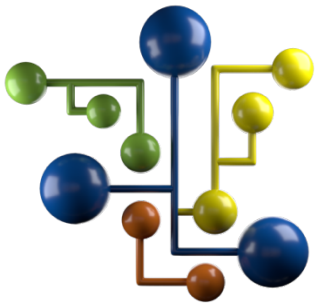


Data Science Academy

www.datascienceacademy.com.br

Big Data Real-Time Analytics com Python e Spark

Árvores de Probabilidade

Algumas decisões de negócios envolvem uma avaliação mais sutil das probabilidades. Dadas as probabilidades de várias circunstâncias que podem afetar os negócios, podemos usar uma imagem chamada “árvore de probabilidade” ou “diagrama de árvore” para ajudar a pensar no processo de tomada de decisão. Uma árvore mostra sequências de eventos como caminhos que parecem ramos de uma árvore. Isso pode nos permitir comparar vários cenários possíveis. Esse é o conceito por trás de uma família de algoritmos de Machine Learning, as Decision Trees e as Random Forests. Aqui está um exemplo.

Dispositivos eletrônicos pessoais, como smartphones e tablets, estão se tornando mais capazes o tempo todo. Fabricar componentes para esses dispositivos é um desafio e, ao mesmo tempo, os consumidores estão exigindo cada vez mais funcionalidades. As leis microscópicas e até submicroscópicas podem se desenvolver durante sua fabricação, o que pode eliminar pixels nas telas ou causar falhas de desempenho intermitentes. Os defeitos sempre ocorrerão, portanto, o engenheiro de qualidade responsável pelo processo de produção deve monitorar o número de defeitos e agir se o processo parecer fora de controle.

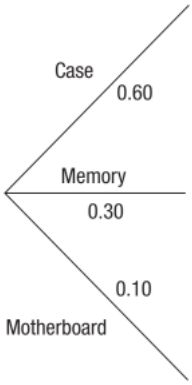
Suponhamos que o engenheiro seja chamado para a linha de produção porque o número de defeitos ultrapassou um limite. Ele deve decidir entre duas ações possíveis. Ele sabe que um pequeno ajuste nos robôs que montam os componentes pode resolver vários problemas, mas para problemas mais complexos, toda a linha de produção precisa ser desligada para identificar a fonte. O ajuste requer que a produção seja interrompida por cerca de uma hora, mas desligar a linha de produção interrompe o trabalho de pelo menos um turno inteiro (oito horas). Naturalmente, seu chefe preferiria que ele fizesse o ajuste simples. Mas sem saber a origem ou a gravidade do problema, ele não pode ter certeza se isso será bem-sucedido.

Se o engenheiro quiser prever se o ajuste menor funcionará, ele poderá usar uma árvore de probabilidade para ajudar a tomar a decisão. Com base em sua experiência, o engenheiro acredita que há três problemas possíveis:

(1) As placas-mãe podem ter conexões defeituosas,
(2) A memória pode ser a fonte das conexões defeituosas ou
(3) Alguns cases das placas-mãe podem estar incorretamente acoplados na linha de montagem.

Ele sabe de dados empíricos anteriores com que frequência esses tipos de problemas surgem e como é provável que apenas fazer um ajuste conserte cada tipo de problema. Os problemas da placa-mãe são raros (10%), os problemas de memória têm aparecido em cerca de 30% do tempo e os problemas de alinhamento dos cases ocorrem com mais frequência (60%).

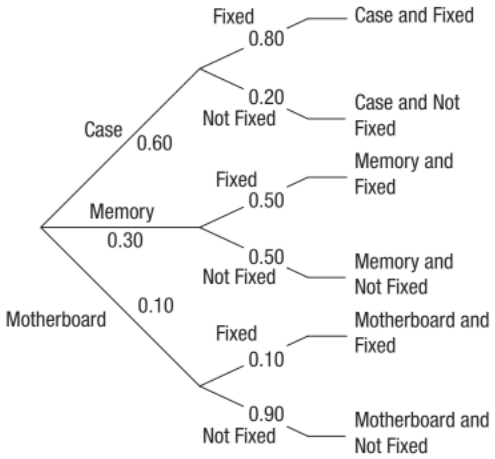
Podemos colocar essas probabilidades no primeiro conjunto de ramificações, como mostra a imagem abaixo.



