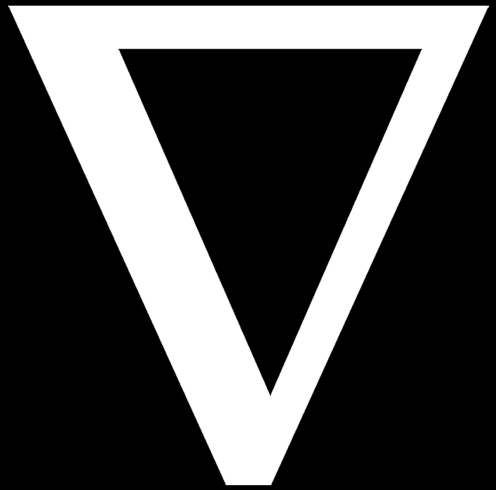


Classificador Naive Bayes

João Marcos Cardoso da Silva
@joaomarcoscsilva

Probabilidades

Variável Aleatória

Uma variável aleatória pode assumir valores diferentes cada vez que for medida, seguindo uma distribuição de probabilidades específica.

Espaço Amostral

O Espaço Amostral de uma variável aleatória é o conjunto de valores que ela pode assumir.

Variável Aleatória Discreta

Uma variável aleatória é **discreta** se o seu espaço amostral tem uma quantidade finita de possibilidades.

Exemplos:

- Um dado de 6 lados (possui 6 possibilidades)
- Ter ou não ter uma doença (possui 2 possibilidades)
- Uma letra aleatória (possui 26 possibilidades)
- Uma palavra aleatória (possui alguns milhares de possibilidades)

Distribuição de Probabilidades Discreta

A **distribuição de probabilidades discreta** é uma função que determina a probabilidade de cada possibilidade do espaço amostral ocorrer.

Escrevemos P(X) como a distribuição de probabilidades da variável X e P(X = x_i) como a probabilidade de X assumir a opção x_i.

Por exemplo, se D é um dado, P(D = 1) = 1/6

Propriedades de Distribuições Discretas

- A soma de todas as possibilidades é 1:

$$\sum P(X = x_i) = 1$$

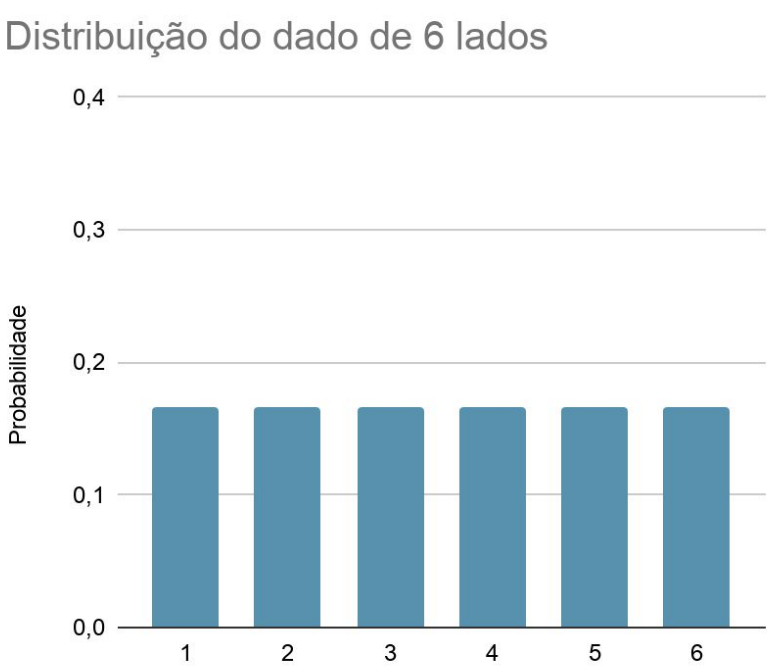
- Nenhuma possibilidade pode ter probabilidade negativa:

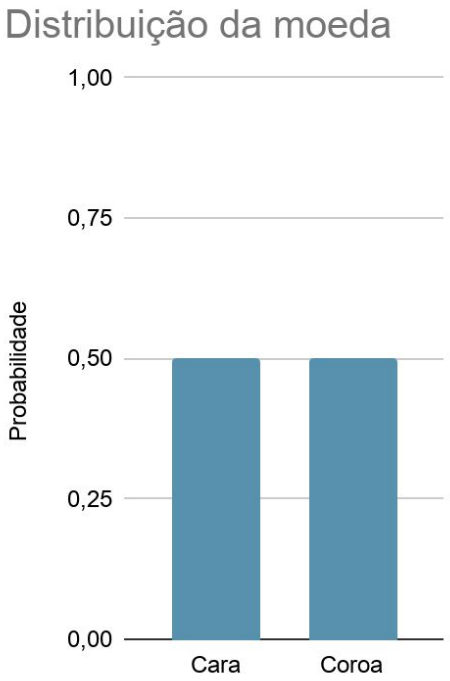
$$P(X = x_i) \geq 0$$

- Nenhuma possibilidade pode ter probabilidade maior que 1:

