

## MPEI 2024/25 - PL 5

# Classificador Naïve Bayes

**Palavras chave:** Regra de Bayes, independência, classificador Naïve Bayes, Naïve Bayes binário, classificação de texto, análise de sentimento, *Machine Learning*, *Bag of Words* e filtragem de SPAM.

Recomenda-se a leitura do capítulo “Naïve Bayes and Sentiment Classification” disponível em <https://web.stanford.edu/~jurafsky/slp3/4.pdf>

Responda às seguintes questões em Matlab:

1. Considere a informação relativa a um exemplo simples de *sentiment analysis* – retirado do capítulo *Naïve Bayes and Sentiment Classification* referido acima – apresentada na Tabela:

|        | Categoria | “Documentos”                          |
|--------|-----------|---------------------------------------|
| Treino | -         | just plain boring                     |
|        | -         | entirely predictable and lacks energy |
|        | -         | no surprises and very few laughs      |
|        | +         | very powerful                         |
|        | +         | the most fun film of the summer       |
| Teste  | ?         | predictable with no fun               |

- (a) Comece por criar um vetor com a lista de palavras únicas no conjunto de treino. Deve ser criado de forma automática, numa função, tendo por entrada um *cell array* com o conteúdo dos documentos.
- (b) Crie uma matriz com informação sobre o número de ocorrências de cada uma das palavras em cada um dos documentos. As linhas devem representar o documento e as colunas a palavra.
- (c) Crie um vetor com informação da categoria correspondente a cada linha da matriz da alínea anterior. Sugestão: usar '+' e '-' para representar as categorias.
- (d) Com base nas variáveis que criou anteriormente, calcule, considerando *Laplace (add 1) smoothing*:
  - $P(-)$  e  $P(+)$
  - $P(\text{“predictable”} | +)$ ,  $P(\text{“no”} | +)$ ,  $P(\text{“fun”} | +)$
  - $P(\text{“predictable”} | -)$ ,  $P(\text{“no”} | -)$ ,  $P(\text{“fun”} | -)$
  - Com base nos valores obtidos anteriormente, determine qual a categoria mais provável para a frase de teste? Sugestão: não considere a palavra “with”.

2. Estamos interessados em avaliar o Classificador Naïve Bayes Binário na **deteção de SPAM**. Como habitual neste problema, consideremos que o sistema se baseia na ocorrência ou não de um conjunto de palavras relevantes. As palavras e a sua ocorrência ou não num conjunto de 20 emails encontram-se na Tabela seguinte (disponível em formato .CSV em <http://bit.ly/4eW4Fgy>).

| Email | Offer | Click | Free | Money | Meeting | Project | Win | Urgent | Prize | Class |
|-------|-------|-------|------|-------|---------|---------|-----|--------|-------|-------|
| 1     | 1     | 1     | 0    | 1     | 0       | 0       | 1   | 1      | 1     | SPAM  |
| 2     | 0     | 0     | 0    | 0     | 1       | 1       | 0   | 0      | 0     | OK    |
| 3     | 1     | 0     | 1    | 1     | 0       | 0       | 1   | 0      | 0     | SPAM  |
| 4     | 0     | 0     | 0    | 0     | 1       | 0       | 0   | 0      | 0     | OK    |
| 5     | 1     | 1     | 0    | 1     | 0       | 0       | 1   | 1      | 1     | SPAM  |
| 6     | 0     | 0     | 0    | 0     | 1       | 1       | 0   | 0      | 0     | OK    |
| 7     | 1     | 0     | 1    | 0     | 0       | 0       | 0   | 0      | 0     | SPAM  |
| 8     | 0     | 0     | 0    | 0     | 1       | 1       | 0   | 0      | 0     | OK    |
| 9     | 1     | 1     | 1    | 1     | 0       | 0       | 1   | 1      | 1     | SPAM  |
| 10    | 0     | 0     | 0    | 0     | 1       | 1       | 0   | 0      | 0     | OK    |
| 11    | 1     | 0     | 0    | 1     | 0       | 0       | 0   | 1      | 0     | SPAM  |
| 12    | 0     | 0     | 0    | 0     | 1       | 1       | 0   | 0      | 0     | OK    |
| 13    | 1     | 1     | 1    | 0     | 0       | 0       | 1   | 1      | 1     | SPAM  |
| 14    | 0     | 0     | 0    | 0     | 1       | 1       | 0   | 0      | 0     | OK    |
| 15    | 1     | 1     | 0    | 1     | 0       | 0       | 1   | 1      | 1     | SPAM  |
| 16    | 0     | 0     | 0    | 0     | 1       | 1       | 0   | 0      | 0     | OK    |
| 17    | 1     | 0     | 1    | 1     | 0       | 0       | 1   | 0      | 1     | SPAM  |
| 18    | 0     | 0     | 0    | 0     | 1       | 1       | 0   | 0      | 0     | OK    |
| 19    | 1     | 1     | 0    | 1     | 0       | 0       | 1   | 1      | 0     | SPAM  |
| 20    | 0     | 0     | 0    | 0     | 1       | 1       | 0   | 0      | 0     | OK    |

