Uso de Reconhecimento de Padrões na Classificação de Documentos de Texto

Trabalho de Conclusão de Curso

André Dieb Martins

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Departamento de Engenharia Elétrica

13 de dezembro de 2011

Sumário

Introdução

- Reconhecimento de Padrões: área de estudo dos sistemas de classificação

Introdução

- Reconhecimento de Padrões: área de estudo dos sistemas de classificação
- Áreas relacionadas: Sistemas de Recuperação da Informação, Aprendizado de Máquina, Inteligência Artificial, Processos Estocásticos

Introdução

- Reconhecimento de Padrões: área de estudo dos sistemas de classificação
- Áreas relacionadas: Sistemas de Recuperação da Informação, Aprendizado de Máquina, Inteligência Artificial, Processos Estocásticos
- Tem como objetivo produzir uma categoria para um dado de entrada

Introdução

- Reconhecimento de Padrões: área de estudo dos sistemas de classificação
- Áreas relacionadas: Sistemas de Recuperação da Informação, Aprendizado de Máquina, Inteligência Artificial, Processos Estocásticos
- Tem como objetivo produzir uma categoria para um dado de entrada
- Histórico: 1960 (trabalhos teóricos) - Atual (implementação de sistemas "inteligentes" e aprendizado de máquina)

Introdução

- Reconhecimento de Padrões: área de estudo dos sistemas de classificação
- Áreas relacionadas: Sistemas de Recuperação da Informação, Aprendizado de Máquina, Inteligência Artificial, Processos Estocásticos
- Tem como objetivo produzir uma categoria para um dado de entrada
- Histórico: 1960 (trabalhos teóricos) - Atual (implementação de sistemas "inteligentes" e aprendizado de máquina)
- Aplicações: Scanners (OCR), Sistemas de Detecção de Face, Estudo de Descargas Parciais, Processamento Digital de Sinais pra medicina, dentre outras;

Objetivos

- Construir um sistema de classificação capaz de determinar a qual curso uma tese pertence

Objetivos

- Construir um sistema de classificação capaz de determinar a qual curso uma tese pertence
- Observar aspectos construtivos do classificador, assim como levantar possíveis melhorias aos algorítmos

Objetivos

- Construir um sistema de classificação capaz de determinar a qual curso uma tese pertence
- Observar aspectos construtivos do classificador, assim como levantar possíveis melhorias aos algorítmos
- Avaliar o classificador utilizando métricas bem estabelecidas (acurácia e complexidade computacional)

Objetivos

- Construir um sistema de classificação capaz de determinar a qual curso uma tese pertence
- Observar aspectos construtivos do classificador, assim como levantar possíveis melhorias aos algorítmos
- Avaliar o classificador utilizando métricas bem estabelecidas (acurácia e complexidade computacional)
- Levantar vantagens e desvantagens dos métodos utilizados

Sistemas de Classificação Artificial (SCA)

Problema de Classificação

Dado um conjunto de classes $c_i \in \mathcal{C}$ e um padrão de entrada p, deseja-se determinar a qual classe c_i o padrão p mais se assemelha.

Sistemas de Classificação Artificial (SCA)

Problema de Classificação

Dado um conjunto de classes $c_i \in \mathcal{C}$ e um padrão de entrada p, deseja-se determinar a qual classe c_i o padrão p mais se assemelha.

- Problema: classificar (ou categorizar) um padrão em classes de forma automática

Sistemas de Classificação Artificial (SCA)

Problema de Classificação

Dado um conjunto de classes $c_i \in \mathcal{C}$ e um padrão de entrada p, deseja-se determinar a qual classe c_i o padrão p mais se assemelha.

- Problema: classificar (ou categorizar) um padrão em classes de forma automática
- Solução:

Sistemas de Classificação Artificial (SCA)

Problema de Classificação

Dado um conjunto de classes $c_i \in \mathcal{C}$ e um padrão de entrada p, deseja-se determinar a qual classe c_i o padrão p mais se assemelha.

- Problema: classificar (ou categorizar) um padrão em classes de forma automática
- Solução:
  - Observar características de vários padrões conhecidos e pré-classificados.

Sistemas de Classificação Artificial (SCA)

Problema de Classificação

Dado um conjunto de classes $c_i \in \mathcal{C}$ e um padrão de entrada p, deseja-se determinar a qual classe c_i o padrão p mais se assemelha.

- Problema: classificar (ou categorizar) um padrão em classes de forma automática
- Solução:
  - Observar características de vários padrões conhecidos e pré-classificados.
  - Ao obter novos padrões, verificar as mesmas características e compará-las com as observações prévias

Projeto de Sistemas de Classificação

- Sensor (aquisição dos padrões)

Projeto de Sistemas de Classificação

- Sensor (aquisição dos padrões)
- Extração das Características e escolha do vetor de características

Projeto de Sistemas de Classificação

- Sensor (aquisição dos padrões)
- Extração das Características e escolha do vetor de características
- Projeto do classificador

Projeto de Sistemas de Classificação

- Sensor (aquisição dos padrões)
- Extração das Características e escolha do vetor de características
- Projeto do classificador
- Avaliação do sistema

Conceitos (Vetor de Características)

Vetor de Características

Dado um padrão p, o vetor de características $\mathbf{x}$ é um vetor composto por medições x_i sobre p, seguindo uma lista de características pré-selecionadas.

$$\mathbf{x} = (x_1, x_2, ..., x_k) \tag{1}$$

Conceitos (Vetor de Características)

Vetor de Características

Dado um padrão p, o vetor de características $\mathbf{x}$ é um vetor composto por medições x_i sobre p, seguindo uma lista de características pré-selecionadas.

$$\mathbf{x} = (x_1, x_2, ..., x_k) \tag{1}$$

- Representa unicamente um padrão

Conceitos (Vetor de Características)

Vetor de Características

Dado um padrão p, o vetor de características $\mathbf{x}$ é um vetor composto por medições x_i sobre p, seguindo uma lista de características pré-selecionadas.

$$\mathbf{x} = (x_1, x_2, ..., x_k) \tag{1}$$

- Representa unicamente um padrão
- Apresenta um conjunto reduzido de características de um padrão

Conceitos (Vetor de Características)

Vetor de Características

Dado um padrão p, o vetor de características $\mathbf{x}$ é um vetor composto por medições x_i sobre p, seguindo uma lista de características pré-selecionadas.

$$\mathbf{x} = (x_1, x_2, ..., x_k) \tag{1}$$

- Representa unicamente um padrão
- Apresenta um conjunto reduzido de características de um padrão
- As características observadas são escolhidas pelo projetista de maneira empírica

Conceitos (Vetor de Características)

Vetor de Características

Dado um padrão p, o vetor de características $\mathbf{x}$ é um vetor composto por medições x_i sobre p, seguindo uma lista de características pré-selecionadas.

$$\mathbf{x} = (x_1, x_2, ..., x_k) \tag{1}$$

- Representa unicamente um padrão
- Apresenta um conjunto reduzido de características de um padrão
- As características observadas são escolhidas pelo projetista de maneira empírica
- Comumente representado na forma de vetor $\mathbf{x}$ como mostrado acima

Exemplo

Exemplo: Problema de Classificação

Dado um objeto o_i, classificá-lo dentre duas classes: $Largo(L)$ ou $Comprido(C)$.

