SPAMta

Miguel Mendes
Jéssica Aparecida
Ricardo Luiz

- Introdução
- Objetivos
- Descrição dos Dados e Modelagem
- Resultados
- Conclusão
- Referências

- **Introdução**
- Objetivos
- Descrição dos Dados e Modelagem
- Resultados
- Conclusão
- Referências

Introdução

- SPAM: **S**tupid **P**ointless **A**nnoying **M**essages
- HAM: Não-spam
- Motivação: separar bons e-mails de e-mails, que não oferecem nenhum valor, automaticamente.

- Introdução
- **Objetivos**
- Descrição dos Dados e Modelagem
- Resultados
- Conclusão
- Referências

Objetivos

- Apresentar um classificador de SPAM.
- Filtrar e-mails automaticamente.
- Construir um classificador capaz de detectar SPAM sem ser explicitamente programado.
- Aplicar um conceito visto em aula em um problema real.
- Comparar o desempenho entre duas das principais técnicas de aprendizado supervisionado.

- Introdução
- Objetivos
- **Descrição dos Dados e Modelagem**
- Resultados
- Conclusão
- Referências

Descrição dos Dados e Modelagem

- Datasets obtidos no projeto SpamAssassin.
- 3900 emails rotulados como não-spam (ham).
- 1896 emails rotulados como spam.
- Emails obtidos entre os anos 1999-2004.
- Todos os emails estão em formato bruto.

Descrição dos Dados e Modelagem

- Contém cabeçalhos, links e tags HTML.
- Todas as palavras foram postas em *lowercase*.
- Foram removidos todas as tags HTML.
- Números foram substituídos por “number”.
- Urls removidas e substituídas por “httpaddr”.

Descrição dos Dados e Modelagem

- Endereços de emails substituídos por “emailaddr”
- Substituição de $ por “dollar”.
- Remoção de todas as pontuações.
- Dados tokenizados e submetidos ao processo de stemming.
- Contagem das palavras realizada no dataset de spams.
- *Bag-of-word* com as 2000 palavras mais frequentes no dataset spam.

Descrição dos Dados e Modelagem

- Foram utilizadas duas técnicas:
  - SVM
  - Redes Neurais
- SVM com kernel linear.
- Rede Neural com 25 neurônios na camada escondida.

Descrição dos Dados e Modelagem

- Divisão dos dados:
  - 70% dos dados para treinamento
  - 15% dos dados para cross-validation
  - 15% dos dados para teste
- Como são apenas duas classes não foi necessário usar one-vs-all.

- Introdução
- Objetivos
- Descrição dos Dados e Modelagem
- **Resultados**
- Conclusão
- Referências

Resultados

- Métricas Utilizadas:
  - Acurácia = $\frac{VP + VN}{P + N}$
  - Recall = $\frac{VP}{VP + FN}$
  - Precisão = $\frac{VP}{VP + FP}$
- VP = Verdadeiro Positivo
- FN = Falso Negativo
- FP = Falso Positivo
- P = Total de positivos
- N = Total de negativos

Resultados

- Usando SVM foram obtidos os seguintes resultados:
  - Treinamento:
    - Acurácia: 100%
  - Cross-validation:
    - Acurácia : 99.65%
  - Teste:
    - Acurácia : 99.19%

Resultados

- Matriz de Confusão no *Training Set* (SVM)

	Previsto	
Atual	2706	0
	0	1351

- Acurácia: 100%
- Recall: 100%
- Precisão: 100%

Resultados

- Matriz de Confusão no *Cross-Validation* (SVM)

	Previsto	
Atual	590	0
	3	276

- Acurácia: 99.65%
- Recall: 100%
- Precisão: 99.49%

Resultados

- Matriz de Confusão no *Test Set* (SVM)

	Previsto	
Atual	600	4
	3	263

- Acurácia: 99.19%
- Recall: 99.33%
- Precisão: 99.50%

Resultados

- Usando SVM foram obtidos os seguintes resultados:
  - Treinamento
    - Acurácia: 100%
  - Cross-validation
    - Acurácia : 100%
  - Teste
    - Acurácia : 99.80%

Resultados

- Matriz de Confusão no *Training Set* (Rede Neural)

	Previsto	
Atual	2703	0
	0	1355

- Acurácia: 100%
- Recall: 100%
- Precisão: 100%

Resultados

- Matriz de Confusão no *Cross-Validation* (Rede Neural)

	Previsto	
Atual	620	0
	0	249

- Acurácia: 100%
- Recall: 100%
- Precisão: 100%

Resultados

- Matriz de Confusão no *Test Set* (Rede Neural)

	Previsto	
Atual	577	0
	2	290

- Acurácia: 99.8%
- Recall: 100%
- Precisão: 99.70%

- Introdução
- Objetivos
- Descrição dos Dados e Modelagem
- Resultados
- **Conclusão**
- Referências

Conclusão

- Embora os resultados sejam parecidos, a Rede Neural teve uma ligeira vantagem sobre o SVM.
- E-mails "antigos" são facilmente classificados em comparação com spam disseminados atualmente.
- Pre-processar os e-mails foi a tarefa mais complexa e que tomou mais tempo.

- Introdução
- Objetos
- Descrição dos Dados e Modelagem
- Resultados
- Conclusão
- **Referências**

Referências

- Slides vistos em aula.
- Han, J.; Kamber, M.; Pei J. Data Mining: Concepts and Techniques, 3rd Edition;
- BEZERRA, E.; GOLDSCHMIDT, R. R. *A Tarefa de Classificação em Text Mining*. Revista de Sistemas de Informação da FSMA n. 5, pp. 42-62, 2010.
- The Apache SpamAssassin Project. Disponível em: http://spamassassin.apache.org/publiccorpus/. Acesso em: 23/09/2013 às 19:26)