UNIVERSIDADE DE ITAÚNA
PRÓ-REITORIA DE ENSINO
COORDENAÇÃO DE CIÊNCIA DA COMPUTACAO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Eugênio Cunha

APRENDIZADO DE MÁQUINA APLICADO À VALORAÇÃO DE REDAÇÕES

ITAÚNA
2017

EUGÊNIO CUNHA

APRENDIZADO DE MÁQUINA APLICADO À VALORAÇÃO DE REDAÇÕES

Projeto submetido à Coordenadoria do Curso de Bacharelado em Ciência da Computação da Universidade de Itaúna - Campus Verde, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Área de pesquisa: Aprendizagem de Máquina

Orientador: Prof. Dr. Marco Túlio Alves N Rodrigues

Itaúna
2017

UNIVERSIDADE DE ITAÚNA
COORDENAÇÃO DE CIÊNCIA DA COMPUTAÇÃO

EUGÊNIO CUNHA

Este projeto foi julgada adequada para a obtenção do Grau de Bacharel em Ciência da Computação, sendo aprovada pela coordenação de ciência da computação do curso de Bacharelado em Ciência da Computação do Campus verde da Universidade de Itaúna e pela banca examinadora:

_____

Orientador: Prof. Dr. Marco Túlio Alves N Rodrigues
Universidade de Itaúna- UIT

_____

Avaliador: Prof. Me. Felipe Domingos da Cunha
Universidade de Itaúna- UIT

Itaúna, 19 de Junho de 2017

Dedico este trabalho a Davi, meu filho, sempre preocupado em proporcionar um minuto de pausa para brincadeiras durante minhas horas de estudo, “meus melhores minutos”!

“Os computadores são incrivelmente rápidos, precisos e burros; os homens são incrivelmente lentos, imprecisos e brilhantes; juntos, seus poderes ultrapassam os limites da imaginação”.

Albert Einstein

Resumo

Este trabalho baseou-se na avaliação das competências exigidas em um texto de redação do tipo dissertativo-argumentativo com temas diversificados de ordem social, científica, cultural ou política. Fundamentou-se no estudo das técnicas de aprendizado de máquina supervisionado que provê uma gama diversificada de algoritmos poderosos para classificações de textos.

O objetivo deste trabalho é classificar as competências exigidas em um texto de redação a partir do treinamento de um modelo de aprendizado de máquina com base em um *corpus* de redações avaliadas seguindo as competências exigidas em uma redação do tipo dissertativa-argumentativa.

A compilação de um *corpus* de redações para treinamento e teste de um modelo de aprendizado de máquina exigiu a prática de extração de informações que compreende técnicas e algoritmos que realisam duas tarefas importantes: a identificação de informações desejadas a partir de documentos estruturados e não-estruturados, e o armazenamento dessas informações em um formato apropriado para uso futuro. Afim de se avaliar a eficácia dos classificadores, vários experimentos foram executados usando-se um *corpus* extraido do banco de redações da UOL (UOL, 2017), um serviço que estimula o estudante a treinar produção de textos, em especial do gênero dissertativo-argumentativo.

O resultado de classificação das competências exigidas em um texto de redação obtidas experimentalmente mostraram que o resultado geral do sistema proposto é comparável à avaliação manual por avaliadores capacitados.

Palavras-chaves: Aprendizado de máquina, Banco de Redações, Classificação, Redação

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

TAG	Rótulo
BOW	Representação da frequência acumulada de palavras em documentos diferentes a partir de um dicionário pré-criado

Lista de Abreviacoes

IA	Inteligência artificial
CSF	Ciência sem fronteiras
UnB	Universidade de Brasília
SISU	Sistema de Seleção Unificada
ENEM	Exame Nacional do Ensino Médio
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais
CEBRASPE	Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos

CAPÍTULO 1

INTRODUÇÃO

O desenvolvimento de uma redação e uma atividade prática presente na cultura civilizada desde a invenção da escrita. Já faz pelo menos uma década que um bom desempenho na redação do Exame Nacional de Ensino Médio - ENEM virou sinônimo de chances maiores para ser aprovado no processo seletivo de acesso a inúmeras universidades públicas (SISU, 2017) e a importantes programas de governo como Ciência sem fronteiras (CSF, 2017).

Em todo processo seletivo é comum o uso de marcações em gabaritos afim de automatizar o processo de correção, uma alternativa rápida e segura, até mesmo aplicações de provas eletrônicas são cada vez mais comum. Um exemplo seria o processo seletivo para avaliador das redações do ENEM que durante a "FASE II" respondem uma prova eletrônica eliminatória (CEBRASPE, 2016a). É notável que todo o processo evoluiu com objetivo de agilidade, confiança e segurança do resultado, entretanto a avaliação das competências de uma redação ainda depende exclusivamente da supervisão de duas ou mais pessoas envolvidas (INEP, 2016).