Exemplo

Exemplo: Problema de Classificação

Dado um objeto o_i, classificá-lo dentre duas classes: $Largo(L)$ ou $Comprido(C)$.

- Suponha um sensor que obtenha a largura (l_i) e comprimento c_i) do objeto

Exemplo

Exemplo: Problema de Classificação

Dado um objeto o_i, classificá-lo dentre duas classes: *Largo*(L) ou *Comprido*(C).

- Suponha um sensor que obtenha a largura (l_i) e comprimento c_i) do objeto
- Suponha um vetor de características definido por: $\mathbf{x} = (l_i, c_i)$

Exemplo

Exemplo: Problema de Classificação

Dado um objeto o_i, classificá-lo dentre duas classes: *Largo*(L) ou *Comprido*(C).

- Suponha um sensor que obtenha a largura (l_i) e comprimento c_i) do objeto
- Suponha um vetor de características definido por: $\mathbf{x} = (l_i, c_i)$
- Considere dois objetos adquiridos: $x_1 = (1, 5)$ e $x_2 = (3, 2)$

Exemplo

Exemplo: Problema de Classificação

Dado um objeto o_i, classificá-lo dentre duas classes: *Largo*(L) ou *Comprido*(C).

- Suponha um sensor que obtenha a largura (l_i) e comprimento c_i) do objeto
- Suponha um vetor de características definido por: $\mathbf{x} = (l_i, c_i)$
- Considere dois objetos adquiridos: $x_1 = (1, 5)$ e $x_2 = (3, 2)$
- Supondo uma regra: L se $l_i > c_i$ e C se $c_i > l_i$, definimos os limites de classificação, denominado *Linha Decisória*

Exemplo

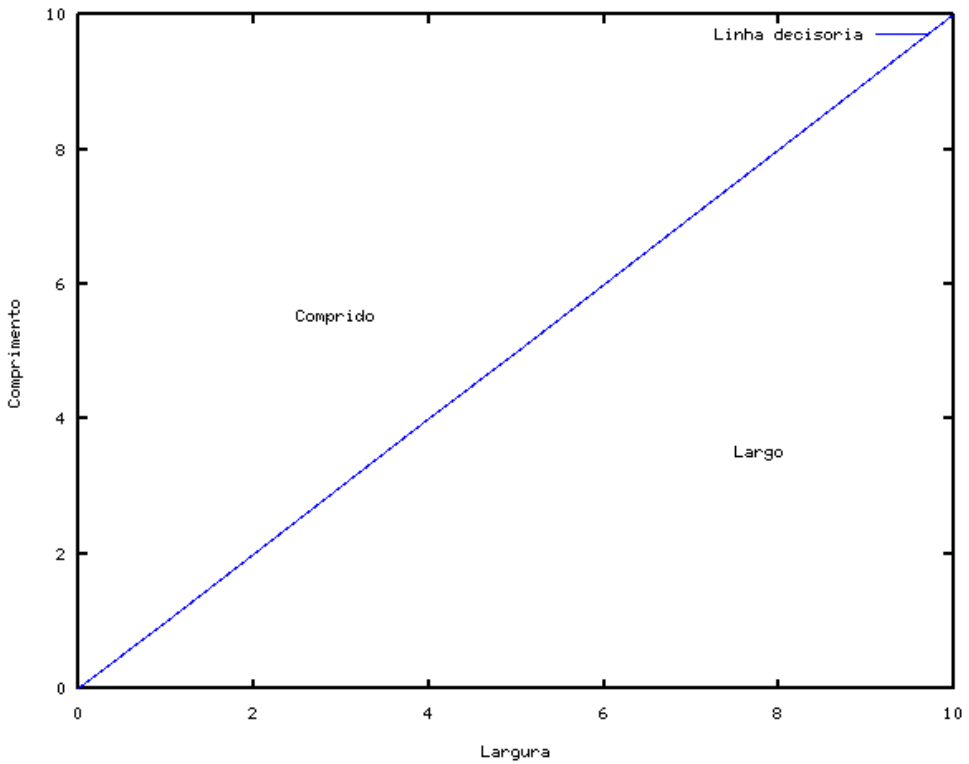

Figura: Linha decisória entre classes Largo e Comprido

Sistemas de Classificação de Texto

Problema da Classificação Textual

Dado um documento d pertencente ao espaço de documentos $\mathcal{D}$, deseja-se determinar a qual classe $c_i \in \mathcal{C}$ o documento pertence.

Sistemas de Classificação de Texto

Problema da Classificação Textual

Dado um documento d pertencente ao espaço de documentos $\mathcal{D}$, deseja-se determinar a qual classe $c_i \in \mathcal{C}$ o documento pertence.

- Solução do problema inicia-se pela escolha das características

Sistemas de Classificação de Texto

Problema da Classificação Textual

Dado um documento d pertencente ao espaço de documentos $\mathcal{D}$, deseja-se determinar a qual classe $c_i \in \mathcal{C}$ o documento pertence.

- Solução do problema inicia-se pela escolha das características
- Converte-se então documentos de treinamento para sua representação em vetores de características

Sistemas de Classificação de Texto

Problema da Classificação Textual

Dado um documento d pertencente ao espaço de documentos $\mathcal{D}$, deseja-se determinar a qual classe $c_i \in \mathcal{C}$ o documento pertence.

