Introdução ao Aprendizado de Máquina

Prof. Erneson A. Oliveira*

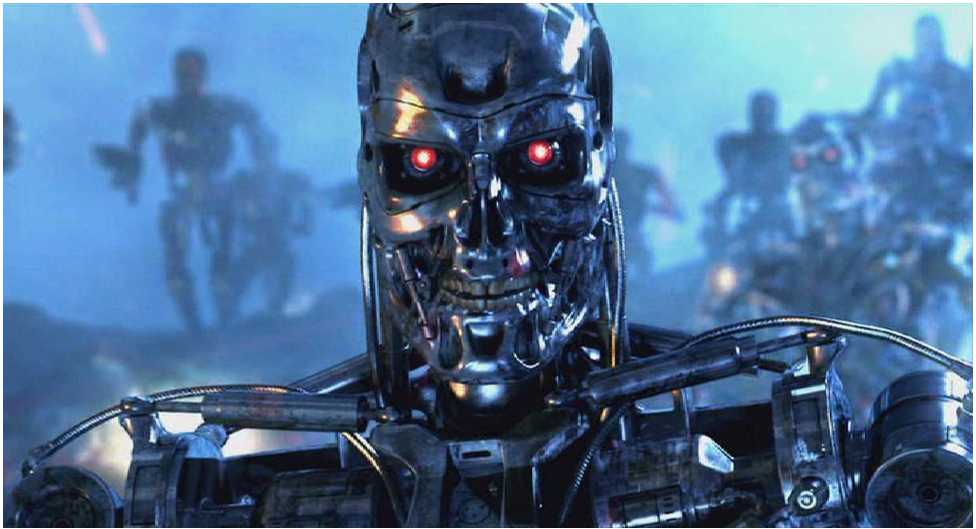
MBA em Ciência de Dados
Universidade de Fortaleza

27 de Março de 2021

FUNDAÇÃO EDSON QUEIROZ
UNIVERSIDADE DE FORTALEZA
ENSINANDO E APRENDENDO

* erneson@unifor.br

Aula 3 - Desafios de Aprendizado de Máquina



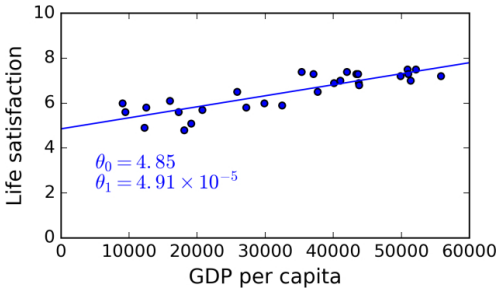
Exemplo

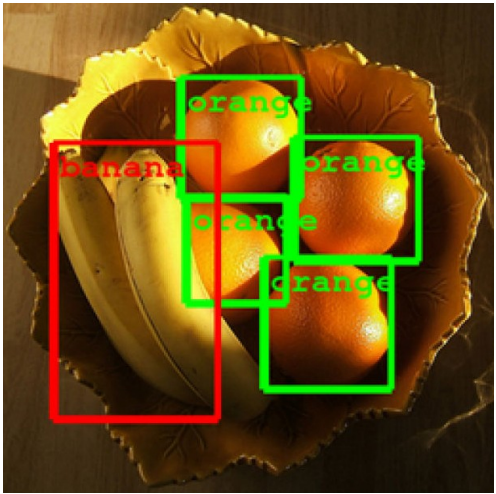
Como prever preços de habitações?

Quais são os principais desafios em AM?