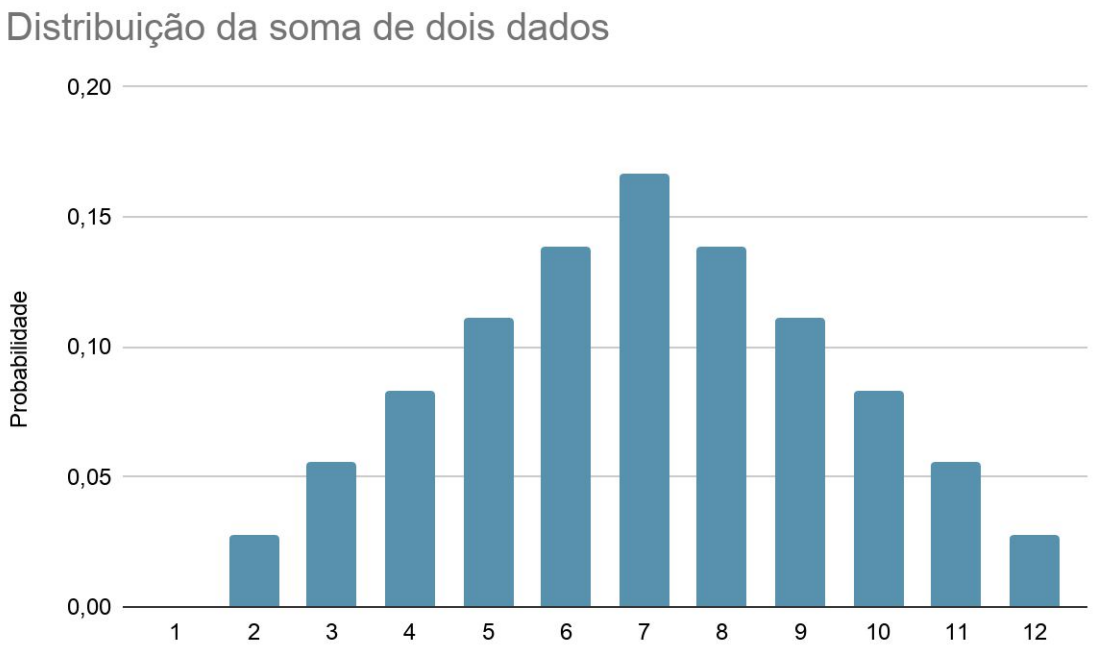
$$P(X = x_i) \leq 1$$

Exemplos gráficos





Exemplos gráficos



Distribuição Conjunta

Nós podemos definir a probabilidade de dois eventos aleatórios ocorrerem ao mesmo tempo com uma distribuição conjunta P(X,Y).

Exemplos:

- A probabilidade de dois dados terem o resultado 6
- A probabilidade de uma pessoa ter uma doença e também apresentar seus sintomas

Distribuição Condicional

Se já sabemos que um evento X ocorreu, isso pode (ou não) mudar a probabilidade de um outro evento Y.

Definimos a probabilidade nova como P(Y|X) (probabilidade de Y dado X).

Ex: P(Ter a doença | Apresentou os sintomas)

Fatoração da Distribuição Conjunta

Nós podemos fatorar a distribuição conjunta P(X,Y):

- P(X,Y) = P(X)P(Y|X) (a chance de ocorrer X, e depois Y)
- P(X,Y) = P(Y)P(X|Y) (a chance de ocorrer Y, e depois X)

Probabilidade Marginal

Às vezes, nós conhecemos P(X,Y) mas queremos encontrar P(X) (ou P(Y)).

Podemos pensar em P(X) como a probabilidade esperada de X dado **qualquer** valor de Y, então basta usar a equação:

$$P(X) = \sum_y P(X,Y) = \sum_y P(X|Y)P(Y)$$

Isso se chama marginalização e P(X) recebe o nome de probabilidade marginal.

Variáveis Independentes

Se duas variáveis são independentes, saber uma não nos diz nada sobre a probabilidade da outra, então a distribuição condicional não muda:

$$P(Y|X) = P(Y)$$

Variáveis Independentes

Se X e Y forem independentes, podemos fatorar a distribuição conjunta:

$P(X,Y) = P(X)P(Y)$

Por exemplo, como cada lançamento de um dado é independente, podemos encontrar a probabilidade de obtermos dois resultados 6:

$P(\text{dois resultados 6}) = P(\text{um resultado 6})*P(\text{um resultado 6}) = \frac{1}{6} * \frac{1}{6} = 1/36$

dúvidas?