- Solução do problema inicia-se pela escolha das características
- Converte-se então documentos de treinamento para sua representação em vetores de características
- Aplica-se um treino supervisionado ao classificador com os vetores acima

Sistemas de Classificação de Texto

Problema da Classificação Textual

Dado um documento d pertencente ao espaço de documentos $\mathcal{D}$, deseja-se determinar a qual classe $c_i \in \mathcal{C}$ o documento pertence.

- Solução do problema inicia-se pela escolha das características
- Converte-se então documentos de treinamento para sua representação em vetores de características
- Aplica-se um treino supervisionado ao classificador com os vetores acima
- Converte-se documentos de teste para vetores de características e avalia-se o desempenho do classificador

Abordagem Adotada

- Subdivide-se o espaço $\mathcal{D}$ em dois: um espaço de treinamento e um de testes

Abordagem Adotada

- Subdivide-se o espaço $\mathcal{D}$ em dois: um espaço de treinamento e um de testes
- Converte-se os documentos do *corpus* de treinamento na representação de vetores de características e alimenta-se o classificador com tais documentos para o aprendizado

Abordagem Adotada

- Subdivide-se o espaço $\mathcal{D}$ em dois: um espaço de treinamento e um de testes
- Converte-se os documentos do *corpus* de treinamento na representação de vetores de características e alimenta-se o classificador com tais documentos para o aprendizado
- Finalizado o aprendizado, utiliza-se o *corpus* de teste para avaliar a qualidade do classificador

Abordagem Adotada

- Subdivide-se o espaço $\mathcal{D}$ em dois: um espaço de treinamento e um de testes
- Converte-se os documentos do *corpus* de treinamento na representação de vetores de características e alimenta-se o classificador com tais documentos para o aprendizado
- Finalizado o aprendizado, utiliza-se o *corpus* de teste para avaliar a qualidade do classificador
- * Literatura também convenciona o nome *corpus* aos espaços de documentos

Modelo de Espaço Vetorial (MEV)

Modelo de Espaço Vetorial

Considere um espaço de documentos $\mathcal{D}$ composto por documentos d_i, indexados por um ou mais termos T_i, que podem ter pesos atribuídos de acordo com uma regra de importância ou função de ranqueamento. O documento d_i pode ser representado como:

$$d_i = (d_{i1}, d_{i2}, \ldots, d_{in}) \tag{2}$$

onde d_{ij} é o peso atribuido ao termo T_j para o documento d_i.

Modelo de Espaço Vetorial (MEV)

Modelo de Espaço Vetorial

Considere um espaço de documentos $\mathcal{D}$ composto por documentos d_i, indexados por um ou mais termos T_i, que podem ter pesos atribuídos de acordo com uma regra de importância ou função de ranqueamento. O documento d_i pode ser representado como:

$$d_i = (d_{i1}, d_{i2}, \ldots, d_{in}) \tag{2}$$

onde d_{ij} é o peso atribuido ao termo T_j para o documento d_i.

- Proposto por Salton, Wong e Yang, 1975

Modelo de Espaço Vetorial (MEV)

Modelo de Espaço Vetorial

Considere um espaço de documentos $\mathcal{D}$ composto por documentos d_i, indexados por um ou mais termos T_i, que podem ter pesos atribuídos de acordo com uma regra de importância ou função de ranqueamento. O documento d_i pode ser representado como:

$$d_i = (d_{i1}, d_{i2}, \ldots, d_{in}) \tag{2}$$

onde d_{ij} é o peso atribuido ao termo T_j para o documento d_i.

- Proposto por Salton, Wong e Yang, 1975
- Define uma maneira simples e genérica de se representar documentos

Modelo de Espaço Vetorial (MEV)

Modelo de Espaço Vetorial

Considere um espaço de documentos $\mathcal{D}$ composto por documentos d_i, indexados por um ou mais termos T_i, que podem ter pesos atribuídos de acordo com uma regra de importância ou função de ranqueamento. O documento d_i pode ser representado como:

$$d_i = (d_{i1}, d_{i2}, \ldots, d_{in}) \tag{2}$$

onde d_{ij} é o peso atribuido ao termo T_j para o documento d_i.

- Proposto por Salton, Wong e Yang, 1975
- Define uma maneira simples e genérica de se representar documentos
- Utiliza-se de um vetor de componentes ponderadas

Modelo de Espaço Vetorial (MEV)

- Cada componente d_{ij} tem seu valor definido por uma função de ranqueamento ou uma função peso

Modelo de Espaço Vetorial (MEV)

- Cada componente d_{ij} tem seu valor definido por uma função de ranqueamento ou uma função peso
- Por exemplo, uma função bastante conhecida é a **TF-IDF**, onde:

$$d_{ij} = tf_{i,j} \log \frac{|\mathcal{D}|}{|\bar{d} \in \mathcal{D} | t \in \bar{d}|} \tag{3}$$

Modelo de Espaço Vetorial (MEV)

- Cada componente d_{ij} tem seu valor definido por uma função de ranqueamento ou uma função peso
- Por exemplo, uma função bastante conhecida é a **TF-IDF**, onde:

$$d_{ij} = tf_{i,j} \log \frac{|\mathcal{D}|}{|\bar{d} \in \mathcal{D} | t \in \bar{d}|} \tag{3}$$

- $tf_{i,j}$ é a frequência do termo j no documento i

Modelo de Espaço Vetorial (MEV)

- Cada componente d_{ij} tem seu valor definido por uma função de ranqueamento ou uma função peso
- Por exemplo, uma função bastante conhecida é a **TF-IDF**, onde:

$$d_{ij} = tf_{i,j} \log \frac{|\mathcal{D}|}{|\bar{d} \in \mathcal{D} | t \in \bar{d}|} \tag{3}$$

- $tf_{i,j}$ é a frequência do termo j no documento i
- A componente logarítmica é denominada *frequência inversa* (*inverse document frequency*): frequência de documentos em que o termo aparece

Limiar de Frequência (DF)

Limiar de Frequência (DF)

Dado um conjunto de documentos $d \in \mathcal{D}$, as características são definidas por:

$$X = \{t \in V(\mathcal{D}) | t_f > t_{lim}\} \tag{4}$$

onde $V(\mathcal{D})$ denota o vocabulário do espaço de documentos $\mathcal{D}$ e t_{lim} é a frequência limiar em que se deve considerar um termo.

Limiar de Frequência (DF)

Limiar de Frequência (DF)

Dado um conjunto de documentos $d \in \mathcal{D}$, as características são definidas por:

$$X = \{t \in V(\mathcal{D}) | t_f > t_{lim}\} \tag{4}$$

onde $V(\mathcal{D})$ denota o vocabulário do espaço de documentos $\mathcal{D}$ e t_{lim} é a frequência limiar em que se deve considerar um termo.