Quantidade insuficiente de dados no conjunto de treinamento

Quantidade insuficiente de dados no conjunto de treinamento





Problemas simples $\times$ Problemas complicados

Quantidade insuficiente de dados no conjunto de treinamento

Scaling to Very Very Large Corpora for Natural Language Disambiguation

Michele Banko and Eric Brill
Microsoft Research
1 Microsoft Way
Redmond, WA 98052 USA
{mbanko,brill}@microsoft.com

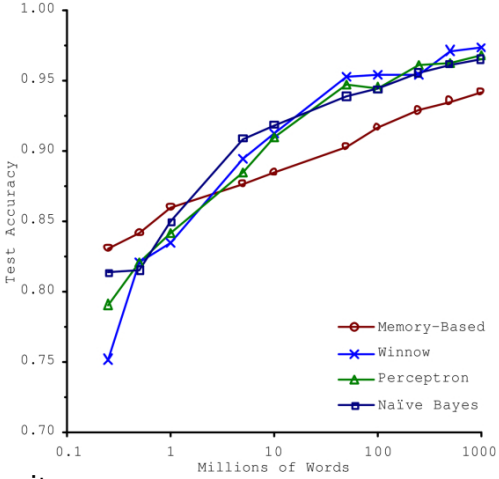
Abstract

The amount of readily available on-line text has reached hundreds of billions of words and continues to grow. Yet for most core natural language tasks, algorithms continue to be optimized, tested and compared after training on corpora consisting of only one million words or less. In this paper, we evaluate the performance of different learning methods on a prototypical natural language disambiguation task

standardization of data sets used within the field, as well as the potentially large cost of annotating data for those learning methods that rely on labeled text.

The empirical NLP community has put substantial effort into evaluating performance of a large number of machine learning methods over fixed, and relatively small, data sets. Yet since we now have access to significantly more data, one has to wonder what conclusions that have been drawn on small data sets may carry over when these learning methods are trained using much larger corpora.

In this paper, we present a study of the



Dados $\times$ Algoritmos

Quantidade insuficiente de dados no conjunto de treinamento

EXPERT OPINION

Contact Editor: **Brian Brannon**, bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"[1] examines why so much of physics can be neatly explained with simple mathematical formulas such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that involve human beings rather than elementary particles have proven more resistant to elegant mathematics. Economists suffer from physics envy over their inability to neatly model human behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

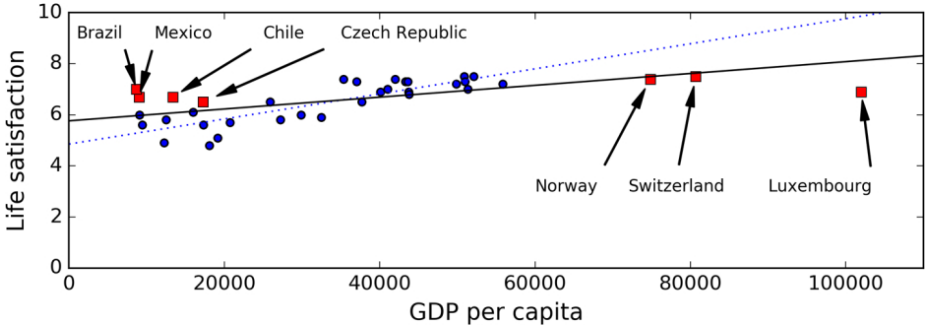
Learning from Text at Web Scale

The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation. The reason for these successes is not that these tasks are

Dados são mais importantes do que algoritmos?

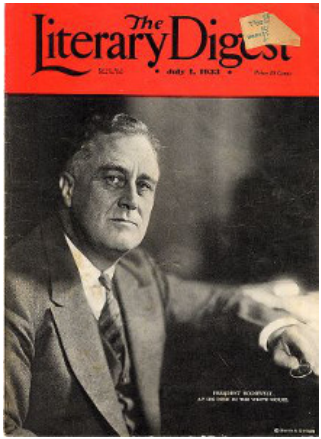
Dados de treinamento não-representativos

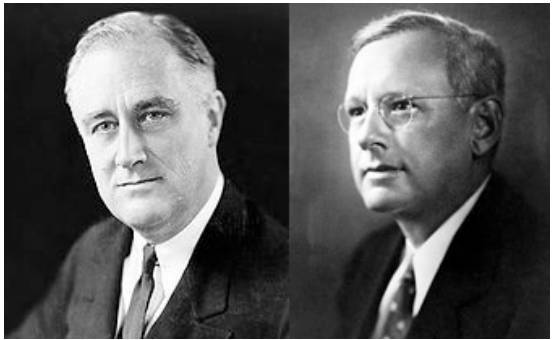
Dados de treinamento não-representativos



Dados de treinamento devem ser representativos para novos casos (ruído na amostragem)!