A redação é aplicada no ENEM desde a primeira edição 1998, hoje o maior exame do Brasil, que na edição de 2016 teve 8.627.195 escritos confirmados, e a participação direta de 11.360 profissionais externos na correção de 5.825.134 redações, entre eles, 378 supervisores e 10.982 avaliadores de acordo com a CEBRASPE (CEBRASPE, 2016b).

Segundo o edital do ENEM 2016 (INEP, 2016) cada redação foi avaliada por, pelo menos, dois avaliadores, de forma independente, contabilizando um número mínimo de 11.650.268 avaliações manuais, das competências exigidas em um texto de redação pelo ENEM.

Machine Learning ou aprendizado de máquina é uma área de IA cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática. Os

algoritmos de aprendizado de máquina procuram padrões dentro de um conjunto de dados (MITCHELL, 1997). Esses algoritmos existem há bastante tempo, uma ciência que não é nova, mas que está ganhando um novo impulso enquanto o processamento computacional cresce e fica mais barato.

A hipótese desta monografia é que um ou mais modelos de aprendizado de máquina na valoração das competências de uma redação pode ser tão eficiênte quanto o processo de avaliação manual.

1.1 Definição do Problema de Pesquisa

Dado um corpus de redações avaliar as competências exigidas em um texto de redação do tipo dissertativo-argumentativo substituindo a etapa de avaliação manual.

1.2 Motivação

Com crescente volume e variedade de dados disponíveis, o processamento computacional que está mais barato e mais poderoso, e o armazenamento de dados de forma acessível, o aprendizado de máquina está no centro de muitos avanços tecnológicos atingindo áreas antes exclusivas de seres humanos. Os carros autônomos do Google são o exemplo de uma atividade antes exclusiva de um humano e hoje exercida e aperfeiçoada por algoritmos de aprendizado de máquina (WAYMO, 2017).

Aplicações de aprendizado de máquina estão presentes na nossa vida cotidiana como, resultados de pesquisa web, análise de sentimento baseado em texto e na detecção de fraudes em operações com cartões de crédito (BATISTA *et al.*, 1999).

As competências exigidas em uma redação podem ser avaliadas por aprendizado de máquina. Diferente de um ser humano um modelo de aprendizado de máquina está livre de ansiedade, fadiga, *stress* entre outro fatores emocionais que afetão uma avaliação. Isto representaria a classificação de um texto imparcial a opinião do autor classificando todas as competências necessárias para definição de uma nota.

1.3 Objetivos Gerais e Específicos

Este trabalho tem como objetivo geral aplicar aprendizado de máquina na avaliação das competências exigidas em um texto de redação do tipo dissertativo-argumentativo.

1.3.1 Objetivos Especificos

Dentro do escopo geral de forma detalhada e refinada as ações que se prentende executar para alcançar o objetivo geral acima, são particularizadas como os seguintes objetivos específicos:

- Coletar, normalizar e compilar um *corpus* de redações com textos em prosa, do tipo dissertativo-argumentativo com temas diversificados, onde todas as competências foram avaliadas;
- Montar um fluxo de trabalho utilizando a ferramenta de *Data Mining* Orange3 com modelos classificadores de multiplas classes;
- Ajustar e treinar os modelos classificadores com o corpus de redações;
- Realizar os seguintes testes;
  - Acurácia ou taxa de erro;
  - *Overfitting* ou super-ajustamento;
  - *Noise* ou ruído;
- Representar graficamente os resultados obtidos.

1.4 Metodologia

Para concluir com êxito os objetivos propostos o método utilizado para solução do problema tem os seguintes passo:

- Linguagens, Frameworks e bibliotecas
  - Definição e justificativa da linguagem utilizada.

  - Definição e justificativa das bibliotecas utilizadas.
  - Definição e justificativa das frameworks utilizadas.

- Coleta de dados
  - A partir de fontes de dados disponíveis na web, coletar e organizar um corpus de redações com textos em prosa, do tipo dissertativo-argumentativo de temas variados, avaliados segundo as competências exigidas pelo ENEM.
  - Dividir 75% do corpus para ser utilizado na etapa de treinamento do modelo e os demais 25% restantes para testes posteriores.
- Algoritmo e modelo
  - Avaliar a viabilidade do uso de um algoritimo supervisionado ou semi-supervisionado.
  - Definição dos modelos que atendem o escopo geral do problema de avaliação de redações.
- Treinamento e teste.
  - Treinamento dos modelos avaliados.
  - Teste de acurácia, *overfitting* e *noise* do modelos avaliados.
- Resultados.
  - Comparar gráficamente o resultado dos testes de cada modelo avaliado entre cada um.
  - Elaboração da conclusão dos teste e viabilidade de uso de cada modelo.
- Aplicação prática.
  - Embasado na conclusão dos testes utilizar os modelos em uma aplicação web.