- Do inglês *document frequency* (DF)

Limiar de Frequência (DF)

Limiar de Frequência (DF)

Dado um conjunto de documentos $d \in \mathcal{D}$, as características são definidas por:

$$X = \{t \in V(\mathcal{D}) | t_f > t_{lim}\} \tag{4}$$

onde $V(\mathcal{D})$ denota o vocabulário do espaço de documentos $\mathcal{D}$ e t_{lim} é a frequência limiar em que se deve considerar um termo.

- Do inglês *document frequency* (DF)
- Método de seleção de características

Limiar de Frequência (DF)

Limiar de Frequência (DF)

Dado um conjunto de documentos $d \in \mathcal{D}$, as características são definidas por:

$$X = \{t \in V(\mathcal{D}) | t_f > t_{lim}\} \tag{4}$$

onde $V(\mathcal{D})$ denota o vocabulário do espaço de documentos $\mathcal{D}$ e t_{lim} é a frequência limiar em que se deve considerar um termo.

- Do inglês *document frequency* (DF)
- Método de seleção de características
- Utiliza-se de conceito similar à frequência inversa definida anteriormente

Limiar de Frequência (DF)

- Segundo Yang e Pedersen (1997), a utilização do método DF produz resultados satisfatórios quando comparado com métodos mais complexos como o IG (*information gain*) e CHI (χ^2).

Limiar de Frequência (DF)

- Segundo Yang e Pedersen (1997), a utilização do método DF produz resultados satisfatórios quando comparado com métodos mais complexos como o IG (*information gain*) e CHI (χ^2).
- Além disso, possui complexidade computacional inferior, sendo de mais simples implementação e execução

Conversão e Tratamento dos Documentos

- Tanto para o treinamento quanto para os testes, os documentos devem ser convertidos em vetores de características, baseados nas características obtidas na extração

Conversão e Tratamento dos Documentos

- Tanto para o treinamento quanto para os testes, os documentos devem ser convertidos em vetores de características, baseados nas características obtidas na extração
- Do tratamento do texto:

Conversão e Tratamento dos Documentos

- Tanto para o treinamento quanto para os testes, os documentos devem ser convertidos em vetores de características, baseados nas características obtidas na extração
- Do tratamento do texto:
  1. Remoção de caracteres indesejados (números, símbolos, etc)

Conversão e Tratamento dos Documentos

- Tanto para o treinamento quanto para os testes, os documentos devem ser convertidos em vetores de características, baseados nas características obtidas na extração
- Do tratamento do texto:
  1. Remoção de caracteres indesejados (números, símbolos, etc)
  2. Remoção das palavras vazias (*stop words*)

Conversão e Tratamento dos Documentos

- Tanto para o treinamento quanto para os testes, os documentos devem ser convertidos em vetores de características, baseados nas características obtidas na extração
- Do tratamento do texto:
  1. Remoção de caracteres indesejados (números, símbolos, etc)
  2. Remoção das palavras vazias (*stop words*)
  3. Marcação do documento (separação de palavras)

Conversão e Tratamento dos Documentos

- Tanto para o treinamento quanto para os testes, os documentos devem ser convertidos em vetores de características, baseados nas características obtidas na extração
- Do tratamento do texto:
  1. Remoção de caracteres indesejados (números, símbolos, etc)
  2. Remoção das palavras vazias (*stop words*)
  3. Marcação do documento (separação de palavras)
- No procedimento implementado, foram removidos todos os números e símbolos indesejados

Conversão e Tratamento dos Documentos

- Tanto para o treinamento quanto para os testes, os documentos devem ser convertidos em vetores de características, baseados nas características obtidas na extração
- Do tratamento do texto:
  1. Remoção de caracteres indesejados (números, símbolos, etc)
  2. Remoção das palavras vazias (*stop words*)
  3. Marcação do documento (separação de palavras)
- No procedimento implementado, foram removidos todos os números e símbolos indesejados
- Um dicionário de palavras vazias foi construído através de várias fontes (como do Apache Lucene)

Probabilidade Condicional

- Considere duas variáveis aleatórias discretas X e Y, cujos valores encontram-se representados por letras minúsculas (e.g. x, y).

Probabilidade Condicional

- Considere duas variáveis aleatórias discretas X e Y, cujos valores encontram-se representados por letras minúsculas (e.g. x, y).
- Notação adotada: $P(x_1|y_1) = P(X = x_1|Y = y_1)$

Probabilidade Condicional

- Considere duas variáveis aleatórias discretas X e Y, cujos valores encontram-se representados por letras minúsculas (e.g. x, y).
- Notação adotada: $P(x_1|y_1) = P(X = x_1|Y = y_1)$
- Por definição:

$$P(x|y) = \frac{P(x,y)}{P(y)} \tag{5}$$

Probabilidade Condicional

- Considere duas variáveis aleatórias discretas X e Y, cujos valores encontram-se representados por letras minúsculas (e.g. x, y).
- Notação adotada: $P(x_1|y_1) = P(X = x_1|Y = y_1)$
- Por definição:

$$P(x|y) = \frac{P(x,y)}{P(y)} \tag{5}$$

- Caso X e Y sejam independentes, então $P(x,y) = P(x)P(y)$ e portanto $P(x|y) = P(x)$

Lei da Probabilidade Total

Lei da Probabilidade Total

Dado um evento A e m diferentes maneiras de ocorrer este evento $A_1, A_2, \ldots, A_m$, caso estes eventos sejam mutuamente exclusivos, a probabilidade $P(A)$ é dada pela soma das probabilidades dos subeventos A_i:

$$P(A) = \sum_i P(A_i) \tag{6}$$

Regra de Bayes

Regra de Bayes

Sejam X, Y variáveis aleatórias discretas e x, y valores que estas podem assumir, respetivamente. A regra de Bayes pode ser escrita como:

$$P(x|y) = \frac{P(y|x)P(x)}{\sum_{x \in X} P(y|x)P(X)} \tag{7}$$

Regra de Bayes

Regra de Bayes

Sejam X, Y variáveis aleatórias discretas e x, y valores que estas podem assumir, respetivamente. A regra de Bayes pode ser escrita como:

$$P(x|y) = \frac{P(y|x)P(x)}{\sum_{x \in X} P(y|x)P(X)} \tag{7}$$

- Deriva-se da Probabilidade Condicional e da Lei da Probabilidade Total ao se considerar uma V.A. Y que assume um valor y de m diferentes maneiras, em função de outra V.A. X

Regra de Bayes (Continuação)

- Uma série de causas x_i ocasionam no evento y, portanto, a observação de y não é útil na determinação da causa.

