FGA0221 - Inteligência Artificial

Portfólio 05

Fevereiro, 2025

Universidade de Brasília

Faculdade UnB Gama

Tema do portfólio:

Quantificando incertezas e redes Bayesianas, raciocínio probabilístico ao longo do tempo e filtros de Kalman.

Aluno: João Matheus de Oliveira Schmitz
Matrícula: 200058525
Turma: T01
Semestre: 2024.2

Fevereiro, 2025

Sumário

Fevereiro, 2025

1. Incertezas

Os agentes de inteligência artificial comentados nos portfólios anteriores são capazes de resolver uma grande gama de problemas. Entretanto, existe um ponto específico que os torna, muitas vezes, incapazes de lidar com situações complexas do mundo real: a **incerteza**.

O agente lógico do mundo de Wumpus, por exemplo, é capaz de lidar com algumas incertezas como "Se tem uma brisa no quadrado atual, então haverá um ou mais poços nos quadrados adjacentes". Porém, o método utilizado só é eficaz em situações simples como as encontradas no micro-mundo, isso ocorre pois o agente usa o conhecimento de todo os estados possíveis do ambiente que ele consegue observar para tentar resolver a incerteza e descobrir a localização do poço que gerou aquela brisa, algo inviável em ambientes mais complexos. Outro ponto a ser levado em conta é que esses agentes sempre tentam evitar as incertezas e são incapazes de lidar de forma inteligente quando as únicas escolhas possíveis são incertas.

Para resumir, podemos elencar os seguintes motivos que tornam inviável o uso da lógica no tratamento de incertezas:

- **Complexidade**: listar todas as possibilidades de causa e efeito pode ser algo inviável;
- **Ignorância teórica**: o conhecimento sobre o domínio pode não ser completo;
- **Ignorância prática**: podem haver informações insuficientes para a tomada de uma decisão 100% precisa;

Fazendo jus a sua características de ser um campo multidisciplinar, os agentes de inteligência artificial utilizam-se da **teoria da probabilidade** para tratar incertezas, atribuindo um valor entre 0 e 1 para cada possibilidade, representando a porcentagem de chances de tal possibilidade ocorrer.

2. Utilidade

Considere a seguinte questão: temos duas escolhas, uma com 90% de chances de alcançar a meta e outra com 80% de chances, podemos dizer que a primeira é sempre a melhor? A resposta é não! Embora possa parecer contra intuitivo, a verdade é que agentes de IA podem ser configurados **não somente para alcançar uma meta**, mas também para **levar em consideração o caminho até ela**.

Podemos visualizar bem isso no caso de um carro autônomo com a meta de deixar seu passageiro no destino A em até 1 hora de viagem. Porém, o passageiro pode requisitar que o carro dê **preferência** a rotas que passem à beira do mar, mesmo que isso gere um pequeno atraso. Neste caso, o carro poderá escolher a rota com 80% de chance de alcançar sua meta, mas que passa à beira do mar, ao invés da rota com 90% de chance.

As preferências de um agente são relativas, podendo variar de agente para agente ou de usuário para usuário. Para que o agente consiga balancear as preferências com o alcance de sua meta, será utilizado a **teoria da utilidade**. Segundo [2](Russell & Norvig, 2009), "A teoria da utilidade diz que **todo estado tem determinado grau de utilidade para um agente** e que o **agente preferirá estados com utilidade mais alta**". Com isso os agentes serão capazes de escolher o caminho mais útil até sua meta, ao invés do mais simples ou mais rápido.

3. Teoria de Decisão

A teoria da decisão é utilizada como a base para um agente determinar suas ações em situações incertas. Ela é formada pela combinação entre a teoria da probabilidade e a teoria da utilidade:

$$Teoria\ da\ decisão\ =\ teoria\ da\ probabilidade\ +\ teoria\ da\ utilidade$$

Segundo [2](Russell & Norvig, 2009), a base da teoria de decisão é que "... um agente é **racional** se e somente se escolhe a ação que resulta na **mais alta utilidade esperada**, calculada como **a média sobre todos os resultados possíveis** da ação". Esse é o princípio da **utilidade máxima esperada**, onde "esperada" tem como significado a média dos resultados ponderada por sua probabilidade.

4. Notação básica de probabilidade

Para programar um agente capaz de lidar com incertezas, precisamos utilizar a teoria da probabilidade e trazer seus conceitos para dentro da área de inteligência artificial. Alguns dos conceitos e representações utilizadas são:

- **Espaço amostral** se refere ao conjunto de todos os estados de mundos possíveis;
- Um **mundo possível** é definido como uma atribuição de valores a todas as variáveis aleatórias em consideração
- Mundos possíveis são **mutuamente exclusivos e exaustivos**, ou seja, dois mundos não podem coexistir e um mundo deve ser sempre válido;
- **Modelo de probabilidade** associa uma probabilidade numérica *P* para cada mundo possível;
- **Axioma básico de probabilidade** diz que todo mundo possível tem uma probabilidade entre 0 e 1 e a probabilidade total do conjunto de mundos possíveis é 1;
- Afirmações e consultas probabilísticas são geralmente realizadas sobre conjuntos de mundos possíveis, esses conjuntos são chamados de **eventos** (e.g. a probabilidade de dois dados rolarem o mesmo número);
- Os conjuntos são sempre descritos por **proposições** e, para cada proposição, o conjunto correspondente contém apenas os mundos onde a proposição é válida. A probabilidade de uma proposição é a soma das probabilidades dos mundos nos quais é válida;
- **Probabilidades incondicionais** ou ***à priori***: são probabilidades já conhecidas que não dependem de nenhuma outra informação (e.g. tirar o número 3 em um dado tem uma probabilidade de 1/6);
- **Probabilidades condicionais** ou ***à posteriori***: são probabilidades que só serão conhecidas quando houver informações já relevadas, as chamadas **evidências** (e.g. considerando que o valor de um dado lançado é 5, qual a probabilidade de um segundo dado tirar o mesmo valor?);
- Para representar a probabilidade de uma **disjunção**, utilizamos a seguinte regra: $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$;

- Para representar a probabilidade de uma **conjunção**, usamos a **regra do produto**: $P(a \wedge b) = P(a \mid b)\, P(b)$ ou $P(a \wedge b) = P(b \mid a)\, P(a)$;

- A **negação de uma proposição** 'a' é dada pela equação $1 - P(a)$. Por exemplo, se $P(a) = 0.3$, então $P(\neg a) = 0.7$.

- **Variáveis aleatórias** são as variáveis na teoria da probabilidade e sempre começam em letra maiúscula (e.g. $Dado$);

- Cada variável aleatória tem seu **domínio**, uma faixa de valores que ela pode receber, sempre em letras minúsculas (e.g. $Dado = \{1, 2, 3, 4, 5, 6\}$ e $Idade = \{criança, adolescente, adulto, idoso\}$). Os domínios podem ser finitos ou infinitos, discretos (números inteiros) ou contínuos (números reais);

- Para representar variáveis contínuas, utilizamos a chamada **função densidade de probabilidade**, onde determinamos a probabilidade de que uma variável aleatória assume algum valor de x através de uma função (e.g. $P(TempMeioDia = x) = Uniforme_{[18C,\, 26C]}(x)$, o que diz que a temperatura varia uniformemente entre 18°C e 26°C no meio dia);

- Uma **distribuição de probabilidade** de uma variável é escrita como **P**(Nome da variável) e representa as probabilidades de uma variável assumir cada valor possível em seu domínio. Por exemplo, levando em conta a variável a seguir: $Idade = \{criança, adolescente, adulto, idoso\}$; podemos dizer que, em um grupo de pessoas, a probabilidade de alguém ter certa idade é determinada por $\boldsymbol{P}(Idade) = < 0.2;\ 0.4;\ 0.3;\ 0.1 >$, onde $P(Idade = criança) = 0.2$, $P(Idade = adolescente) = 0.4$, e assim sucessivamente, seguindo a ordem em que o domínio foi escrito;

- Há também a **distribuição de probabilidade conjunta** , a qual funciona de forma idêntica a explicada acima, mas representando as probabilidades de mais de uma variável ao mesmo tempo utilizando a regra do produto (uma de suas utilidades é nos permitir extrair a probabilidade incondicional de uma variável ao fixar o valor das outras);

- Um modelo de probabilidade é completamente determinado pela distribuição conjunta de todas as variáveis aleatórias - a chamada **distribuição de probabilidade conjunta completa**;

- Utilizamos α como símbolo de **normalização**, ou seja, para que a soma das probabilidades seja igual a 1 (e.g. $\boldsymbol{P}(a) = \alpha < 0.2;\ 0.3 > = < 0.4;\ 0.6 >$).

5. Independência

Quando observamos as probabilidades de mais de uma variável ao mesmo tempo, podemos notar casos onde não importa o valor que uma variável X assuma, a probabilidade da variável Y permanecerá constante ou vice-versa. Com isso podemos concluir que as variáveis X e Y são independentes uma da outra, o que pode vir a diminuir muito a complexidade e a quantidade de informações presentes em uma distribuição de probabilidade conjunta.

Como exemplo, podemos utilizar as variáveis:

- $Casaco = \{sim, não\}$;
- $Temperatura = \{quente, morno, frio\}$;
- $Celular = \{verdadeiro, falso\}$;

Nesse caso, estamos avaliando as probabilidades de alguém vir de casaco para um local, a probabilidade de estar uma certa temperatura e a probabilidade dessa pessoa trazer um celular com ela. Ao analisar as variáveis, podemos perceber que a variável Celular é independente das outras, portanto, podemos concluir três coisas:

- $P(verdadeiro \mid sim, frio) = P(verdadeiro)$;
- $P(sim, frio, verdadeiro) = P(não, frio)P(verdadeiro)$;
- $\boldsymbol{P}(Casaco, Temperatura, Celular) = \boldsymbol{P}(Casaco, Temperatura)\boldsymbol{P}(Celular)$;