Invertendo uma probabilidade condicional

Em geral, probabilidades condicionais não são simétricas: P(A|B) != P(B|A)

Mas em alguns casos nós sabemos P(A|B), mas estamos interessados em P(B|A).

Invertendo uma probabilidade condicional

Em um exame médico, normalmente sabemos a probabilidade de uma pessoa ter um resultado positivo, dado que ela tem a doença: P(resultado positivo|doença). Isso é a taxa de acerto do exame e normalmente é bem alta.

Mas o médico está realmente interessado na probabilidade inversa: ele tem um paciente que teve um teste positivo, e quer saber qual a probabilidade de o paciente ter a doença: P(doença|resultado positivo)

Fatoração da Distribuição Conjunta

Nós podemos fatorar a distribuição conjunta P(X,Y) de duas maneiras:

- P(X,Y) = P(X)P(Y|X) (a chance de ocorrer X, e depois Y)
- P(X,Y) = P(Y)P(X|Y) (a chance de ocorrer Y, e depois X)

Teorema de Bayes

Podemos usar essas duas fatorações para encontrar uma maneira de “inverter” uma probabilidade condicional:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

dúvidas?

Naive Bayes

Classificação

Vamos definir o problema de classificação:

- Para cada entrada **x**, predizer a classe **y** em que aquele valor provavelmente pertence

Por exemplo, em um detector de spam:

- **x** é o texto do email
- **y** indica se o email é ou não é spam

Distribuições de Probabilidades

Para treinar o classificador nós temos um dataset de **N** pares (**x**,**y**) de exemplos resolvidos (por exemplo, pares de (email, spam ou não spam)).

Todos esses pares vêm de uma mesma distribuição de probabilidades, a distribuição dos dados, que indica quais pares (**x**, **y**) são mais prováveis de aparecerem no dataset:

$(\mathbf{x},\mathbf{y}) \sim P(\mathbf{x},\mathbf{y})$

Fatorando a distribuição dos dados

Nós podemos fatorar a distribuição em dois termos:

$P(\mathbf{x},\mathbf{y}) = P(\mathbf{x})\, P(\mathbf{y}|\mathbf{x})$

- $P(\mathbf{x})$ é a distribuição dos dados entrada (por exemplo, quais emails são mais prováveis)
- $P(\mathbf{y}|\mathbf{x})$ é a distribuição condicional da saída, dada a entrada (por exemplo, se é provável ou não um determinado email $\mathbf{x}$ ser spam).

Aprendizado com as Probabilidades

Se nós soubéssemos exatamente o valor dessas distribuições, todos os problemas estariam resolvidos:

Com P(**y**|**x**), nós já temos exatamente a probabilidade de cada valor da saída **y** para entrada **x**!

Só precisamos então ver qual valor é mais provável para criar o melhor classificador possível.

Aproximação das distribuições

Infelizmente, nós não sabemos essas distribuições. Em vez disso, nós precisamos aproximá-las de alguma forma (**quase todos** os classificadores estão de algum modo aproximando essa distribuição para fazer uma predição).

O método mais direto para aproximar uma distribuição é contar quantas vezes uma coisa acontece no dataset. Vamos fazer isso em um problema simplificado:

Detector de Spam com uma única palavra

Vamos criar uma versão simplificada de um detector de spam: em vez de receber como entrada o texto inteiro, ele só recebe uma única informação: se a palavra “free” apareceu no texto.

Caso “free” apareça:

No dataset que usaremos, a palavra “free” aparece em 265 mensagens, das quais 199 foram marcadas como spam. Assim:

P(**y** = spam | **x** = “free” apareceu) = 199/265 = 75%

Logo, apenas por ter a palavra “free” uma mensagem qualquer já tem 75% de chance de ser spam.

Caso “free” não apareça

A palavra “free” não aparece em 5307 outras mensagens. das quais apenas 548 eram spam. Assim:

P(**y** = spam | **x** = “free” não apareceu) =548/5307 = 10%

Então uma mensagem que não tem a palavra “free” possui apenas 10% de chance de ser spam.

Expandindo a capacidade do modelo

Nós também podemos considerar sequências de palavras. Por exemplo, para a frase “for free”:

P(**y** = spam | **x** = “for free” apareceu) 16/17 = 94%

P(**y** = spam | **x** = “for free” não apareceu) 731/5555 = 13%

Um problema...

