EDITAL 436 DE 07 DE JUNHO DE 2022 CONCURSO PÚBLICO PARA DOCENTE - VAGA RT 004 PROGRAMA DE ENGENHARIA DE SISTEMAS E COMPUTAÇÃO SETORIZAÇÃO: ENGENHARIA E CIÊNCIA COMPUTACIONAL	Nº 13	Folha: 1

Questão 13)

O algoritmo CART ("Classification and Regression Tree") consiste em particionar o espaço de entrada, definido pelos atributos ("features") da tarefa, em diferentes regiões. Para isso, o algoritmo busca, para cada atributo, um valor de divisão ("split-point"), o qual divide os dados de entrada em duas regiões. O objetivo do algoritmo é encontrar partições idealmente "puras", no sentido de agrupar apenas dados de entrada que sejam similares sob algum aspecto.

Uma das principais vantagens do algoritmo CART é que o modelo de regressão ou classificação treinado consiste em uma árvore de decisões em que é fácil de interpretar todas as etapas que levaram o modelo a retornar um determinado valor numérico ou rótulo de classificação. Os nós terminais da árvore, chamados de "nós folha", são os que representam a saída do modelo. Os demais nós intermediários representam atributos que foram escolhidos, em cada etapa de decisão, para particionar o espaço, e os ramos representam os critérios de divisão, baseado nos valores de divisão selecionados para um determinado atributo em uma etapa de decisão. O processo de decisão começa do nó raiz e vai seguindo por ramos dessa árvore, até se alcançar um nó folha, que apontará a decisão final da árvore.

Para avaliar qual é a melhor partição do espaço de entrada a ser feita, é preciso definir algum critério que quantifique a "pureza" das partições que serão obtidas. Esse critério depende do tipo de problema (regressão ou classificação) sendo tratado. Para problemas

de regressão, é comum utilizar o erro quadrático médio entre o valor predito pelo modelo e o valor-alvo real da amostra em questão. O valor predito pelo modelo corresponde à média dos valores-alvo reais de cada amostra presentes em uma partição. Dessa forma, a árvore de regressão buscará por atributos e pontos de divisão desses atributos que levem à maior redução no erro quadrático médio. Para problemas de classificação, o critério de análise da "pureza" das partições se baseia no conceito de ganho de informação. A quantidade de informação presente em uma partição pode ser calculada de diferentes formas. Duas funções bastante utilizadas para esse propósito são a entropia-cruzada e o coeficiente de Gini. O algoritmo buscará por divisões do espaço de entrada que maximizem o ganho de informação, definido como a quantidade de informação presente originalmente (antes da divisão) subtraída da quantidade de informação presente em cada uma das partições obtidas após a divisão. Partições "puras" são aquelas que apresentem amostras de uma única classe. Se um nó folha não for puro, a saída da árvore de decisão será a classe da maioria das amostras presentes no nó folha "impuro".

Uma grande limitação das árvores de decisão é que elas são muito suscetíveis a "overfitting" (ou seja, ficarem altamente ajustadas ao conjunto de dados sendo avaliado). Além disso, são suscetíveis a "outliers" no conjunto de dados, o que tende a piorar seu desempenho. Essas questões ocorrem, porque a árvore pode crescer tanto quanto houver dados no conjunto, se adequando, no caso extremo, a cada amostra de entrada presente nos dados (ou seja, as partições ficam reduzidas a um único elemento). Consequentemente, a fronteira de decisão no espaço de entrada fica altamente descontínua, pouco suave (muito "quebrada"). Nesses casos, a generalização da árvore fica comprometida. Uma forma de evitar esse problema consiste em limitar a profundidade máxima da árvore. Outra solução consiste em realizar técnicas de "poda", as quais colapsam ramos da árvore, de modo a reduzir o modelo final. Uma outra abordagem consiste no treino de múltiplas árvores de decisão e utilizar a média das saídas de cada árvore (regressão) ou a classe mais votada dentre todas as árvores (classificação) como saída final.

