Statistical Methods: Portfolio 1

By Henry Bourne

Abstract

In this document we will summarise content from the first 4 lectures from the Statistical methods course (at Compass, University of Bristol). These lectures cover content on using statistical methods for decision making.

1 Introduction to Decision-Making

Computers are often used to answer complex questions, however, these methods are often a "black-box". We would like to have methods which let us conduct **rational decision-making**:

1. Predictions should be precise (no gibberish)
2. They should be data driven
3. They should take cost (of making the wrong decision) into consideration
4. They should take the random nature of data into consideration

In a **regression problem** we want to predict an outcome given some known inputs. We can use the following as an objective function:

Definition 1.1 ***Least Squares (LS):***

$$\min_f \sum_{i \in D_0} (y_i - f(x_i))^2 \tag{1}$$

where f is the function that gives our prediction for x, where x_i is the i-th input, y_i is the i-th (observed) output and $D_0 \subseteq D$ is the training dataset.

By minimizing this objective function wrt. f, we obtain a function f that minimizes the squared difference between its predictions and our observed values of the target variable.

Let $\boldsymbol{w}$ be a vector parameterising f [1], then the LS solution is $\boldsymbol{w}_{LS} := \text{argmin}_{\boldsymbol{w}} \sum_{i \in D_0}(y_i - f(\boldsymbol{x}_i; \boldsymbol{w}))^2$. We can prove that:

$$\boldsymbol{w}_{LS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{2}$$

The proof can be found in the appendix (CITE).

Alternatively we can find $\boldsymbol{w}$ in another data-driven way but also whilst taking the randomness of the data into account by using a probabilistic approach. We define a new objective function:

Definition 1.2 ***Maximum likelihood estimation:***

$$\max_{\boldsymbol{w}} \log \mathbb{P}(D|\boldsymbol{w}) \tag{3}$$

*we denote the parameters that maximize this as $\boldsymbol{w}_{ML}$, this is called the **Maximum Likelihood Estimator (MLE)**, we can write,*

$$\boldsymbol{w}_{ML} := \underset{\boldsymbol{w}}{\text{argmax}} \log \mathbb{P}(D|\boldsymbol{w}) \tag{4}$$

*where D is the dataset and $\mathbb{P}(D|\boldsymbol{w})$ is called the **likelihood**.*

[1] In the case of linear LS we have: $f(\boldsymbol{x}; \boldsymbol{w}) := \langle \boldsymbol{w_1}, \boldsymbol{x} \rangle + w_0$.

Note that we can show $\boldsymbol{w}_{ML} = \boldsymbol{w}_{LS}$[2]. We can also show that $\sigma^2_{ML} = \frac{1}{n}||\boldsymbol{y} - f(\boldsymbol{x}; \boldsymbol{w}_{ML})||^2$

Appendices

A Proofs

A.1 Proof of Equation (2)

Let $\boldsymbol{X}$ be,

$$\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{5}$$

we can write our linear model in matrix form,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon} \tag{6}$$

note that,

$$\sum_{i \in D_0} (y_i - f(\boldsymbol{x}_i; \boldsymbol{w}))^2 = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2 \tag{7}$$

we can find the minimum by differentiating wrt. $\boldsymbol{w}$ and finding the solution when the gradient equals zero. However, first we expand our expression,

$$\begin{aligned} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2 &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) && (8) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} && (9) \end{aligned}$$

we now can find the derivative wrt. $\boldsymbol{w}$,

$$\frac{\partial}{\partial \boldsymbol{w}}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2 = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} \tag{10}$$

now setting this to zero and solving,

$$\begin{aligned} &\Rightarrow \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y} && (11) \\ &\Rightarrow \boldsymbol{w} = \boldsymbol{X}^{-1}\boldsymbol{X}^{-T}\boldsymbol{X}^T\boldsymbol{y} && (12) \\ &\Rightarrow \boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} && (13) \end{aligned}$$

hence proven.

References

[2]This is true in cases where the underlying data-generating process is normal, ie. error terms Gaussian and iid.