Regra de Bayes (Continuação)

- Uma série de causas x_i ocasionam no evento y, portanto, a observação de y não é útil na determinação da causa.
- Utilidade da regra de Bayes aparece na determinação de $P(x|y)$, isto é, na probabilidade de uma causa x ocorrer, uma vez observado o efeito y

Regra de Bayes (Continuação)

- Uma série de causas x_i ocasionam no evento y, portanto, a observação de y não é útil na determinação da causa.
- Utilidade da regra de Bayes aparece na determinação de $P(x|y)$, isto é, na probabilidade de uma causa x ocorrer, uma vez observado o efeito y
- Utiliza-se da semelhança $P(y|x)$ e da probabilidade *a priori* da causa $P(x)$

Regra de Bayes (Continuação)

- Uma série de causas x_i ocasionam no evento y, portanto, a observação de y não é útil na determinação da causa.
- Utilidade da regra de Bayes aparece na determinação de $P(x|y)$, isto é, na probabilidade de uma causa x ocorrer, uma vez observado o efeito y
- Utiliza-se da semelhança $P(y|x)$ e da probabilidade *a priori* da causa $P(x)$
- Costuma-se chamar $P(x|y)$ de probabilidade *a posteriori*

Regra de Bayes (Continuação)

- Uma série de causas x_i ocasionam no evento y, portanto, a observação de y não é útil na determinação da causa.
- Utilidade da regra de Bayes aparece na determinação de $P(x|y)$, isto é, na probabilidade de uma causa x ocorrer, uma vez observado o efeito y
- Utiliza-se da semelhança $P(y|x)$ e da probabilidade *a priori* da causa $P(x)$
- Costuma-se chamar $P(x|y)$ de probabilidade *a posteriori*
- Pode-se entender $P(x|y)$ como as mudanças causadas na distribuição de $P(x)$ posterior ao evento y

Teoria Decisória de Bayes

- Trata-se de uma abordagem estatística ao problema de classificação

Teoria Decisória de Bayes

- Trata-se de uma abordagem estatística ao problema de classificação
- Considere o problema de determinar a classe $c_i \in \mathcal{C}$ de um documento d. Pela regra de Bayes, temos:

$$P(c_i|d) = \frac{P(d|c_i)P(c_i)}{P(d)} \tag{8}$$

Teoria Decisória de Bayes

- Trata-se de uma abordagem estatística ao problema de classificação
- Considere o problema de determinar a classe $c_i \in \mathcal{C}$ de um documento d. Pela regra de Bayes, temos:

$$P(c_i|d) = \frac{P(d|c_i)P(c_i)}{P(d)} \tag{8}$$

- $P(d|c_i)$ é a probabilidade de obtermos um documento d dado que o mesmo é da classe c_i

Teoria Decisória de Bayes

- Trata-se de uma abordagem estatística ao problema de classificação
- Considere o problema de determinar a classe $c_i \in \mathcal{C}$ de um documento d. Pela regra de Bayes, temos:

$$P(c_i|d) = \frac{P(d|c_i)P(c_i)}{P(d)} \tag{8}$$

- $P(d|c_i)$ é a probabilidade de obtermos um documento d dado que o mesmo é da classe c_i
- $P(c_i)$ é a probabilidade de obtermos um documento da classe c_i

Teoria Decisória de Bayes

- Trata-se de uma abordagem estatística ao problema de classificação
- Considere o problema de determinar a classe $c_i \in \mathcal{C}$ de um documento d. Pela regra de Bayes, temos:

$$P(c_i|d) = \frac{P(d|c_i)P(c_i)}{P(d)} \tag{8}$$

- $P(d|c_i)$ é a probabilidade de obtermos um documento d dado que o mesmo é da classe c_i
- $P(c_i)$ é a probabilidade de obtermos um documento da classe c_i
- $P(d)$ é denominado por *evidência* e é definido por:

$$P(d) = \sum_{c \in \mathcal{C}} P(d|c)P(c) \tag{9}$$

Teoria Decisória de Bayes (Continuação)

- Em nosso modelo, o documento $\mathbf{d}$ já encontra-se representado em vetor de características.

Teoria Decisória de Bayes (Continuação)

- Em nosso modelo, o documento $\mathbf{d}$ já encontra-se representado em vetor de características.
- Sendo assim, pode-se considerar cada característica como uma variável aleatória, e d como sendo uma variável aleatória conjunta de todas as características

Teoria Decisória de Bayes (Continuação)

- Em nosso modelo, o documento **d** já encontra-se representado em vetor de características.
- Sendo assim, pode-se considerar cada característica como uma variável aleatória, e d como sendo uma variável aleatória conjunta de todas as características
- Desta forma, supondo $d = (d_1, d_2, \ldots, d_n)$ para n características, temos:

$$P(d|c_i) = P(d_1, d_2, \ldots, d_n|c_i) \tag{10}$$

Teoria Decisória de Bayes (Continuação)

- Em nosso modelo, o documento **d** já encontra-se representado em vetor de características.
- Sendo assim, pode-se considerar cada característica como uma variável aleatória, e d como sendo uma variável aleatória conjunta de todas as características
- Desta forma, supondo $d = (d_1, d_2, \ldots, d_n)$ para n características, temos:

$$P(d|c_i) = P(d_1, d_2, \ldots, d_n|c_i) \tag{10}$$

- Supondo as características independentes entre si (Bayes Ingênuo), temos:

$$P(d|c_i) = P(d_1|c_i)P(d_2|c_i)\ldots P(d_n|c_i) = \prod_{i=1,\ldots,n} P(d_i|c_i)$$

Teoria Decisória de Bayes (Continuação)

- Uma vez calculados $P(c_i)$, $P(d|c_i)$ e $P(d)$, pode-se obter para cada classe $c_i \in \mathcal{C}$ a probabilidade $P(c_i|d)$.

Teoria Decisória de Bayes (Continuação)

- Uma vez calculados $P(c_i)$, $P(d|c_i)$ e $P(d)$, pode-se obter para cada classe $c_i \in \mathcal{C}$ a probabilidade $P(c_i|d)$.
- Adota-se então a maior $P(c_x|d)$, sendo então c_x a classe que melhor configura o documento d.

