Elements of statistical learning: Chapter 2

July 10, 2020

Content

1. Introduction
2. Linear models and least squares
   - Linear regression
   - Linear classification
3. Nearet-neigbour algorithm

Introduction

Supervised learning

Goal is to use *inputs* to predict *outputs*.

- inputs are also referred to as *predictors*, *features* or *independent variable*
- outputs are also referred to as *response* or *dependent variables*

The outputs may be

- quantitative (takes values in $\mathbb{R}$)
- qualitative (also known as categorical or discrete)
  - Ordered (e.g. small, medium or large)
  - Unordered (e.g. pass or fail, on or off)

Regression vs classification

We use *regression* to predict quantitative outputs and *classification* to predict qualitative outputs.

Notations

For different predictors $\{X_k\}_{k=1}^{p}$ across different observations $i = 1, \cdots, n$ denote

$X_{i,k}$ for random variable and $x_{i,k}$ for an observation

X_i for a $p \times 1$ vector of variables $- i.e. X_i = [X_{i,1}, \cdots, X_{i,p}]'$

$\boldsymbol{X}$ for a $N \times p$ matrix of variables across different obvs

In other words

$$\boldsymbol{X} = \begin{bmatrix} X_{11} & \cdots & X1p \\ X_{21} & \cdots & X2p \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix} = \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{bmatrix}$$

Therefore, denotes (X_i, Y_i) as the random variables and (x_i, y_i) as the observed values at i.

Linear models and least squares

To predict Y (which would be denoted by $\hat{Y}$), we use the linear regression model, which may be expressed as

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i, \quad i = 1, \cdots, n \tag{1}$$

or as

$$Y_i = \beta_0 + \sum_{k=1}^{P} \beta_k X_{i,k} + \epsilon_i, \quad i = 1, \cdots, n \tag{2}$$

or as

$$Y_i = X_i' \beta + \epsilon_i, \quad , i = 1, \cdots, n \tag{3}$$

where $X_i = [1, X_{i,1}, \cdots, X_{i,p}]'$ and $\beta = [\beta_0, \beta_1, \cdots, \beta_p]'$ are $(p+1) \times 1$ vectors.

In matrix notation, the above can be expressed as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon} \tag{4}$$

which if expanded is expressed as follows

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{5}$$

Least squares estimation

The least squares estimation in one approach to fit the model. In essence, we find the coefficiens β that minimize the sum of squared residuals. Thus, we wish to minimize $RSS(\beta)$

$$\arg\min_{\beta} RSS(\beta) \text{ or } (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

The above can be rearranged and expanded to (which is not necessary, as chain rule can be used)

$$\begin{aligned}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{y}' - \boldsymbol{\beta'}\boldsymbol{X'})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta'}\boldsymbol{X'}\boldsymbol{y} + \boldsymbol{\beta'}\boldsymbol{X'}\boldsymbol{X}\boldsymbol{\beta}\end{aligned}$$

Differentiating with respect to β yields

$$\begin{aligned} \frac{\partial RSS(\beta)}{\partial \beta} &= 0 - \boldsymbol{X'y} - \boldsymbol{X'y} + 2\beta \boldsymbol{X'X} \\ &= -2\boldsymbol{X'y} + 2\beta \boldsymbol{X'X} \end{aligned}$$

and as this is a minimization problem

$$\begin{aligned} \frac{\partial RSS(\beta)}{\partial \beta} &= 0 \\ -2\boldsymbol{X'y} + 2\beta \boldsymbol{X'X} &= 0 \\ \beta \boldsymbol{X'X} &= \boldsymbol{X'y} \end{aligned}$$

and thus so long as $\boldsymbol{X'X}$ is non-singular, the LS estimator of $\hat{\beta}$ is

$$\hat{\beta} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \tag{6}$$

Logit models

Now let us consider the case , where the dependent variable Y_i can assume only two categories (say win or lose), and hence two discrete values (i.e. $Y_i = 0$ or $Y_i = 1$), where as the vector of independent variables are continuous, say $X_i \in \mathbb{R}^p$.

