MACHINE LEARNING READING GROUP, LINEAR REGRESSION

PAWEŁ NOSAL

1. Introduction

These notes are meant to serve as equivalent form of the talk given by me on the Machine Learning reading group. They should not add anything to the talk and neither the other way around, but in case you have any questions please feel free to contact me or the Reading Group organizers. The talk is mostly based on [1] and [2]. If something is not included in these notes one should consult these sources.

2. Linear Least-squares regression

The focus of this whole talk is to introduce the participants to possibly the simplest regression model, which is the linear least-squares regression. The framework is that we have some data pairs $(\underline{x}, y)_{1\leq i\leq n}$ where we assume for simplicity that $\underline{x}_i \in X, y_i \in \mathbb{R}$ that our loss function of choice is the square loss function.

$$\ell(y, z) := (y - z)^2.$$

The goal, is to find a function $f : X \to \mathbb{R}$ of $\underline{x}$ that represents the dependence of y on $\underline{x}$ through minimizing the empirical risk:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\underline{x}))^2.$$

In the set up of linear regression, we choose to find the best representation of the aforementioned function in a family of prediction functions $f_{\underline{\theta}}$ (often called models in statistics) that are parametrized by $\underline{\theta} \in \Theta$. This idea does not necessarily force us to work with functions that are linear in $\underline{\theta}$, as in the case e.g. in neural networks, but this is the case that we shall focus on in this talk, so we assume that $\underline{\theta} \in \mathbb{R}^d, d \in \mathbb{N}_{>0}$.

Note however, that even though we want the regression to be linear in $\underline{\theta}$, this doesn't say anything about linearity in $\underline{x}$. Especially, many models that can be modelled using

linear regression can have pretty much polynomial structure, where we are optimizing over the linear coefficients $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_d)$. In fact, if X is not a vector space, the concept of linearity might not even make sense there. Therefore, the family of functions that we are optimizing over has the shape

$$f_{\underline{\theta}}(\underline{x}) = \sum_{i=1}^{d} \alpha_i(\underline{x})\theta_i,$$

for functions $\alpha_i : X \to \mathbb{R}, i \leq d$. If we write $\underline{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_d)$, we can rewrite the above as

$$f_{\underline{\theta}}(\underline{x}) = \underline{\alpha}(\underline{x})^T\underline{\theta}.$$

The vector $\underline{\alpha}$ is often called the feature vector, and is assumed to be known. Therefore, given the feature vector $\underline{\alpha}$, the original problem becomes that of choosing $\underline{\theta}$ that minimizes the empirical risk in the following form

$$\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \underline{\alpha}(\underline{x}_i)^T\underline{\theta})^2.$$

When $X \subseteq \mathbb{R}^d$ we can make f_θ affine by adjoining 1 to all vectors $\underline{x}_i$.

For the rest of this talk we shall switch to matrix notation. Let $\underline{\mathbf{y}} = (\mathbf{y}_1, ..., \mathbf{y}_n) \in \mathbb{R}^n$ be the vector of outputs (response vector) and $\Phi \in \mathbb{R}^{n\times d}$ the matrix of inputs whose rows are $\underline{\alpha}(\underline{x}_i)$. We call it the data matrix. In this notation, the empirical risk can be written as

$$\hat{R}(\underline{\theta}) = \frac{1}{n}||\underline{y} - \Phi\underline{\theta}||_2^2,$$

with obvious notation.

3. Ordinary Least-Square Estimator

Notice that in geometry terms, $\hat{R}(\underline{\theta})$ is just a squared Euclidean length of the vector $\underline{y} - \Phi\underline{\theta}$, its value shall remain the same if rotated or reflected. This observation will serve as a basis for practical method for finding the estimator for $\underline{\theta}$. We use it in the following way: As for any real matrix, we can use the procedure called the QR decomposition, that allows us to represent Φ as $\Phi = \mathcal{Q}\begin{bmatrix} R \\ 0 \end{bmatrix} = QR$, where $\mathcal{Q}$ is an orthogonal $n \times n$ matrix, whose first p columns form Q, and R is a $d \times d$ upper triangular matrix. Now, we use the previous

observation and the fact that orthogonal matrices only rotate or reflect vectors, but to not change their length and also $\mathcal{Q}^T\mathcal{Q} = \mathcal{Q}\mathcal{Q}^T = I_n$. Therefore, multiplying $||\underline{y} - \Phi\theta||_2$ by $\mathcal{Q}^T$ implies

$$||\underline{y} - \Phi\underline{\theta}||_2^2 = ||\mathcal{Q}^T\underline{y} - \mathcal{Q}^T\Phi\underline{\theta}||_2^2 = \left|\left|\mathcal{Q}^T\underline{y} - \begin{bmatrix} R \\ 0 \end{bmatrix}\underline{\theta}\right|\right|_2^2$$

In this form, we can notice that if we write $\mathcal{Q}^T\underline{y} = \begin{bmatrix} a \\ b \end{bmatrix}$, where $a \in \mathbb{R}^d, b \in \mathbb{R}^{n-d}$ the above equation gives

$$||\underline{y} - \Phi\underline{\theta}||_2^2 = ||a - R\underline{\theta}||^2 + ||b||^2.$$

The length of the second vector doesn't depend on $\underline{\theta}$, and the first one can be reduced to zero by choosing $\underline{\theta} = R^{-1}a$. Therefore, provided X (and so R) have full column rank, this gives us the least squares estimator

$$\hat{\underline{\theta}} = R^{-1}a, \tag{1}$$

which tracking back our notation can be also written as $\hat{\underline{\theta}} = (\Phi^T\Phi)^{-1}\Phi^T\underline{y}$.

There is also an additional geometric interpretation of the prediction vector $\Phi\hat{\underline{\theta}}$: it is the orthogonal projection of $\underline{y} \in \mathbb{R}^n$ onto $\mathrm{im}(\Phi)$, the column space of Φ.

References

1. Francis Bach, *Learning theory from first principles.*
2. Simon N. Wood, *Core statistics*, Institute of Mathematical Statistics Textbooks, Cambridge University Press, 2015.

University of Warwick, Mathematics Institute, Zeeman Building, University of Warwick, Coventry CV4 7AL

Email address: Pawe.Nosal@warwick.ac.uk

URL: https://sites.google.com/view/pawelnosalmaths