MULTINOMIAL LOGISTIC REGRESSION – BIS

Classification and Prediction

Lê Hồng Phương

<*phuonglh@gmail.com*>
Vietnam National University of Hanoi
Hanoi University of Science

March 2015

Content

Multinomial Logistic Regression

- The multinomial logistic regression (MLR) model is defined as

$$P(y = k|\,\mathbf{x}; \theta_k) = \frac{1}{Z} \exp(\theta_k^T\, \mathbf{x}),$$

where Z is the normalization term:

$$Z = \sum_{k=1}^{K} P(y = k|\,\mathbf{x}; \theta_k) = \sum_{k=1}^{K} \exp(\theta_k^T\, \mathbf{x}).$$

- Parameter of the model:

$$\theta = \begin{pmatrix} \theta_{10} & \theta_{11} & \cdots & \theta_{1D} \\ \theta_{20} & \theta_{21} & \cdots & \theta_{2D} \\ \cdots & \cdots & \cdots & \cdots \\ \theta_{K0} & \theta_{K1} & \cdots & \theta_{KD} \end{pmatrix}.$$

Cost Function

Given a training set of N examples $(\mathbf{x}_i, y_i)$, the cost function of unregularized MLR is:

$$
\begin{aligned}
J(\theta) &= -\frac{1}{N} \sum_{i=1}^{N} \log P(y_i | \, \mathbf{x}_i; \theta) \\
&= -\frac{1}{N} \sum_{i=1}^{N} \left\{ \theta_{y_i}^T \, \mathbf{x}_i - \log \left(\sum_{k=1}^{K} \exp(\theta_k^T \, \mathbf{x}_i) \right) \right\}.
\end{aligned}
$$

Gradient Matrix

The gradient of the cost is a matrix of the same size as θ:

$$\frac{\partial}{\partial \theta_{kj}} J(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \left\{ \delta(y_i = k) x_{ij} - \frac{\exp(\theta_k^T \mathbf{x}_i)}{\sum_{k=1}^{K} \exp(\theta_k^T \mathbf{x}_i)} x_{ij} \right\}$$

$$= -\frac{1}{N} \sum_{i=1}^{N} [\delta(y_i = k) - P(y = k | \mathbf{x}_i; \theta)] x_{ij}.$$

In row vectors of θ_k:

$$\frac{\partial}{\partial \theta_k} J(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \underbrace{[\delta(y_i = k) - P(y = k | \mathbf{x}_i; \theta)]}_{w_i} \mathbf{x}_i$$

$$= -\frac{1}{N} \sum_{i=1}^{N} w_i \, \mathbf{x}_i \, .$$

Cost Function – L_2 Regularization

The cost function of L_2-regularized MLR is:

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log P(y_i|\,\mathbf{x}_i;\theta) + \frac{\lambda}{2N}\sum_{k=1}^{K}\sum_{j=1}^{D}\theta_{kj}^2$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\left\{\theta_{y_i}^T\,\mathbf{x}_i - \log\left(\sum_{k=1}^{K}\exp(\theta_k^T\,\mathbf{x}_i)\right)\right\} + \frac{\lambda}{2N}\sum_{k=1}^{K}\sum_{j=1}^{D}\theta_{kj}^2.$$

Note that we do not regularize the intercept parameters $\theta_{k0}, \forall k$.

Gradient Matrix – L_2 Regularization

The gradient of the cost is a matrix of the same size as θ:

$$\frac{\partial}{\partial \theta_{kj}} J(\theta) = -\frac{1}{N} \sum_{i=1}^{N} [\delta(y_i = k) - P(y = k | \, \mathbf{x}_i; \theta)] x_{ij} \underbrace{+ \frac{\lambda}{N} \theta_{kj}}_{\text{if } j>0} .$$

In row vectors of θ_k:

$$\frac{\partial}{\partial \theta_k} J(\theta) = -\frac{1}{N} \sum_{i=1}^{N} w_i \, \mathbf{x}_i \underbrace{+ \frac{\lambda}{N} \theta_k}_{\theta_{k0} \equiv 0} .$$