Dados de treinamento não-representativos





Eleições dos EUA (1936): Roosevelt $\times$ Landon

- Viés na amostragem e viés de não-resposta.

Qualidade baixa dos dados

Qualidade baixa dos dados

Modelo	**Quilometragem (km)**	**Ano**	**Marca**	...	**Valor (R$)**
COROLLA	1.000	2019	Toyota	...	60 000 000
Fusquinha	100 000	1984	V0lkswagem	...	2 000
Onix 2017	15,000	2017	Chevrolet	...	30 000
⋮	⋮	⋮	⋮	⋮	⋮
Duster	10 000	2218	Renault	...	45.000

- Mineração, higienização e padronização dos dados.

Características irrelevantes

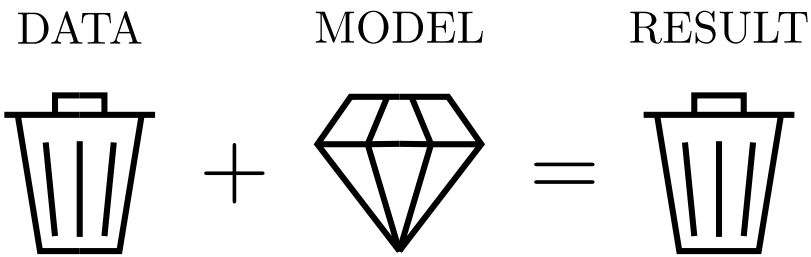
Características irrelevantes

Modelo	**Quilometragem (km)**	~~**Ano**~~	**Marca**	...	**Valor (R$)**
Corolla	1 000	~~2019~~	Toyota	...	60 000
Fusca	100 000	~~1984~~	Volkswagem	...	2 000
Onix	15 000	~~2017~~	Chevrolet	...	30 000
⋮	⋮	⋮	⋮	⋮	⋮
Duster	10 000	~~2018~~	Renault	...	45 000

- Engenharia de características: Seleção, extração, criação de características.

Em suma, para dados...

Em suma, para dados...

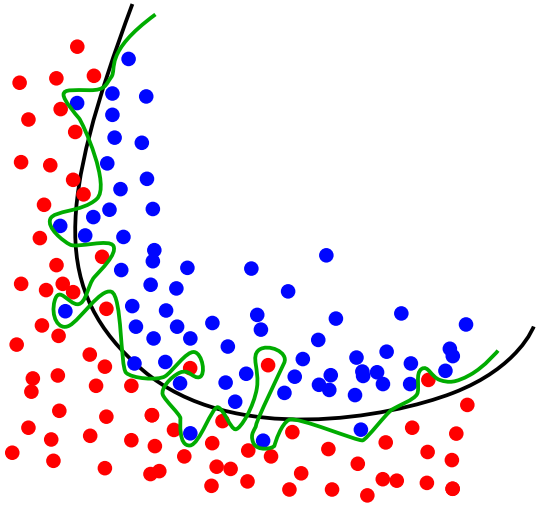


Garbage in, garbage out!

