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.

Answers

Answer of exercise 2.1

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).

Answer of exercise 2.2

1. OK: this is one of the most formal definition for ML, provided by Mitchell, highlighting the central role of experience to solve a problem;
2. KO: many application fields uses models and relationships to solve problems. For instance, some techniques uses physical model to provide predictions;
3. OK: here we underline that the experience is able to generate some models which are used to perform induction and use the learned models to generalize;
4. KO: ML does not generate new information. The ML techniques are only able to gather the most interesting information from raw data;
5. OK: this is one of the more informal definition for ML, provided by Samuel, highlighting that with ML we are able to make inference from data without an explicit implementation of the procedure.

Answer of exercise 2.3

1. The problem of dividing into categories the employee of a company can either be a machine learning problem (e.g., if we do not know the criterion used to partition) or not (e.g., if we are given the criteria). In the first case, we would use some unsupervised ML techniques (e.g., clustering) by considering for instance the personal data of each employee.
2. In principle, fortune-telling is not science. If you consider it as a process where a fortune-teller is able to infer information about a person by looking at her/him,

we could use a ML algorithm to determine the important features to infer information from a person picture (feature selection) and another algorithm able to couple the appropriate phrase to provide to each person (classification). For instance a "magic" methodology has been used in the past to predict what a person is thinking `http://www.google.com/patents/US20060230008`;

3. In the case we have a logical formula and we simply want to derive if it is true or false, it would be unnecessarily complicate to use a ML algorithm, since it exists a deterministic procedure to provide the answer in a fast way.

4. The computation of the stress consists in the application of the mathematical or physical model to a specific case. Here, we might resort to analytical solutions, in the case they exists, or techniques in the field of scientific computing, where an approximation is given by using a discretization scheme.

5. In this case, since we want to predict a continuous value (e.g., temperature), we are considering a regression problem. One might consider the use of linear regression models, regression trees or neural networks. If we consider these methodology and take into account the temperature records of each day as independent from the previous one, then the considered methodologies are consistent with the problem. If we want to take into account the fact that the temperature is a time series (there exists correlation among days) we should resort to different techniques, e.g., time series prediction techniques like AutoRegressive (AR) models.

Answer of exercise 2.4

1. **Regression problem** One might consider as features the amount of rooms in a house, its distance to the city center, the presence of nearby facilities (primary school, church, underground), age of the building.

2. **Classification problem** In the specific, this problem can be categorized as an anomaly detection one. We can consider as interesting variables the amount of stock actions of a firm exchanged by traders, their affiliation and information about their money transaction.

3. **Feature extraction/selection** The interesting features in a image recognition problem are often application dependent. One of the most popular approach considered nowadays is the one offered by deep networks which automatically identifies the interesting features from an image for a given task.

4. **Classification problem** In this case, we would need a dataset composed by the couples audio recording/species from all the species we would like to discriminate.

5. **Reinforcement learning problem** One of the techniques we might consider is

q-learning and as features we might consider records from air hockey matches played by humans, for instance where you have information about the position of the hockey mallet and the forces used to kick the puck.

6. **Multiple interpretation** In this kind of problem one might consider the problem as a missing data reconstruction problem, a classification problem or a clustering one. In all the previous interpretation the features we would like to have are records (evaluations) coming from previously seen users.

7. **Classification problem** Here some of the solutions performing really well make use of deep neural networks, but also Support Vector Machine (SVM) might be an option. If we want to categorize a handwritten digit into the classes $0, \dots, 9$ we need images of handwritten digits as features, coupled with the correct class. One might also consider it as a twofold problem of identifying the correct features in an image and then classification into 10 classes.

8. **Online learning problem** Here we might consider algorithm coming from the Multi-Armed Bandit (MAB) framework. The problem of pricing a good, in the case we do not have information about the user, is a problem of selecting the best action in the shortest time. Here, we simply need the outcome of newly seen buyers. In the case we have information about users we might resort to techniques coming from the recommendation field.

Answer of exercise 2.5

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 w.r.t. 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 in an effective way. That is not always true and depends also on the loss function we want to minimize.

