



Pedro Oguri

Aprendizado de Máquina para o Problema de Sentiment Classification

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Mestrado em Informática do Departamento de Informática da PUC-Rio

Orientador : Prof. Ruy Luiz Milidiú

Co-Orientador: Prof. Raúl Rentería

Rio de Janeiro
outubro de 2006



Pedro Oguri

Aprendizado de Máquina para o Problema de Sentiment Classification

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Mestrado em Informática do Departamento de Informática do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Ruy Luiz Milidiú

Orientador

Departamento de Informática — PUC-Rio

Prof. Raúl Rentería

Co-Orientador

Departamento de Informática — PUC-Rio

Prof. Marcus Vinícius Soledade Poggi de Aragão

Departamento de Informática — PUC-Rio

Prof. Marco Antonio Casanova

Departamento de Informática — PUC-Rio

Prof. José Eugênio Leal

Coordenador Setorial do Centro

Técnico Científico — PUC-Rio

Rio de Janeiro, 25 de outubro de 2006

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

Pedro Oguri

Graduou-se Bacharel em Informática pela PUC-Rio em 2003 tendo cursado parte de sua graduação na Universidad Autónoma de Madrid. Trabalhou como Engenheiro de Software na Accenture e na Fast Search and Transfer. Tem interesse na pesquisa de algoritmos, especialmente nas áreas de aprendizado de máquina e otimização combinatória.

Ficha Catalográfica

Oguri, Pedro

/ Pedro Oguri; orientador: Ruy Luiz Milidiú; co-orientador: Raúl Rentería. — Rio de Janeiro : PUC-Rio, Departamento de Informática, 2006.

v., 54 f: il. ; 29,7 cm

1. Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui referências bibliográficas.

1. Informática – Tese. I. Milidiú, Ruy Luiz. II. Renteria, Raúl. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

Agradecimentos

À minha esposa Vanessa e ao meu filho Breno, pelo apoio incondicional e compreensão.

Aos meus pais e à minha irmã que tanto me apoiaram e incentivaram.

Ao meu orientador Ruy Milidiú por ter me despertado tamanho fascínio pelo estudo de algoritmos através de seus ensinamentos.

Ao meu co-orientador e amigo Raul Rentería pelo apoio, incentivo e valiosas discussões.

Ao professor Marcus Poggi pelos ensinamentos, incentivo e entusiasmo contagiante.

Aos brasileiros que pagam impostos, por me sustentarem durante o primeiro ano do meu mestrado.

Resumo

Oguri, Pedro; Milidiú, Ruy Luiz; Renteria, Raúl. **Aprendizado de Máquina para o Problema de Sentiment Classification**. Rio de Janeiro, 2006. 54p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Sentiment Analysis é um problema de categorização de texto no qual deseja-se identificar opiniões favoráveis e desfavoráveis com relação a um tópico. Um exemplo destes tópicos de interesse são organizações e seus produtos. Neste problema, documentos são classificados pelo sentimento, conotação, atitudes e opiniões ao invés de se restringir aos fatos descritos neste.

O principal desafio em *Sentiment Classification* é identificar como sentimentos são expressados em textos e se tais sentimentos indicam uma opinião positiva (favorável) ou negativa (desfavorável) com relação a um tópico. Devido ao crescente volume de dados disponível na Web, onde todos tendem a ser geradores de conteúdo e expressarem opiniões sobre os mais variados assuntos, técnicas de Aprendizado de Máquina vem se tornando cada vez mais atraentes.

Nesta dissertação investigamos métodos de Aprendizado de Máquina para *Sentiment Analysis*. Apresentamos alguns modelos de representação de documentos como saco de palavras e N-grama. Testamos os classificadores SVM (Máquina de Vetores Suporte) e Naive Bayes com diferentes modelos de representação textual e comparamos seus desempenhos.

Palavras-chave

Aprendizado de Máquina. Sentiment Analysis. Classificação de Texto. Classificadores Bayesianos. Support Vector Machines.

