



**ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS
(BIG DATA, MACHINE LEARNING E APLICAÇÕES)
DISCIPLINA: PROBABILIDADE**

PROJETO AVALIATIVO

Apresentação

Nesse projeto, nós construiremos um classificador Naive Bayes, usando Python. Para isso, realizaremos três tarefas:

Tarefa 1

O objetivo dessa tarefa é entendermos como funciona um classificador Naive Bayes. Para isso, implementaremos ele “no cru”, sem usar bibliotecas com funções pré-implementadas, usando como referência o tutorial disponível nesse [link](#), e na seção 18.3 do capítulo 18 do livro *Probability for Machine Learning*, disponibilizado no github do professor.

Execute os tutoriais, entendendo o funcionamento de cada parte envolvida e explicando, com suas palavras, cada passo implementado e os testes de funcionamento do classificador, considerando os exemplos dos dois tutoriais (como cada tutorial traz um exemplo diferente, execute as duas implementações, para os dois exemplos, para fins de comparação).

Tarefa 2

Nessa tarefa, nós veremos como funciona o classificador implementado na biblioteca `scikit-learning`, a principal biblioteca de Python para aprendizagem de máquina.

Para isso, estude o exemplo da página 154/155 do livro *Probability for Machine Learning (Listing 18.13)*, executando ele para os dois exemplos trabalhados acima.

Tarefa 3

Agora vamos fazer um teste de desempenho aplicado, dos três algoritmos: os dois implementados na tarefa 1, e o exemplo executado na tarefa 2.

Para isso, usaremos para treinamento do classificador os dados usados no exemplo do link da tarefa 1, que usa informações de peso (em libras; para converter kilogramas em libras, use essa [calculadora](#)), altura (em pés; ; para converter de centímetros para pés, use essa [calculadora](#)) e numeração do pé (em polegadas, de acordo com a numeração americana, que pode ser comparada com a nacional [aqui](#)) para classificar classificação de gênero.

O conjunto de dados que será usado na classificação será montado com os alunos presentes na aula. Para isso faremos uma breve pesquisa usando um [Google Forms](#), e esses dados serão testados e avaliados pelo nossos classificadores.

Iremos computar a taxa de erros e acertos para verificar a eficácia das implementações.

DATA DE ENTREGA: 30 de outubro de 2019

OBSERVAÇÕES

- o trabalho é individual;
- somente serão considerados os resultados apresentados em notebooks jupyter;
- o projeto é o principal componente da nota, portanto, deixar de fazê-lo implica em perder, literalmente, a disciplina;
- obviamente, é **mandatório** que todas as implementações feitas sejam discutidas, com explicações claras e diretas sobre como foi feito cada passo constante em seu trabalho. Projetos que contenham apenas o código serão considerados incompletos, assim como projetos com comentários iguais, **gramatica ou semanticamente**, serão considerados **PLÁGIO**. E plágio é crime!