No caso anterior, a expressão “for free” só apareceu 17 vezes no total! Então o número que nós encontramos (94%) não é tão confiável quanto no caso anterior, onde “free” tinha aparecido mais de 200 vezes.

Se nós quisermos usar expressões com ainda mais palavras ou ainda o texto inteiro, esse problema fica cada vez pior: nunca teremos dados suficientes para ter uma estimativa confiável das probabilidades.

A maldição da dimensionalidade

Suponha um dicionário de 10000 palavras diferentes. Existem:

- 10^8 sequências de duas palavras distintas
- 10^{12} sequências de três palavras distintas
- 10^{16} sequências de quatro palavras distintas
- 10^{4n} sequências de n palavras distintas

O número de possibilidades cresce exponencialmente! Como podemos usar o texto inteiro?

A maldição da dimensionalidade: probabilidades

Vamos dividir a mensagem $\mathbf{x}$ nas palavras $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$.

Podemos então transformar a distribuição $P(\mathbf{x}|\mathbf{y})$ (invertemos a ordem, mas podemos usar o Teorema de Bayes depois) em uma distribuição conjunta de cada palavra:

$$P(\mathbf{x}|\mathbf{y}) = P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \mid \mathbf{y})$$

Se as palavras fossem independentes...

Se todas as variáveis $\mathbf{x}_i$ fossem independentes, poderíamos fatorar a distribuição:

$$P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \mid \mathbf{y}) = P(\mathbf{x}_1|\mathbf{y}) * P(\mathbf{x}_2|\mathbf{y}) * \ \ldots * \ P(\mathbf{x}_n|\mathbf{y})$$

Conseguimos calcular cada termo da direita com precisão já que eles só contém uma palavra, e com isso poderíamos calcular o valor da esquerda que considera o texto inteiro!

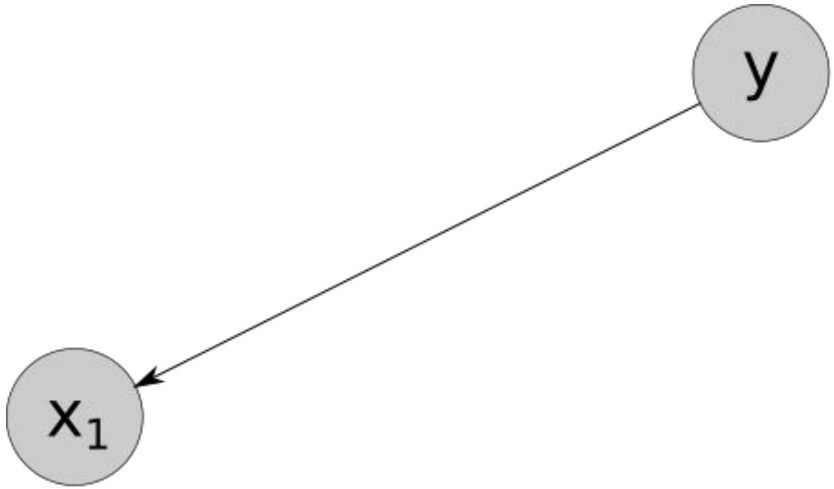
Mas qual é o problema em fazer isso?

Grafo Probabilistico das Dependências

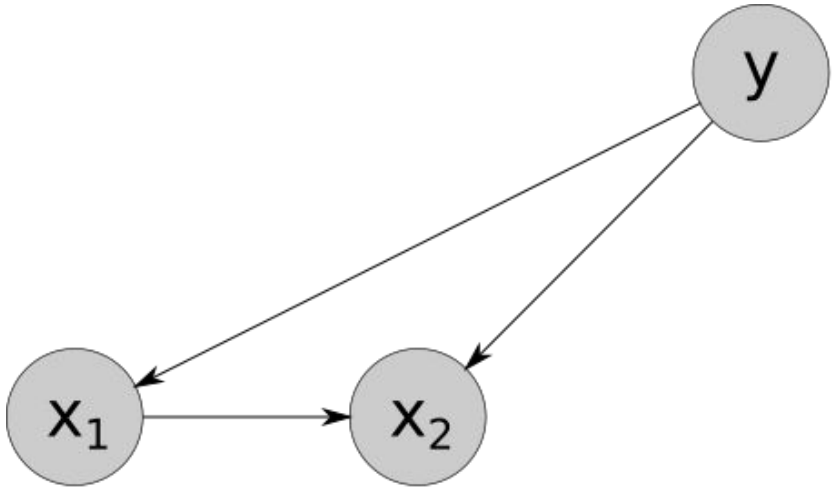
Vamos simular o processo de geração de um texto em um grafo probabilístico.

Dois nós serão conectados se um afeta a distribuição do outro.