A independência entre proposições e a independência entre variáveis podem ser escritas da seguinte forma, respectivamente:

- $P(a \mid b) = P(a)$ **ou** $P(b \mid a) = P(b)$ **ou** $P(a \wedge b) = P(a)P(b)$;
- $\boldsymbol{P}(X \mid Y) = \boldsymbol{P}(X)$ ***ou*** $\boldsymbol{P}(Y \mid X) = \boldsymbol{P}(Y)$ ***ou*** $\boldsymbol{P}(X \wedge Y) = \boldsymbol{P}(X)\boldsymbol{P}(Y)$;

Existe também a chamada **independência condicional**, onde uma variável X é independente de uma variável Y se e somente se a variável Z que é a causa de ambas tiver seu valor definido. Nesse caso, podemos dizer que a variável Z faz uma conexão entre as probabilidades de ambas as variáveis, algo que não ocorre mais quando Z se torna fixa. Outro ponto importante é que, segundo [2](Russell & Norvig, 2009), "as asserções de independência condicional podem permitir o aumento da escala de sistemas probabilísticos; além disso, elas são muito mais comuns que as asserções de independência absoluta"

6. Regra de Bayes, aplicações e modelo ingênuo

Na seção 4, definimos que é utilizada a regra do produto para determinar a probabilidade de conjunção entre proposições. Tomando essa regra como base, é possível derivar outra: a **regra de Bayes**.

$$P(b \mid a) = \frac{P(a \mid b)P(b)}{P(a)}$$

Essa equação, mesmo sendo simples, é extremamente importante para a área de inteligência artificial. Isso ocorre pois ela é a base para todos os sistemas modernos de IA para **inferência probabilística**.

O seu uso tem como principal objetivo **calcular uma causa dado um determinado efeito**, como em um diagnóstico médico, onde um paciente exibe sintomas variáveis e queremos determinar qual a causa desses sintomas. Nesse caso, para facilitar a compreensão, podemos reescrever a regra de Bayes como:

$$P(causa \mid efeito) = \frac{P(efeito \mid causa)P(causa)}{P(efeito)}$$

Portanto, para determinar a probabilidade de uma causa X gerar o efeito Y, devemos considerar a probabilidade de Y ser causado por X, a probabilidade de X ocorrer e a probabilidade de Y ocorrer. Como exemplo, podemos considerar o sintoma 'tosse' e a doença 'gripe' em um paciente doente de um hospital:

$$P(gripe \mid tosse) = \frac{P(tosse \mid gripe)P(gripe)}{P(tosse)}$$

- $P(tosse \mid gripe)$ representa a probabilidade de um paciente apresentar tosses dado que ele está com gripe;
- $P(gripe)$ representa a probabilidade da doença de um paciente ser gripe;
- $P(tosse)$ representa a probabilidade de um paciente ter tosse;

É importante dizer que essas probabilidades são calculadas com base em **todos** os pacientes que já foram nesse hospital no passado, independente do sintoma ou doença que ele apresentava. Aqui usaremos os valores fictícios:

- $P(tosse \mid gripe) = 0.7$;
- $P(gripe) = 0.08$;
- $P(tosse) = 0.2$;

Com base nisso, podemos então calcular a probabilidade de um paciente ter gripe caso esteja tossindo:

$$P(gripe \mid tosse) = \frac{0.7 * 0.08}{0.2} = 0.28$$

Ou seja, 28% dos pacientes que apresentam tosse estão gripados.

O exemplo acima demonstra uma probabilidade incondicional, mas e se tivermos evidência que um paciente X chegou ao hospital com os sintomas **condicionalmente independentes** tosse e catarro? Nesse caso, podemos dizer que $P(efeito) = 1$. Para completar ainda mais o exemplo, podemos obter a distribuição de probabilidade da variável *Gripe* com $domínio = \{verdadeiro, falso\}$, o qual será representado como $\{gripe, \neg gripe\}$.

$$\boldsymbol{P}(gripe \mid tosse \wedge catarro) =$$
$$\alpha < \boldsymbol{P}(tosse \wedge catarro \mid gripe); \boldsymbol{P}(tosse \wedge catarro \mid \neg gripe) >$$

A distribuição de probabilidade conjunta imaginária dessas variáveis é:

–	$tosse$		$\neg tosse$	
	$catarro$	$\neg catarro$	$catarro$	$\neg catarro$
$gripe$	0.12	0.08	0.05	0
$\neg gripe$	0.05	0.1	0.1	0.5

Tabela 01 - Distribuição de probabilidade conjunta de um diagnóstico hospitalar

Com esses dados, podemos então calcular a resposta:

$$\boldsymbol{P}(Gripe \mid tosse \wedge catarro) = \alpha < 0.12\,;\ 0.05 >$$
$$\boldsymbol{P}(Gripe \mid tosse \wedge catarro) = < 0.705\,;\ 0,295 >$$

Portanto, podemos dizer que um paciente que apresenta tosse e catarro tem uma probabilidade de ~70,5% de ter gripe e ~29,5% de não ter gripe.

