Aprendizagem Automática

Assignment 1 - Classifiers

Andrea Mascaretti N52222 Daniel Pimenta N45404

October 21, 2017

Abstract

The main goal of this assignment was the parameterization, fitting and comparation of Logistic Regression, K-nearest Neighbours and Naive Bayes classifiers.
The data set used was the banknote authentication which was obtained from the UCI machine learning repository.
To achieve our goals we used the Spyder IDE with programming language Python 3.6 and some of its modules such as *Panda* to load and manage our data, *NumPy* which is the fundamental package for scientific computing with Python and was used for various mathematical calculations and managing multidimensional arrays, *sklearn (scikit-learn)* which has efficient tools for data mining and data analysis used in our assignment to preprocess data (shuffle, standardize, split), fold our data into k stratified folds, calculate cross validation score and fit our data on Logistic Regression and K-nearest Neighbours classifiers. Another imported module was Matplotlib.Pyplot which provides a MATLAB-like plotting framework used to build the error plots.

1 Introduction

WHAT WE INITIALLY HAD,
WHAT WAS ASKED TO DO,
HOW WE PLAN ON DOING IT

2 Classifiers

2.1 Logistic Regression

The logistic regression is a generalised linear model. Before considering a more classic machine learning framework, we provide a brief introduction to the main idea behind the model.

The logistic regression is composed, as every linear model, of three part:

- A sequence of *response variables* $Y_1, \ldots, Y_n$. Such response variables are called random components. The main assumption we shall make regarding those variables is that they are all independent random variables. Moreover, each one of them will have a distribution from the exponential family. Bear in mind that we are not imposing that the various Y_is are identically distributed.
- The *systematic component* is our model. It is a function of some predictors (also known as regressors or covariates), linear in the parameter and related to the mean of Y_i.
- The *link function* $g(\mu_i)$ will allow us to link the two components, allowing us to say that

$$g(\mu_i) = \beta_0 + \sum_{i=1}^{r} \beta_i x_i \tag{1}$$

where $\mu_i = \mathbb{E}[Y_i]$. In the machine learning framework, it is usually more common to consider the inverse of the link function.

In our model we will assume the following:

$$Y_i \overset{ind}{\sim} \mathsf{Bernoulli}\left(\pi_i\right) \tag{2}$$

and therefore we will have that

$$\mathbb{E}\left[Y_i\right] = \pi_i. \tag{3}$$

Now, in this model we will assume a relationship between the logarithm of the odds of success for Y_i (the log odds or *logit*) and the predictor $\mathbf{x} = (1, x_1, \ldots, x_r)$. Mathematically, this entails

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta'\mathbf{x} \tag{4}$$

and $\beta = (\beta_0, \ldots, \beta_r)$.

Rewriting in order to obtain the exponential family form we have

$$\pi^y \left(1-\pi\right)^{1-y} = \left(1-\pi\right) \exp\left\{y \ln\left(\frac{\pi}{1-\pi}\right)\right\}, \tag{5}$$

where the term $\ln\left(\frac{\pi}{1-\pi}\right)$ is the natural parameter of the exponential family and shows why we will implement the link function

$$g\left(\pi\right) = \ln\left(\frac{\pi}{1-\pi}\right). \tag{6}$$

Rewriting 4, we obtain

$$\pi\left(x\right) = \frac{\exp\left\{\beta'\mathbf{x}\right\}}{1+\exp\left\{\beta'\mathbf{x}\right\}}. \tag{7}$$

What we have written is equivalent to saying that

$$\mathsf{p}\left(\mathcal{C}_1 | \mathbf{x}\right) = \pi\left(\mathbf{x}\right) = \sigma\left(\beta'\mathbf{x}\right) \tag{8}$$

where $\sigma\left(\cdot\right)$ is the logistic sigmoid and $\mathcal{C}_1$ is to identify class 1. Notice that it holds $\mathsf{p}\left(\mathcal{C}_2 | \mathbf{x}\right) = 1 - \mathsf{p}\left(\mathcal{C}_1 | \mathbf{x}\right) = 1 - \sigma\left(\beta'\mathbf{x}\right)$.

