BCAM & UPV/EHU Course

Regression

Introduction to Machine Learning

Carlos Cernuda

Carmen Ana Domínguez-Bravo

basque center for applied mathematics

Outline

What is regression?

Now we are interested on predicting a **quantitative** response variable. Similar to classification but where the response is continuous.

We want to build a model from observed data to predict a quantitative response Y on the basis of one or some predictor variables $\boldsymbol{X}$.

What is regression?

Now we are interested on predicting a **quantitative** response variable. Similar to classification but where the response is continuous.

We want to build a model from observed data to predict a quantitative response Y on the basis of one or some predictor variables $\boldsymbol{X}$.

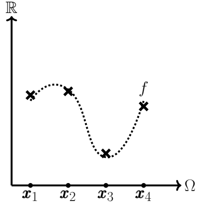
- Y Response (e.g. sales) (variable to be predicted)
  It can take any real (continuous) value, quantitative.
- $\boldsymbol{X}$ Predictors (e.g. advertising budget)

What is regression?

We assume that the true relationship between predictors and response is

$$Y = f(\boldsymbol{X}) + \epsilon \qquad\qquad Y \sim f(\boldsymbol{X})$$

- f unknown function (regression model),
- ϵ mean-zero random noise.



What is regression?

Given observed data $(\boldsymbol{x}_i, y_i)$ for $i = 1, \ldots, N$

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n), (\boldsymbol{x}_{n+1}, y_{n+1}), \ldots, (\boldsymbol{x}_{N-n}, y_{N-n})$$

How to obtain the model that best describe them in order to make a good prediction at x_0?

What is regression?

Given observed data $(\boldsymbol{x}_i, y_i)$ for $i = 1, \ldots, N$

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n), (\boldsymbol{x}_{n+1}, y_{n+1}), \ldots, (\boldsymbol{x}_{N-n}, y_{N-n})$$

How to obtain the model that best describe them in order to make a good prediction at x_0?

Select:

1. structure of the model,
2. fitting method to build the model,
3. train (build the model),
4. test (measure the accuracy of the model).

1. **Structure of the model:** $Y = f(\boldsymbol{X}) + \epsilon$
   - Parametric approach (assumptions over the shape of f)
   - Non-parametric approach (does not assume a form over f)

Parametric regression

- **E.g. linear regression:**

$$Y_\beta = \beta_0 + \beta_1 \boldsymbol{X} + \epsilon$$

Parametric regression

- **E.g. linear regression:**

$$Y_\beta = \beta_0 + \beta_1 \boldsymbol{X} + \epsilon$$

- Simplify the model function to a known form.

Parametric regression

- **E.g. linear regression:**

$$Y_\beta = \beta_0 + \beta_1 \boldsymbol{X} + \epsilon$$

- Simplify the model function to a known form.
- The model requires the estimation of a finite number of parameters.

Parametric regression

- **E.g. linear regression:**

$$Y_\beta = \beta_0 + \beta_1 \boldsymbol{X} + \epsilon$$

- Simplify the model function to a known form.
- The model requires the estimation of a finite number of parameters.
- No matter how much data you use, the model will not change its mind about the parameters it needs.

Non-parametric regression

- **E.g. K nearest neighbour (KNN regression)**

Non-parametric regression

- **E.g. K nearest neighbour (KNN regression)**
- Do not make strong assumptions about the form of the function.

Non-parametric regression

- **E.g. K nearest neighbour (KNN regression)**
- Do not make strong assumptions about the form of the function.
- Free to learn “any functional form” from the training data.

Non-parametric regression

- **E.g. K nearest neighbour (KNN regression)**
- Do not make strong assumptions about the form of the function.
- Free to learn "any functional form" from the training data.
- Good when you have a lot of data and no prior knowledge.

