6.867: Homework 3

Anonymous authors

***Abstract*— This is Machine Learning homework 3. Topics include neural networks, conv nets, and LDA. All work has been done in Python.**

I. Convolutional Neural Network (CNN)

In this section, we consider the Convolutional Neural Network (CNN) to perform artist identification on paintings.

A. Part 1

In a convolutional filter, if the first layer applies a 5x5 patch to the image to generate feature Z_1, and the second layer applies a 3x3 patch to feature Z_1 to generate feature Z_2, the receptive field of Z_2 (or dimensions of image that affect the node) is 7x7. That is, a window of 49 neighboring pixels in the original image affects a single node at the output of the filter. This allows the network to learn spatial features from the original image. If the conv net becomes deeper (more layers), the network can use larger and more complex combinations of features/regions of the image.

B. Part 2

We are provided with a conv net (conv.py). In total, there are [todo] layers: 2 convolutional layers, 1 flatten layer, and 2 dense layers. The output is the maximum logit of the final dense layer. [todo] confirm activation function. The hidden layer activation function is relu. The loss function is softmax cross entropy with logits. Loss is minimized with Gradient Descent (with a tunable learning rate parameter).

The provided network took about 45 seconds to train on a Macbook CPU. After 1500 steps, the training accuracy is 87.4%, and the validation accuracy is 57.5%. These numbers suggest overfitting, because the model does not generalize to unseen data very well.

C. Part 3

II. LDA

In this section, we use latent Dirichlet allocation (LDA) to perform topic modeling on a corpus of text.

A. Part 1

The first part is very straightforward. We installed the Mallet implementation of LDA and ran it with the default parameters. The corpus of text is 19,997 news articles in 20 categories from the Usenet newsgroups dataset.

B. Part 2

During training, the model discovered 100 topics, which include several keywords related to the topic. Two example topics are:

- health cancer disease medical aids hiv number research patients page april volume newsletter study hicnet drug children risk years age
- image data graphics package images program processing line format ray code objects analysis points center it's postscript sgi formats ftp

The first topic seems to be related to medicine/health and the second is related to image processing/computers. The topics aren't extremely specific, but each component seems roughly connected. However, if a human were to develop a list of keywords related to health, it probably would not include words like "newsletter" or "april", so there are some strange components of the topic lists.

Of the 20 article categories, we chose to analyze two in more detail: "sci.med" and "rec.sport.hockey", which include articles about Medicine and Hockey, respectively. The top topics related to Medicine were:

1) article pain doctor writes gordon banks patients good surgery treatment disease patient medicine blood intellect chronic skepticism surrender shameful chastity
2) food msg eat diet article people fat writes foods chinese body eating effects taste reaction i'm effect apr steve meat
3) candida yeast medical patients medicine symptoms treatment infection quack infections body steve doctors article clinical disease writes studies nutrition jon

and the top topics related to Hockey were:

1) period game play goal pts power shots puck blues goals flyers leafs team detroit win wings scoring penalty season goalie
2) game games team win espn won year lost baseball fans boston division series hockey teams chicago pittsburgh fan season cubs
3) team game players hockey writes play player article roger games apr league baseball good time year teams bob nhl ice

At first glance, these topics seem appropriate for the categories. Medicine's top topics roughly relate to medicine, nutrition, and disease, whereas hockey's top topics include rules of the game and team names. Upon closer look, there are some bizarre elements of the chosen topics, like "i'm" and "jon" for medicine, and "bob" and "apr" for hockey. But, of the 60 keywords for hockey, only about 3 of them are ridiculous whereas the other 57 are very "on-topic."