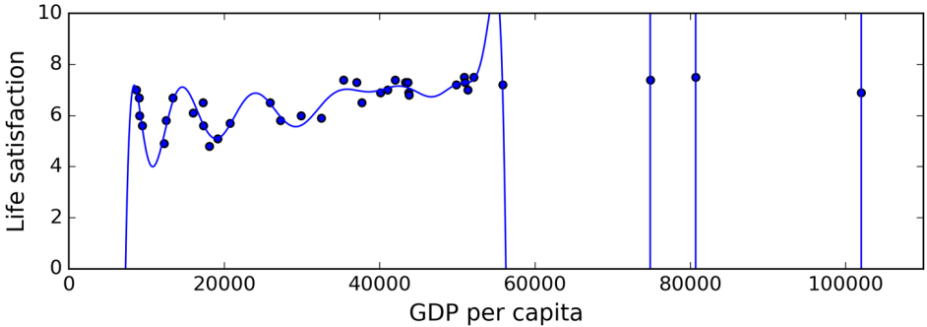
Sobreajuste do conjunto de treinamento

Sobreajuste do conjunto de treinamento

- Sobreajuste (ou *Overfitting*): Acontece quando o SAM tem bom desempenho apenas no conjunto de treinamento, *i.e.* quando o SAM é muito complexo em relação ao conjunto de treinamento.



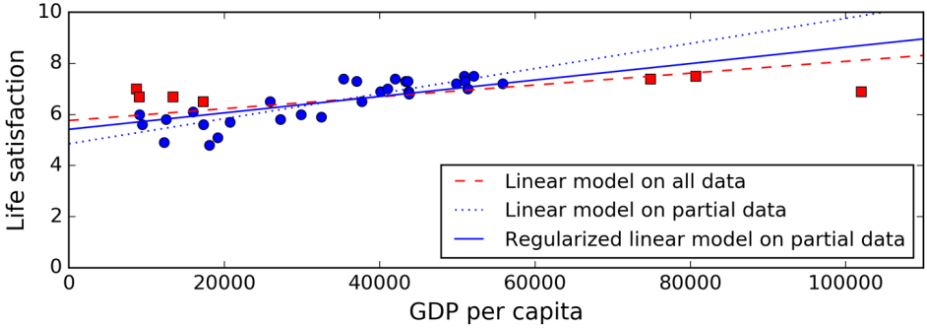
Sobreajuste do conjunto de treinamento



Sobreajuste do conjunto de treinamento!

Sobreajuste do conjunto de treinamento

- Regularização: Imposição de restrições, controladas pelos hiperparâmetros, ao AA para tornar o SAM mais simples.



Regularização reduz os riscos de sobreajustes!

Sobreajuste do conjunto de treinamento

Como resolver o sobreajuste?

Sobreajuste do conjunto de treinamento

1. Utilizar um modelo mais simples;

Sobreajuste do conjunto de treinamento

1. Utilizar um modelo mais simples;
2. Reduzir o número de atributos;

Sobreajuste do conjunto de treinamento

1. Utilizar um modelo mais simples;
2. Reduzir o número de atributos;
3. Restrigir o modelo;

Sobreajuste do conjunto de treinamento

1. Utilizar um modelo mais simples;
2. Reduzir o número de atributos;
3. Restrigir o modelo;
4. Adquirir mais dados;

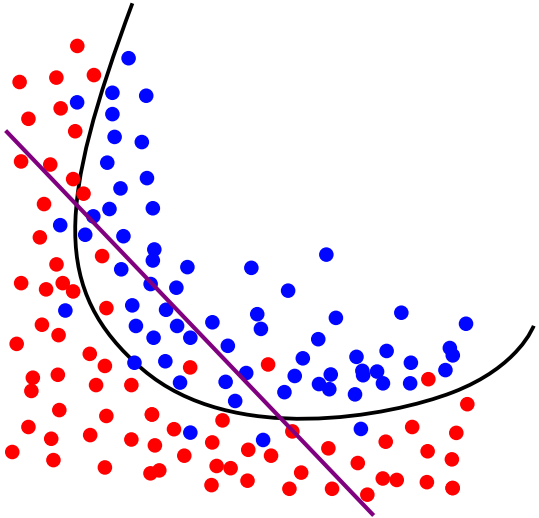
Sobreajuste do conjunto de treinamento

1. Utilizar um modelo mais simples;
2. Reduzir o número de atributos;
3. Restrigir o modelo;
4. Adquirir mais dados;
5. Reduzir o ruído do conjunto de treinamento.