2. Fitting method to determine the model parameters: Y_β

- Given data $(\boldsymbol{x}_i, y_i)$ for $i = 1, \ldots, n$,
- Determine the model parameters by minimizing a cost function.
- Example:
  - Error or residual value at $\boldsymbol{x}_i$: $e_i = y_i - y_{\beta,i}$
  - Cost function: Squared Error or Residual Sum of Squares

$$RSS = \sum_{i=1}^{n} e_i^2$$

  - Select $\widehat{\beta}$ that minimizes the cost function.
  - Model obtained: $\widehat{Y} = Y_{\widehat{\beta}}$

Outline

Linear regression model

$$Y = f(\boldsymbol{X}) + \epsilon = \beta_0 + \beta_1 \boldsymbol{X} + \epsilon$$

- It assumes that there is approximately a linear relationship between a predictor $\boldsymbol{X}$ and the response Y.
- Model parameters or coefficients: β_0 and β_1.
- Valid for some practical problems and used as base by other more sophisticated models.

Linear regression $Y = \beta_0 + \beta_1 \boldsymbol{X} + \epsilon$

Given data $(\boldsymbol{x}_i, y_i)$ for $i = 1, \ldots, n$,

How to calculate the model parameters **?**

Estimate the parameters from the data such that
at each point $\boldsymbol{x}_i$ the model predicted value is similar to the true response value observed.

Linear regression $Y = \beta_0 + \beta_1 \boldsymbol{X} + \epsilon$

Given data $(\boldsymbol{x}_i, y_i)$ for $i = 1, \ldots, n$,

How to calculate the model parameters **?**
Estimate the parameters from the data such that
at each point $\boldsymbol{x}_i$ the model predicted value is similar to the true response value observed.
Example:

$$\widehat{\beta}_0, \widehat{\beta}_1 = \arg\min_{\beta} \sum_{i=1}^{n} e_i^2 = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \boldsymbol{x}_i)^2$$

Choose β_0, β_1 that minimize RSS (Least Squares Fitting).

Example: Simple Linear regression

course_regression_1_SimpleLinear.ipynb

K Nearest Neighbour regression

Given data $(y_i, \boldsymbol{x}_i)$ for $i = 1, \ldots, n$,
Predict at point $\boldsymbol{x}_0$ (unobserved) the response value y_0.

The method works as follows:

K Nearest Neighbour regression

Given data $(y_i, \boldsymbol{x}_i)$ for $i = 1, \ldots, n$,
Predict at point $\boldsymbol{x}_0$ (unobserved) the response value y_0.

The method works as follows:

1. Identify the K data point closest to $\boldsymbol{x}_0$, that is, the local neighbourhood $\{(y_i, \boldsymbol{x}_i)\}_{i \in I_0}$ with $|I_0| = K$

K Nearest Neighbour regression

Given data $(y_i, \boldsymbol{x}_i)$ for $i = 1, \ldots, n$,
Predict at point $\boldsymbol{x}_0$ (unobserved) the response value y_0.

The method works as follows:

1. Identify the K data point closest to $\boldsymbol{x}_0$, that is, the local neighbourhood $\{(y_i, \boldsymbol{x}_i)\}_{i \in I_0}$ with $|I_0| = K$
2. Estimate the response value y_0 at $\boldsymbol{x}_0$ by using the average of the neighbours: $y_0 = \dfrac{1}{K} \sum_{i \in I_0} y_i$

K Nearest Neighbour regression

Given data $(y_i, \boldsymbol{x}_i)$ for $i = 1, \ldots, n$,
Predict at point $\boldsymbol{x}_0$ (unobserved) the response value y_0.

The method works as follows:

1. Identify the K data point closest to $\boldsymbol{x}_0$, that is, the local neighbourhood $\{(y_i, \boldsymbol{x}_i)\}_{i \in I_0}$ with $|I_0| = K$
2. Estimate the response value y_0 at $\boldsymbol{x}_0$ by using the average of the neighbours: $y_0 = \dfrac{1}{K} \sum_{i \in I_0} y_i$

K Nearest Neighbour regression

Given data $(y_i, \boldsymbol{x}_i)$ for $i = 1, \dots, n$,
Predict at point $\boldsymbol{x}_0$ (unobserved) the response value y_0.

The method works as follows:

1. Identify the K data point closest to $\boldsymbol{x}_0$, that is, the local neighbourhood $\{(y_i, \boldsymbol{x}_i)\}_{i \in I_0}$ with $|I_0| = K$
2. Estimate the response value y_0 at $\boldsymbol{x}_0$ by using the average of the neighbours: $y_0 = \dfrac{1}{K} \displaystyle\sum_{i \in I_0} y_i$

- non-parametric regression method,
- it is a piecewise constant approximation,
- the model is smoother as the number of neighbours increases,
- there are methods to identify the optimal value for K.