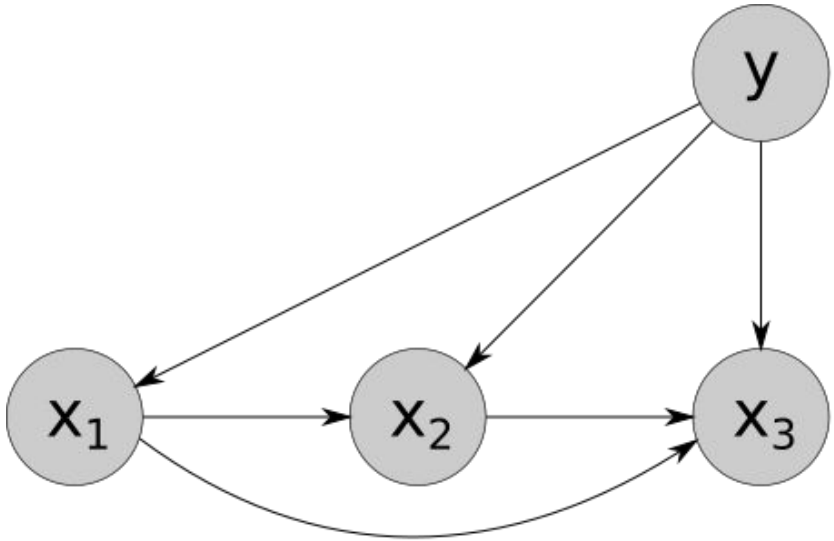
A primeira coisa decidida pelo autor é o conteúdo da mensagem, então já é definido é se a mensagem será legítima ou spam.



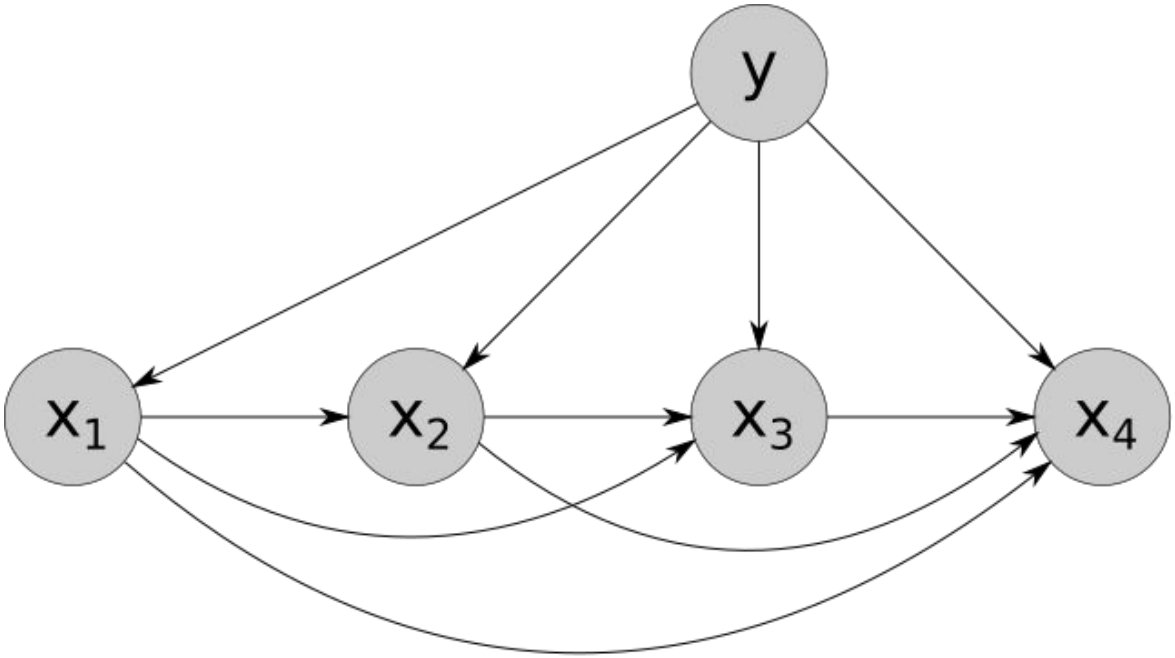
Logo depois, é escrita a primeira palavra. Ela depende da variável **y**, pois mensagens que são spam normalmente usam palavras diferentes.

