Aprendizagem Probabilística – Classificação Usando Naive Bayes

ESTAT0016 – Tópicos Especiais em Estatística (Introdução à Apredizagem de Máquina)

Prof. Dr. Sadraque E.F. Lucena

Introdução

- O algoritmo Naive Bayes, também conhecido como Bayes ingênuo, faz parte da família dos *métodos bayesianos*.
- Ele utiliza dados de treino para calcular a probabilidade de um resultado pertencer a uma classe, com base nos dados de entrada.
- Classificadores bayesianos podem ser usados para:
  - Classificação de texto. Exemplo: detecção de e-mail de spam.
  - Segurança de Rede. Exemplo: detecção de intrusos em uma rede de computadores.
  - Diagnóstico médico. Exemplo: identificação de condições médicas com base em sintomas observados.
- Classificadores bayesianos são ideais para problemas que requerem a consideração simultânea de vários atributos.
  - Mesmo atributos com efeitos relativamente pequenos podem ter um impacto significativo quando combinados em um modelo bayesiano.

Base: Teorema de Bayes

- O teorema de Bayes é uma ferramenta que nos permite atualizar probabilidades de eventos com base em novas evidências.
- Suponha que conhecemos a probabilidade de uma instância pertencer à classe C, denotada por $P(C)$.
- Ao obtermos uma informação adicional sobre um atributo A dessa instância, podemos ajustar a probabilidade de C levando em consideração a informação do atributo A:

$$P(C|A) = \frac{P(C \cap A)}{P(A)} = \frac{P(A|C)P(C)}{P(A)}. \qquad (1)$$

- A última parte da Equação 1 vem do resultado:

$$P(A|C) = \frac{P(C \cap A)}{P(C)} \Rightarrow P(C \cap A) = P(A|C)P(C).$$

Base: Teorema de Bayes

Terminologia:

- $P(C)$: probabilidade ***a priori*** (*prior probability*).
  - Conhecimento inicial sobre C.
  - Ou seja, conhecimento sobre a aparição de exemplos da classe C no problema.
- $P(C|A)$: probabilidade ***a posteriori*** (*posterior probability*).
  - Probabilidade da instância pertencer à classe C após consideração do atributo A.
- $P(A|C)$: verossimilhança (*likelihood*).
- $P(A)$: probabilidade marginal (*marginal likelihood*).

- O uso do teorema de Bayes permite então recalcular a probabilidade de uma instância pertencer a uma classe após a obtenção de novas evidências sobre essa instância.

Exemplo 7.1

Suponha que queremos calcular a probabilidade de um email ser classificado como *spam* quando a palavra “viagra” está presente. Calcule a probabilidade a partir dos dados apresentados na tabela abaixo.

	viagra		
Frequência	**Sim**	**Não**	**Total**
spam	4	16	20
não *spam*	1	79	80
Total	5	95	100

Exemplo 7.1

$$
\begin{aligned}
P(spam|viagra) &= \frac{P(spam \cap \text{viagra})}{P(\text{viagra})} \\
&= \frac{P(\text{viagra}|spam)P(spam)}{P(\text{viagra})} \\
&= \frac{\frac{4}{20}\ \frac{20}{100}}{\frac{5}{100}} \\
&= \frac{4}{5} \\
&= 0,80
\end{aligned}
$$

O classificador Naive Bayes

- Esse método é chamado de “ingênuo” (*naive*) devido à sua suposição simplificada de que todos os atributos são igualmente importantes e independentes, o que geralmente não é verdadeiro na prática.
  - Na realidade, os atributos raramente são independentes entre si.
- Vamos considerar um exemplo prático, como classificar emails. Na prática, a importância dos atributos pode variar.
  - Por exemplo, o remetente do email pode ter mais peso do que o próprio conteúdo.
  - Algumas palavras não são independentes; por exemplo, se encontrarmos a palavra “viagra”, isso pode sugerir a presença de outras palavras como “prescrição” ou “remédio”.
- **Observação:** Apesar de as suposições do método Naive Bayes não refletirem completamente a complexidade da realidade, o método ainda demonstra um desempenho razoável em muitos casos práticos.

O classificador Naive Bayes

- Vamos considerar agora que temos mais informações sobre uma intância (mais atributos coletados) para calcular a probabilidade *a posteriori* de que ela pertença à classe C.
- Por simplicidade, vamos considerar apenas três atributos A_1, A_2 e A_3. Então

$$P(C|A_1 \cap A_2 \cap A_3) = \frac{P(A_1 \cap A_2 \cap A_3|C)P(C)}{P(A_1 \cap A_2 \cap A_2)}. \qquad (2)$$

- Essa fórmula é computacionalmente muito difícil de resolver.
  - À medida que são adicionados mais atributos, são necessárias quantidades enormes de memória para armazenar as probabilidades de todas as intereseções entre os eventos possíveis.