In order to restrict Y_i to 0 and 1. In this case it would make sense to make the probability of $Y_i = 1$ and not the value of Y_i itself. This leads to a probability model, which specifies the the probbility of the outcome as a function of the predictor:

$$\begin{aligned} P[Y_i = 1] &= P[X_i, \beta] \quad (7) \\ P[Y_i = 0] &= 1 - P[X_i, \beta] \quad (8) \end{aligned}$$

Since P is a probability, it is bounded between 0 and 1. The regression equation may be revived by briefly denoting

$$P(X_i, \beta) = X_i' \beta$$

As we wish the pobability to vary monotically with X, we may use a *sigmoid* function:

$$P(X_i, \beta) = \frac{\exp(\beta' X_i)}{1 + \exp(\beta' X_i)} \tag{9}$$

Let us denote $Z_i = \beta' X_i$, then

$$\lim_{z \to \infty} \frac{\exp(z)}{1 + \exp(z)} = 1$$

and

$$\lim_{z \to -\infty} \frac{\exp(z)}{1 + \exp(z)} = 0$$

Therefore,

$$P[Y_i = 1] = \frac{\exp(\beta' X)}{1 + \exp(\beta' X)}$$

and

$$P[Y_i = 0] = \frac{1}{1 + \exp(\beta' X)}$$

Alternatively, one could look at the odd $P[Y_i = 1]/P[Y_i = 0]$, which may be expressed as

$$\begin{aligned} \frac{P[Y_i = 1]}{P[Y_i = 0]} &= \frac{\exp(\beta' X)}{1 + \exp(\beta' X)}[1 + \exp(\beta' X)]. \\ &= \exp(\beta' X). \end{aligned}$$

now taking the logarithm from both sides will yield

$$\log(odds) = \beta' X \tag{10}$$

where now $\log(odds)$ is no longer bounded by 0 and 1.

Nearet-neighbour algorithm

- A non-parametric approach used for both *classification* and *regression*
- Input consists of the k closest training examples in the feature space.
- Output depends on whether $K - NN$ is used or classification or regression.
- For classification, the output is a class membership
- For regression, this value is the average of the k nearest neighbbours

Dependence modelling using copulas

- As the mediangale assumption allows for non-linear serial dependence, testing assumption A1 by considering linear correlation is inappropriate.
- we suggest fitting copula models, which provide the means of separating the marginal distribution of the process from their respective dependence structure.
- For instance, the marginals can be assumed to possess standard normal distributions, while the nonlinear dependency is modeled using rank correlation measures (e.g. Kendall Tau) and copulas, where the latter are invariant under monotonic transformations; hence, they are not affected by the marginal distributions [see].

- A special case is where $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{T-1}, \varepsilon_T$ are distributed according to $N(0,1)$. As suggested before, since the form of the serial dependence of the errors is non-linear, we may estimate the bivariate probabilities using copulas.
- The latter stem from a Theorem brought forward by, which states that there is a copula C such that

$$F_\varepsilon(-\beta' x_0, \cdots, -\beta' x_{n-1}) = C(F_1(-\beta' x_0), \cdots, F_n(-\beta' x_{n-1})).$$

- Therefore, as noted earlier, the bivariate probabilities can be calculated using copulas, which separate the marginal distribution of the error terms from their dependence structure. We may consider a rank correlation measure (e.g. Kendall Tau) as the measure of non-linear dependency and choose the Gaussian copula for evaluating the bivariate probabilities $P[\, \varepsilon_{t-1} < \cdot,\ \varepsilon_t < \cdot \mid \mathcal{I}_{t-1}]$.

