Machine Learning to Deep Learning

Palash Chauhan

May 9, 2017

1 Multinomial Logistic Classification

1.1 Softmax

- Scores → probabilities
- Multiply by 10 → close to 0/1
- Divide by 10 → close to uniform

1.2 Cross Entropy

- $D(S, L) = -\sum_i L_i log(S_i)$
- L are true one hot labels, S are output of softmax from the model
- Minimize average cross entropy (loss) w.r.t parameters and biases to learn

1.3 Numerical Stability

- Loss function should never get too big or too small
- We want variables to aleays have 0 mean and equal variances
- For images (0-255), subtract 128 and divide by 128
- **Initialization**: Draw weights and biases from a gaussian with mean μ and small variance σ.

1.4 Measuring Performance

- Train ,Test, Validation
- Use a validation set to prevent overfitting on test set
- A change that affects 30 examples in the validation set is significant and can be trusted
- Therefore, validation set should be greater than 30K examples. Accuracy figures are then significant to the first decimal place ($> 0.1\%$)
- These heuristics are true only if classes are balanced. Otherwise, get more data!

1.5 SGD

- Normal GD has scaling issues
- Calculate the estimate of the loss using some random batch of data and use this to get gradients
- Scales well both with data and model size
- **Momentum**: Keep a running average of the gradients ($M \leftarrow 0.9M + \Delta L$) and use this instead of the current batch average.
- **Learning Rate Decay**: Make the steps smaller and smaller as you train (eg. exponential decay)

Classifier	Accuracy
Logistic Regression	0.85
3-NN	0.88
SVM	0.9
Decision Tree	0.757
Random Forest	0.748
Adaboost	0.793
GaussianNB	0.81
QDA	0.6

Table 1: Accuracies using various shallow classifiers

1.6 Parameter Hyperspace

- Many many hyperparameters to select - Initial learning rate, learning rate decay, momentum, batch size, weights initialization etc
- **KEEP CALM and LOWER your LEARNING RATE**
- **AdaGrad**: Modification of SGD, implicitly does momentum and learning rate decay and makes models less sensitive to hyperparameters

2 Assignment-1

2.1 Dataset

- notMNIST dataset of alphabets A-J in various fonts, tougher dataset than MNIST
- A subset (8000 train. 1000 test) evaluated using logistic regression and other classifiers present in sklearn with their default settings. Refer to Table 1