Subajuste do conjunto de treinamento

Subajuste do conjunto de treinamento

- Subajuste (*Underfitting*): Acontece quando o SAM não tem bom desempenho nem mesmo no conjunto de treinamento, *i.e.* quando o SAM é muito simples em relação ao conjunto de treinamento.



Subajuste do conjunto de treinamento

Como resolver o subajuste?

Subajuste do conjunto de treinamento

1. Utilizar um modelo mais poderoso;

Subajuste do conjunto de treinamento

1. Utilizar um modelo mais poderoso;
2. Escolher atributos melhores;

Subajuste do conjunto de treinamento

1. Utilizar um modelo mais poderoso;
2. Escolher atributos melhores;
3. Reduzir o número de restrições.

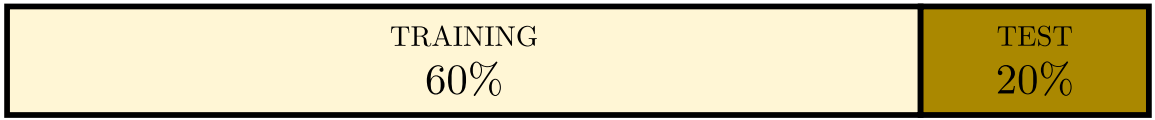
Conjunto de teste e conjunto de validação

Conjunto de teste e conjunto de validação

- Conjunto de treinamento: Dados usados para ajustar os parâmetros do SAM;
- Erro de treinamento.

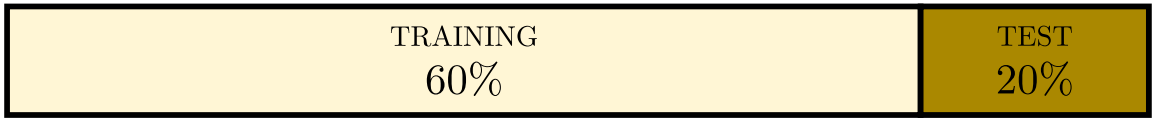
Conjunto de teste e conjunto de validação

- Conjunto de teste: Dados usados para fazer a avaliação final do SAM;
- Erro de generalização.



Conjunto de teste e conjunto de validação

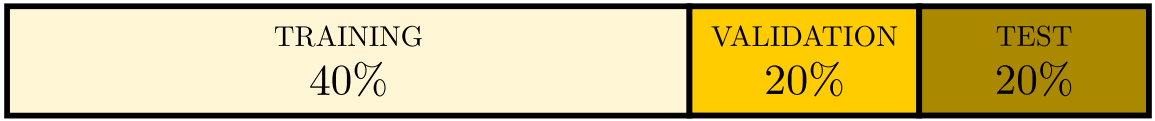
- Conjunto de teste: Dados usados para fazer a avaliação final do SAM;
- Erro de generalização.



- Se o erro de treinamento é baixo e erro de generalização é alto, então o SAM está sobreajustado.

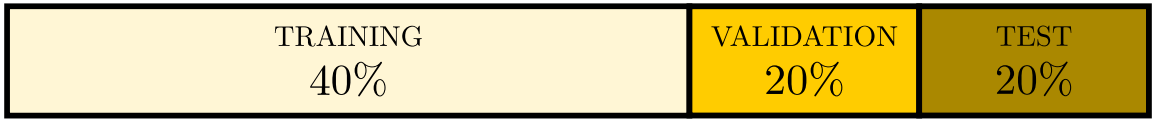
Conjunto de teste e conjunto de validação

- Conjunto de validação: Dados usados para ajustar os hiperparâmetros do AA e fazer avaliações intermediárias do SAM;
- Erro de validação.