K Nearest Neighbour regression

Possible variants ...

- KNN (basic or uniform): all the points in the local neighbourhood has the same weight.
- KNN (distance): the points in the neighbourhood have a weight proportional to the inverse of the distance from $\boldsymbol{x}_0$.
- KNN (radius): the neighbours are the points within a fixed radius of the input point.
- More ...

Example: K Nearest Neighbour regression

course_regression_2_KNN.ipynb

Outline

Multiple linear regression

In practice, p predictor variables!
Extend the linear model to accommodate multiple predictors.

$$\begin{aligned} Y &= \beta_0 + \quad \beta_1 \boldsymbol{X}_1 + \ldots + \beta_p \boldsymbol{X}_p \quad + \epsilon \\ &= \beta_0 + \quad \sum_{j=1}^{p} \beta_j \boldsymbol{X}_j \quad + \epsilon \end{aligned}$$

Coefficient β_j: measure the average effect on the response y when the predictor variable $\boldsymbol{x}_j$ increase one unit and all the other predictors are fixed.

Multiple Linear regression $Y = \beta_0 + \beta_1 \boldsymbol{X}_1 + \ldots + \beta_p \boldsymbol{X}_p + \epsilon$

Given n data with p predictors and one response values:

- $\boldsymbol{x}_{i\,j}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$
- y_i for $i = 1, \ldots, n$

the design matrix will look like

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \sim \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

Multiple Linear regression $Y = \beta_0 + \beta_1 \boldsymbol{X}_1 + \ldots + \beta_p \boldsymbol{X}_p + \epsilon$

Given the data $(\boldsymbol{x}_{ij}, y_i)$, estimate the regression coefficients from the data.

Choose the $\beta = (\beta_1, \ldots, \beta_p)$ that minimizes the RSS:

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \boldsymbol{x}_{i1} - \ldots - \beta_p \boldsymbol{x}_{ip})^2$$

Obtain the model to make new predictions $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \boldsymbol{X}_1 + \ldots + \widehat{\beta}_p \boldsymbol{X}_p$.

Outline

Linear regression but non-additive

Linear model two main restrictive assumptions: **additive** and **linear**.

Additive assumption: the effect of changes in a predictor $\boldsymbol{X}_j$ on the response Y is independent on the values of the other predictors.

$$Y = \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \epsilon$$

Remove additive assumption: introduce an interaction term to allow interactions.

Linear regression but non-additive

Linear model two main restrictive assumptions: **additive** and **linear**.

Additive assumption: the effect of changes in a predictor $\boldsymbol{X}_j$ on the response Y is independent on the values of the other predictors.

$$Y = \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \epsilon$$

Remove additive assumption: introduce an interaction term to allow interactions.

$$\begin{aligned} Y &= \beta_0 + \quad \beta_1 \boldsymbol{X}_1 + \beta_3 \boldsymbol{X}_1 \boldsymbol{X}_2 \quad + \beta_2 \boldsymbol{X}_2 + \epsilon \\ &= \beta_0 + \quad (\beta_1 + \beta_3 \boldsymbol{X}_2) \boldsymbol{X}_1 \quad + \beta_2 \boldsymbol{X}_2 + \epsilon \end{aligned}$$

The effect on $\boldsymbol{X}_1$ is no longer constant, adjusting $\boldsymbol{X}_2$ will change the impact of $\boldsymbol{X}_1$ on Y.

Example: Multiple Linear regression

course_regression_3_MultipleLinear.ipynb

Outline

Non-linear regression but still additive

Linear model two main restrictive assumptions: **additive** and **linear**.

Linear assumption: the change in the response Y due to one-unit change in $\boldsymbol{X}$ is constant, independently of the $\boldsymbol{X}$ value.

$$Y = \beta_0 + \beta_1 \boldsymbol{X} + \epsilon$$

Remove linear assumption: introduce non-linear relationships.

Non-linear regression but still additive

Linear model two main restrictive assumptions: **additive** and **linear**.

