On Parameter Tying by Quantization

Vibhav Goagate, Samdeb Sarkhel, Nick Ruozzi, Li Chou - University of Texas at Dallas, 2016

I. Introduction

A major problem in machine learning is over-fitting. The paper proposes a solution to this problem, using quantization. The experiments show that the performance is better when there are a large number of feature, compared to the number of training data points.

II. Notation

x=$\{x_1, x_2, \ldots, x_n\}$ denote the variables. $\bar{x} = \{\bar{x_1}, \ldots, \bar{x_n}\}$ denote an assignment of values to the variables and y denote the target values.

Let $g_\theta : \mathbb{R}^n \rightarrow \mathbb{R}$ be a function that maps the input vector x to the Real space. For example, for a linear decision boundary, $g_\theta(x) = \theta^T x + b$ where $\theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$ are the parameters to be learnt and b is the bias.

Let D be the number of samples in the training data denoted by $\bar{X}_{D \times n}$, the rows of which be denoted by $\bar{x}^{(i)}$, $1 \leq i \leq D$. Let $\bar{Y}$ be a column vector of the target values $\{\bar{y}^{(1)}, \ldots, \bar{y}^{(D)}$

Define the cost function $J : \mathbb{R}^m \rightarrow \mathbb{R}$. For example, for Logistic Regression it will be

$$J(\theta) = \sum_{i=1}^{D} \bar{y}^{(i)} log(\sigma_\theta(\bar{x}^{(i)}) + (1 - \bar{y}^{(i)}) log(1 - \sigma_\theta(\bar{x}^{(i)})$$

where $\sigma_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$

III. The algorithm

- We start off by learning the weights using any method and any optimization technique. For example, we may use Logistic Regression with any standard minimization algorithm like BFGS.
- The weights so obtained may fit the input data very well, however these may be prone to over-fitting.
- Regularization may be used to overcome this problem, instead we impose certain equality constraints (through equivalence relations) on the parameters and re-learn the model.
- The basic idea is to group together similar weights and make them all equal.
- The paper uses one dimensional k-means clustering on the parameters. k is not learned but is specified apriori.
- This clustering will give us k classes of parameters that are tied together.
- An equality constraint is imposed on these parameters so that all parameters in the same cluster have equal weight.
- The model relearned using any method and optimization technique but this time using these constraints.

In a more formal way, one dimensional k means clustering on the m parameters, where k is specified apriori (i.e. k is not learnt). This can be achieved in $O(m^2 k)$ time using dynamic programing (Wang and Song, 2011).

Let $S_1, S_2, \ldots, S_k$ be the k clusters thus obtained. Define a quantization $\mu = \{\mu_1, \mu_2, \ldots, \mu_k\}$ of θ and $\mathbb{Q} : \bar{\theta} \rightarrow \mu$ be the quantizer between x and μ, so that $\mathbb{Q}(\theta_j) = \mu_i$ iff $\theta_j \in S_i$. Further, let the notation $\mathbb{Q}(\theta)$ denote the m dimensional vector $\{\mathbb{Q}(\theta_1), \mathbb{Q}(\theta_2), \ldots, \mathbb{Q}(\theta_m)\}$ A new cost function $J_2(\theta)$ can then be defined based on these constraints. For example for Logistic Regression:

$$J_2(\theta) = \sum_{i=1}^{D} \bar{y}^{(i)} log(\sigma_{\mathbb{Q}(\theta)}(\bar{x}^{(i)}) + (1 - \bar{y}^{(i)}) log(1 - \sigma_{\mathbb{Q}(\theta)}(\bar{x}^{(i)})$$

The k parameters $\{\mu_1, \ldots, \mu_k\}$ are to be learnt so that $J_2(\theta)$ is minimized and this can be done using any standard minimization technique like BFGS. The initial values for $\{\mu_1, \ldots, \mu_k\}$ can be taken as the means of clusters $S_1, \ldots, S_k$.

It is better to normalize the data appropriately so that the exponential term in the sigmoid function does not become very large or very small. An easy way to normalize is to divide $\theta^T \bar{x}^{(i)}$ by $max_{1 \leq i \leq D}(\theta^T \bar{x})$.

IV. Experiments and Results

The model was tested on 4 datasets - bn2o (20k parameters), students (1300 parameters), grid (1100 parameters), friends(4k parameters). The accuray increase by approximately 11.7% for bn2o, 0.7% for students, 2.7% for grid, 12.5% for friends.

V. References

Optimal k-means Clustering in ONe Dimension by Dynamic Programming - Wang and Song. 2011. The R Journal 3(2):29-33

On Parameter Tying by Quantization - Gogate, Ghou, Sarkhel, Ruozzi, 2016, University of Texas at Dallas