Introduction to Machine Learning
Homework 1

October 6, 2021

Academic integrity

Our lesson cares much more on academic integrity. No matter who should do our utmost to handle the establishment of academic integrity standard including the host teacher and assistants of this lesson. We hope you will have the same faith with us.

(1) Discussion between students is allowing. The work named by yourself must be completed by your own hands. Any kind of Copying from existing documents will be seen as illegal.

(2) Any kind of Copying from other people's fruits of labour(Publication or Internet documents) will be accused of plagiarism. The score of plagiarists will be canceled. Please mark the authors if you cited any public documents of them;

(3) Highly resemble homework will be seen as Coping. No matter who you are, the one who copy or the one who is copied, both of your score will be canceled. Please protect your homework not to be copied by others actively.

Homework submission notes

(1) Please follow the submission methods on the website;

(2) If you are not follow the methods or your submission format are not correct. We will deduct some score of your homework;

(3) Unless some special cases, the submission over deadline will not be accepted and your score will be set as zero.

1 [30pts] Basic concepts

1.1 Probabiliy

Suppose Bob has been tested with a terrible disease. The event T and D represent a person has been tested positive for this disease and actually has this disease, respectively. According to statistics, we know:

$$\begin{cases} \Pr(\mathrm{T}|\mathrm{D}) & = 0.98 \\ \Pr(\mathrm{T}|\neg\mathrm{D}) & = 0.10 \\ \Pr(\mathrm{D}) & = 0.01 \end{cases} \tag{1.1}$$

He wants you to help him calculate the probability that he actually has the disease?

1.2 Maximum likelihood estimation

We have an uneven coin, and the probability of tossing it heads up at random is p. Suppose you toss this coin 10 times, 8 of which are heads up. Please estimate p based on the existing information using MLE.

1.3 Performance meause

We have a set of samples with binary classes (denoted as 0 and 1) and two classifers C_1 and C_2. For each sample, the classifier gives a score to measure the confidence that the classifier believes that the sample belongs to class 1. Below are the predicted results of two classifiers (C_1 and C_2) for 8 samples, their ground truth labels (y), and the scores for both classifiers (y_{C_1} and y_{C_2}).

y	1	0	1	1	1	0	0	1
y_{C_1}	0.62	0.39	0.18	0.72	0.45	0.01	0.32	0.93
y_{C_2}	0.34	0.12	0.82	0.89	0.17	0.75	0.36	0.97

(1) Calculate the area under the ROC curve (AUROC) for both classifiers C_1 and C_2.

(2) For the classifier C_1, we select a decision threshold $th_1 = 0.40$ which means that C_1 classifies a sample as class 1, if its score $y_{C_1} > th_1$, otherwise it classifies this sample as class 0. Calculate the confusion matrix and the F_1 score. Do the same thing for the classifier C_2 using a threshold value $th_2 = 0.90$.

2 [30pts] Linear model

Suppose you are given a data set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$. We want to use a regularized linear regression model to fit this data set, that is, to solve the following minimization problem:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2 \tag{2.1}$$

where, $\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in R^{n\times d}$. Assume that $\mathbf{X}$ is column full-rank matrix.

1. Please give the closed-form solution for Eq.(2.1). You need to give your solution in detail.

2. The data set D is shown in the Table 1, where each sample has 3 dimensions (F_1, F_2, F_3). Please calculate the optimal solution for $\mathbf{w}$ when $\lambda = 1$.

F_1	2	9	8	8	2	8	4	1	3	5
F_2	9	3	3	8	1	4	3	8	3	3
F_3	1	1	1	1	1	1	1	1	1	1
y	290	1054	944	964	246	948	488	167	370	598

Table 1: Training set for linear regression

3 [40pts] Logistic Regression

In a binary classification problem, each instance $\mathbf{x}_i$ in a data set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ has a label $y_i \in \{0, 1\}$. We have already known that the logistic regression model Eq.(3.1) is a powerful tool to handle this kind of problem.

$$y = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}} \tag{3.1}$$

To simplify this problem, we assume that $\boldsymbol{\beta} = (\mathbf{w}; b)$, $\hat{\mathbf{x}}_i = (\mathbf{x}_i; 1)$. Because its negative log-likelihood function Eq.(3.2) is convex, we can optimize it efficiently with Gradient Descent method, Newton method, and so on.

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(-y_i \boldsymbol{\beta}^\top \hat{\mathbf{x}}_i + \ln \left(1 + e^{\boldsymbol{\beta}^\top \hat{\mathbf{x}}_i} \right) \right) \tag{3.2}$$

1. Prove the Eq.(3.2) is convex.

2. Suppose we are facing a multi-class classification problem instead of a binary classification problem, where $y_i \in \{1, 2, \ldots, K\}$. Please expand the logistic regression model Eq.(3.1) to a multi-class version and give the log-likelihood function of this multi-class logistic regression model.

3. Use out-of-the-box machine learning tools (e.g., scikit-learn, ...) to build your logistic regression model and comprehensively evaluate your results on Yeast[1] data set. You are recommended to try different techniques (e.g., OvO, OvR, multi-class logistic regression) for solving this multi-class problem. Briefly showing your analysis, experimental results, and conclusions.

[1] http://archive.ics.uci.edu/ml/datasets/Yeast