Linear assumption: the change in the response Y due to one-unit change in $\boldsymbol{X}$ is constant, independently of the $\boldsymbol{X}$ value.

$$Y = \beta_0 + \beta_1 \boldsymbol{X} + \epsilon$$

Remove linear assumption: introduce non-linear relationships.

$$Y = \beta_0 + \beta_1 \boldsymbol{X} + \beta_2 \boldsymbol{X}^2 + \epsilon$$

The effect on $\boldsymbol{X}$ over Y is now quadratic (polynomial regression).

Non-linear regression but additive

Remove linear assumption: introduce non-linear relationships by using a family of transformations that can be applied to the predictor variable $\boldsymbol{X}$.

$$Y \sim \beta_0 + \beta_1 b_1(\boldsymbol{X}) + \beta_2 b_2(\boldsymbol{X}) + \ldots + \beta_K b_K(\boldsymbol{X})$$

We choose the basis functions $b_k(\cdot)$ and given the data we can estimate the vector of coefficients.

E.g. In the polynomial case $Y \sim \beta_0 + \beta_1 \boldsymbol{X} + \beta_2 \boldsymbol{X}^2$, the basis functions are $\{\boldsymbol{X}, \boldsymbol{X}^2\}$.

Non-linear regression but additive

Possible models to relax linearity assumptions and maintain at the same time interpretability:

- Polynomial
- Splines
- Generalized Additive Models

Polynomial regression

For polynomial regression,

$$Y = \beta_0 + \beta_1 \boldsymbol{X} + \beta_2 \boldsymbol{X}^2 + \beta_3 \boldsymbol{X}^3 + \ldots + \beta_d \boldsymbol{X}^d + \epsilon$$

the basis functions are:

$$b_k(\boldsymbol{X}) = \boldsymbol{X}^k$$

We can produce non-linear curves (polynomial of d-degrees).

Example: Polynomial regression

course_regression_4_polynomial.ipynb

Splines regression

Instead of fitting a high-degree polynomial over the entire predictor space, fit low-degree polynomials over different regions of the predictor space.

Splines regression

Instead of fitting a high-degree polynomial over the entire predictor space, fit low-degree polynomials over different regions of the predictor space.

- Regions are determined by some points (called knots) where we want to change the fitting model.
- For instance, a piecewise cubic polynomial with a single knot at point c,

$$Y = \begin{cases} \beta_{01} + \beta_{11}\boldsymbol{X} + \beta_{21}\boldsymbol{X}^2 + \beta_{31}\boldsymbol{X}^3 + \epsilon & \text{if } \boldsymbol{X} < c \\ \beta_{02} + \beta_{12}\boldsymbol{X} + \beta_{22}\boldsymbol{X}^2 + \beta_{32}\boldsymbol{X}^3 + \epsilon & \text{if } \ Xpoints \geq c \end{cases}$$

Cubic spline regression

$$Y \sim \beta_0 + \quad \beta_1 \boldsymbol{X} + \beta_2 \boldsymbol{X}^2 + \beta_3 \boldsymbol{X}^3 + \quad \beta_{3+1} h(\boldsymbol{X}, c_1) + \ldots + \beta_{3+k} h(\boldsymbol{X}, c_k)$$

where $h(\boldsymbol{X}, c)$ are truncated power basis,

$$h(\boldsymbol{X}, c) = \begin{cases} (\boldsymbol{X} - c)^3 & \text{if } \boldsymbol{X} > c \\ 0 & \text{if } \boldsymbol{X} \leq c \end{cases}$$

Cubic spline regression

$$Y \sim \beta_0 + \quad \beta_1 \boldsymbol{X} + \beta_2 \boldsymbol{X}^2 + \beta_3 \boldsymbol{X}^3 + \quad \color{red}{\beta_{3+1} h(\boldsymbol{X}, c_1) + \ldots + \beta_{3+k} h(\boldsymbol{X}, c_k)}$$

where $h(\boldsymbol{X}, c)$ are truncated power basis,

$$h(\boldsymbol{X}, c) = \begin{cases} (\boldsymbol{X} - c)^3 & \text{if } \boldsymbol{X} > c \\ 0 & \text{if } \boldsymbol{X} \leq c \end{cases}$$

- Piecewise polynomials of degree $d = 3$ continuous and smooth!
- Impose some constraints at the knots (up to degree $d - 1$): continuous, 1st and 2nd derivatives continuous.
- Number of coefficients: 1 (β_0) + 3 (degree) + k (knots)

Splines regression

Cubic splines: lowest order for which the discontinuity at the knots cannot be noticed by the human eye.