1.5 Contribuicoes

1.6 Organizacao do trabalho

Capitulo 2: descricao...

Capitulo 4: descricaoo...

Capitulo 5: descricao...

Capitulo 6: descricao...

CAPÍTULO 2

TRABALHOS RELACIONADOS

CAPÍTULO 3

FUNDAMENTAÇÃO TEÓRICA

3.1 Métodos de Kernel

Este capítulo tem como objetivo

3.2 Kernel em análise de padrões

Em análise de padrões, temos como objetivo detectar automaticamente padrões em um conjunto de dados de um determinado problema. Por padrões, podemos entender qualquer relação ou regularidades inerentes à alguma estrutura em uma fonte de dados. Essa análise geralmente é feita a partir dos valores de entrada e suas respectivas saídas(no caso da aprendizagem supervisionada) fornecidas no problema. Essas informações podem formar padrões em que se torna possível verificar o valor de uma saída dada uma nova entrada fornecida pelo usuário.

Diversos problemas podem ser resolvidos utilizando esta abordagem, categorização de textos, análise de sequências de DNA, reconhecimento de escrita, por exemplo.

A abordagem de análise de padrões utilizando métodos de kernel se baseia em adaptar os dados de entrada em um espaço característico adequado e nos algoritmos usados para descobrir os padrões do problema. Levando em conta isso, podemos pensar que qualquer solução com métodos de kernel é composta por estas duas partes: uma em que é feito o mapeamento nesse espaço característico e a outra em que é executado o algoritmo de aprendizagem para detectar os padrões neste espaço. A ideia por trás desta abordagem é poder mapear os dados em um espaço em que possamos

Uma das principais caracerísticas desses métodos é o atalho computacional que pode ser utilizado, tal atalho é conhecido como função de kernel.

O Kernel é uma função de mapeamento de dados em dimensões superiores com a motivação de torná-los mais fáceis de separar ou estruturá-los de maneira mais adequada. Essas funções podem ser utilizadas nas tarefas de reconhecimento de padrões.

$$Z = X \cdot Y, \tag{3.1}$$

em que Z, X e Y são variáveis complexas. A referenciação à Equação (3.1) é feita por meio do comando “ref”. O mesmo vale para outros tipos de elementos.

3.2.1 Exemplo de uma equação mais complexa

Equações mais complexas podem ser mais facilmente escritas com uso do programa TexAide. Como, por exemplo,

$$f_{\Gamma^{1/2}}(x;\alpha,\lambda) = \frac{2\lambda^{\alpha}}{\Gamma(\alpha)} x^{2\alpha-1} \exp\left(-\lambda x^2\right). \tag{3.2}$$
$$\alpha,\lambda > 0.$$

em que $\Gamma(\cdot)$ é a função Gama. O programa TexAide é semelhante ao *MathType* do Office, porém ao copiar e colar a equação em um arquivo tex, é gerado o código em LaTex referente a esta equação.