Teoria Decisória de Bayes (Continuação)

- Uma vez calculados $P(c_i)$, $P(d|c_i)$ e $P(d)$, pode-se obter para cada classe $c_i \in \mathcal{C}$ a probabilidade $P(c_i|d)$.
- Adota-se então a maior $P(c_x|d)$, sendo então c_x a classe que melhor configura o documento d.
- O cálculo de $P(c_i)$ se dá pelo Estimador de Máxima Verossimilhança:

$$P(c_i) = \frac{N_{c_i}}{N} \tag{12}$$

Teoria Decisória de Bayes (Continuação)

- Uma vez calculados $P(c_i)$, $P(d|c_i)$ e $P(d)$, pode-se obter para cada classe $c_i \in \mathcal{C}$ a probabilidade $P(c_i|d)$.
- Adota-se então a maior $P(c_x|d)$, sendo então c_x a classe que melhor configura o documento d.
- O cálculo de $P(c_i)$ se dá pelo Estimador de Máxima Verossimilhança:

$$P(c_i) = \frac{N_{c_i}}{N} \tag{12}$$

- Onde N_{c_i} é o número de documentos de treinamento da classe c_i e N é o número total de documentos

Teoria Decisória de Bayes (Continuação)

- Similarmente, o cálculo de $P(d_i|c_i)$ se dá por:

$$P(d_i|c_i) = \frac{N_{d_i,c_i} + 1}{N_{dco} + |d|} \tag{13}$$

Teoria Decisória de Bayes (Continuação)

- Similarmente, o cálculo de $P(d_i|c_i)$ se dá por:

$$P(d_i|c_i) = \frac{N_{d_i,c_i} + 1}{N_{dco} + |d|} \tag{13}$$

- Onde N_{d_i,c_i} é o número de ocorrências da característica d_i em documentos da classe c_i, e N_{dco} é o total de ocorrências de características em documentos da classe c_i., $|d|$ é o tamanho do vetor de características

Teoria Decisória de Bayes (Continuação)

- Similarmente, o cálculo de $P(d_i|c_i)$ se dá por:

$$P(d_i|c_i) = \frac{N_{d_i,c_i} + 1}{N_{dco} + |d|} \tag{13}$$

- Onde N_{d_i,c_i} é o número de ocorrências da característica d_i em documentos da classe c_i, e N_{dco} é o total de ocorrências de características em documentos da classe c_i., $|d|$ é o tamanho do vetor de características
- Notar a aplicação de uma suavização de Laplace para eliminação de divisões por zero

Classificação

- Formalmente, a classificação é definida por:

$$c_i = \arg\max_{c \in C} P(c|\mathbf{d}) \tag{14}$$

onde

$$P(c|\mathbf{d}) = \frac{P(c)\prod_{i=1,\ldots,n} P(d_i|c)}{P(\mathbf{d})} \tag{15}$$

Extração das Características

- Programa criado na linguagem de programação Python

Extração das Características

- Programa criado na linguagem de programação Python
- Aplicado o método do Limiar de Frequência (DF)

Extração das Características

- Programa criado na linguagem de programação Python
- Aplicado o método do Limiar de Frequência (DF)
- Mapeamento de termos para suas frequências inversas de documento

Extração das Características

- Programa criado na linguagem de programação Python
- Aplicado o método do Limiar de Frequência (DF)
- Mapeamento de termos para suas frequências inversas de documento
- Criada uma função de instrumentação capaz de controlar o tamanho do vetor de características

Treinamento

- Algorítmo proposto por Manning (2008) e modificado para atingir maior eficiência

Treinamento

- Algorítmo proposto por Manning (2008) e modificado para atingir maior eficiência
- No treinamento, são calculados os valores de $P(c_i)$ e $P(d_i|c_i)$, armazenados nos mapas *prior* e *conditional*, respectivamente

Treinamento

- Algorítmo proposto por Manning (2008) e modificado para atingir maior eficiência
- No treinamento, são calculados os valores de $P(c_i)$ e $P(d_i|c_i)$, armazenados nos mapas *prior* e *conditional*, respectivamente
- Aplicação direta da fórmula do Estimador de Máxima Verossimilhança

Treinamento

- Algorítmo proposto por Manning (2008) e modificado para atingir maior eficiência
- No treinamento, são calculados os valores de $P(c_i)$ e $P(d_i|c_i)$, armazenados nos mapas *prior* e *conditional*, respectivamente
- Aplicação direta da fórmula do Estimador de Máxima Verossimilhança
- Contagem de termos utilizando uma concatenação de todos os documentos do espaço de treinamento, reduzindo o tempo gasto com operações de E/S no banco de dados

Treinamento

- Algorítmo proposto por Manning (2008) e modificado para atingir maior eficiência
- No treinamento, são calculados os valores de $P(c_i)$ e $P(d_i|c_i)$, armazenados nos mapas *prior* e *conditional*, respectivamente
- Aplicação direta da fórmula do Estimador de Máxima Verossimilhança
- Contagem de termos utilizando uma concatenação de todos os documentos do espaço de treinamento, reduzindo o tempo gasto com operações de E/S no banco de dados
- Utilização de técnicas de otimização como inversões de laços e memoização

Treinamento

- Algorítmo proposto por Manning (2008) e modificado para atingir maior eficiência
- No treinamento, são calculados os valores de $P(c_i)$ e $P(d_i|c_i)$, armazenados nos mapas *prior* e *conditional*, respectivamente
- Aplicação direta da fórmula do Estimador de Máxima Verossimilhança
- Contagem de termos utilizando uma concatenação de todos os documentos do espaço de treinamento, reduzindo o tempo gasto com operações de E/S no banco de dados
- Utilização de técnicas de otimização como inversões de laços e memoização
- Cálculo prévio dos logarítmos utilizados na classificação

Classificação

- Algorítmo também proposto por Manning (2008) e modificado para atingir maior eficiência

Classificação

- Algorítmo também proposto por Manning (2008) e modificado para atingir maior eficiência
- A fim de evitar *integer overflow*, aplicamos a função log ao produtório de $P(d|c_i)$, obtendo a seguinte soma:

$$c_i = \arg\max_{c \in C} \left(log(P(c)) + \sum_i log(P(d_i|c)) \right) \qquad (16)$$