Outro ponto importante é que a implementação de um sistema probabilístico completo tem complexidade $O(2^n)$, porém, segundo [2](Russell & Norvig, 2009), "... as asserções de independência condicional podem permitir o aumento da escala de sistemas probabilísticos; além disso, elas são muito mais comuns que as asserções de independência absoluta".

Isso ocorre pois a **independência de variáveis diminui o tamanho da distribuição de probabilidade conjunta**, transformando a complexidade do algoritmo de $O(2^n)$ em $O(n)$, permitindo que os sistemas de probabilidade para inteligência artificial consigam lidar com um número extremamente maior de variáveis em comparação ao que seria possível sem levar o conceito de independência condicional em consideração. De acordo com [2](Russell & Norvig, 2009): "A decomposição de grandes domínios probabilísticos em subconjuntos conectados livremente por meio de independência condicional é um dos desenvolvimentos mais importantes na história recente da IA".

Devido a independência condicional de variáveis ser algo muito comum em exemplos reais, foi encontrado um padrão que pode ser exemplificado pelo exemplo anterior com as variáveis Gripe, Tosse e Catarro, no qual **uma única causa influencia de maneira direta vários efeitos, todos condicionalmente independentes, dada a causa**. Em casos como esse, a distribuição conjunta total pode ser escrita como:

$$\boldsymbol{P}(Causa, Efeito_1, Efeito_2, \ldots, Efeito_n) = \boldsymbol{P}(Causa) \prod_i \boldsymbol{P}(Efeito_i \mid Causa)$$

Essa distribuição é chamada de **modelo bayesiano ingênuo**, pois embora seu uso seja próprio para o padrão descrito anteriormente, ele é usado com frequência (como hipótese simplificadora) mesmo em casos onde as variáveis "efeito" não são condicionalmente independentes dada a causa.

7. Redes Bayesianas

Uma rede bayesiana é uma estrutura de dados que visa **representar as dependências entre as variáveis aleatórias**, sendo capazes de representar qualquer distribuição de probabilidade conjunta completa e, geralmente, de forma concisa.

Mas então, como funciona essa estrutura de dados? Ela tem as seguintes características:

- É um **grafo orientado** em que cada nó é identificado com informações de probabilidade quantitativa;
- Cada **nó** corresponde a uma variável aleatória, que pode ser discreta ou contínua;
- Um conjunto de vínculos orientados ou setas conecta pares de nós;
- Se houver uma seta do nó X até o nó Y, X será denominado **pai** de Y. Nesse caso, também pode ser dito que **X será uma causa com efeito Y**;
- O grafo **não tem ciclos orientados** (grafo acíclico);
- Cada nó X_i tem uma distribuição de probabilidade condicional $P(X_i \mid Pais(X_i))$ que quantifica o efeito dos pais sobre o nó;
- A topologia da rede especifica os relacionamentos de independência condicional que são válidos no domínio;

Quando falamos de independência condicional em redes bayesianas, podemos afirmar que:

- Um nó é condicionalmente independente de seus **predecessores, dados seus pais**;
- Um nó é condicionalmente independente de seus **não descendentes, dados seus pais**;
- Um nó é condicionalmente independente de **todos os outros nós na rede, dados seus pais, filhos e pais dos filhos**, o que é chamado de **Cobertor de Markov**;

Um exemplo de rede bayesiana se encontra na figura a seguir, onde uma casa tem um alarme instalado que pode ser acionado por um terremoto ou um

roubo. Se o alarme for ativado, os vizinhos da casa podem ouvir e ligar para o dono da casa, avisando sobre o alarme. Porém, há uma chance dos vizinhos se equivocarem (um deles mais do que outro) ou do alarme tocar mesmo quando não ocorre roubo ou terremoto.

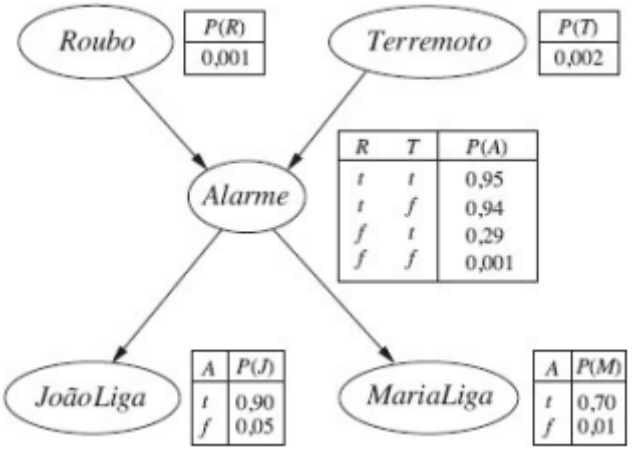

Figura 01 - Rede bayesiana básica

Como podemos ver, existem variáveis que não dependem de nenhuma outra e tem sua probabilidade fixa, e variáveis que dependem de outras (tem uma causa). No caso dessas últimas, a probabilidade delas depende do valor atribuído à(s) variável(is) que as causam. Além disso, todas as variáveis 'efeito' são condicionalmente independentes umas das outras dado o valor de sua causa.

