Machine Learning
Exercises

Christian Rossi

Academic Year 2023-2024

Abstract

The course topics are:

- Introduction: basic concepts.
- Learning theory:
  - Bias/variance tradeoff. Union and Chernoff/Hoeffding bounds.
  - VC dimension. Worst case (online) learning.
  - Practical advice on how to use learning algorithms.
- Supervised learning:
  - Supervised learning setup. LMS.
  - Logistic regression. Perceptron. Exponential family.
  - Kernel methods: Radial Basis Networks, Gaussian Processes, and Support Vector Machines.
  - Model selection and feature selection.
  - Ensemble methods: Bagging, boosting.
  - Evaluating and debugging learning algorithms.
- Reinforcement learning and control:
  - MDPs. Bellman equations.
  - Value iteration and policy iteration.
  - TD, SARSA, Q-learning.
  - Value function approximation.
  - Policy search. Reinforce. POMDPs.
  - Multi-Armed Bandit.

Contents

CHAPTER 1

Exercise session I

1.1 Exercise one

Given the relationship:

$$S = f(TV, R, N)$$

where S is the amount of sales revenue, TV, R and N are the amount of money spent on advertisements on TV programs, radio and newspapers, respectively, explain what are the:

1. Response.
2. Independent variables.
3. Features.
4. Model.

Which kind of problem do you think it is trying to solve?

1.1.1 Solution

In the proposed relationship we have:

1. The response (or target or output) is the amount of sales S.
2. The independent variables (or input) are TV, R and N
3. The features (or input) are TV, R and N.
4. The model is identified by the function $f(\cdot)$.

Since the amount of sales S is a continuous and ordered variable, we are trying to solve a regression problem (supervised learning).

1.2 Exercise two

Why is linear regression important to understand? Select all that apply and justify your choice:

1. The linear model is often correct.
2. Linear regression is extensible and can be used to capture nonlinear effects.
3. Simple methods can outperform more complex ones if the data are noisy.
4. Understanding simpler methods sheds light on more complex ones.
5. A fast way of solving them is available.

1.2.1 Solution

1. False: it rarely happens that the problem we are modeling has linear characteristics.
2. True: it is true that this model is easy to interpret and can be extended to also consider nonlinear relationships among variables, e.g., using basis functions.
3. True: since we are able only to minimize the discrepancy between the considered function and the real one, and we can not reduce the variance introduced by noise, the use of linear model might be a better choice with respect to more complex ones since they usually are prone to overfitting, i.e., they try to model also the noise of the considered process.
4. True: they are easy to interpret and might give suggestions on more sophisticated techniques which can be used to tackle specific problems.
5. True/False: for some loss functions we have a closed form solution for linear model (LS method), thus we can guarantee that we are able to find the parameters minimizing the loss function effectively. That is not always true and depends also on the loss function we want to minimize.

1.3 Exercise three

Consider a linear regression with input x, target y and optimal parameter θ^*.

1. What happens if we consider as input variables x and $2x$?
2. What we expect on the uncertainty about the parameters we get by considering as input variables x and $2x$?
3. Provide a technique to solve the problem.
4. What happens if we consider as input variables x and x^2?

Motivate your answers.

1.3.1 Solution

The original formulation is:

$$y = \theta^* x$$

1. In this scenario, the formulation simplifies to:

   $$y = \theta_1 x + \theta_2 2x$$

   As the two variables are dependent, it can alternatively be expressed as:

   $$y = \underbrace{(\theta_1 + 2\theta_2)}_{\theta^*} x$$

   Thus, yielding the original formulation:

   $$y = \theta^* x$$

   Moreover, for computing the closed-form optimization, the formula employed is:

   $$\omega = \left(x^T x\right)^{-1} xt$$

   However, in this particular case, the matrix x takes the form:

   $$x = \begin{bmatrix} x_1 & 2x_1 \\ x_2 & 2x_2 \\ \vdots & \vdots \\ x_n & 2x_n \end{bmatrix}$$

   Hence, $x^T x$ becomes singular, rendering the previous formula inapplicable.

   In general, if x lacks full rank, then $x^T x$ becomes singular, and its inverse cannot be computed.

2. The parameter we get have a high variance, since we have an infinite number of couples of parameters minimizing the loss of the samples in the considered problem. Indeed, if the parameters of the two inputs are w_1 and w_2 we would have that the true relationship would be:

   $$y = \theta_1 x + \theta_2 2x = (\theta_1 + 2\theta_2)\, x$$

   Which can be satisfied by an infinite number of solutions.

3. In this case, the use of ridge regression is able to partially cope with the influence of using highly linearly correlated features. Another viable option is to remove the variables which are linearly dependent, for instance by checking if they have correlation equal to 1 or -1.

4. In this case we do not have a badly conditioned matrix since x and x^2 are not linearly dependent and the corresponding design matrix would not be ill-conditioned. As a result, we can find a closed form optimization.

1.4 Exercise four

After performing Ridge regression on a dataset with $\lambda = 10^{-5}$ we get one of the following one set of eigenvalues for the matrix $\left(\Phi^T\Phi + \lambda I\right)$:

1. $\Lambda = \{0.00000000178, 0.014, 12\}$
2. $\Lambda = \{0.0000178, -0.014, 991\}$
3. $\Lambda = \{0.0000178, 0.014, 991\}$
4. $\Lambda = \{0.0000178, 0.0000178, 991\}$

Explain whether these sets are plausible solutions or not.

1.4.1 Solution

Since the matrix $\left(\Phi^T\Phi + \lambda I\right)$ is definite positive and its eigenvalues should all be greater than $\lambda = 10^{-5}$, we have:

1. Not plausible: one eigenvalue is smaller than 10^{-5}.
2. Not plausible: one eigenvalue is negative.
3. Plausible: all positive and greater than 10^{-5}.
4. Plausible: all positive and greater than 10^{-5}.