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. Neural Networks

In this problem, we implement a simple neural network in Python. The network is trained with backpropogation and stochastic gradient descent.

A. Part 1

In this problem, we are mainly interested in classification tasks. Thus, we pass the output of the final hidden layer through a softmax layer to generate a probability (outputs are positive and sum to 1) vector where each element k is the probability of the input being in class k. While ReLU has a simple activation function and derivative ($f(z) = max(0, z)$ and $f'(z) = 1[x > 0]$), the softmax activation and derivative are slightly more complicated ($f(z)_i = \frac{e^{z_i}}{\sum e^{z_j}}$ and $f'(z)_i = [todo]$).

II. 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.

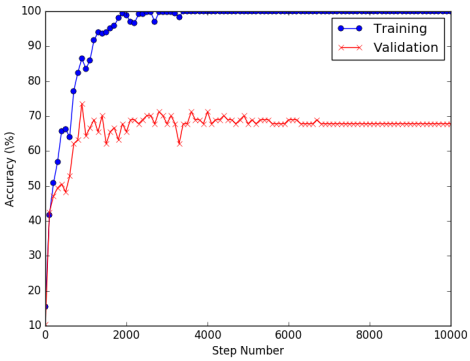

Fig. 1: Training and validation accuracy are plotted over 10,000 training steps. Model performance levels off after 1,000 training steps.

C. Part 3

Next, we try two common techniques to improve CNN performance. The first is early stopping. Early stopping is when training is stopped after validation accuracy levels out, before it starts to decrease, to avoid overfitting. In this scenario, Fig. 1 shows that early stopping does not make much difference, because the validation accuracy levels off but doesn't decline. It does have the benefit of shorter training time, though.

[todo] pooling layers performance seems to make no difference.

D. Part 4

Finally, we use the network on a transformed version of the dataset. Results in Table I show that the performance is not the same for every transormation type. For example, the original CNN only gets 10% accuracy on inverted images, compared to 66.7% on low contrast images (all relative to a 70.1% accuracy on normal images).

Transformation	Validation Acc (%)
Normal	70.1
Translated	29.9
Brightened	47.1
Darkened	49.4
High Contrast	63.2
Low Contrast	66.7
Flipped	41.4
Inverted	10.3

TABLE I: Accuracy of CNNs on transformed dataset.

III. 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 on the corpus of text of 19,997 news articles in 20 categories from the Usenet newsgroups dataset.

B. Part 2

During training, we asked the model to discover 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."

C. Part 3

Then, we lowered the number of topics down to 20. We hypothesized that this would lead to broader topics, because the same number of articles and words now would need to be binned into a smaller number of categories. Interestingly, the number of article classes is now equal to the number of topics.

Two of the resulting topics were:

- writes article people don't objective moral truth true question evidence god point good i'm morality it's religion exist claim reason
- israel president jews israeli writes jewish war article people arab state peace stephanopoulos arabs world policy land years country american

These topics seem slightly more broad than before, but it's hard to measure/quantify this concept.

The top topics most related to Medicine were:

1) article medical writes disease study health patients food cancer number treatment men msg medicine research doctor sex drug aids studies
2) don't it's people i'm good make time writes article that's i've things thing can't doesn't lot point you're find work
3) writes article theory science fact universe evolution bill book university black water time don't apr michael physical scientific hole books

and the top topics most related to Hockey were:

1) game team year games writes players article play hockey season baseball win good league player apr teams hit fans time
2) don't it's people i'm good make time writes article that's i've things thing can't doesn't lot point you're find work
3) san period convention pit party institute pts political adl det van april bos cal los committee college state vote karl

The first topic for each article category still seems appropriate, but the second and third are not very related. This result matches with our original intuition that that there would be fewer specific topics to pair with each article type as the number of topics decreases.