Projeto Final - Aprendizagem por Máquinas

Eduardo F. Mendes e Leonardo Pinheiro

October 22, 2015

1 Objetivo

O principal objetivo deste trabalho é avaliar a habilidade de utilizar técnicas de machine learning para solucionar um problema real. Para isso, utilizaremos a competição do kaggle **"Classify handwritten digits using the famous MNIST data"**. A descrição da competição é apresentada abaixo e também pode ser lida no website Kaggle: Digit Recognizer

Os alunos deverão se dividir em grupos de 2 ou 3 (não serão aceitos trabalhos individuais exceto em casos específicos). O trabalho será dividido em duas partes: competição e relatório. A primeira é uma competição interna, entre os alunos da turma, para ver quem consegue a melhor performance preditiva do modelo. A segunda parte, que corresponde a maior parte da nota, é um relatório descrevendo técnicas, modelo e o caminho feito até chegar lá. O grupo será arguido sobre o relatório e a nota será dada baseada na arguição.

1.1 Descrição (retirada do site da competição)

The goal in this competition is to take an image of a handwritten single digit, and determine what that digit is. As the competition progresses, we will release tutorials which explain different machine learning algorithms and help you to get started.

The data for this competition were taken from the MNIST dataset. The MNIST ("Modified National Institute of Standards and Technology") dataset is a classic within the Machine Learning community that has been extensively studied. More detail about the dataset, including Machine Learning algorithms that have been tried on it and their levels of success, can be found at http://yann.lecun.com/exdb/mnist/index.html.

1.2 Dados (retirado do site da competição)

URL: https://www.kaggle.com/c/digit-recognizer/data

The data files train.csv and test.csv contain gray-scale images of hand-drawn digits, from zero through nine.

Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive.

The training data set, (train.csv), has 785 columns. The first column, called "label", is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image.

Each pixel column in the training set has a name like pixelx, where x is an integer between 0 and 783, inclusive. To locate this pixel on the image, suppose that we have decomposed x as $x = i * 28 + j$, where i and j are integers between 0 and 27, inclusive. Then pixel x is located on row i and column j of a 28 x 28 matrix, (indexing by zero).

The test data set, (test.csv), is the same as the training set, except that it does not contain the "label" column.

2 Avaliação

O procedimento de avaliação é descrito abaixo:

1. A nota total dada será entre 0-100*(número de integrantes) e grupo será responsável por determinar a divisão de pontos entre os membros;

2. A nota possui dois componentes: performance no ranking público do kaggle e um relatório sobre a condução do trabalho e resultados alcançados.

3. O componente de performance corresponderá a 30% da nota. Para definir o valor exato dos pontos, olharemos para a distribuição do poder preditivo dos envios ao kaggle, porém centrada na mediana da turma (manteremos a variância). O poder preditivo alcançado deverá ser superior ao *benchmark* apresentado pelo professor.[1]

[1]Construiremos uma distribuição normal truncada, onde o truncamento inferior será a performance do benchmark e o superior será maior performane dentre os grupos. A nota do grupo i será dada por $nota_i = 30\,P(Z_{[a,b]} < performance_i) \times no\_integrantes_i$, onde $A_{[a,b]}$ possui a distribuição normal truncada descrita acima.

4. O componente de Relatório corresponderá a 70% da nota e deverá conter os seguintes elementos:

   (a) Descrição do processo de *feature engineering* e de quais *features* from criadas e/ou transformadas;

   (b) Descrição de quais algoritmos foram utilizados e quais levaram a melhor ganho de performance;

   (c) Descrição do processo de seleção de modelo e seleção de parâmetros;

   (d) Descrição do modelo final e variáveis utilizadas;

   (e) Resultado in-sample e out-of-sample; e

   (f) Referências utilizadas na construção dos modelos.

   Haverá ainda uma arguição oral com relação ao resultado do relatório com cada grupo.

3 Prazos

- Submissão de resultados ao kaggle: 4/12/2015.
- Entrega do relatório final: 6/12/2015 (online).
- Arguição Oral: Será sorteado dia 6/12/2015.