Conjunto de teste e conjunto de validação

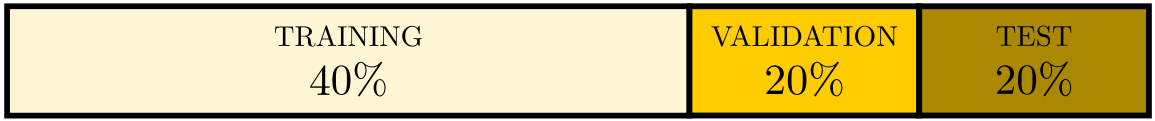
- Conjunto de validação: Dados usados para ajustar os hiperparâmetros do AA e fazer avaliações intermediárias do SAM;
- Erro de validação.



- Se o erro de treinamento é baixo e erro de validação é alto, então o SAM está sobreajustado;

Conjunto de teste e conjunto de validação

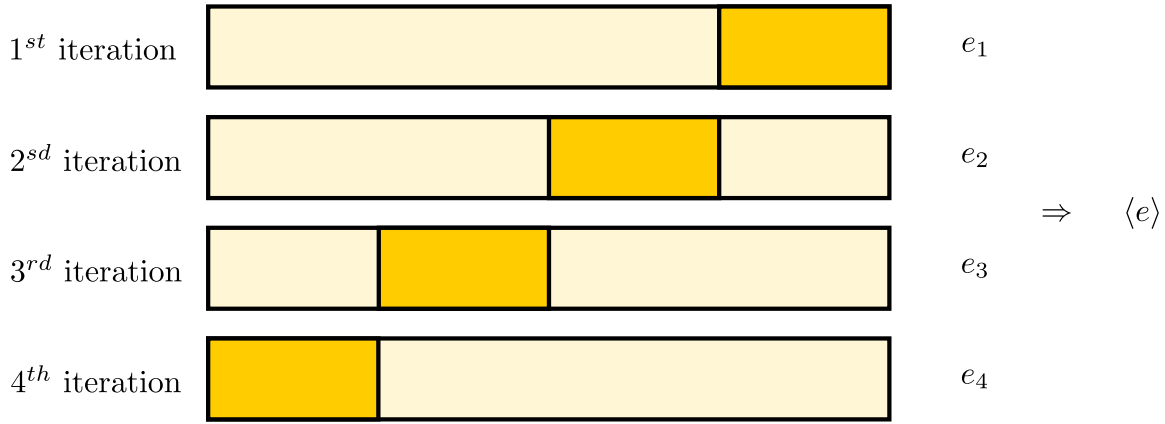
- Conjunto de validação: Dados usados para ajustar os hiperparâmetros do AA e fazer avaliações intermediárias do SAM;
- Erro de validação.



- Se o erro de treinamento é baixo e erro de validação é alto, então o SAM está sobreajustado;
- Se os erros de treinamento e validação são baixos e erro de generalização é alto, então o SAM está sobreajustado.

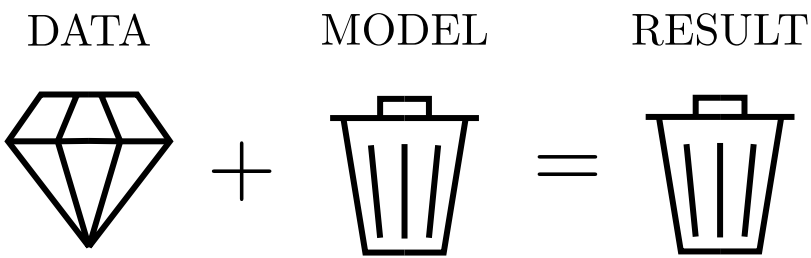
Conjunto de teste e conjunto de validação

- Validação cruzada: Avalia a capacidade de generalização de uma análise estatística a partir de um conjunto de dados (*e.g.* k-*Fold*).



Em suma, para modelos...

Em suma, para modelos...



Garbage in, garbage out!


Hasta la vista, baby!