Machine Learning

Homework #2 – Spam Classification

B02901080 電機四 董皓文

Logistic Regression Function

Logistic Regression Model:

$$\boldsymbol{y} = \boldsymbol{b} + \boldsymbol{\theta} \cdot \boldsymbol{X}$$

$\boldsymbol{X}$: feature array $\boldsymbol{b}$: bias

$\boldsymbol{y}$: label array $\boldsymbol{\theta}$: array of the weight of each feature

All 57 features are used in this work.

Function Set:

$$f_{w,b}(C|x) = \sigma\left(\sum_i \theta_i x_i + b\right), \qquad \sigma(z) = \frac{1}{1+e^{-z}}$$

Goodness of Function:

$$Loss(f) = -\sum_n \left(\hat{y}^n lnf(x^n) + (1-\hat{y}^n)ln\big(1-f(x^n)\big)\right)$$

Gradient Descent Method:

SGD:

$$\theta_t = \theta_{t-1} - \eta g_t$$

AdaGrad:

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{\sum_{i=0}^{t}(g_i)^2 + \varepsilon}} g_t$$

AdaDelta:

$$E[g^2]_t = \gamma E[g^2]_{t-1} + (1-\gamma)$$

$$E[\Delta\theta^2]_t = \gamma E[\Delta\theta^2]_{t-1} + (1-\gamma)\Delta{\theta_t}^2$$

$$\theta_t = \theta_{t-1} - \frac{\sqrt{E[\Delta\theta^2]_{t-1} + \varepsilon}}{\sqrt{E[g^2]_t + \varepsilon}} g_t$$

Alternative Method – Naïve Bayes Classifier

Given a message **m**, we may extract its feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_m)$, where x_i is 1 if **m** contains a certain word v_i and 0, otherwise. Then we may calculate the following likelihood function of **x**. See [1] for detail.

$$\Lambda(\mathbf{x}) = \frac{P(\mathbf{x}|Spam)}{P(\mathbf{x}|Not\_Spam)}, \qquad P(\mathbf{x}|c) = \prod_{i=1}^{m} P(x_i|c)\,, c = Spam, Not\_Spam$$

Then we may make the following prediction.

If $\Lambda(\mathbf{x}) > \lambda \frac{P(Not\_Spam)}{P(Spam)}$, predict **m** is a spam.

If $\Lambda(\mathbf{x}) < \lambda \frac{P(Not\_Spam)}{P(Spam)}$, predict **m** is a NOT a spam.

Comparison

Logistic Regression(LR): Accuracy = 69.08%

Naïve Bayes Classifier(NBC): Accuracy = 61.55%

LR has a higher accuracy over NBC, while NBC is much easier to implement. No iteration is required when training a NBC, thus NBC runs faster than LR.

NBC is a generative classifier and LR is a discriminative classifier. We can find consistency in our works and the result in [2] that

> ***while a discriminative learning has lower asymptotic error, a generative classifier may also approach its (higher) asymptotic error much faster.***

Reference

[1] K. Tretyakov, Machine Learning Techniques in Spam Filtering, 2004

[2] A. Y. Ng, M. I. Jordan, On Discriminative vs. Generative classifiers: A comparison of logistic regression and naïve Bayes, 2002