Classification Rule

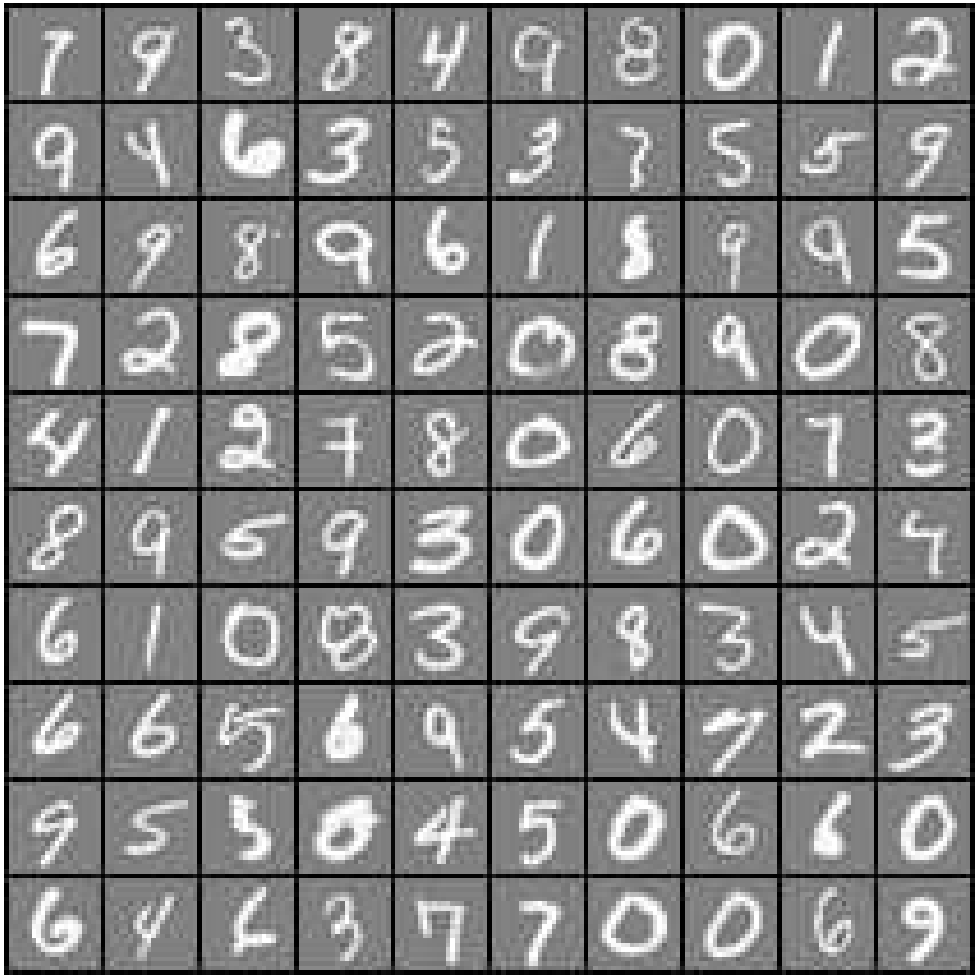
When parameters θ are known, the classification rule of MLR for a given $\mathbf{x}$ is:

$$\widehat{y} = \underset{k=1,\ldots,K}{\arg\max}(\theta_k^T \mathbf{x}).$$

Handwritten Digit Recognition

- Application of MLR on a dataset of handwritten digits. The set contains 5000 training examples.
- This is a subset of the MNIST handwritten digit dataset:
  - *http://yann.lecun.com/exdb/mnist/*
- Each training example is a 20x20 pixels grayscale image of the digit.
- Each pixel is represented by a floating point number indicating the grayscale intensity at that location.
- The 20x20 grid of pixel is unrolled into a 400-dimensional vector.

Handwritten Digit Recognition



Handwritten Digit Recognition

- L_2 regularization with $\lambda = 0.1$
- Logistic regression using one-versus-all classification:
  - Train 10 binary logistic regression classifiers
  - Training accuracy: 94.98%
- MLR: 96.72%
  - $J(0) = 2.302585, J(\theta^*) = 0.147061$

Wine Origin

- Data of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars.
- 13 constituents found in each of the three types of wines:
- *http://archive.ics.uci.edu/ml/datasets/Wine*

- Training accuracy: 100%.
  - No regularization
  - $J(0) = 1.098612, J(\theta^*) = 0.007770$

Wine Quality

- Two datasets related to red and white vinho verde wine samples, from the north of Portugal.
- The goal is to model wine quality based on physicochemical tests.
- *http://archive.ics.uci.edu/ml/datasets/Wine+Quality*

- Number of instances: red wine = 1599; white wine = 4898; $D = 11$.

Wine Quality

No regularization:

	Red Wine	**White Wine**
Training accuracy	60.85%	54.35%
$J(0)$	1.791759	1.945910
$J(\theta^*)$	0.921343	1.093946

Exercises

1. Implement multiple binary logistic regression classifiers using one-versus-all classification.
2. Implement multinomial logistic regression classifier
3. Test these classifiers on different datasets and report the results.