Preparação

- Objetivo: observar a relação entre o tamanho do vetor de características e a qualidade resultante do classificador (taxa de acertos)

Preparação

- Objetivo: observar a relação entre o tamanho do vetor de características e a qualidade resultante do classificador (taxa de acertos)
- Características do experimento:

Preparação

- Objetivo: observar a relação entre o tamanho do vetor de características e a qualidade resultante do classificador (taxa de acertos)
- Características do experimento:
  - Total de teses: 647 teses;

Preparação

- Objetivo: observar a relação entre o tamanho do vetor de características e a qualidade resultante do classificador (taxa de acertos)
- Características do experimento:
  - Total de teses: 647 teses;
  - Cursos escolhidos ao acaso, considerando a inclusão de dois cursos com alto volume e os demais com volume moderado a fim de observar as diferenças de performance;

Preparação

- Objetivo: observar a relação entre o tamanho do vetor de características e a qualidade resultante do classificador (taxa de acertos)
- Características do experimento:
  - Total de teses: 647 teses;
  - Cursos escolhidos ao acaso, considerando a inclusão de dois cursos com alto volume e os demais com volume moderado a fim de observar as diferenças de performance;
  - Utilizados 75% de teses de cada curso para treinamento, escolhidos aleatoriamente dentro do *corpus*;

Preparação

- Objetivo: observar a relação entre o tamanho do vetor de características e a qualidade resultante do classificador (taxa de acertos)
- Características do experimento:
  - Total de teses: 647 teses;
  - Cursos escolhidos ao acaso, considerando a inclusão de dois cursos com alto volume e os demais com volume moderado a fim de observar as diferenças de performance;
  - Utilizados 75% de teses de cada curso para treinamento, escolhidos aleatoriamente dentro do *corpus*;
  - Utilizados os restantes 25% de teses para testes e avaliação;

Preparação

Curso	Número de Teses
Engenharia Mecânica de Energia de Fluídos	51
Biotecnologia	203
Geotectônica	54
Processamento de Sinais e Instrumentação	55
Genética	52
Enfermagem Psiquiátrica	164
Engenharia de Sistemas	68
Total:	647

Tabela: Cursos e Teses

Resultados

- Maior taxa de acertos com aumento do tamanho do vetor de características, devido a maior densidade de informação absorvida pelo vetor

size divider	Características $\lvert\mathbf{c}\rvert$	Acurácia Global (%)
2	730	$17,18\,\%$
4	2453	$43,56\,\%$
6	4184	$61,35\,\%$
8	5724	$65,64\,\%$
10	7016	$68,71\,\%$
12	8378	$74,23\,\%$
14	9892	$76,07\,\%$
16	11065	$77,30\,\%$
18	12359	$84,66\,\%$

Resultados

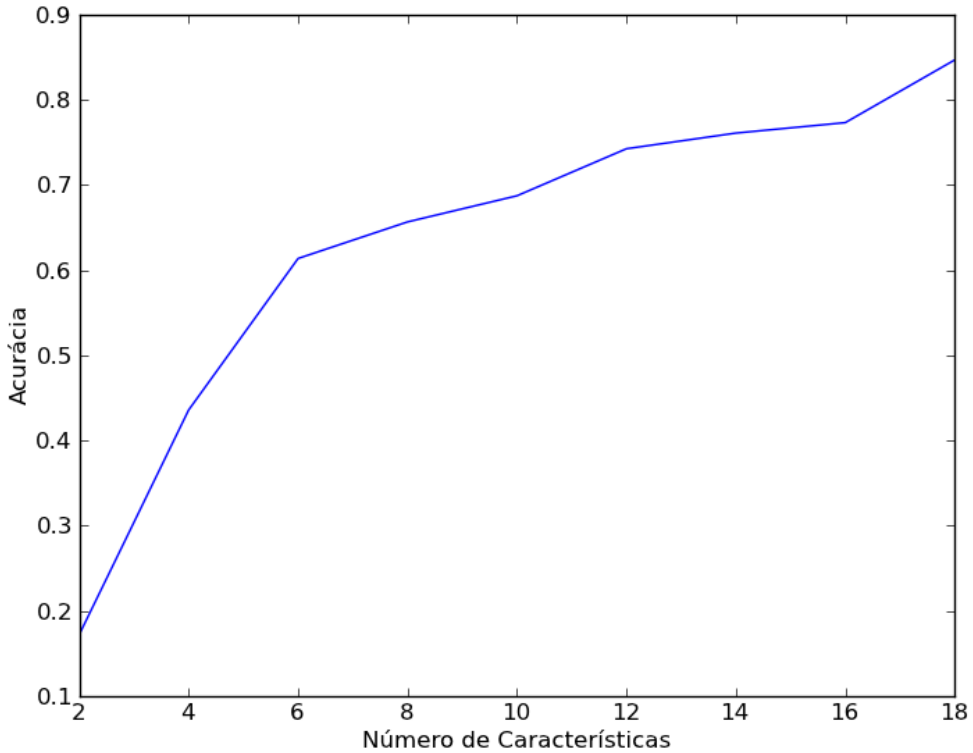

Figura: Acurácia $\times$ Número de características

Resultados

- Com o aumento do vetor de características, observou-se também um aumento no tempo de execução de cada etapa.

$\|\mathbf{c}\|$	$t_{extracao}$ (s)	$t_{treinamento}$ (s)	$t_{classificacao}$ (s)	t_{total} (s)
730	19,005	20,587	6,774	46,366
2453	19,224	22,065	16,725	58,014
4184	40,791	49,013	17,227	107,031
5724	40,947	48,837	18,232	108,016
7016	40,130	48,803	19,084	108,017
9892	39,339	53,127	23,227	115,693
11065	39,954	50,699	20,994	111,617
12359	39,530	49,789	20,374	109.693

Tabela: Número de Características e Tempos de Execução das Etapas

Resultados

- Segundo Manning (2008), a complexidade de tempo do treinamento é

$$O(|\mathcal{D}|L_{avg} + |C||V|) \tag{17}$$

Resultados

- Segundo Manning (2008), a complexidade de tempo do treinamento é

$$O(|\mathcal{D}|L_{avg} + |C||V|) \tag{17}$$

- onde $|D|$ é o número de documentos, L_{avg} é o tamanho médio dos documentos, $|C|$ é o número de classes e $|V|$ é o tamanho do vetor de características. De fato, com o aumento de $|V|$, podemos constatar da tabela 3 um tempo predominantemente crescente, como esperado.