As tabelas da figura são chamadas de **tabelas de probabilidade condicional**, onde cada linha contém a probabilidade daquela variável ser verdadeira em um **caso de condicionamento**, que, por sua vez, nada mais é que uma atribuição possível de valores para as variáveis causa da variável que a tabela se refere.

É importante dizer que, como as variáveis da figura tem domínio booleano, só representamos a probabilidade dela ser verdadeira nessas tabelas, pois a probabilidade dela ser falsa é facilmente descoberta usando a fórmula $1 - P(X)$, então omitimos esta segunda. Porém, no caso de variáveis como Dado, que tem $domínio = \{1, 2, 3, 4, 5, 6\}$, **a tabela teria um número de colunas igual ao número de valores possíveis em seu domínio**, no caso de Dado, seriam 6 colunas. Além disso, a soma dos valores de uma linha na tabela sempre deve ser 1 (tirando domínios booleanos como explicado acima).

Uma rede bayesiana pode ter sua semântica compreendida de dois modos válidos:

- Uma representação da **distribuição de probabilidade conjunta**;
- Uma codificação de uma **coleção de declarações de independência condicional**;

Outra característica das redes bayesianas é sua capacidade de ser mais **compacta** que a distribuição conjunta total, o que torna possível seu uso em situações com um número muito grande de variáveis aleatórias. A densidade das redes bayesianas é um exemplo de propriedade de **sistemas localmente estruturados**, também chamados de **sistemas esparsos**.

Estes sistemas são caracterizados por cada componente interagir com um número bem limitado de outros componentes. Um exemplo de sistema esparso é uma rede social, onde cada usuário é um nó no grafo e cada ligação entre nós significa se um usuário segue outro ou não. Podemos concluir que, nesse caso, a quantidade de usuários que um usuário X está conectado pode estar na casa das dezenas, centenas ou até milhares, mas quando comparada às milhões ou até bilhões de usuários cadastrados naquela rede social, esse número é ínfimo.

Essa característica é importante pois, ao modelar uma rede desse modo, a quantidade de probabilidades que necessitam ser calculadas é tremendamente menor do que se fossemos calcular a distribuição conjunta total. Um exemplo seria em uma rede com 30 nós $(n = 30)$, onde cada nó tem 5 pais $(k = 5)$. O número de probabilidades das duas opções será:

- Rede bayesiana: $n2^k = 30 * 2^5 = 30 * 32 = 960$ probabilidades;
- Distribuição conjunta total: $2^n = 2^{30} = 1.073.741.824$ probabilidades;

Como podemos ver, a complexidade de crescimento de uma é linear enquanto de outra é exponencial. Consequentemente, a diferença de poder computacional e tempo necessários para rodar as duas opções é enorme - dependendo do valor de n, a segunda opção é completamente inviável.

Existem alguns outros conceitos que podem ajudar a representar uma rede bayesiana de forma ainda mais eficiente, eles são:

- **Nós determinísticos**: Quando um nó tem seu valor especificado exatamente pelos valores de seus pais, sem incerteza;

- **Independência de contexto específico (CSI)**: Uma distribuição condicional exibe CSI se uma variável é condicionalmente independente de alguns de seus pais dados certos valores de outros;

- **OR-ruidoso**: O modelo OR ruidoso permite incerteza sobre a capacidade de cada genitor de causar o filho ser verdadeiro;

Entretanto, tudo o que foi discutido acima leva como base que os valores das variáveis são discretos. Para representar variáveis contínuas em redes bayesianas, existem algumas alternativas:

- **Discretização**: Dividir os valores possíveis em intervalos fixos, por exemplo, faixas de temperatura (<0°C, 0°C-100°C, e >100°C);

- **Distribuições probabilísticas**: Outra abordagem é definir uma variável contínua usando uma das famílias padrão de funções de densidade de probabilidade, como a distribuição Gaussiana ou Uniforme, por exemplo, a temperatura varia uniformemente entre 18°C e 26°C no meio dia - $P(TempMeioDia = x) = Uniforme_{[18C,\, 26C]}(x)$;

- **Não paramétrica**: Define a distribuição condicional implicitamente com uma coleção de instâncias, cada uma contendo valores das variáveis genitoras e filho.

As redes bayesianas com variáveis discretas e contínuas são conhecidas como **redes bayesianas híbridas**. Elas trazem duas novas considerações:

- A distribuição condicional para uma variável contínua dados pais discretos ou contínuos;
- A distribuição condicional para uma variável discreta dados pais contínuos.

Essas distribuições geralmente são resolvidas utilizando **distribuições gaussianas lineares**, **distribuições gaussianas condicionais**, **distribuições probit** ou **distribuições logit**.

A inferência de uma rede bayesiana pode ser exata ou aproximada. A primeira só é viável quando a rede tem um tamanho menor, o que torna necessário o uso da segunda nesses casos. Para inferências aproximadas, é comum o uso de **algoritmos de Monte Carlo**, que fornecem respostas aproximadas cuja exatidão depende do número de amostras geradas.