Suppose we are given a dataset $\{\mathbf{x}_i, y_i\}$ where now $y_i \in \{-1, 1\}$ and $\mathbf{x}_i \in \mathbb{R}^r$ for $i = 1, \ldots, n$. Given our hypotheses, we can write the negative log-likelihood function as

$$\sum_{i=1}^{n} \ln\left(1 + \exp\left\{-y_i \ \left(\beta'\mathbf{x}_i\right)\right\}\right). \tag{9}$$

To this term we added a $\mathsf{L2}$ penalization term to obtain a more regularized result, preferring this to its $\mathsf{L1}$ version as the latter usually provides sparser solutions and we were working in an environment with only four covariates. This lead to the following problem:

$$\min_{\beta} f\left(\beta\right) = \frac{1}{2}\beta'\beta + C\sum_{i=1}^{n} \ln\left(1 + \exp\left\{-y_i \ \left(\beta'\mathbf{x}_i\right)\right\}\right), \tag{10}$$

where $C > 0$ is a parameter that can be assigned in order to balance the two terms. To tune it, we relied on a stratified 5-fold cross-validation on the training set, picking the value that yielded the lowest value. From a numerical point of view, the trust region Newton method built in the $\mathsf{ScikitLearn}$ class for $\mathsf{L2}$-logistic regression was used.

2.2 k-Nearest Neighbours

Suppose that we have some classes $\mathcal{C}_k$ such that each class contains N_k points and let $N = \sum_k N_k$ be the total number of points. Let us suppose that we are also given a distance function $\rho : \mathbb{R}^r \to \mathbb{R}^+$. To classify a new data point $\mathbf{x}$, we draw a sphere (according to ρ) such that exactly k points,

regardless of their class, fall into it. If we let V be the volume of such sphere and K_k be the number of points of class k contained in it, we see that

$$\mathsf{p}(\mathbf{x}|\,\mathcal{C}_k) = \frac{K_k}{N_k V}, \tag{11}$$

the marginal probability is given by

$$\mathsf{p}(\mathbf{x}) = \frac{K}{NV} \tag{12}$$

and the prior distributions are

$$\mathsf{p}(\mathcal{C}_k) = \frac{N_k}{N}. \tag{13}$$

Therefore, applying Bayes' theorem we have

$$\mathsf{p}(\mathcal{C}_k|\mathbf{x}) = \frac{\mathsf{p}(\mathbf{x}|\,\mathcal{C}_k)\,\mathsf{p}(\mathcal{C}_k)}{\mathsf{p}(\mathbf{x})} = \frac{K_k}{K}. \tag{14}$$

Comparing the postierior probabilities, we classified each data picking the maximum *a posteriori*. We consider the usual Euclidean distance for our distance function and trained only for odds value of k (between 1 and 39), picking the one that yielded the best result on a Stratified 5-fold cross validation on our data set.

2.3 Kernel Naive Bayes Classifier

Let $\mathcal{C}_k$ be a class and let $\mathbf{x}$ our data vector. As done before, we can apply Bayes' theorem to provide the *a posteriori* probability of belonging to such class by stating 14. Naive Bayes' classifier is based on the hypothesis that the attributes of our data are conditionally independent, allowing us to rewrite the conditional likelihood as

$$\mathsf{p}(\mathbf{x}|\,\mathcal{C}_k) = \prod_{i=1}^{r} \mathsf{p}(x_i|\,\mathcal{C}_k)\,. \tag{15}$$

This assumption, that is in fact *naive*, does not usually hurt too much as we are interested in picking the maximum probability a *posteriori*.

Naive Bayes is usually implemented by considering a probability density function, that is fitted to the feature being taken into account. However, we decided to model the conditional distribution with a more complex kernel density estimation to allow for more parameter freedom. As far as the prior are concerned, we decided to consider the relative frequencies of the two classes.

Explain what are the three parameters that were optimized and their effects on their respective classifiers. This will also require a brief explanation of how each classifier works.
Explain the method by which the optimal values were found, noting the differences between the errors and the importance of leaving out a test set for the final evaluation
The report should show the error plots but no other plots are necessary

Logistic Regression

K-nearest Neighbours

Naive Bayes

3 Comparation

McNemar's test TALK ABOUT WHAT IS MCNEMAR TEST AND HOW IT WAS IMPLEMENTED AND USED
Estimate the true error for each classifier, compare the classifiers with McNemar's test and discuss which, if any, is better for this application

4 Conclusion

WHAT WAS ACHIEVED AND WHAT WE LEARNT

5 Bibliography

- ...