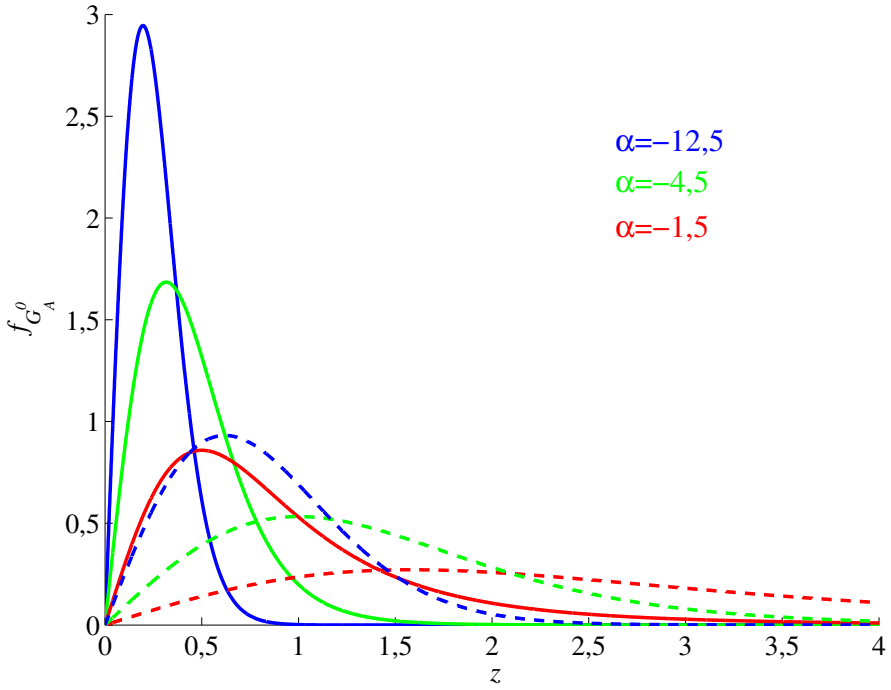
3.3 Tabelas

Tabelas são essenciais na apresentação de dados. A Tabela 1 mostra um exemplo do uso deste tipo de elemento. Vale ressaltar que não é aconselhável o uso de linhas verticais em trabalhos acadêmicos e de pesquisa.

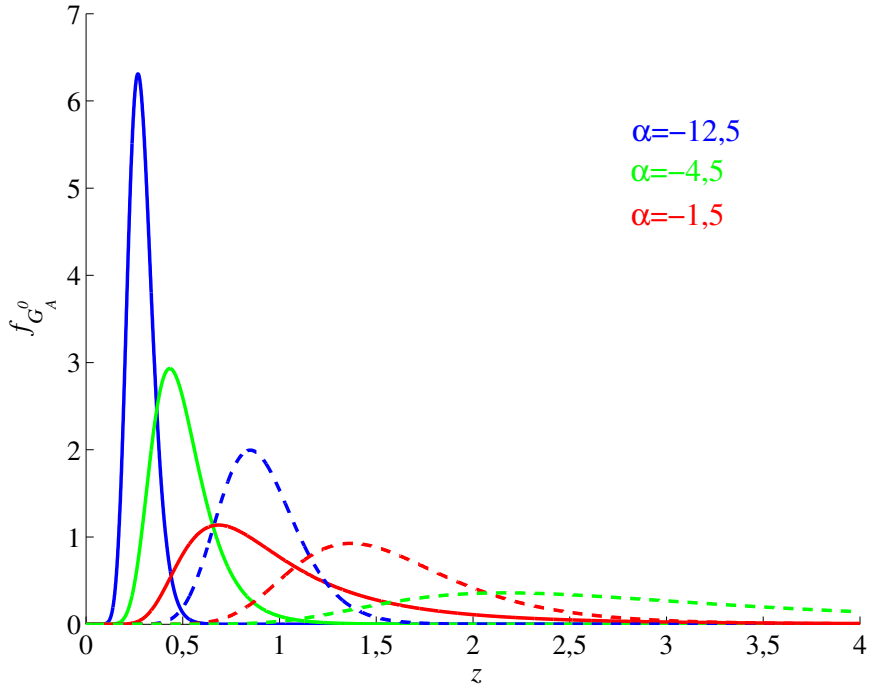
A Figura 1 mostra o exemplo do uso do comando “subfigure”. Apesar de aceitar diferentes tipos de imagens. É preferível que as imagens estejam no formato .eps. Isso garante que a imagem impressa seja exatamente aquela visualizada, como acontece com arquivos pdf.

Tabela 1: Modelos estatísticos e suas relações.

$\mathscr{N}^{-1/2}(x;\alpha,\gamma,\lambda)$	$\alpha,\lambda > 0$		$\alpha,\lambda \to \infty$	Homogêneo
	$\gamma \to 0$	Heterogêneo	$\alpha/\lambda \to \beta$	
	$\xrightarrow{D}$	$\sqrt{\Gamma}(\alpha,\lambda)$	$\xrightarrow{P}$	$\sqrt{\beta}$
	$\xrightarrow{D}$	$\Gamma^{-1/2}(\alpha,\gamma)$	$\xrightarrow{P}$	$\sqrt{\zeta^{-1}}$
	$\lambda \to 0$	Extremamente	$-\alpha/\gamma \to \zeta^{-1}$	
	$-\alpha,\gamma > 0$	Heterogêneo	$-\alpha,\gamma \to \infty$	Homogêneo
$\mathscr{G}_A(z;\alpha,\gamma,\lambda,n)$	$\alpha,\lambda > 0$		$\alpha,\lambda \to \infty$	Homogêneo
	$\gamma \to 0$	Heterogêneo	$\alpha/\lambda \to \beta$	
	$\xrightarrow{D}$	$\mathscr{K}_A(\alpha,\lambda,n)$	$\xrightarrow{P}$	$\sqrt{\Gamma}(n,n/\beta)$
	$\xrightarrow{D}$	$\mathscr{G}_A^0(\alpha,\gamma,n)$	$\xrightarrow{P}$	$\sqrt{\Gamma}(n,n\zeta)$
	$\lambda \to 0$	Extremamente	$-\alpha/\gamma \to \zeta$	
	$-\alpha,\gamma > 0$	Heterogêneo	$-\alpha,\gamma \to \infty$	Homogêneo



(a)



(b)

Figura 1: Curvas de funções de probabilidade: (a) exemplo 1, (b) exemplo 2.