Observe que cobrimos todas as possibilidades e, portanto, as probabilidades somam um. Para este diagrama, podemos agora adicionar as probabilidades condicionais de que um ajuste menor conserte cada tipo de problema. Dados anteriores indicam que:

- P (correção | placa-mãe) = 0.10,
- P (correção | memória) = 0.50 e
- P (correção | alinhamento do case) = 0.80.

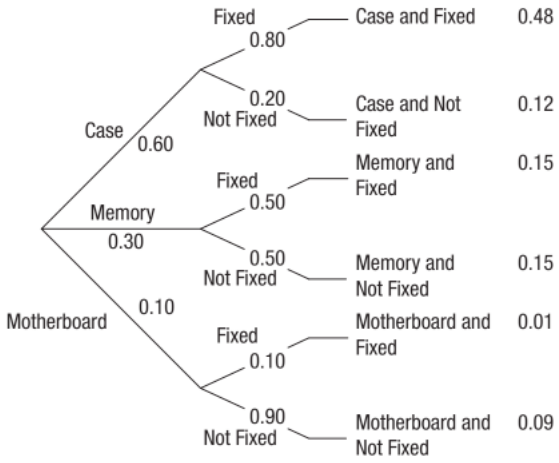
No final de cada ramo que representa o tipo de problema, desenhamos duas possibilidades (Fixed ou Not Fixed – Corrigido ou Não Corrigido) e escrevemos as probabilidades condicionais nas ramificações, como mostra a imagem abaixo.



No final de cada segundo ramo, escrevemos o evento conjunto correspondente à combinação dos dois ramos. Por exemplo, o ramo superior é a combinação do problema sendo o alinhamento do case, e o resultado do pequeno ajuste é que o problema agora está corrigido. Para cada um dos eventos conjuntos, podemos usar a Regra Geral de Multiplicação para calcular sua probabilidade conjunta. Por exemplo:

P (case e corrigido) = P (case) * P (corrigido | case) = 0.60 * 0.80 = 0.48

Escrevemos essa probabilidade ao lado do evento correspondente. Fazendo isso para todas as combinações de ramificações, é apresentada a imagem abaixo.



Todos os resultados na extrema direita são desarticulados, ou seja, eles não se sobrepõem, pois em cada nó, todas as opções são alternativas desarticuladas. E essas alternativas são todas as possibilidades, então as probabilidades na extrema direita devem se somar a um. Como os resultados finais são desarticulados, podemos adicionar qualquer combinação de probabilidades para encontrar probabilidades de eventos compostos.

Em particular, o engenheiro pode responder sua pergunta: qual é a probabilidade de o problema ser corrigido por um simples ajuste? Ele encontra todos os resultados na extrema direita em que o problema foi corrigido. Há três (um correspondente a cada tipo de problema) e ele soma suas probabilidades: 0,48 + 0,15 + 0,01 = 0,64. Portanto, 64% de todos os problemas são corrigidos pelo simples ajuste. Os outros 36% exigem uma investigação importante. Baseado em dados, agora o engenheiro pode tomar uma decisão.

Nesta aula, desenhamos nossas árvores de probabilidade da esquerda para a direita. Também podemos desenhar na vertical, de cima para baixo.

Referências:

Probability Theory: The Logic of Science: Principles and Elementary Applications Vol 1
E. T. Jaynes

* Aqui encerra a parte de Probabilidade deste capítulo. Passaremos agora à Estatística Inferencial.