Exercise for Chapter 5 —— Machine Learning Basics

Xiao Zhang & Fnu Suya

1. To better understand the machine learning concepts in Chapter 5, let's consider another simple machine learning algorithm (beyond linear regression): **logistic regression**, which is widely used for classification. In particular, we focus on binary classification problem in which the output $y \in \{0, 1\}$. As introduced in Section 5.7.1, the relationship between the input feature vector $\boldsymbol{x} \in \mathbb{R}^p$ and the output class y is modeled as

$$\mathbb{P}(y = 1 \mid \boldsymbol{x}; \boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}), \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ is the model parameter we are going to estimate, and $\sigma(z) = \frac{1}{1+\exp(-z)}$ is the logistic sigmoid function.

   (a) Given a set of m training examples $\{\boldsymbol{x}_i, y_i\}_{i=1}^m$, construct the maximum likelihood estimator $\boldsymbol{\theta}_{\text{ML}}$ for logistic regression. (you may want to review Section 5.5.1)

   (b) Another estimator for logistic regression can be obtained by minimizing the mean square error (as in the case of linear regression), namely

$$\boldsymbol{\theta}_{\text{MSE}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{m} \sum_{i=1}^{m} \big(\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}) - y_i\big)^2.$$

   But in practice, no one use $\boldsymbol{\theta}_{\text{MSE}}$ for logistic regression. Can you explain the reasons why $\boldsymbol{\theta}_{\text{ML}}$ is preferred over $\boldsymbol{\theta}_{\text{MSE}}$?

   (c) Different from linear regression, maximum likelihood estimator for logistic regression doesn't have explicit analytical form. Thus, people turn to iterative algorithm such as gradient descent to approximate $\boldsymbol{\theta}_{\text{ML}}$. Consider the negative loglikelihood as the cost function, write down the gradient descent algorithm. (you may want to review Section 4.5)

   (d) In practice, people add an additional ℓ_2 regularizer $\lambda \|\boldsymbol{\theta}\|_2^2 = \lambda \sum_{i=1}^p \theta_i^2$ to the cost function to achieve better performance, where $\lambda > 0$ is the regularization parameter. Can you briefly explain why this regularization term can help reduce overfitting? (you may refer to Section 5.2.2)

(e) Besides ℓ_2 regularizer, another widely-used regularization technique for logistic regression is ℓ_1 regularization, where $\|\boldsymbol{\theta}\|_2^2$ is replaced by $\|\boldsymbol{\theta}\|_1 = \sum_{i=1}^{p} |\theta_i|$ in the objective function. Explain the difference between ℓ_1 and ℓ_2 regularization.

(f) The machine learning course taught by Andrew Ng on Coursera is provided with programming exercises, which are available here: https://github.com/rieder91/MachineLearning. Exercise 2 would be a good practice for (regularized) logistic regression. The solution written in python can be found in this url: https://github.com/nex3z/machine-learning-exercise.