CAPÍTULO 4

MÉTODO PROPOSTO

CAPÍTULO 5

DESENVOLVIMENTO

CAPÍTULO 6

RESULTADOS EXPERIMENTAIS

CAPÍTULO 7

CONCLUSÃO E TRABALHOS FUTUROS

Referências Bibliográficas

BATISTA, GEAPA; CARVALHO, ACPLF; MONARD, Maria C; BRASIL, Silicon Graphics. Aplicando seleção unilateral em conjuntos de exemplos desbalanceados: Resultados iniciais. In: **XIX CONGRESSO NACIONAL DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO "EDUCAÇÃO E APRENDIZAGEM NA SOCIEDADE DA INFORMAÇÃO**. [S.l.: s.n.], 1999. v. 20, p. 327–340.

CEBRASPE. **CENTRO BRASILEIRO DE PESQUISA EM AVALIAÇÃO E SELEÇÃO E DE PROMOÇÃO DE EVENTOS (CEBRASPE) PROGRAMA DE ATUALIZAÇÃO, QUALIFICAÇÃO E SELEÇÃO DE AVALIADORES DAS REDAÇÕES DO ENEM 2016** . 2016. Online; acessado 07 Abril 2017. Disponível em: <http://www.cespe.unb.br/colaboradores/ENEM_16_AVALIADOR_REDACAO/arquivos/ENEM_REDA____ES_2016___AVALIADOR___REGULAMENTO.PDF>.

CEBRASPE, CESPE UNB. **Relatório de Gestão CEBRASPE**. 2016. 1–20 p. Online; acessado 07 Abril 2017. Disponível em: <http://www.cespe.unb.br/cebraspe/arquivos/Relatorio_de_Gestao_2016.pdf>.

CSF, Ciência sem Fronteiras. **Estudante de Graduação**. 2017. Online; acessado 07 Abril 2017. Disponível em: <http://www.cienciasemfronteiras.gov.br/web/csf/estudante>.

INEP. **EDITAIS**. 2016. Online; acessado 07 Abril 2017. Disponível em: <http://download.inep.gov.br/educacao_basica/enem/edital/2016/edital_enem_2016.pdf>.

MITCHELL, Tom M. Mitchell. **Machine Learning**. [S.l.]: Book News, Inc., 1997. ISBN 0070428077.

SISU, Sistema de seleção unificada. **O que é o Sisu**. 2017. Online; acessado 07 Abril 2017. Disponível em: <http://sisu.mec.gov.br/>.

UOL. **Banco de redações**. 2017. Online; acessado 01 Junho 2017. Disponível em: <https://educacao.uol.com.br/bancoderedacoes/>.

WAYMO. **We're building a safer driver for everyone**. 2017. Online; acessado 07 Abril 2017. Disponível em: <https://waymo.com/>.

APÊNDICE A – Título do Apêndice

APÊNDICE B – Exemplo do pacote Algorithm

Algoritmo 1 Estimador ML otimizado.

Inicializar o contador: $j \leftarrow 1$;
Fixar o limiar de variação das estimativas: $e_{\text{out}} \leftarrow 10^{-4}$;
Fixar o número máximo de iterações: $N \leftarrow 1000$;
Computar o ponto inicial: $\hat{\gamma}(0)$;
Determinar o limiar inicial: $e_1 \leftarrow 1000$;
Estabelecer o valor inicial de α: $\hat{\alpha}(0) \leftarrow -10^{-6}$;
enquanto $e_j \geq e_{\text{out}}$ e $j \leq M$ **fazer**
    Solucionar $\hat{\alpha}_j \leftarrow \arg\max_{\alpha} l_1(\alpha; \gamma_{j-1}, \mathbf{z}, n)$;
    Solucionar $\hat{\gamma}_j \leftarrow \arg\max_{\gamma} l_2(\gamma; \alpha_j, \mathbf{z}, n)$;
    $j \leftarrow j + 1$
    Computar o critério de convergência: e_j;
fim enquanto