2 Linear Regression

Exercise 2.1

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?

Exercise 2.2

Which one/ones of the following is a good definition of Machine Learning? Motivate your answer.

1. A computer program is said to learn from experience with respect to some class of tasks and performance measure, improves with experience;
2. Machine Learning are all the techniques using relationship to provide prediction or to suggest the action to perform in practical situations;
3. Machine Learning is the sub-field of Artificial Intelligence where the knowledge comes from the use of experience to perform induction;
4. Machine Learning is the process of creating new information to provide meaningful suggestions to human beings;
5. Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.

Exercise 2.3

Explain why the following problems can or cannot be addressed by Machine Learning (ML) techniques:

1. Partition a set of employees of a large company;
2. Fortune-telling a person information about her/his personal life;
3. Determine the truthfulness of a first order logic formula;
4. Compute the stress on a structure given its physical model;
5. Provide temperature predictions.

In the case the problem can be addressed by ML, provide a suggestion for the technique you would use to solve the problem. (hint: wait until the end of this course to answer these questions)

Exercise 2.4

Categorize the following ML problems:

1. Predicting housing prices for real estate;
2. Identify inside trading among stock market exchange;
3. Detect interesting features from an image;
4. Determine which bird species is/are on a given audio recording;
5. Teach a robot to play air hockey;
6. Predicting tastes in shopping/streaming;
7. Recognise handwritten digits;
8. Pricing goods for an e-commerce website.

For each one of them suggest a set of features which might be useful to solve the problem and a method to solve it.

Exercise 2.5

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.

Exercise 2.6

Consider a generic regression model. Tell if one should consider the LS method as a viable option in each one of the following 4 different situations. Motivate your answer.

1. Small number of parameters;
2. The loss function is $L(\mathbf{w}|x_n, t_n) = |y(x_n, \mathbf{w}) - t_n|$;
3. Huge number of samples;
4. The loss function is $L(\mathbf{w}|x_n, t_n) = \begin{cases} (y(x_n, \mathbf{w}) - t_n)^2 & \text{if} |y(x_n, \mathbf{w}) - t_n| < \delta \\ |y(x_n, \mathbf{w}) - t_n| & \text{if } |y(x_n, \mathbf{w}) - t_n| > \delta \end{cases}$

Exercise 2.7

Consider a linear regression with input x, target t 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.

* Exercise 2.8

Consider a data set in which each data point $(\mathbf{x}_n, t_n)$ is associated with a weighting factor $r_n > 0$, so that the error function becomes:

$$J(\mathbf{w}) = \frac{1}{2N} \sum_{n=1}^{N} r_n (\mathbf{w}^\top \mathbf{x}_n - t_n)^2$$

Find an expression for the solution that minimizes this error function. Give two alternative interpretations of the weighted sum-of-squares error function in terms of (i) data dependent noise variance and (ii) replicated data points.

Exercise 2.9

Consider an initial parameter $\mathbf{w}^{(0)} = [0\ 0\ 1]^\top$ and a loss function of the form:

$$J(\mathbf{w}) = \frac{1}{2N} \sum_{n=1}^{N} (\mathbf{w}^\top \mathbf{x}_n - t_n)^2.$$

Derive the update given from the gradient descent for the datum $\mathbf{x}_1 = [1\ 3\ 2]^\top,\ t_1 = 4$, and a learning rate $\alpha = 0.3$.

What changes if we want to perform a batch update with $K = 10$ data?

Exercise 2.10

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

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.

Exercise 2.11

We run a linear regression and the slope estimate is $\hat{w}_k = 0.5$ with estimated standard error of $\hat{\sigma}\ v_k = 0.2$. What is the largest value of w for which we would NOT reject the null hypothesis that $\hat{w}_1 = w$? (hint: assume normal approximation to t distribution, and that we are using the $\alpha = 5\%$ significance level for a two-sided test).

Exercise 2.12

Which of the following statements are true? Provide motivations of your answers.

1. The estimate w_1 in a linear regression for many variables (i.e., a regression with many predictors in addition to x_1) is usually a more reliable measure of a causal relationship than w_1 from a univariate regression on X_1;
2. One advantage of using linear models is that the true regression function is often linear;

3. If the F-statistic is significant, all of the predictors have statistically significant effects;

4. In a linear regression with several variables, a variable has a positive regression coefficient if and only if its correlation with the response is positive.

Exercise 2.13

Let us assume that the solution with the LS method of a regression problem on a specific dataset has as a result:

$$\hat{t} = 5 + 4x.$$

We would like to repeat the same regression with a Gaussian Bayesian prior over the parameter space $[w_0,\ w_1]$ with mean $\mu = [3,\ 2]^T$ and covariance matrix $\sigma^2 = I_2$ (i.e., an identity matrix of order 2).

Which one/ones of the following paramters $\mathbf{w}$ is/are consistent solution/solutions to the previous regression problem with the previously specified Bayesian prior?

1. $\mathbf{w} = [5,\ 4]$;
2. $\mathbf{w} = [4,\ 3]$;
3. $\mathbf{w} = [6,\ 5]$;
4. $\mathbf{w} = [3,\ 2]$.

* Exercise 2.14

Derive the analytical solution for the *Ridge Regression*. We remember that it considers as loss function the following one:

$$J(\mathbf{w}) = \sum_{n=1}^{N} (\mathbf{w}^T x_n - t_n)^2 + \lambda ||\mathbf{w}||_2^2$$

Derive the gradient descent scheme for the Ridge Regression.