Revisão da prova

Marcos Vinicius - T17 - G1

Dissertativa 1

Perguntas

1. O que é acurácia em um modelo de classificação? Explique o que significa uma acurácia de 100% no conjunto de treinamento.
2. Qual a diferença entre a ácuracia no conjunto de treinamento e a acurácia no conjunto de teste?
3. Qual é o problema que pode ocorrer ao obter uma acurácia de 100% no conjunto de treinamento com Naive Bayes? Explique como você chegou a essa conclusão e quais características do modelo podem indicar esse problema.
4. Como suposição de independência das características no Naive Bayes pode influenciar o desempenho do modelo? O que pode acontecer se essas suposição não for verdadeira, especialmente no contexto de análise de spam?

Respostas

1. Acurácia é a porcentagem de instâncias classificadas corretamente pelo modelo. Uma acurácia de 100% no conjuto de treinamento, indica que o modelo aprendeu exatamente como funcionam os dados deste conjunto.
2. A acurácia de 100% no conjunto de teste prevê que o modelo se adaptou perfeitamente os dados daquele conjunto, portanto, para ele será mais difícil interpetar dados novos, como os do conjunto de teste.
3. Como mencionado antes, devido a alta adaptação aos dados já conhecidos, o modelo tem dificuldade em interpretar novos dados, por isso, encontramos aqui um caso de Overfitting. É possível de perceber isso graças a acuráci perfeita nos dados de treinamento e os valores distoantes entre treino e teste
4. O Naive Bayes é um algoritmo de classificação, que trata cada coluna como independente para classificar os dados. É um algoritmo muito bom quando se trata de performance, consegue aprender rápido e até mesmo com uma baixa quantidade de dados. É um algoritmo utilizado muito comumente com análise de spam, por sua habilidade de aprender com poucos dados e conseguir trabalhar bem com PLN (Processamento de Lingugagem Natural).

Dissertativa 2

Resposta O SHAP é um modelo que ajuda a identificar o valor de contribuição de cada atributo para uma determinada previsão, é especialmente útil quando se trata de um algoritmo de Random Forest, pois é um algoritmo considerado "caixa preta", onde só enviamos dados e recebemos resultados, sem entender

muito do processo. Com o SHAP, podemos análisar todas as predições realizadas e entender exatamente quais são os valores que mais afetaram a decisão do modelo. Isso pode ajudar a entender melhor qual sintoma do paciente está mais relacionado ao problema, tendo assim uma explicabilidade maior para auxiliar quem esstá atendendo o paciente a ter uma analise mais aprofundada do problema.

Dissertativa 3

Perguntas

1. O que é acurácia em um modelo de classificação? Explique o que significa acurácia de 100% no conjunto de treinamento.
2. Qual a diferença entre a acurácia no conjunto de treinamento e a acurácia no conjunto de teste? Como essa diferença pode afetar a capacidade do modelo de generalizar para novos dados?
3. Qual problema está ocorrendo o usar um K muito baixo? Explique como você chegou a essa conclusão e quais características do modelo indicam esse problema.
4. Como a escolha da métrica de distância pode influenciar o desempenho do modelo KNN?

Respostas

1. Acurácia é o valou que indica a quantidade de classificações corretas em relação a quantidade de predições realizadas. Uma acurácia de 100% significaria que o modelo aprendeu perfeitamente como classificar os dados nos quais ele treinou.

2. A acurácia no conjunto de treinamento constuma ser maior do que a no conjunto de teste, afinal, o modelo já sabe o que ele deve prever nos dados que ele aprendeu, já os dados de teste funcionam como um novo mundo para o modelo, com várias diferenças daquilo que ele já conhece, então ele tentará prever com o que ele já sabe a partir de similaridade