

Inteligência Computacional - Denoising Dirty Documents

Bernardo D. C. G. de Amorim & Pedro V. Nacif

14 de Outubro de 2015

1 Introdução

OBS: As sub-seções do documento são preliminares e baseadas no modelo de relatório e devem mudar ao longo do projeto.

1.1 O Problema

O problema do denoising (remoção de ruído) de imagens consiste de desenvolver algoritmos que “limpem” imagens deterioradas por uma série de fatores, desde o processo de escaneamento até manchas de fluidos e dobras no papel.

Nossa tarefa é limpar imagens: remover as manchas; remover as dobras de papel; melhorar o contraste; e deixar somente o texto, o mais legível possível. Temos uma série de imagens de treinamento, constituídas de uma imagem “limpa” e uma onde um ruído foi adicionado artificialmente. No nosso caso, queremos usar aprendizado de máquina para gerar um algoritmo que consiga limpar essas imagens, treinando-o de alguma forma com as imagens dadas.

1.2 Bibliografia

Blog do Colin Priest¹, segundo colocado no contest do Kaggle²

1.3 Dados

O dataset³ consiste de cerca de 200 imagens de treinamento sujas e a mesma quantidade de treinamento limpas entre 40 e 50 Mb, em formato png.

Um dos maiores desafios desse problema, assim como em grande parte dos problemas de aprendizado de máquina são:

- Definição do tipo de problema: Define qual vai ser a saída. Será um problema de regressão que dirá o valor para cada pixel da imagem de saída? Será um problema de classificação, que usará cores fixas para aumentar o contraste? Podemos tomar diversos caminhos; um dos mais simples é a regressão, onde a saída é o brilho de cada pixel.
- Extração de Características: A imagem só nos dá a informação base, isto é, o valor de brilho para cada pixel. Podemos usar isto diretamente como uma característica para treinar e utilizar nosso algoritmo, isto é: determinar o valor do brilho esperado a partir do brilho dado. Podemos pensar um pouco a mais, como tentar extrair informações dos arredores de cada pixel, como o brilho de cada pixel num raio qualquer, ou a variancia dos pixels num outro raio, ou qualquer outra informação que possa ser extraída e que tenha algum valor para “entender” a imagem.

Entretanto, deve-se tomar cuidado para não adicionar muitas features sem pensar sobre o que é realmente útil para o modelo funcionar, e por isso vamos ter que explorar cautelosamente como analisar os dados.

¹<http://colinpriest.com/>

²<https://www.kaggle.com/>

³Obtido em: <https://www.kaggle.com/c/denoising-dirty-documents>

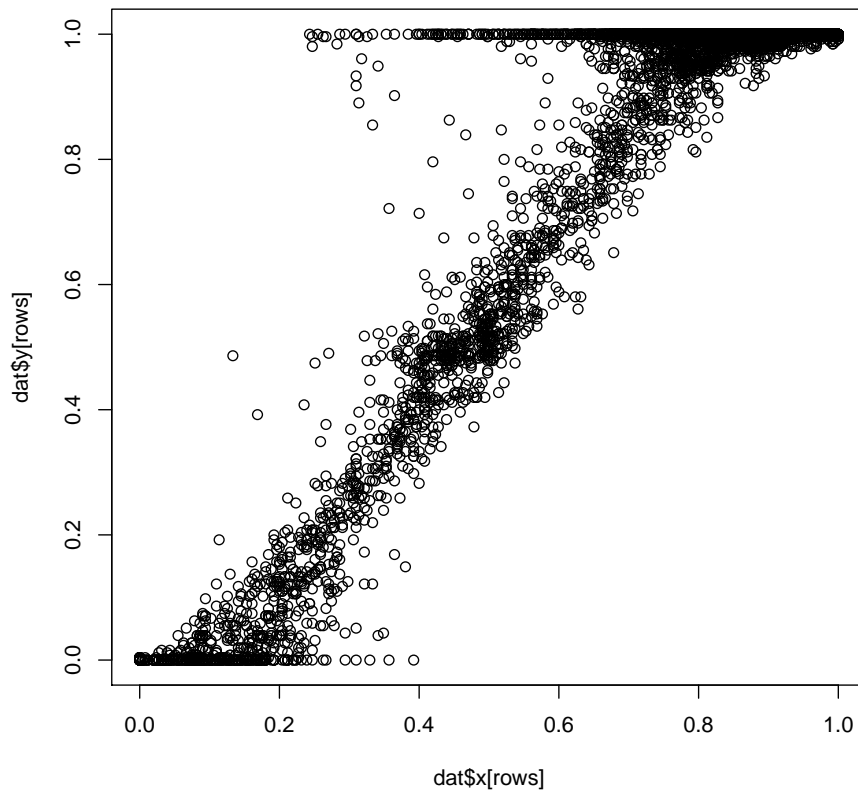
2 Tecnologia

A ferramenta selecionada para realizar este projeto é a linguagem R⁴, que é uma implementação da linguagem de programação de estatística S.

A decisão por esta ferramenta está baseada no fato de ser um projeto de código aberto e de software livre; por conter, já incluso na linguagem, diversas ferramentas estatísticas; e por ter um repositório de pacotes (CRAN⁵) com diversos pacotes de aprendizado de máquina e de visualização de dados.

2.1 Avaliação Preliminar dos Dados

Para começar com o modelo mais simples, vamos utilizar o modelo onde realizamos uma regressão tentando relacionar o brilho da imagem desejada com o brilho da imagem com ruído, ou seja, um problema de $y = f(x)$, onde x é o brilho da imagem com ruído e y é o brilho da imagem limpa. Ao olhar a relação entre todos os x e todos os y de todas as imagens, temos a seguinte relação:



Ou seja, para grande parte dos casos, podemos representar $y = \alpha x + \beta$, obtendo o α e o β a partir de um algoritmo de regressão linear. Entretanto, isto não é verdade para todas as imagens e todos os pixels, portanto devemos

⁴<https://www.r-project.org/>

⁵<https://cran.r-project.org/>

pensar em modelos diferentes, ou numa combinação linear de potencias de x , ou até mesmo adicionando outras características como variáveis a nossa função de regressão: $y = f(a, b, c, d, e, \dots)$.

3 Metodologia

3.1 Descrição da Metodologia

3.2 Descrição da Solução do Problema Proposto

3.3 Resultados Preliminares

4 Resultados