Answer of exercise 2.6

1. YES: since the most computationally complex part of the method implies the inversion of a design matrix, if we have only few parameter we only need invert a small matrix.
2. NO: it does not exists a closed form solution for the so called Laplace Loss function. Differently from the squared loss, this loss is not derivable in the origin, thus the definition of its derivative is not unique there.
3. YES/NO: in principle the inversion of the LS matrix is linear in the number of samples considered, therefore it does not depend on the number of samples. Nonetheless, to compute the design matrix $X^\top X$ it is required to perform a matrix multiplication, which is linear in the number of samples. Therefore for extremely large dataset this operations might be unfeasible.
4. NO: again the LS method only minimizes the RSS. The so called Huber Loss can be minimized by an iterative method. (FYI It is robust to noise, but differently from the squared one does not have a unique solution).

Answer of exercise 2.7

1. We are getting a badly conditioned design matrix.
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:

   $$t = w_1 x + w_2 2x = (w_1 + 2w_2)x$$

   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.

Answer of exercise 2.9

The update for a given parameter for the gradient descent for a single datum is:

$$w_j^{(1)} = w_j^{(0)} - \alpha(\mathbf{x}_1^T \mathbf{w}^{(0)} - t_1)x_{1j},$$

thus, we have:

$$w_0^{(1)} = 0 - 0.3(2-4) \cdot 1 = 0.6,$$
$$w_1^{(1)} = 0 - 0.3(2-4) \cdot 3 = 1.8,$$
$$w_2^{(1)} = 1 - 0.3(2-4) \cdot 2 = 2.2,$$

and the final coefficient is $\mathbf{w} = [0.6,\ 1.8,\ 2.2]$.

In the case we want to consider a set of $K = 10$ samples we could use the batch update formulation:

$$\mathbf{w}^{(1)} = \mathbf{w}^{(0)} - \frac{\alpha}{K} \sum_{n=1}^{K} (\mathbf{x}_n^T \mathbf{w}^{(0)} - t_n)\mathbf{x}_n$$

Answer of exercise 2.10

Since the matrix $\Phi^T\Phi + \lambda I$ is semi-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}.

Answer of exercise 2.11

We know that the variable $S = \frac{\hat{w}_k - w_k}{\hat{\sigma}\ v_k}$ is distributed as a t-student distribution with $N - M - 1$ degree of freedom. If we use the Gaussian approximation we might say that $S \sim \mathcal{N}(0, 1)$. We are performing a two sided test, thus, we do not reject the null hypothesis if we have:

$$\left| \frac{\hat{w}_k - w_k}{\hat{\sigma} v_k} \right| \leq z_{1-\alpha/2}$$
$$-z_{1-\alpha/2} \leq \frac{\hat{w}_k - w_k}{\hat{\sigma}\ v_k} \leq z_{1-\alpha/2}$$
$$\hat{w}_k - z_{1-\alpha/2}\,\hat{\sigma}\ v_k \leq w_k \leq \hat{w}_k + z_{1-\alpha/2}\,\hat{\sigma}\ v_k,$$

where we are using the symmetry properties of the Gaussian distribution. The maximum value for which we would not reject the null hypothesis is:

$$w_k = \hat{w}_k + z_{1-\alpha/2}\,\hat{\sigma}\ v_k = 0.5 + 1.96 \cdot 0.2 \approx 0.9\ \ (0.892).$$

Answer of exercise 2.12

1. FALSE: Adding extra predictors to the model can obfuscate the interpretation of w_1 in the model since there might be some interdependencies among features in the real process generating data.
2. FALSE: Even if the process considered is often non-linear, a linear model might give interesting insight on it.
3. FALSE: The significance of the F-statistic means that at least one of the considered variables are meaningful to model the considered relationship.
4. FALSE: Positive correlation and a positive coefficients are equivalent only in a univariate regression.

Answer of exercise 2.13

1. NOT CONSISTENT: Since the prior does not coincide with the MLE solution it is not possible that the solution does not change.
2. CONSISTENT: The effect of the prior moves the solution towards the areas in the parameter space where the prior has higher probability.
3. NOT CONSISTENT: The effect of the prior does not allow the solution to move further to the region of prior maximal probability, i.e., the point $[3,\ 2]$
4. NOT CONSISTENT: Since the prior does not coincide with the MLE solution it is not possible that the solution coincides with the prior.