- (a) Comece por ler a informação do ficheiro .CSV.

Sugestão 1: Experimente usar `readcell()` e `cell2mat()`.

Sugestão 2: Adote uma matriz e um vetor para guardar a informação, adaptando o exemplo a seguir:

```

1 % Features / Características
2 % Colunas = [offer, click, free, money, meeting, project]
3 X = [1 1 0 1 0 0;
4       0 0 0 0 1 1;
5       1 0 1 1 0 0;
6       0 0 0 0 1 0];
7
8 Y = categorical({'spam'; 'ok'; 'spam'; 'ok'}); % categorias

```

- (b) Divida os dados da tabela em 2 conjuntos de forma aleatória: um de treino, com 70 % das linhas; outro de teste, com as restantes linhas. Sugestão: use `randperm()`.
- (c) Calcule as probabilidades *a priori* das duas classes (SPAM e OK). Apenas pode utilizar os emails do conjunto de treino.
- (d) Obtenha todas as probabilidades necessárias para a definição de um classificador Naïve Bayes binário. Mais uma vez, só pode usar os emails dos conjunto de treino.
- (e) Estime a classe de cada um dos emails do seu conjunto de teste.
- (f) Considerando que a última coluna da tabela contém a verdadeira classe do email, determine a Precisão, Recall e  $F_1$  do seu classificador (no conjunto de teste).
- (g) (TPC) Execute o processo várias vezes (por exemplo 10 vezes), guardando os valores de Precisão e Recall. No final calcule a média dos valores obtidos.
- (h) (TPC) Repita o processo anterior para com o junto de treino sendo apenas 50% das linhas. O que aconteceu ao desempenho.

3. Consideremos agora um exemplo simples para prever o **interesse de um filme** utilizando um classificador Naïve Bayes e características baseadas em atributos do filme e avaliações dos espetadores.

Pode-se criar um classificador simples para prever se os espetadores terão interesse num filme com base no género, duração e algumas avaliações, nomeadamente:

- Se é um filme de Ação (1 para Ação, 0 para outros géneros);
- Se é um filme do género Comédia (1 para Comédia, 0 para outros géneros);
- Duração: Duração do filme (0 para menos de 2 horas; 1 para durações de 2 horas ou mais).
- Número de avaliações positivas de utilizadores (0 para até 5 avaliações positivas; 1 para mais de 5 avaliações positivas).

Considerando os seguintes dados de treino:

```

1 %% Dados de Treino:
2
3 % Características
4 % [Ação?, Comédia?, Duração, Num Avaliações positivas]
5 X = [1 0 1 1;
6      0 1 0 0;
7      0 1 0 1;
8      1 0 1 1;
9      1 1 0 1;
10     0 1 0 0];
11
12 classes = categorical({'Interessa', 'Não Interessa', 'Interessa', ...
13                      'Interessa', 'Interessa', 'Não Interessa'});

```

Usando Matlab, qual a classe mais provável para:

- Filme representado por [1, 0, 1, 1] ?
- Filme representado por [0, 1, 0, 0] ?
- Filme de Ação e Comédia, com 100 minutos e 4 comentários positivos?