EDITAL 436 DE 07 DE JUNHO DE 2022 CONCURSO PÚBLICO PARA DOCENTE - VAGA RT 004 PROGRAMA DE ENGENHARIA DE SISTEMAS E COMPUTAÇÃO SETORIZAÇÃO: ENGENHARIA E CIÊNCIA COMPUTACIONAL	Nº 13	Folha: 2

Questão (6.3)

Em análise multivariada, busca-se encontrar relações entre os atributos que definem o espaço de entrada, de modo a obter uma descrição estatística dos dados de entrada. A grande importância é que a caracterização da distribuição dos dados de entrada permite o entendimento e uma análise mais profunda do problema em questão. Conhecendo-se a distribuição dos dados pode-se, por exemplo, analisar casos pouco frequentes (pontos de "outliers") e como o modelo de aprendizado de máquina reage a eles. Além disso, novos dados podem ser gerados sintéticamente, o que contribui para um aumento da base de dados existente, a qual é usada para treinar modelos. Conhecendo-se a distribuição dos dados, também é possível obter as distribuições condicionadas dos dados (isto é, a distribuição de classe ou de dados numéricos, dada uma determinada entrada, para os casos de classificação e regressão, respectivamente). Com tais distribuições, é possível prever valores de saída com mais certeza e menos chances de erro.

Um modelo multivariado bastante conhecido é a mistura de gaussianas. Nele, assume-se que a distribuição de entrada pode ser modelada como uma combinação linear de múltiplas gaussianas, cada uma com um vetor de médias e matriz de correlações próprios. O objetivo deste modelo consiste em descobrir quais são as gaussianas que compõem a mistura (isto é, quais as médias e matriz de correlações associados a cada uma) e qual a contribuição de cada uma

delas na distribuição de dados sendo considerada. Matematicamente:

$$p(\underline{x}) = \sum_{K=1}^{K} \pi_K \mathcal{N}(\underline{x}, \mu_K, \Sigma_K)$$

O coeficiente π_K define a probabilidade a priori de $\underline{x}$ pertencer à K-ésima gaussiana. O processo de otimização é feito maximizando o negativo do logaritmo da função de verossimilhança em função de cada parâmetro (μ, Σ e π).

Outro modelo que visa obter a caracterização estatística dos dados através de análise multivariada é o classificador bayesiano. Nele, busca-se modelar a probabilidade a posteriori de uma classe, dado um vetor de entrada. O classificador rotula a amostra como sendo da classe que apresenta a maior ~~distribuição a~~ probabilidade a posteriori para a entrada em questão. Um modelo simples do classificador bayesiano é o classificador bayesiano "ingênuo" ("naive" Bayes), no qual assume-se que os atributos que compõem o vetor de entrada são provenientes de distribuições gaussianas idênticas e independentes. Sendo assim, o modelo busca encontrar os parâmetros do modelo gaussiano que maximizam o negativo do logaritmo da função de verossimilhança.

O conceito de mistura de modelos pode ser estendido para outros modelos tradicionais, como regressores lineares e não-lineares. Nestes casos, busca-se encontrar a distribuição dos alvos condicionados a entrada observada e os modelos regressores considerados. A premissa é similar ao caso de misturas gaussianas, no sentido de que tenta-se encontrar, além dos parâmetros ótimos de cada função de regressão, a contribuição de cada ~~regressor~~ modelo de regressão àquela distribuição sendo modelada.

Questão 8.3)

A tarefa de regressão consiste em encontrar uma função que mapeia dados de entrada em um espaço N-dimensional em um vetor de valores reais $\underline{t}$. Como os vetores de entrada $\underline{x}$ são conhecidos, a regressão busca encontrar os coeficientes que escalarão os atributos que definem o vetor de entrada. A função de regressão, portanto, depende destes coeficientes (chamados de parâmetros ou pesos) e dos vetores de entrada.