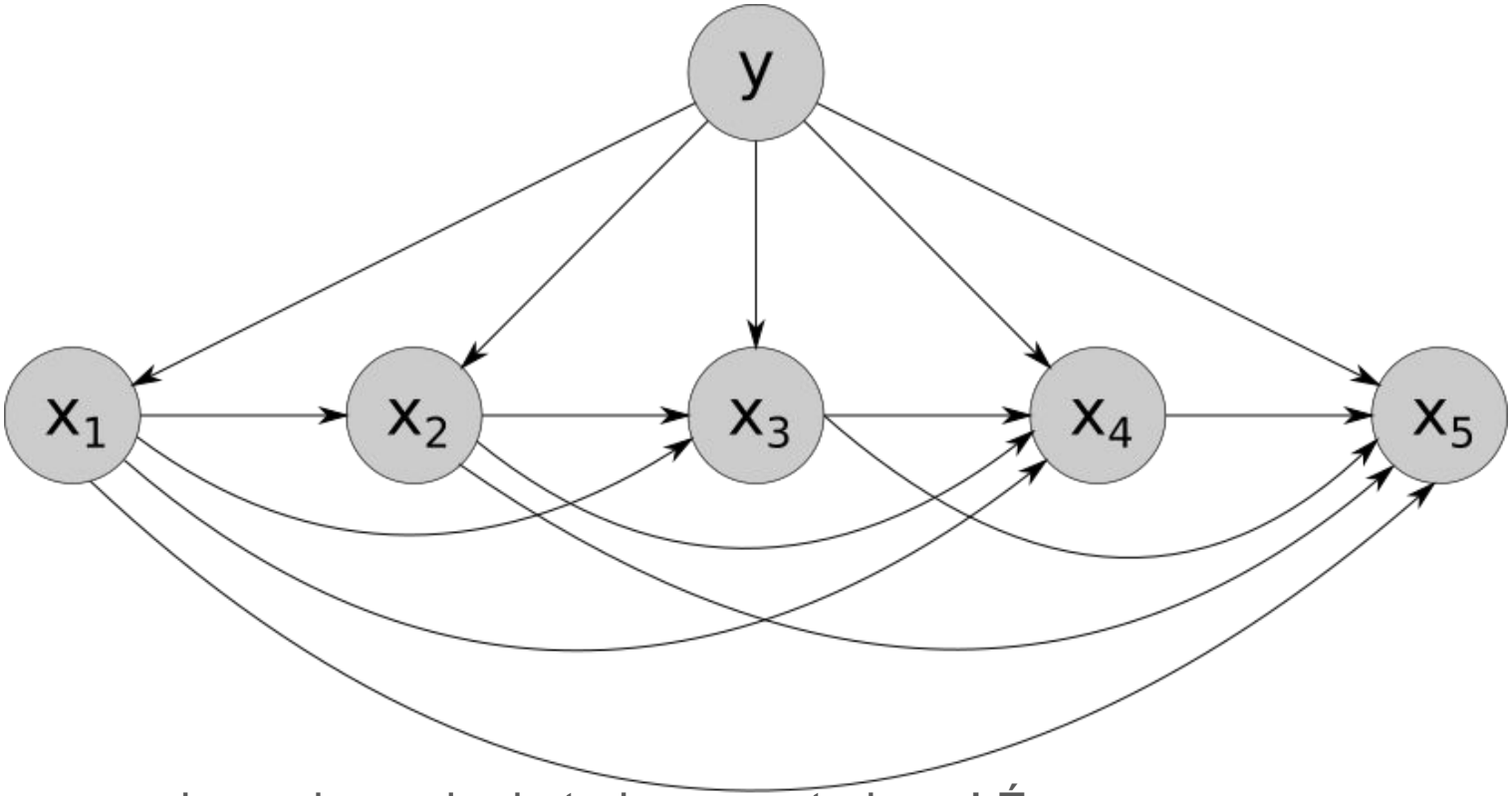


A segunda palavra depende tanto da palavra anterior quanto da classificação do texto.