Abstract

Oguri, Pedro; Milidiú, Ruy Luiz; Renteria, Raúl. **Machine Learning for Sentiment Classification**. Rio de Janeiro, 2006. 54p. MsC Thesis — Department of Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Sentiment Analysis is a text categorization problem in which we want to identify favorable and unfavorable opinions towards a given topic. Examples of such topics are organizations and its products. In this problem, documents are classified according to their sentiment, connotation, attitudes and opinions instead of being limited to the facts described in it.

The main challenge in *Sentiment Classification* is identifying how sentiments are expressed in texts and whether they indicate a positive (favorable) or negative (unfavorable) opinion towards a topic. Due to the growing volume of information available online in an environment where we all tend to be content generators and express opinions on a variety of subjects, Machine Learning techniques have become more and more attractive.

In this dissertation, we investigate Machine Learning methods applied to Sentiment Analysis. We present document representation models such as bag-of-words and N-grams. We compare the performance of the Naive Bayes and the Support Vector Machine classifiers for each proposed model.

Keywords

Machine Learning. Sentiment Analysis. Text Classification. Bayesian Classifiers. Support Vector Machines.

Sumário

1	Introdução	11
1.1	Objetivos da Dissertação	11
1.2	Organização da Tese	12
2	Sentiment Analysis	13
2.1	Definição do Problema	13
2.2	Trabalhos Relacionados	14
3	Modelagem Textual	18
3.1	Saco de Palavras	19
3.2	N-gramas	19
3.3	Part of Speech Tagging	20
3.4	Filtro de Subjetividade	20
3.5	Seleção de Features	21
3.5.1	Informação Mútua Média	23
4	Classificador Naive Bayes	25
4.1	Fundamentos Teóricos	26
4.1.1	Modelo Binário	27
4.1.2	Modelo Multinomial	27
5	Classificador SVM	29
5.1	Fundamentos Teóricos	29
5.1.1	A dimensão VP e o limite do risco esperado	30
5.2	Máquinas de Vetores Suporte Lineares	30
5.2.1	Formulação Lagrangeana	32
5.3	Máquinas de Vetores Suporte Não-Lineares	33
6	Experimentos	35
6.1	Corpus	36
6.2	Metodologia de Teste	36
6.3	Descrição dos Experimentos	37
6.3.1	Corpus com Subjetividade Filtrada	38
6.4	Resultados	38
6.4.1	Discussão dos Resultados	40
7	Conclusão	44
A	Resultados Detalhados	46
B	Fluxograma do Ambiente de Experimentação	50
	Referências Bibliográficas	52

Lista de figuras

2.1	Processo de Classificação de Sentimento	14
5.1	Solução para um cenário de 2 dimensões linearmente separável. Os pontos pelos quais a linha pontilhada passa são chamados de <i>vetores suporte</i>	31
6.1	Validação Cruzada com 3 folds	37
6.2	Desempenho do Classificador NB	39
B.1	Fluxograma da Solução	51

Lista de tabelas

2.1	Resultados reportados em Pang et al. (19)	15
3.1	15 features de IMM mais alto	24
6.1	Estatísticas do <i>movie review data set</i>	36
6.2	Estatísticas do <i>corpus</i> com subjetividade filtrada	38
6.3	Melhores precisões atingidas por cada classificador para cada modelo de representação de documentos	43
A.1	Resultados para unigramas do corpus original	46
A.2	Resultados para bigramas do corpus original	47
A.3	Resultados para unigramas com classe gramatical (<i>part of speech</i>)	47
A.4	Resultados para bigramas com classe gramatical (<i>part of speech</i>)	48
A.5	Resultados para unigramas com subjetividade filtrada	48
A.6	Resultados para bigramas com subjetividade filtrada	49

Science is what we understand well enough to explain to a computer. Art is everything else we do.

Donald Knuth, *Prefácio do livro A=B.*