- First let us forget about the Theorem mentioned earlier, where $s(\varepsilon_1), \cdots, s(\varepsilon_n)$ are i.i.d conditional on $\mathcal{I}_{t-1}$. Let us explore this from an earlier point and what led to this Theorem in . We know from the mediangale assumption that

$$P[s(\varepsilon_t) = s_t \mid s(\varepsilon_{t-1}) = s_{t-1}, \cdots, s(\varepsilon_1) = s_1] = \\ P[\varepsilon_t \geq 0 \mid \varepsilon_{t-1}, \cdots, \varepsilon_1]^{s(\varepsilon_t)} P[\varepsilon_t < 0 \mid \varepsilon_{t-1}, \cdots, \varepsilon_1]^{1-s(\varepsilon_t)} = \frac{1}{2}$$

- As this corresponds only to the median, it is only true *iff* both sides of the inequality are zero. SInce the probabilities $P[\varepsilon_t \geq 0 \mid \varepsilon_{t-1}, \cdots, \varepsilon_1]$ and $P[\varepsilon_t < 0 \mid \varepsilon_{t-1}, \cdots, \varepsilon_1]$ are the same and equal to $1/2$ for all $t = 1, \cdots, n$. Thus, it can be argued that $s(\varepsilon_1), \cdots, s(\varepsilon_n)$ which are defined by the aformentioned probabilties can be considered i.i.d, as their distribution is determined by the said probabilities, which are identical over time.

- And here is the proposition (pay careful attention to the last paragraph):

- Now let us instead assume that we are interested in the signs $\tilde{s}(\varepsilon + c_t)$, where c_t for $t = 1, \cdots, n$ are some constants. Then the likelihood function in terms of the signs $\tilde{s}_1, \cdots, \tilde{s}_n$

$$L(\tilde{s}_1, \cdots, \tilde{s}_n \mid X) = \\ \prod_{t=1}^{n} P[\varepsilon_t \geq -c_t \mid \varepsilon_1, \cdots, \varepsilon_n]^{\tilde{s}_t} P[\varepsilon_t < -c_t \mid \varepsilon_1, \cdots, \varepsilon_n]^{1-\tilde{s}_t}$$

- Now the joint p.m.f. depnends on all the past signs and and thus is no longer *i.i.d*, as it violates the Mediangale assumption (particularly the point about permutations mentioned in Proposition 1 of). In other words, we no longer have that the p.m.fs $P[\varepsilon_t < -c_t \mid \varepsilon_{t-1}, \cdots, \varepsilon_1]$ are constant and identical over time, so the distribution of the signs $\tilde{s}(\varepsilon_1 + c_1), \cdots, \tilde{s}(\varepsilon_n + c_n)$, now dependens on the joint p.m.fs.

Alternative proof

- Under the null hypotheis, it is evident that

$$P[\varepsilon_t \geq 0 \mid X] = P[\varepsilon_t < 0 \mid X], \quad t = 1, \cdots, n$$

is equivalent to

$$P[y_t \geq 0 \mid X] = P[y_t < 0 \mid X], \quad t = 1, \cdots, n$$

- Therefore, the mediangale assumption can be extended to $y_1, \cdots, y_n$:

$$P[y_t \geq 0 \mid y_1, \cdots, y_n, X] = P[y_t < 0 \mid y_1, \cdots, y_n, X] = \frac{1}{2}$$

Alternative proof

- Then the variables $s(y_1), \cdots, s(y_n)$ are i.i.d conditional on X according to the distribution

$$P[s(y_t) = 1 \mid X] = P[s(y_t) = 0 \mid X] = \frac{1}{2}, \quad t = 1, \cdots, n$$

- Under the alternative hypothesis, however, the mediangale property does not hold for permutations $\pi : i \to j$ as noted in the Proposition 3.1 of , as the conditional distribution of the signs vary across observations.
- Therefore, the signs $s(u_1), \cdots, s(y_n)$ can no longer be assumed to be i.i.d.