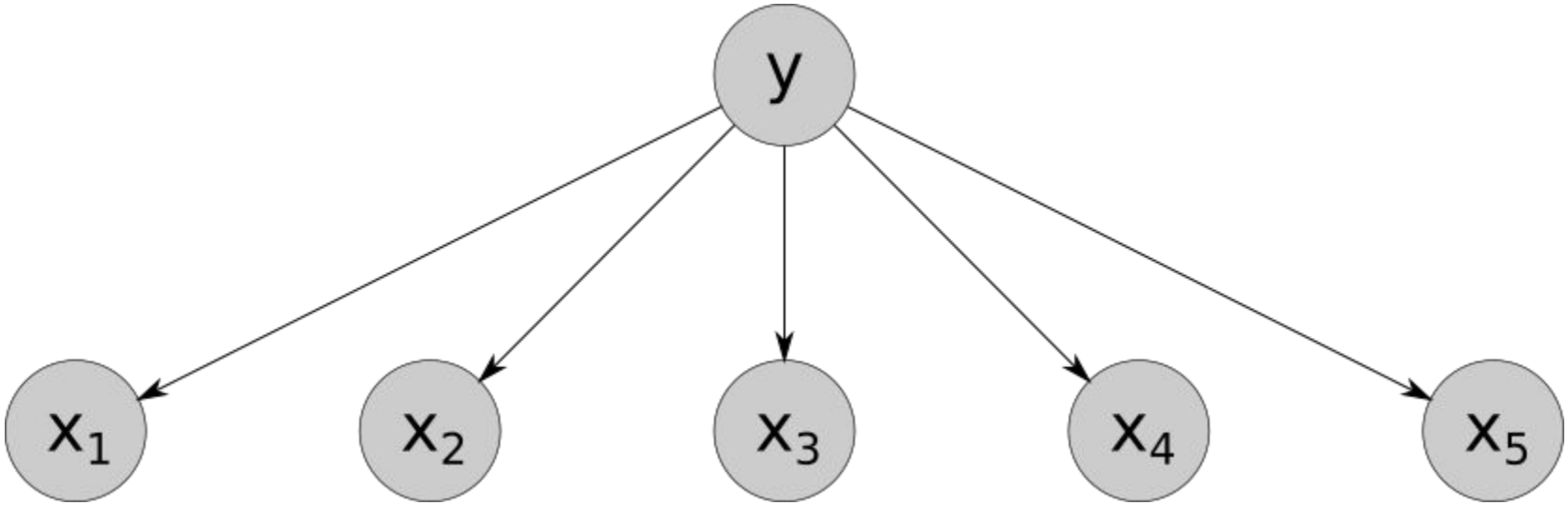
A terceira palavra depende de todas as palavras anteriores, além da classificação.


y
x1
x2
x3
x4



Cada nova palavra depende de todas as anteriores! É por isso que ocorre a explosão combinatória e a maldição da dimensionalidade.

Mas como seria o grafo se as palavras fossem independentes entre si, dada a classificação **y**?



O número de dependências fica muito menor! As palavras só dependem entre si através de **y**.

A suposição ingênua (Naive)

Para criar um classificador, nós iremos simplesmente fingir que as palavras são independentes entre si!

Nós podemos calcular para cada palavra a probabilidade $P(\mathbf{x}_i|\mathbf{y})$ simplesmente contando as ocorrências no dataset e podemos usar o texto inteiro com a fatoração:

$$P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \mid \mathbf{y}) = P(\mathbf{x}_1|\mathbf{y}) * P(\mathbf{x}_2|\mathbf{y}) * \ \ldots * \ P(\mathbf{x}_n|\mathbf{y})$$

Classificador Naive Bayes

Com a suposição “Naive”, podemos calcular P(**x** | **y**) simplesmente dividindo **x** em cada palavra e usando a fatoração.

Podemos então encontrar a probabilidade de classificação P(**y** | **x**) simplesmente usando o Teorema de Bayes!

$$P(\mathbf{y}|\mathbf{x}) = \frac{P(\mathbf{x}|\mathbf{y})P(\mathbf{y})}{P(\mathbf{x})}$$

O Classificador Ótimo

Se os atributos realmente fossem independentes, o classificador Naive Bayes é o melhor classificador possível!

Infelizmente, para a maior parte dos problemas os atributos **não** são independentes, como é o caso de textos. Isso leva a algumas consequências:

A Desvantagem da Suposição

Por causa da suposição, o classificador perdeu toda a informação sobre a ordem das palavras, já que nós retiramos do grafo as arestas que levavam da palavra anterior para a próxima. Por esse motivo, em NLP esse modelo também é chamado Bag-of-Words.

Apesar dessa desvantagem, esse tipo de classificador pode funcionar muito bem para algumas tarefas de classificação, como é o caso da detecção de spam.

dúvidas?

Variáveis Contínuas

Distribuições de Variáveis Contínuas

Variáveis aleatórias contínuas podem assumir infinitos valores diferentes (por exemplo, um valor real aleatório).

Apesar disso, a soma das probabilidades de cada possibilidade precisa ser igual a 1. Para isso, usaremos integrais.