A definição de linearidade ou não-linearidade da regressão leva em conta a relação entre a função de regressão e os pesos que a compõem, e não relação entre a função e os vetores de entrada. Dessa forma a regressão é dita linear se a função define uma combinação linear entre seus parâmetros. Os coeficientes desta combinação são obtidos a partir dos atributos que definem o vetor de entrada, podendo ser os próprios valores dos atributos (caso em que o regressor linear também é linear em relação ao vetor de entrada) ou transformações não-lineares dos mesmos (caso em que o regressor linear é não-linear em relação ao vetor de entrada). Um exemplo deste último caso é um polinômio de ordem n: $f(\underline{w}, x) = w_0 + w_1 x + w_2 x^2 + \dots + w_n x^n$. Ele é linear em relação aos pesos $w_0, w_1, \dots w_n$, mas não-linear em relação à entrada x.

A regressão é dita não-linear se a relação da função com seus parâmetros for não-linear. Um exemplo deste caso é a regressão logística: $f(w_0, w_1, x) = \frac{1}{1+e^{-(w_0+w_1 x)}}$. (Caso em que a entrada é escalar). A regressão logística, embora tenha esse nome, é usada para tarefas de classificação.

O caso mais simples de regressão é quando a função é linear nos parâmetros e no vetor de entrada. Neste caso, ela é escrita como: $f(\theta, x) = \theta_0 + \sum_{i=1}^{M-1} \theta_i x_i$, onde

"M-1" é a dimensão do vetor de entrada. Os coeficientes $\theta_0, \dots \theta_{M-1}$ são obtidos através de um processo de otimização, que requer o uso de uma função de avaliação (chamada de função-custo) que diz o quão bem os pesos encontrados para a função regressora aproximam os valores-alvo das amostras de entrada. No caso da regressão linear, o erro quadrático (ou sua variante, o erro quadrático médio) é utilizado. A ideia é encontrar o valor "ótimo" dos parâmetros, no sentido de minizar a função-custo. Para o caso da regressão logística, uma função-custo muito utilizada é a entropia cruzada.

O valor "ótimo" dos pesos é obtido derivando a função-custo e a igualando a 0. Na regressão linear, assumindo que o erro de predição tem uma distribuição gaussiana de média nula e desvio-padrão unitário, a minimização do erro quadrático médio leva às chamadas equações normais, as quais têm como solução o valor "ótimo" dos pesos. Matricialmente, a equação normal é escrita como:

$$\theta^* = (\Phi^T\Phi)^{-1}\Phi^T \cdot Y \text{, onde } \Phi = \begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) & \cdots & \phi_{M-1}(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{M-1}(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \cdots & \phi_{M-1}(x_n) \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \text{ e } \theta^* = \begin{bmatrix} \theta_0^* \\ \theta_1^* \\ \vdots \\ \theta_{M-1}^* \end{bmatrix}$$

e $\phi_K(x)$ é a K-ésima função-base que atua sobre o K-ésimo atributo do vetor de entrada $\underline{x} = [x_0\ x_1\ x_2 \cdots x_K \cdots x_{M-1}]$ $(x_0 = 1)$.

A regressão é simples quando a variável predita é um único escalar. A regressão é múltipla quando existem múltiplos valores sendo preditos para uma entrada (ou seja, o alvo é um vetor, e não um escalar). Entretanto, a regressão múltipla pode ser tratada como múltiplas regressões simples independentes, em se tratando de regressão linear. Portanto, o procedimento para se encontrar os pesos ótimos vale tanto para regressões lineares simples quanto para regressões lineares múltiplas.

Para se definir a complexidade de modelos de regressão, deve-se levar em conta a quantidade de dados disponíveis. Modelos altamente complexos para poucos dados disponíveis sofrem com uma alta variância e baixo poder de generalização ("overfitting"). Modelos pouco complexos sofrem com alto "bias" e não conseguem prever corretamente os valores-alvo dos dados de entrada ("underfitting"). Para evitar esses casos, técnicas de regularização são usadas, visando encontrar uma boa complexidade para o modelo, como "weight decay" (l_2-norm).