springer, 2006.

BOX, George E. P; HUNTER, William G; HUNTER, J. Stuart. *Statistics for Experimenters:* An Introduction to Design, Data Analysis, and Model Building. John Wiley & Sons, 1978.

BUSSAB, Wilton de Oliveira; BOLFARINE, Heleno. *Elementos de Amostragem*. 2. ed. Edgard Blucher, 2005.

CHAMBERS, R. L.; SKINNER, C. J. *Analysis of Survey Data*. Wiley, 2003.

CHU, Robert; DULING, David; THOMPSON, Wayne. *Best Practices for Managing Predictive Models in a Production Environment*. Disponível em: <http://mwsug.org/proceedings/2007/saspres/MWSUG-2007-SAS02.pdf>. Acesso em: 31 mar. 2020.

COCHRAN, Willaim G. *Técnicas de amostragem*. Ed. Fundo de Cultura, 1965.

DIETTERICH, Thomas G*. Ensemble Methods in Machine Learning.* Springer, Berlin, Heidelberg, 2000.

Dow, Steven P. et al. *Parallel prototyping leads to better design results, more divergence, and increased self-efficacy*. ACM Trans. Comput.-Hum. Interact, 2010.

FREITAS, Henrique; MOSCAROLA, Jean. *Da Observação à Decisão:* Métodos de Pesquisa e de Análise Quantitativa e Qualitativa de Dados. RAE-eletrônica, 2002.

GEISSER, Seymour. *Predictive Inference:* An Introduction. New York: Chapman & Hall, 1993.

Gopinathan, K.M. et al. *Fraud detection using predictive modeling*. U.S. Patent, 1992.

GRAVETTER, Frederick J.; WALLNAU, Larry B. *Statistics for the Behavioral Sciences*. Cengage Learning, 2012.

HAND, David; MANNILA, Heikki; SMYTH, Padhraic. *Principles of Data Mining*. The MIT Press, 2001.

HEY, Tony; TANSLEY, Stewart; TOLLEN, Kristin. *The Fourth Paradigm:* Data-intensive scientific discovery. Ed. Tony Hey. Vol. 1. Redmond, WA: Microsoft research, 2009.

HUFF, Darrell; GEIS, Irving. *How to Lie with Statistics*. W. W. Norton & Company, 1993

JAIN. Raj. *The Art of Computer Systems Performance Analysis:* techniques for experimental design, measurement, simulation and modeling. John Wiley, 1991.

KONONENKO, Igor; KUKAR, Matjaz. *Machine learning and Data Mining:* introduction to principles and algorithms. Woodhead Publishing, 2007.

KOTSIANTIS, Sotiris; KANELLOPOULOS, Dimitris; PINTELAS, P. E. *Data Preprocessing for Supervised Learning*. International Journal of Computer Science, 2006.

KUHN, Max; KJELL, Johnson. *Applied Predictive Modeling*. Vol. 26. New York: Springer, 2013.

KUHN, Max; KJELL, Johnson. *Applied Predictive Modeling*. Vol. 810. New York: Springer, 2013.

MCGEOCH, Catherine C. *A Guide to Experimental Algorithmics*. Cambridge University Press, 2012.

MYATT, Glenn J. *Making Sense Of Data:* a practical guide to exploratory data analysis and data mining. John Wiley & Sons, 2007.

NASRABADI, Nasser M. *Pattern recognition and machine learning*. Journal of Electronic Imaging Volume 16, Issue 4, 2007.

PYLE, D. *Data Preparation for Data Mining*. Morgan Kaufmann Publishers, 1999.

ROBERTS, Ryan. *Machine Learning:* The Ultimate Beginners Guide For Neural Networks, Algorithms, Random Forests and Decision Trees Made Simple. CreateSpace Independent Publishing Platform, 2017.

SALE, Joanna E. M; LOHFELD, Lyenne H; BRAZIL, Kevin. *Revisiting the quantitative-qualitative debate: Implications for mixed-methods research*. Quality and quantity, 2002.

SCHÖLKOPF, Bernhard; SMOLA, Alexander J. *Learning with kernels:* support vector machines, regularization, optimization, and beyond. MIT press, 2002.

SHAW, Mary. *What Makes Good Research in Software Engineering?* © Springer-Verlag, 2002

Shmueli, Galit, and Otto R. Koppius. *Predictive analytics in information systems research*. Mis Quarterly, 2011.

SIEGEL, Eric. *Predictive Analytics:* The power to predict who will click, buy, lie, or die. Wiley, 2016.

STUART, Alan*. Basic Ideas of Scientific Sampling*. New York: Griffin, 1976.

TAYLOR, Steven. *Applied Predictive Modeling:* An Overview of Applied Predictive Modeling. Amazon Digital Services LLC, 2017

TICHY, Walter F*. Should Computer Scientists Experiment More?* Germany: University of Karlsruhe, 1997.

WALLER, Matthew A; FAWCETT, Stanley E. *Click here for a data scientist:* Big data, predictive analytics, and theory development in the era of a maker movement supply chain. Journal of Business Logistics, 2013.

WALLER, Matthew A; FAWCETT, Stanley E. *Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management*. Journal of Business Logistics, 2013.

WITTEN, Ian H; FRANK, Eibe. *Data Mining:* Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

ZINKERVICH, Martin. *Rules of Machine Learning:* Best Practices for ML Engineering. 2017.

ZINS, Chaim. *Conceptual approaches for defining data, information, and knowledge*. Journal of the Association for Information Science and Technology, 2007.