O classificador Naive Bayes

- Para contornar esse problema, o método Naive Bayes assume independência entre os atributos condicionados à mesma classe. Isto leva ao resultado:

$$P(A_1 \cap A_2 \cap A_3 | C) = P(A_1 | C) P(A_2 | C) P(A_3 | C).$$

Assim, a Equação 2 fica

$$P(C | A_1 \cap A_2 \cap A_3) = \frac{P(A_1 | C) P(A_2 | C) P(A_3 | C)\, P(C)}{P(A_1 \cap A_2 \cap A_3)}.$$

O classificador Naive Bayes

O algoritmo de classificação Naive Bayes pode ser definido da seguinte forma:

- Sejam $A_1, A_2, \ldots, A_n$ os valores dos n atributos de uma instância. A probabilidade de classificar essa instância como pertencente à classe C usando naive Bayes é

$$P(C|A_1, \ldots, A_n) = \frac{P(C) \prod_{i=1}^{n} P(A_i|C)}{P(A_1, \ldots, A_n)}$$

em que:

  - $P(C)$ é a probabilidade de uma instância pertencer à classe C;
  - $P(A_i|C)$ é a probabilidade de ser observado A_i dada a ocorrência da classe C;
  - $P(A_1, \ldots, A_n)$ é a probabilidade conjunta de $A_1, \ldots, A_n$.

O classificador Naive Bayes

- Suponha que temos duas classes C_1 e C_2 para classificar uma intânccia considerando três atributos, A_1, A_2 e A_3.
- Então temos:

$$P(C_1|A_1 \cap A_2 \cap A_3) = \frac{P(A_1|C_1)P(A_2|C_1)P(A_3|C_1)P(C_1)}{P(A_1, A_2, A_3)}$$

e

$$P(C_2|A_1 \cap A_2 \cap A_3) = \frac{P(A_1|C_2)P(A_2|C_2)P(A_3|C_2)P(C_2)}{P(A_1, A_2, A_3)}$$

- Na prática, como o denominador para ambas as equações é o mesmo, os ignoramos para simplificar nossos cálculos e nos concentramos apenas nos numeradores.

O classificador Naive Bayes

- Então a probabilidade de uma instância ser classificada na classe C_1 é a probabilidade de ser C_1 dividida pela probabilidade de que seja tanto C_1 quanto C_2.
- Vejamos um eexmplo para melhor compreensão.

Exemplo 7.2

Vamos expandir nosso filtro de spam do Exemplo 7.1, adicionando mais termos a serem monitorados: “viagra”, “dinheiro”, “compra” e “descadastrar”.

Verossimilhança	**viagra** (W_1)		**dinheiro** (W_2)		**compra** (W_3)		**descadastrar** (W_4)		**Total**
	Sim	**Não**	**Sim**	**Não**	**Sim**	**Não**	**Sim**	**Não**	
spam	4/20	16/20	10/20	10/20	0/20	20/20	12/20	8/20	20
não *spam*	1/80	79/80	14/80	66/80	8/80	72/80	23/80	57/80	80
Total	5/100	95/100	24/100	76/100	8/100	91/100	35/100	65/100	100

Um novo email é recebido e possui os termos “viagra” e “descadastrar”. Calcule a probabilidade do email ser *spam* e não *spam* usando naive Bayes.

Exemplo 7.2

- Probabilidade inicial para *spam*:

$$P(W_1|spam)P(\neg W_2|spam)P(\neg W_3|spam)P(W_4|spam)P(spam) = \\ = (4/20)(10/20)(20/20)(12/20)(20/100) \\ = 0{,}012$$

- Probabilidade inicial para não *spam*:

$$P(W_1|\neg spam)P(\neg W_2|\neg spam)P(\neg W_3|\neg spam)P(W_4|\neg spam)P(\neg spam) = \\ = (1/80)(60/80)(71/80)(23/80)(80/100) \\ = 0{,}002$$

- As probabilidades finais para *spam* e não *spam* são, respectivamente:

$$0{,}012/(0{,}012 + 0{,}002) = 0,857 \qquad 0{,}002/(0{,}012 + 0{,}002) = 0{,}143$$

O estimador de Laplace

- Suponha agora que recebemos um email contendo os quatro termos do Exemplo 7.2: “viagra”, “dinheiro”, “compra” e “descadastrar”.
- Usando naive Bayes, a probabilidade de ser *spam* é

$$(4/20)(10/20)(0/20)(12/20)(20/100) = 0$$

- Enquanto a de não *spam* é

$$(1/80)(14/80)(8/80)(23/80)(80/100) = 0{,}00005$$

- Logo, a probabilidade final para *spam* e não *spam* ficam, respectivamente:

$$0/(0 + 0{,}00005) = 0 \qquad 0{,}00005/(0 + 0{,}00005) = 1$$

O estimador de Laplace

- O resultados indicam probabilidade zero de ser *spam* e 100% de não ser *spam*, mas isso parece improvável.
- Isso ocorreu porque nos dados de treino não foi observada a palavra “compra” em emails de *spam*.
- A solução é usar o estimador de Laplace, que adiciona pequenos valores às contagens para evitar probabilidades nulas e melhorar a precisão da classificação.
  - Em geral somamos 1 às frequências de cada atributo para evitar frequência zero.
  - Então, ao invés de calcularmos a probabilidade de um atributo A_i como a_i/N, calculamos

$$(a_i + 1)/(N + 1 \times d),$$

em que d é o número de atributos totais considerados.

