Machine Learning Notes

Nhan Trong

July 4, 2016

> Detective Spooner. *Can a robot write a symphony? Can a robot turn a canvas into a beautiful masterpiece?*
> Robot. *Can you?*
>
> I, Robot

Note 1. Refer to assignment PDF's. We'll use the usual subscript indexing notation instead of superscript like the lecture.

Part I
Ex 8. Anomaly Detection and Recommender Systems

Keywords. Collaborative filtering, cost function, gradient, regularization, null reduction, Frobenius inner product.

1 Collaborative Filtering Learning Algorithm

Let n_m be the number of movies, n_u be the number of users. Given rating matrix Y and a number n, we want to find a feature matrix X of size $n_m \times n$ and parameter matrix Θ of size $n_u \times n$, where the i-th row of X represents the feature vector for the i-th movie, and the j-th row of Θ represents the parameter vector for the j-th user. In this context, n represents the number of hidden dimensions of a movie, e.g. X_{ik} could refer to say how much action movie i has, X_{il} could refer to how much romance it has, and so on. Similarly, Θ_{jk} would refer to how much user j likes action, Θ_{jl} how much they like romance.

Note 2. These are only example features, since in fact we don't know what features the algorithm will pick up given rating matrix Y. The features learned might have nothing to do with common movie genres, for example.

Question 3. *Can we cross validate to choose the best value n for the number of hidden features?*

2 Cost Function and Gradient

Definition 4. Define the collaborative filtering cost function to be the squared error over all parameters Θ and features X:

$$J(X,\Theta) = \frac{1}{2}\sum_{i,j:R_{ij}=1} (\Theta_j \cdot X_i - Y_{ij})^2.$$

Then the partial derivatives of J with respect to X and Θ are:

$$\frac{\partial J}{\partial X_{ik}} = \sum_{j:R_{ij}=1} (\Theta_j \cdot X_i - Y_{ij})\Theta_{jk}$$
$$\frac{\partial J}{\partial \Theta_{jk}} = \sum_{i:R_{ij}=1} (\Theta_j \cdot X_i - Y_{ij})X_{ik}.$$

The vectorized forms are surprisingly simple:

$$D \stackrel{\text{def}}{=} R * (X\Theta^T - Y)$$
$$J = \frac{1}{2} D \cdot D$$
$$\frac{\partial J}{\partial X} = D\Theta$$
$$\frac{\partial J}{\partial \Theta} = D^T X,$$

where $\cdot$ denotes the Frobenius inner product (just a natural extension of the vector dot product to matrices), and $*$ denotes element-wise multiplication. We need to multiply element-wise by R to reduce $X\Theta^T - Y$ to elements where the corresponding entries in Y are nonzero, because the summation is only over i, j such that $R_{ij} = 1$, i.e. where Y_{ij} is nonzero.

3 Cost Function and Gradient with Regularization

With regularization, the cost function and partials are:

$$
\begin{aligned}
D &\stackrel{\text{def}}{=} R * (X\Theta^T - Y) \\
J &= \frac{1}{2} D \cdot D + \frac{\lambda}{2} X \cdot X + \frac{\lambda}{2} \Theta \cdot \Theta \\
\frac{\partial J}{\partial X} &= D\Theta + \lambda X \\
\frac{\partial J}{\partial \Theta} &= D^T X + \lambda \Theta.
\end{aligned}
$$

4 Some useful matrix derivative formulae

Proposition 5. *If D is a matrix and $J = \frac{1}{2} D \cdot D$, then $\frac{\partial J}{\partial D} = D$.*

Part II
Week 10. Large Scale Machine Learning

Keywords. Online learning, predicted CTR / click through rate, map-reduce, Hadoop.

Question 6. *Can you use, say, Hadoop map-reduce for any generic processing task, e.g. parallel video transcoding?*

Part III
Week 11. Application: Photo OCR

Keywords. Photo OCR vs scanned text OCR, text detection, bounding box expansion, character segmentation, character classification, pedestrian detection, sliding window detection, step size / stride, artificial data synthesis, data amplification, ceiling analysis.

5 Photo OCR Pipeline