Machine Learning 2016/2017, Assignment 4

Yevgeny Seldin, Christian Igel
Department of Computer Science, University of Copenhagen

The deadline for this assignment is **12:00 pm (noon, not midnight) 19/12/2016**. You must submit your individual solution electronically via the Absalon home page.

A solution consists of:

- A PDF file with detailed answers to the questions, which may include graphs and tables if needed. Do *not* include your source code in this PDF file.
- Your solution source code (Matlab / R / Python scripts or C / C++ / Java code) with comments about the major steps involved in each question (see below). Source code must be submitted in the original file format, not as PDF.
- Your code should be structured such that there is one main file that we can run to reproduce all the results presented in your report. This main file can, if you like, call other files with functions, classes, etc.
- Your code should also include a README text file describing how to compile and run your program, as well as list of all relevant libraries needed for compiling or using your code.

1 Finite Hypothesis Space

We would like to derive a prediction rule for predicting whether a person has minors (children under 18 years old). The input space is the space of possible age and gender values $\mathcal{X} = \{0, \dots, 100\} \times \{\text{male}, \text{female}\}$. The output space is the space of positive and negative answers $\mathcal{Y} = \{\pm 1\}$.

We could try to estimate the probability of having minors for each (age, gender) pair individually, but we could also try to estimate the age ranges for males and females when they are likely to have minors. In the latter case it is natural to have a model that has a single age range for men in which the answer is more likely to be positive and a single age range for women in which the answer is more likely to be positive.

1. What is the size of the hypothesis space in the first and in the second approach?

2. Write down a high-probability bound (a bound that holds with high probability) on $L(h)$ in terms of $\hat{L}(h, S)$ and the complexity of $\mathcal{H}$ for the two cases.

3. What are the relative advantages/disadvantages of the two hypothesis spaces? Can you make up some other prediction problem, where you would expect the first hypothesis class to be preferable over the second? (For example, assume that you would like to predict whether a person is likely to visit a dentist in 2017. Assume that people with "new teeth", either baby teeth or permanent teeth [molars] are less likely to visit a dentist than people with "old teeth" and assume that gender has no influence on teeth quality. How would you construct the hypothesis space?)

2 Occam's Razor

We want to design an application for bilingual users. The application should detect the language in which the person is typing based on the first few letters typed. In other words, we want to design a classifier that takes a short string (that may be less than a full word) as input and predicts one of two languages, say, Danish or English. For simplicity we will assume that the alphabet is restricted to a set Σ of 26 letters of the Latin alphabet plus the white space symbol (so in total $|\Sigma| = 27$). Let Σ^d be the space of strings of length d. Let $\mathcal{H}_d$ be the space of functions from Σ^d to $\{0, 1\}$, where Σ^d is the input string and $\{0, 1\}$ is the prediction (Danish or English). Let $\mathcal{H} = \bigcup_{d=0}^{\infty} \mathcal{H}_d$ be the union of $\mathcal{H}_d$-s.

1. Derive a high-probability bound on $L(h)$ that holds for all $h \in \mathcal{H}_d$.

2. Derive a high-probability bound on $L(h)$ that holds for all $h \in \mathcal{H}$.

3. Explain the trade-off between picking short strings (small d) and long strings (large d). Which terms in the bound favor small d (i.e., they increase with d) and which terms in the bound favor large d (i.e., they decrease with d)?

Optional, not for submission You are very welcome to experiment with the influence of the string length d on the performance. You can find a lot of texts in different languages here `http://www.gutenberg.org/catalog/`. Do you observe the effect of initial improvement followed by overfitting as you increase d?

3 Logistic regression

3.1 Cross-entropy error measure

Read section 3.3 in the course textbook [1]. Solve exercise 3.6 on page 92 in the course textbook. The *in-sample error* E_{in} corresponds to what we call the empirical risk (or training error).

3.2 Logistic regression loss gradient

Solve exercise 3.7 on page 92 in the course textbook.

3.3 Logistic regression implementation

Implement logistic regression as presented in the lectures, see also Example 3.3 on page 95 in [1].

Apply logistic regression to the Iris flower data set from Home Assignment 2. Remove class 2 from the training and test data sets. The resulting data sets should contain 62 and 26 patterns, respectively. That is, you are supposed to solve a binary classification task.

Report the training an test error as measured by the 0-1 loss. Furthermore, report the three parameters of the (affine) linear model.

References

[1] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin. *Learning from Data.* AMLbook, 2012.