8. Tempo e incerteza

Anteriormente, trabalhamos com variáveis com valor fixo durante todo o processo de um diagnóstico, por exemplo, um paciente que apresenta tosse e catarro não vai simplesmente parar de tossir enquanto é atendido no hospital. Porém, existem situações em que as variáveis mudam de valor ao longo do tempo, e isso precisa ser considerado. Podemos então dividir os problemas em duas categorias:

- **Problemas estáticos**, quando as variáveis permanecem com um valor definido ao longo do tempo;
- **Problemas dinâmicos**, quando as variáveis podem ter seu valor alterado ao longo do tempo;

9. Estados e observações

Para modelar incertezas através do tempo, podemos utilizar as seguintes afirmações:

- O tempo é **discreto**;
- Cada **instante de tempo** contém um conjunto de variáveis aleatórias, umas com valor observável e outras não;
- Para facilitar, o **intervalo** entre instantes é fixo em um determinado problema;
- O **intervalo** entre instantes pode mudar de um problema para outro;
- Para simplificar, as **variáveis observáveis** são as **mesmas** em todos os instantes;
- Para representar as **variáveis observáveis**, utilizamos E_t ;
- Para representar as **variáveis não observáveis** - também chamadas de **variáveis de estado** -, utilizamos X_t ;
- O tempo é **iniciado** com $t = 0$;
- As **primeiras observações** de variáveis são feitas no momento $t = 1$;
- A **notação a:b** é utilizada para denotar todos os momentos entre o momento 'a' e o momento 'b', além deles próprios, por exemplo, $E_{1:3} = E_1, E_2, E_3$;

10. Modelo de transição e modelo de sensores

Após decidir as variáveis observáveis e não observáveis de um modelo, o próximo passo é ensinar ao agente como o mundo evolui e como ele obtém os valores das variáveis observáveis (evidências). Para isso devemos desenvolver, respectivamente, um modelo de transição e um modelo de sensor.

O **modelo de transição** de um agente que lida com incertezas é diferente do modelo de transição de agentes discutidos em portfólios anteriores. A diferença está no fato de que antes o modelo de transição ditava com certeza qual seria o estado do mundo após o agente tomar determinada ação, agora ele deverá **especificar a distribuição de probabilidade** das variáveis de estado mais recentes, dados os valores anteriores, ou seja $P(X_t \mid X_{0:t-1})$.

O problema com um modelo de transição completo, nesse caso, é que o valor de t aumenta conforme o tempo vai passando, o que gera um aumento na complexidade e exige um maior poder computacional. Para resolver esse problema, nós utilizamos uma **suposição de Markov**, ou seja, fixamos o número de estados anteriores que afetam diretamente o estado atual.

Os processos que utilizam essas suposições são chamados de **processos de Markov** ou **cadeias de Markov**, em que o mais simples deles é a **cadeia de Markov de 1ª ordem** - onde o estado atual só depende do estado anterior a ele ($P(X_t \mid X_{t-1})$). Seguindo a mesma convenção, em uma cadeia de Markov de 2ª ordem o estado atual depende somente dos dois últimos estados, e assim por diante. Para facilitar, dizemos que as regras que definem qual a probabilidade condicional de uma variável não irá mudar ao longo do tempo - o que chamamos de **processo estacionário**.

Já para o **modelo de sensor**, utilizamos a equação $P(E_t \mid X_t)$, ou seja, as evidências coletadas a cada instante são independentes das evidências de instantes anteriores.

Ao unir o que foi dito até agora, podemos modelar a distribuição de probabilidades conjunta completa de todas as variáveis de uma cadeia de Markov de 1ª ordem na equação:

$$P(X_{0:t} \mid E_{1:t}) = P(X_0) \prod_{i=1}^{t} P(X_i \mid X_{i-1}) P(E_i \mid X_i)$$

Observe que o **modelo do estado inicial** $P(X_0)$ está separado dos modelos de transição e de sensor. Isso ocorre pois ele não tem estado anterior para que sua transição leve isso em conta e as evidências só começam a ser coletadas no instante $t = 1$.

Com essa equação, podemos encontrar as probabilidades para, como exemplo, descobrir se choveu em um dia levando em conta a evidência de um funcionário X de uma empresa ter ou não levado guarda-chuva naquele dia. Essa probabilidade pode ser **exata** (0 ou 1) se a variável Chuva for determinística, ou **aproximada** se não. Para uma probabilidade aproximada, pode ocorrer dela não estar muito precisa. Para melhorar essa precisão, podemos utilizar dois métodos:

- Aumentar a **ordem da cadeia de Markov**;
- Aumentar o **conjunto de variáveis observáveis**;

Para cada problema distinto, a ordem ideal da cadeia de Markov e o número ideal de variáveis observáveis também será distinto, o que torna essencial a compreensão das regras que regem tal problema. Caso as regras não sejam bem conhecidas, a descoberta desses valores pode ser realizada através de aprendizado, que será discutido na seção 11.5.