Resultados

- Manning (2008) também mostra que a complexidade de tempo da classificação é dada por:

$$O(|C|M_a) \tag{18}$$

Resultados

- Manning (2008) também mostra que a complexidade de tempo da classificação é dada por:

$$O(|C|M_a) \tag{18}$$

- onde $|C|$ é o número de classes e M_a é o número de termos no documento a ser classificado.

Resultados

- Manning (2008) também mostra que a complexidade de tempo da classificação é dada por:

$$O(|C|M_a) \tag{18}$$

- onde $|C|$ é o número de classes e M_a é o número de termos no documento a ser classificado.
- Tal complexidade está presente algorítmo de classificação, uma vez que o vetor de características é utilizado para verificar se um termo do documento é uma característica.

Resultados

- Manning (2008) também mostra que a complexidade de tempo da classificação é dada por:

$$O(|C|M_a) \tag{18}$$

- onde $|C|$ é o número de classes e M_a é o número de termos no documento a ser classificado.
- Tal complexidade está presente algorítmo de classificação, uma vez que o vetor de características é utilizado para verificar se um termo do documento é uma característica.
- Pode-se observar que a complexidade é majorada pela quantidade de termos, mesmo que ainda possua operações relativas ao tamanho do vetor. Portanto, podemos observar uma relativa independência na coluna $t_{classificacao}$, que varia em proporções bem menores do que das demais colunas.

Considerações Finais

- Conhecimento adquirido nas áreas de mineração de dados, sistemas de recuperação da informação, aprendizado de máquina e sistemas de classificação artificial

Considerações Finais

- Conhecimento adquirido nas áreas de mineração de dados, sistemas de recuperação da informação, aprendizado de máquina e sistemas de classificação artificial
- O sistema construído conseguiu alcançar uma acurácia de $84,66\%$ (taxa de acertos absolutos) dentre teses de sete cursos distintos. Aliado a esse alto desempenho, tem-se uma desvantagem de um longo tempo de execução ($t_{max} = 115,69 segundos$)

Considerações Finais

- Conhecimento adquirido nas áreas de mineração de dados, sistemas de recuperação da informação, aprendizado de máquina e sistemas de classificação artificial
- O sistema construído conseguiu alcançar uma acurácia de $84,66\%$ (taxa de acertos absolutos) dentre teses de sete cursos distintos. Aliado a esse alto desempenho, tem-se uma desvantagem de um longo tempo de execução ($t_{max} = 115,69 segundos$)
- Os algorítmos implementados inicialmente foram propostos por Manning et. al. (2008)

Considerações Finais

- Conhecimento adquirido nas áreas de mineração de dados, sistemas de recuperação da informação, aprendizado de máquina e sistemas de classificação artificial
- O sistema construído conseguiu alcançar uma acurácia de $84,66\%$ (taxa de acertos absolutos) dentre teses de sete cursos distintos. Aliado a esse alto desempenho, tem-se uma desvantagem de um longo tempo de execução ($t_{max} = 115,69 segundos$)
- Os algorítmos implementados inicialmente foram propostos por Manning et. al. (2008)
- Modificações próprias foram feitas nos algorítmos, incluíndo otimizações de laço, cálculo prévio de funções matemáticas e aplicação da técnica de memoização para calculos repetitivos

Considerações Finais

- As modificações se mostraram extremamente úteis, reduzindo em mais de 100% o tempo de execução do algorítmo original

Considerações Finais

- As modificações se mostraram extremamente úteis, reduzindo em mais de 100% o tempo de execução do algorítmo original
- Ao fim das otimizações, o sistema alcançou uma alta velocidade de classificação, levando de 10 a 20 segundos para classificar uma tese típica de 32 mil palavras (1600 palavras por segundo).

Considerações Finais

- As modificações se mostraram extremamente úteis, reduzindo em mais de 100% o tempo de execução do algorítmo original
- Ao fim das otimizações, o sistema alcançou uma alta velocidade de classificação, levando de 10 a 20 segundos para classificar uma tese típica de 32 mil palavras (1600 palavras por segundo).
- Para o treinamento, o sistema alcançou a faixa de 30 a 50 segundos totais para processar um total de 485 documentos, com uma média de 30 mil palavras por documento, totalizando aproximadamente 14 milhões de palavras e uma taxa de 280 mil palavras processadas por segundo.

Considerações Finais

- As marcas de desempenho citadas são características fundamentais do classificador Bayes Ingênuo, visto que suas complexidades de tempo são lineares com as entradas

Considerações Finais

- As marcas de desempenho citadas são características fundamentais do classificador Bayes Ingênuo, visto que suas complexidades de tempo são lineares com as entradas
- Pela facilidade de construção e baixa complexidade temporal, esses classificadores se mostram extremamente úteis para tarefas de classificação rotineiras, alcançando desempenhos razoáveis

Trabalhos Futuros

- Função modificada de classificação:

$$c_i = \arg\max_{c \in C} \left(log(P(c)) + \sum_i N_{d,d_i} \times log(P(d_i|c)) \right) \quad (19)$$

Trabalhos Futuros

- Função modificada de classificação:

$$c_i = \arg\max_{c \in C} \left(log(P(c)) + \sum_i N_{d,d_i} \times log(P(d_i|c)) \right) \quad (19)$$

- N_{d,d_i} denota a frequência do termo d_i no documento d.

Trabalhos Futuros

- Função modificada de classificação:

$$c_i = \arg\max_{c \in C} \left(log(P(c)) + \sum_i N_{d,d_i} \times log(P(d_i|c)) \right) \quad (19)$$

- N_{d,d_i} denota a frequência do termo d_i no documento d.
- Modificação causa uma acurácia de $42,00\%$ para 758 características, enquanto que a função de classificação tradicional apresenta $17,80\%$ para um número similar de características.

Trabalhos Futuros

- Função modificada de classificação:

$$c_i = \arg\max_{c \in C} \left(log(P(c)) + \sum_i N_{d,d_i} \times log(P(d_i|c)) \right) \quad (19)$$

- N_{d,d_i} denota a frequência do termo d_i no documento d.
- Modificação causa uma acurácia de $42,00\%$ para 758 características, enquanto que a função de classificação tradicional apresenta $17,80\%$ para um número similar de características.
- Baseado nesses resultados, torna-se interessante investigar a causa de tal melhora, assim como seu rigor matemático ou até como formalizar tal modificação.