Probabilidades em intervalos

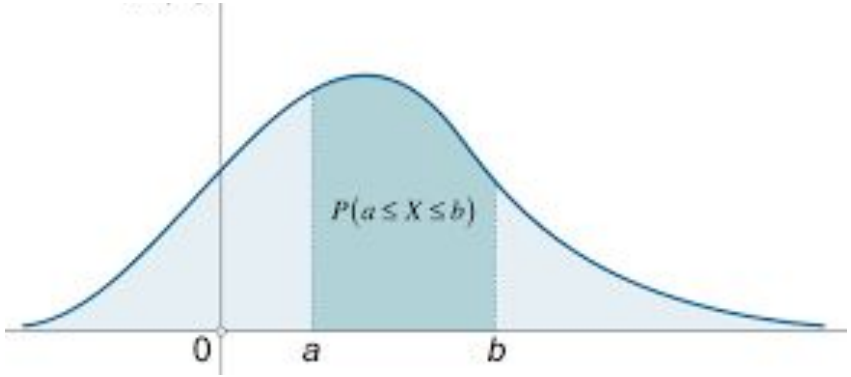
Em uma distribuição contínua, não faz sentido atribuir uma probabilidade a um valor específico, já que ela sempre vai ser infinitesimal. Em vez disso definimos probabilidades em intervalos.

Por exemplo: seja x um número aleatório em uma distribuição uniforme em [0,1]. Qual a probabilidade de x ser menor que 0.5?

Densidade de probabilidades

Vamos definir uma função p(x) que indicará a densidade das probabilidades próxima ao valor x.

Podemos encontrar a probabilidade em um intervalo integrando essa função:



Integral da densidade

Como a probabilidade total precisa ser um, a área total embaixo do gráfico precisa ser igual a 1. Temos então a seguinte propriedade, para qualquer distribuição p(x):

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

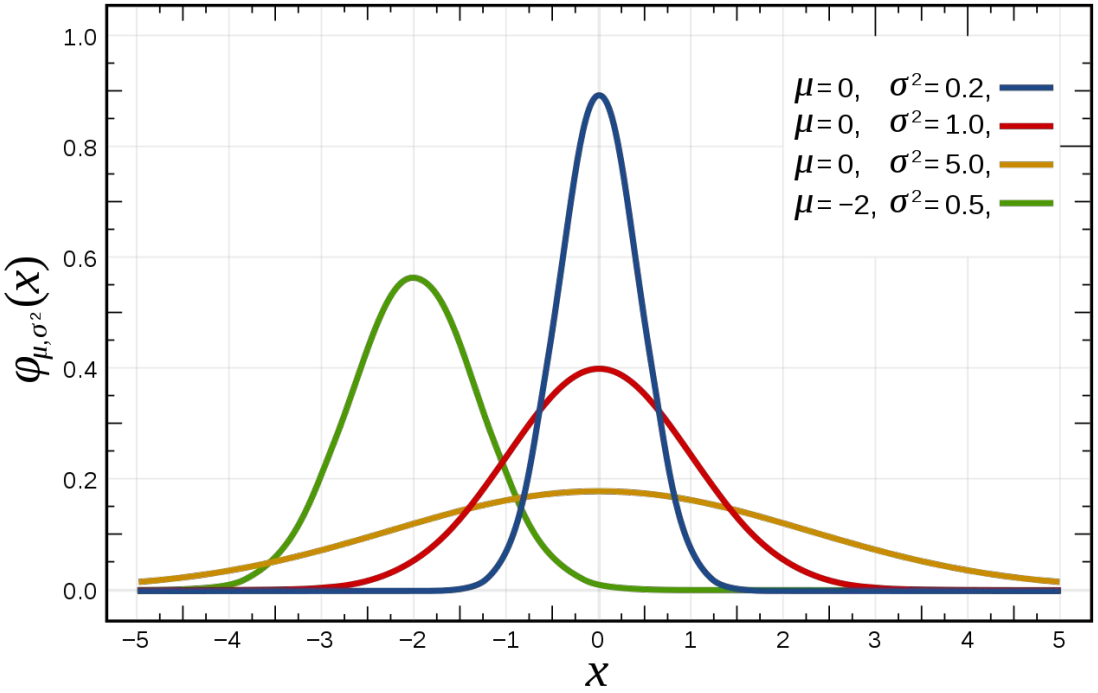
Distribuição Gaussiana

Uma distribuição contínua extremamente comum e a mais usada é a distribuição Gaussiana:

$$\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Onde temos dois parâmetros que podemos controlar para criar variações da distribuição: a média μ da distribuição e o desvio padrão σ.

Gráfico da Distribuição Gaussiana



Naive Bayes para variáveis contínuas

Para variáveis contínuas, não podemos aproximar suas distribuições simplesmente contando as ocorrências no dataset, já que existem infinitas possibilidades.

Em vez disso, podemos aproximar as distribuições $p(\mathbf{x_i}|\mathbf{y})$, se $\mathbf{x_i}$ é contínua, com uma Gaussiana. Para fazer isso, basta calcular a média e o desvio padrão.

Também podemos usar um estimador de densidade com kernels (KDE), que pode funcionar melhor quando os dados não são aproximadamente gaussianos.

dúvidas?