Spline vs polynomial: With polynomials higher degrees are needed to obtain better fitting in some cases and there are also problems at the edges of the data.

Spline vs polynomial: Using splines we can introduce flexibility by increasing the number of knots but keeping the degree fixed.

Example: Splines regression

course_regression_5_Splines.ipynb

Generalized Additive Model

Extend to multiple linear regression model

$$Y \sim \beta_0 + \beta_1 \boldsymbol{X}_1 + \ldots + \beta_p \boldsymbol{X}_p$$

to allow for non-linear relationships between each predictor and the response, for instance, replace each linear component with a smooth non-linear function

$$Y \sim \beta_0 + f_1(\boldsymbol{X}_1) + \ldots + f_p(\boldsymbol{X}_p)$$

A different function f_j for each predictor $\boldsymbol{X}_j$, and then add all the contributions.

Outline

Alternative fitting procedures

Fitting a multiple linear regression model:

$$Y = \beta_0 + \beta_1 \boldsymbol{X}_1 + \ldots + \beta_p \boldsymbol{X}_p + \epsilon$$

$$(\widehat{\beta}_0, \ldots, \widehat{\beta}_p) = \arg\min_{\beta} \sum_{i=1}^{n} e_i^2 \quad \longrightarrow \quad \widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \boldsymbol{X}_1 + \ldots + \widehat{\beta}_p \boldsymbol{X}_p$$

In order to estimate the model parameters given a data set, replace the least squares fitting with some alternative fitting procedures.
Why?

- Obtain better prediction accuracy.
- Improve model interpretability.

Alternative fitting procedures

Alternatives to improve the fitting technique:

- **Subset selection**; identify the subset of predictors that we believe to be related with the response.
- **Dimension reduction**; projecting the predictors into a M-dimensional subespace, where $M < p$, and use these projections as predictors (PCA).
- **Shrinkage**; fitting the model with all the predictors and shrink the estimated coefficients towards zero.
  Example: Lasso

Lasso regression

Fitting a multiple linear regression model

$$Y = \beta_0 + \beta_1 \boldsymbol{X}_1 + \ldots + \beta_p \boldsymbol{X}_p + \epsilon$$

Given data $(\boldsymbol{x}_i, y_{ij})$ estimate the model coefficients that:

$$(\widehat{\beta}_0, \ldots, \widehat{\beta}_p) = \arg\min_{\beta} \sum_{i=1}^{n} e_i^2 + \alpha \sum_{j=1}^{p} |\beta_j|$$

Lasso regression

Fitting a multiple linear regression model

$$Y = \beta_0 + \beta_1 \boldsymbol{X}_1 + \ldots + \beta_p \boldsymbol{X}_p + \epsilon$$

Given data $(\boldsymbol{x}_i, y_{ij})$ estimate the model coefficients that:

$$(\widehat{\beta}_0, \ldots, \widehat{\beta}_p) = \arg\min_{\beta} \sum_{i=1}^{n} e_i^2 + \alpha \sum_{j=1}^{p} |\beta_j|$$

- the penalty term uses the norm l_1 of the coefficient vector $\beta = (\beta_1, \ldots, \beta_p)$,
- when α is large, some coefficients are forced to be zero.

Model fitting

How good is the model obtained?

$$(\widehat{\beta}_0, \ldots, \widehat{\beta}_p) = \arg\min_{\beta} \sum_{i=1}^{n} e_i^2 + \alpha \sum_{j=1}^{p} |\beta_j| \longrightarrow \widehat{Y} = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j \mathbf{X}_j$$

Quantify if the response value predicted by the model $\widehat{Y}$ is close to the true model Y.

Example: mean squared prediction error

$$MSE = \frac{1}{n} \sum_{i=1}^{n} e_i^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

Example: Lasso regression

course_regression_6_Lasso.ipynb

matematika mugaz bestalde