11. Inferência em modelos temporais

Assim como os modelos de transição e de sensor, o processo de inferência também deve ser modificado para se adequar ao tempo. Os processos que podemos utilizar na inferência de modelos temporais são:

11.1. Filtragem

Consiste em calcular a **distribuição *a posteriori* sobre o estado atual**, dada toda a evidência até o momento. Ou seja, $P(X_t \mid E_{1:t})$. Segundo [2](Russell & Norvig, 2009), "A filtragem é o que um agente racional precisa fazer, a fim de manter o controle do estado atual, de forma que possam ser tomadas decisões racionais".

11.2. Predição

Consiste em calcular a **distribuição *a posteriori* sobre um estado futuro**, dada toda a evidência até o momento. Ou seja, $P(X_{t+k} \mid E_{1:t}),\ para\ k > 0$.

11.3. Suavização

Consiste em calcular a **distribuição *a posteriori* sobre um estado passado**, dada toda a evidência até o momento. Ou seja, $P(X_k \mid E_{1:t}),\ para\ t > k \geq 0$. Devido a maior presença de evidências, a suavização gera uma probabilidade mais precisa do que a inicialmente calculada no instante $t = k$.

11.4. Explicação mais provável

Dada uma sequência de observações, poderíamos desejar encontrar a sequência de estados que mais provavelmente gerou tais observações. Isso é muito usado, por exemplo, em reconhecimento de fala, onde precisamos descobrir a sequência mais provável de palavras dado uma sequência de sons.

11.5. Aprendizado

Como dito anteriormente, nem sempre é possível saber qual a melhor ordem de uma cadeia de Markov e/ou a quantidade ideal de variáveis observáveis para resolver um problema. Isso pode ser resolvido através de aprendizado, onde o agente melhora gradualmente esses valores, **aprimorando assim seus modelos de transição e de sensor**.

Segundo [2](Russell & Norvig, 2009), "A inferência fornece uma estimativa de quais transições realmente ocorreram e de quais estados geraram as leituras de sensores, e essas estimativas podem ser usadas para atualizar os modelos. Os modelos atualizados fornecem novas estimativas, e o processo iterage para a convergência. O processo global é uma instância da maximização de expectativas, ou algoritmo EM".

12. Modelo oculto de Markov

Os modelos ocultos de Markov (HMM, do inglês Hidden Markov Model) são **modelos probabilísticos temporais** onde o estado do mundo em cada instante de tempo 't' é **descrito por uma única variável aleatória discreta**, ou seja, é representado por uma única variável de estado.

Um exemplo é o modelo onde queremos descobrir se a variável Chuva é verdadeira ou falsa em determinado dia, levando em conta se um empregado X de uma empresa saiu de casa nesse dia com ou sem guarda-chuva. Nesse caso, também **podemos adicionar outras variáveis observáveis** ao modelo sem nenhum problema, pois não há limites no número destas em um HMM.

De forma contrária as variáveis observáveis, **um HMM não aceita a adição de mais variáveis de estado**. Entretanto, como isso limita muito o escopo de problemas que podem ser modelados como modelos de Markov ocultos, é utilizada uma forma de burlar essa restrição: **juntar as variáveis que descrevem o estado do mundo em uma única megavariável**, onde cada valor do domínio representa uma combinação possível de valores das variáveis que a compõem.

Devido a característica que define os modelos ocultos de Markov - onde só existe uma variável de estado X_t -, os **modelos de transição e de sensor** podem ser representados através de outra forma: através de **matrizes**. Ambos os modelos são representados por matrizes de tamanho $S \, x \, S$ - onde S representa a quantidade de valores possíveis no domínio de X_t.

O seu **modelo de transição é representado por uma única matriz completa** T. Ela contém em cada célula **a probabilidade** T_{ij} **de uma transição do estado** i **para o estado** j, e é construída com base na seguinte equação:

$$T_{ij} = P(X_{t\,=\,j} \mid X_{t-1\,=\,i})$$

Já o seu **modelo de sensor é representado por uma matriz diagonal** O_t **para cada instante de tempo** t. Como ela é uma matriz diagonal todas as células onde $i \neq j$ tem valor igual a zero. Em contrapartida, os outros valores indicam **a probabilidade da variável** E_t **ter obtido seu valor observado levando em conta que a variável de estado está atualmente no estado** i. Isso é representado pela equação:

$$P(e_t \mid X_{t\,=\,i})$$

Como exemplo, podemos utilizar o modelo do guarda-chuva descrito acima. O modelo de transição T da variável Chuva - onde $1 = verdadeiro$ e $2 = falso$ - poderia ser igual a:

$$\mathbf{T} = \mathbf{P}(X_t \mid X_{t-1}) = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}$$

Figura 02 - Matriz de transição do mundo do guarda-chuva

Já o modelo de sensor O_t para os instantes $t = 1$ e $t = 3$ - com a evidência de que o funcionário X trouxe guarda-chuva no dia 1 e não trouxe no dia 3 - poderia ser igual a:

$$\mathbf{O}_1 = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.2 \end{pmatrix}; \qquad \mathbf{O}_3 = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.8 \end{pmatrix}$$

Figura 03 - Matriz de sensor do mundo do guarda-chuva nos instantes t=1 e t=3

13. Filtros de Kalman

Os modelos ocultos de Markov se provam muito úteis quando falamos de variáveis de estado discretas, porém, se a variável que precisamos descobrir é **contínua**, ele se torna incapaz de lidar com o problema da melhor forma possível. Outro ponto que os HMM não consideram é a possibilidade de **ruídos** na observação de evidências.