Exemplo 7.3

Calcule as probabilidades do email ser *spam* ou não usando o estimador de Laplace.

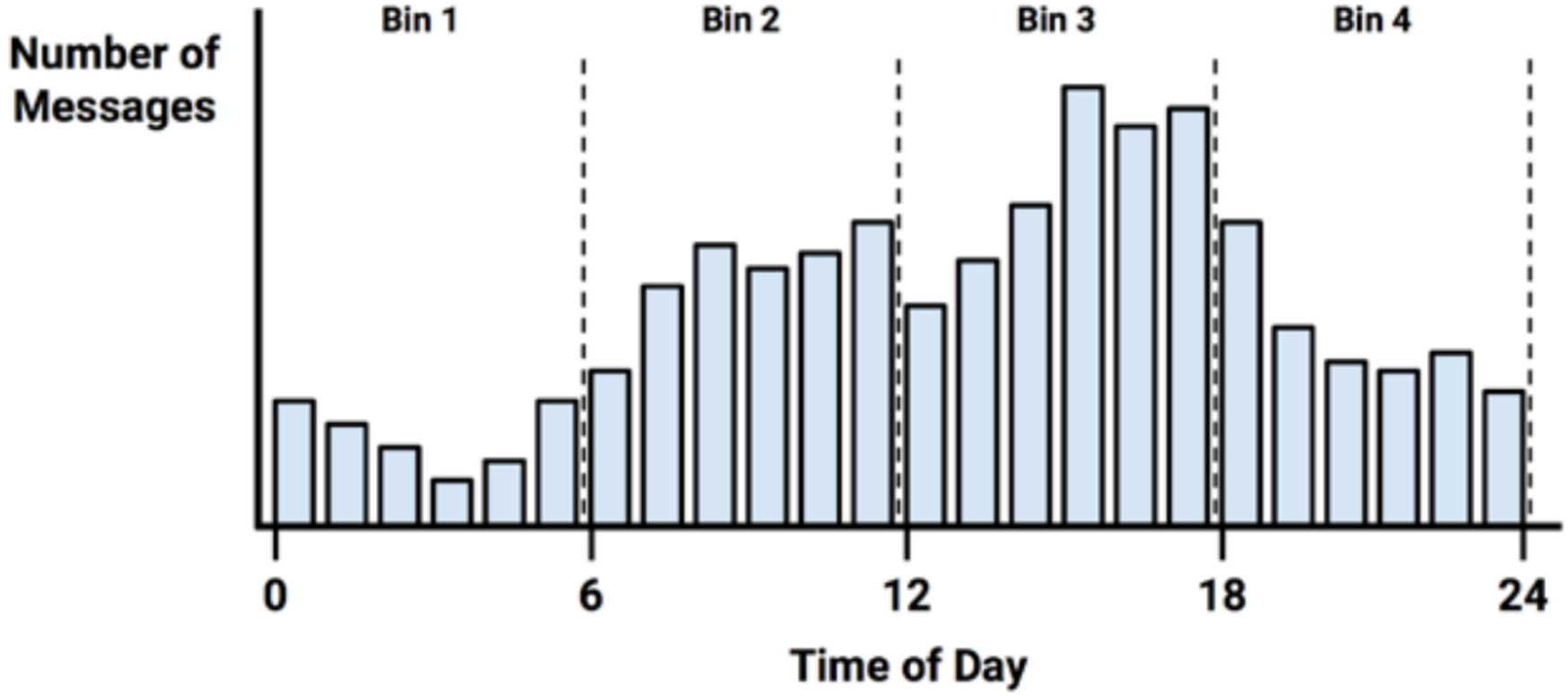
Exemplo 7.3

- Probabilidade inicial de *spam*: $(5/24)(11/24)(1/24)(13/24)(20/100) = 0{,}0004$
- Probabilidade inicial de não *spam*: $(2/84)(15/84)(9/84)(24/84)(80/100) = 0{,}0001$

- Probabilidade final de *spam*: $0{,}0004/(0{,}0004 + 0{,}0001) = 0{,}80$
- Probabilidade final de não *spam*: $0{,}0001/(0{,}0004 + 0{,}0001) = 0{,}20$

Atributos numéricos

- O método Naive Bayes requer que os atributos sejam categóricos.
- Se um atributo for numérico, devemos discretizá-lo, criando categorias.
- Para isso, devermos explorar os dados.



Vantagens e desvantagens

Vantagens

- Simples, rápido e muito efetivo.
- Bom desempenhocom dados ruidosos e ausentes.
- Necessidade de poucos exemplos para o treinamento, mas também funciona bem com um grande número de exemplos.
- Fácil de obter a probabilidade estimada para uma previsão.

Desvantagens

- Baseia-se frequentemente em uma suposição falha de características igualmente importantes e independentes.
- Não é ideal para conjuntos de dados com muitas características numéricas.
- As probabilidades estimadas são menos confiáveis do que as classes previstas.

FIM