

Como escolher a métrica certa para seu problema

A avaliação de modelos de aprendizado supervisionado não é apenas um detalhe técnico; trata-se de uma escolha que influencia diretamente a interpretação dos resultados e, sobretudo, as decisões que serão tomadas a partir deles. Uma mesma rede neural, árvore de decisão ou regressão logística pode parecer adequada ou ineficaz dependendo da métrica escolhida. Por isso, a definição da métrica correta é um passo fundamental no ciclo de aprendizado de máquina.

O primeiro aspecto a considerar é o entendimento do problema e dos dados. É necessário compreender qual é o objetivo final da aplicação, quais são os custos associados a erros e se o conjunto de dados apresenta ou não equilíbrio entre as classes. Essas questões fornecem a base para decidir quais indicadores são mais relevantes.

Em cenários em que as classes são relativamente balanceadas, como no caso da classificação de flores no famoso dataset Iris, a acurácia costuma ser suficiente, pois mede de forma direta a proporção de acertos globais. No entanto, em situações desbalanceadas, como na detecção de fraudes em que apenas uma pequena fração das transações corresponde a casos fraudulentos, a acurácia se torna enganosa. Nesses casos, métricas como precisão e revocação ganham destaque. A precisão mede a proporção de acertos entre as previsões positivas e é essencial quando o custo de um falso positivo é alto, como ao acusar injustamente um cliente de fraude. Já a revocação mede a proporção de casos positivos efetivamente detectados e torna-se central quando o custo de um falso negativo é elevado, como no diagnóstico de uma doença grave que não pode ser ignorada. O F1-score, por sua vez, funciona como um equilíbrio entre precisão e revocação, reunindo ambas em uma média harmônica, sendo bastante útil quando se deseja balancear os dois aspectos.

Quando o modelo retorna probabilidades, métricas como a curva ROC e a área sob a curva (AUC) oferecem uma visão mais ampla da capacidade de separação entre classes em diferentes limiares de decisão. Esses indicadores são relevantes em contextos em que os limiares podem ser ajustados conforme a política ou a necessidade da aplicação, como ocorre em sistemas de apoio a diagnósticos médicos. Embora a ênfase desta aula seja a classificação, vale lembrar que em problemas de regressão as métricas mais comuns incluem o erro quadrático médio, o erro absoluto médio e o coeficiente de determinação R^2 , cuja escolha depende da sensibilidade desejada em relação a erros grandes ou pequenos.

O ponto essencial é que não existe uma métrica universalmente superior. Cada escolha implica em trade-offs que devem ser avaliados de acordo com o impacto no negócio. Em marketing, por exemplo, pode ser aceitável lidar com falsos positivos ao enviar promoções para clientes que não precisariam delas, desde que isso aumente a capacidade de recuperar consumidores em risco de cancelamento. Na segurança bancária, por outro lado, pode ser preferível maximizar a precisão para reduzir o bloqueio indevido de clientes legítimos. Já na medicina, a prioridade recai quase sempre sobre a revocação, pois deixar de detectar um caso crítico pode ter consequências irreversíveis.

Assim, a seleção da métrica envolve uma sequência lógica: compreender os objetivos, analisar os dados, identificar os erros mais custosos, avaliar indicadores adequados e, por fim, conectar a escolha ao contexto da aplicação. Em muitos casos, é recomendável observar múltiplas métricas em conjunto, construindo uma visão mais completa do desempenho do modelo.

Portanto, escolher a métrica certa significa medir aquilo que realmente importa. A decisão deve ser técnica e, ao mesmo tempo, estratégica, de modo que o cientista de dados consiga justificar a métrica adotada, demonstrar sua adequação ao problema e reconhecer suas limitações.