Para lidar com esses dois pontos - variáveis contínuas e ruídos nas observações -, utilizamos os **filtros de Kalman.** Sua aplicabilidade vai desde o acompanhamento por radar de aeronaves e mísseis, sistemas de controle de resfriamento em reatores nucleares e até navegação autônoma de carros e foguetes. Seu uso é tão extenso devido ao filtro de Kalman ser um **algoritmo ótimo de estimação de estados**, sendo capaz de estimar variáveis desconhecidas com exatidão, mesmo com ruídos sensoriais, além de conseguir prever estados futuros do sistema com base em estimativas anteriores.

Entretanto, o filtro de Kalman tradicional não é suficiente em alguns casos, como em problemas onde o sistema modelado não é linear. Portanto, foram criadas versões desse algoritmo capazes de lidar com essa não linearidade, como o **filtro de Kalman estendido** e o **filtro de Kalman de comutação**. É importante dizer que, para lidar com a não linearidade, esses filtros realizam aproximações, ou seja, não são algoritmos 'ótimos'.

Um exemplo de como o filtro de Kalman é utilizado é no rastreamento de aeronaves por um radar. O radar envia um feixe cônico na direção do alvo a cada 5 segundos, recebendo evidências e calculando a posição e velocidade atual do alvo, utilizando estas para prever a posição seguinte do alvo. Nesse caso, o **estado do sistema** é definido como $[x, y, z, v_x, v_y, v_z, a_x, a_y, a_z]$ e é utilizado como entrada para prever o próximo estado:

- x, y, z se referem a posição tridimensional do alvo;
- v_x, v_y, v_z se referem a velocidade tridimensional do alvo;
- a_x, a_y, a_z se referem a aceleração tridimensional do alvo;

Já os ruídos podem ser divididos em duas categorias: **ruídos de medição**, como calibração do radar; e **ruídos de processo**, como vento e turbulência do ar. O papel do filtro de Kalman, nesse caso, é prever o próximo estado do sistema com exatidão levando em conta ambos os tipos de ruído, capacidade essa que o torna indispensável para sistemas como esse.

14. Impressões sobre o conteúdo

Ao incorporar incertezas em suas considerações, os agentes de inteligência artificial são capazes de resolver problemas muito mais complexos do que aqueles discutidos em portfólios anteriores, os tornando realmente viáveis nas mais diversas situações do mundo real. O exemplo de usabilidade que realmente me demonstrou isso foi a utilização do filtro de Kalman como algoritmo responsável pela navegação do foguete Apollo 11, o qual levou o homem à lua.

Inicialmente, pensei que o uso de probabilidade seria algo mais simples. Porém, me surpreendi com a profundidade das teorias utilizadas na criação desses algoritmos. Um dos pontos altos foi entender as diferentes formas que estes algoritmos utilizam para aproximar o máximo possível as probabilidades de seus valores reais, como a suavização.

Antes de iniciar meus estudos dos conteúdos destes portfólios, se me perguntassem sobre o assunto os primeiros agentes de IA que viriam na minha mente seriam aqueles apresentados nesse portfólio - capazes de resolver situações como pilotar um carro de maneira automática. Porém, agora tenho uma visão muito mais aberta da amplitude de problemas para os quais podemos utilizar inteligência artificial; conclusão a qual o conteúdo atual me ajudou a perceber.

15. Referências

[1] SOARES, Fabiano Araujo. Slides da aula 14 à 20. Apresentação do PowerPoint.

[2] Russell, S. & Norvig, P. **Artificial Intelligence - A Modern Approach**. 3ª ed. Pearson Education, Inc. 2009.

[3] COUTINHO, Thiago. Entenda o que é o Teorema de Bayes e veja aplicações e exemplos. **Voitto**, 2020. Disponível em: https://voitto.com.br/blog/artigo/teorema-de-bayes. Acesso em: 08 fev. 2025.

[4] AUGUSTO, Felipe. Redes Bayesianas. O que são? Como funcionam? Onde vivem?? **Medium**, 2020. Disponível em: https://medium.com/oiluna/redes-bayesianas-fcf35516dedb. Acesso em: 08 fev. 2025.

[5] BECKER, Alex. Tutorial sobre o Filtro de Kalman. **KalmanFilter.NET**, ©2024. Disponível em: https://www.kalmanfilter.net/PT/default_pt.aspx. Acesso em: 09 fev. 2025.