CQF Exam 3
June cohort 2021

Andrea Russo

November 2021

Part B: Mathematics of Supervised Learning

Cost function and MLE for logistic classifier

Consider a classification problem. In particular a binary classification where the target takes values in the binary distribution $y \in \{0,1\}$. In this case, linear regression algorithms are not a good choice of model to construct a precise classifier. Among many reasons for why this is true, one of the most relevant is that we know that the outputs of the model should be binary numbers, which makes it complicated to interpret values smaller than 0 or greater than 1.

The need of a better algorithm can be satisfied by **logistic regression** models. The logistic regression algorithm h_β, where $\beta \in \mathbb{R}^n$ are the parameters, will consistently output results in the range $h_\beta(x) \in [0,1]$.

The function of choice is called the **sigmoid** or **logistic** function. Given the vector of data $x = \vec{x} = (1, x_1, x_2, ..., x_n)$ and the coefficient vector $\beta = \vec{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_n)$, the logistic function takes the form

$$h_\beta(x) = g(\beta \cdot x) = \frac{1}{1 + e^{-\beta \cdot x}} \tag{1}$$

where a constant bias β_0 is inserted through the 1st element of the dot product $\beta \cdot x = \beta_0 + \sum_i \beta_i x_i$.
Given the construction of the sigmoid function and the binary property of the target y, we know that the following is true about the conditional probability $p(y|x;\beta)$:

$$p(y=1|x;\beta) = h_\beta(x) \tag{2}$$

$$p(y=0|x;\beta) = 1 - h_\beta(x) \tag{3}$$

These equations can be put together such that the full conditional probability can be expressed as:

$$p(y|x;\beta) = h_\beta(x)^y (1 - h_\beta(x))^{1-y} \tag{4}$$

Therefore, we can now proceed to write the likelyhood function

$$\begin{aligned} \mathcal{L}(\beta) &= p(y|x;\beta) && (5) \\ &= \prod_{i=1}^{m} p(y_i|x_i;\beta) && (6) \\ &= \prod_{i=1}^{m} h_\beta(x_i)^{y_i} (1 - h_\beta(x_i))^{1-y_i} && (7) \end{aligned}$$

To simplify the algebra, we then construct the log-likelyhood function with the intent of exploiting the property of the log

$$\begin{aligned} l(\beta) &= \log(\mathcal{L}(\beta)) && (8) \\ &= \sum_{i=1}^{m} \Big[y_i \log(h_\beta(x_i)) + (1 - y_i) \log(1 - h_\beta(x_i)) \Big] && (9) \end{aligned}$$

where we recall that β, x and y are vectors. This function is log loss form of the MLE log-likelyhood function. The aim is to now choose $\vec{\beta}$ to maximise $l(\beta)$. If we differentiate this with respect to β_i, we arrive at

$$\frac{\partial l(\beta)}{\partial \beta_i} = x_i(y - h_\beta(x)) \tag{10}$$

which is the fundamental ingredient needed to apply numerical methods capable of finding the maximum of $l(\beta)$.

MSE and regression models

In context of regression methods, can there exist an estimator with the smaller MSE than minimal least squares?

The answer to this question is not as straight forward as it first appears to be, as it depends depends on if we are discussing about *unbiased* or *biased* estimators.

In the case of unbiased estimators, the answer is clearly **NO**. This has been formally proven in the **Gauss-Markov theorem**, which states that the ordinary least squares estimator has the lowest sampling variance within the class of linear unbiased estimators, if the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero, which are assumptions usually made in the context of generalised linear regression.

If we are instead talking about biased estimators, then there are examples of estimators with lower variances. If we stick to linear models, then an example is supplied by Ridge regression, which helps particularly when the predictor variables carry some form of multicollinearity.

For a prediction, does the MSE measure an irreducible error or model error?

The MSE should be written as

$$\text{MSE}(\beta) = \left(\text{Bias}[\hat{\beta}]\right)^2 + \text{Var}[\hat{\beta}] + \sigma^2 \tag{11}$$

where σ^2 represents the irreducible variance of the intrinsic noise in present in the data. This equation represents the Bias-Variance tradeoff, since all three terms are non-negative, the irreducible error forms a lower bound on the expected error on unseen samples. Given that the Bias$[\hat{\beta}]$ and Var$[\hat{\beta}]$ are model dependent but σ^2 is not, the correct answer to the question is the following:

”The MSE is a measure of the sum of the model error and the intrinsic irreducible error. The model error can be minimised, but it will never be possible to reduce the MSE more than the irreducible error”