Lecture 6: Variable selection and handling missing values

Stella Hadjiantoni

Department of Mathematical Sciences
University of Essex

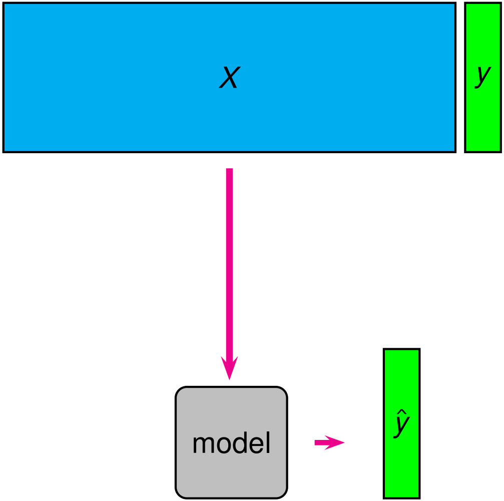
Variable selection: an introduction

Let the dataset be $D = \{\boldsymbol{X}, \boldsymbol{y}\}$:

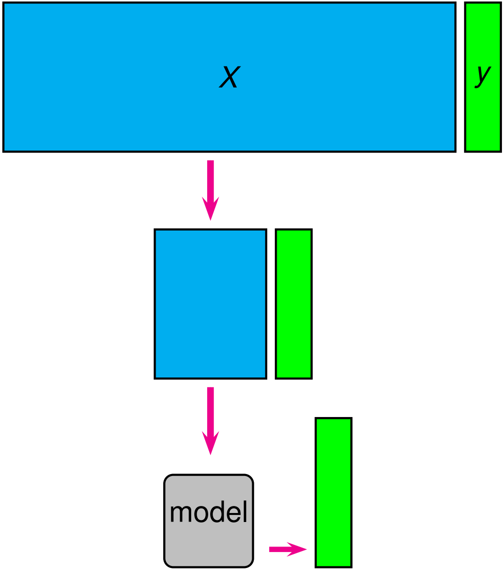
$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & & x_{nd} \end{pmatrix} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_d^T \end{pmatrix}^T, \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- There are n observations (or examples or samples) and d variables (predictors or features).
- Each row of $\boldsymbol{X}$ is an example.
- Each column of $\boldsymbol{X}$ is a variable denoted by x_j, $j = 1, ..., d$.
- $\boldsymbol{y}$ is the vector of responses (or labels).

Variable selection: an illustration



Variable selection: an illustration



Variable selection: an example

Imagine you are trying to guess the price of a smartphone.

Features:

battery power	bluetooth	clock speed
dual sim	front camera pixels	4G
internal memory	number of cores	ram
pixel resol height	pixel resol width	screen height
screen width	talk time	3G

Variable selection: an example

Imagine you are trying to guess the price of a smartphone.

Features:

battery power	bluetooth	clock speed
dual sim	front camera pixels	4G
internal memory	number of cores	ram
pixel resol height	pixel resol width	screen height
screen width	talk time	3G

What are the relevant factors?

Variable selection: an example

Imagine you are trying to guess the price of a smartphone.

Features:

battery power	bluetooth	clock speed
dual sim	front camera pixels	4G
internal memory	number of cores	ram
pixel resol height	pixel resol width	screen height
screen width	talk time	3G

What are the relevant factors?

- **Relevant:** ram, battery power, internal memory, pixel resolution height, pixel resolution width, mobile weight.
- **Irrelevant:** mobile depth.
- **Redundant:** pixel resolution Height/pixel resolution width.

An application: Forecasting spot price in the UK natural gas market

Aim: to analyse and forecast UK gas spot prices using penalized regression, stepwise regression and principal components regression.

- Energy commodity: Brent, Carbon, Coal.
- Technical analysis indicators: moving-average (MA).
- Financial: S&P 500, FTSE 100.
- Metal: Gold, Platinum.
- Agriculture: Wheat, Coffee
- Interest rates: LIBOR, UK government bond yield.

Variable selection: why?

Variable (or feature) selection problem: *To find the subset of variables (features) which are important for predicting y.*

- What are the relevant features?
- Computational complexity?

Goals are:

1. to avoid overfitting.
2. to construct a model that is interpretable.
3. to reduce computational cost.

Variable selection: different procedures

- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim (0, \sigma^2 \boldsymbol{I}_n).$
- $\boldsymbol{X} \in \mathbb{R}^{n \times p}, \qquad \boldsymbol{\beta}^T = (\beta_1 \ \beta_2 \ \ldots \ \beta_p)^T.$

Three categories of variable/feature selection methods:

- Wrapper

  e.g $\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2$

- Filter (not examined)

- Embedded (not examined)

  e.g Least absolute shrinkage and selection operator (LASSO)

  $\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda \sum_{i=1}^{p} |\beta_i|$

Wrapper methods: forward, backward, stepwise selection

In wrapper methods variable selection is a search problem.

What about an ***exhaustive*** search?

Wrapper methods: forward, backward, stepwise selection

In wrapper methods variable selection is a search problem.

What about an ***exhaustive*** search?

For a total of d variables, there are 2^d possible variable sets.

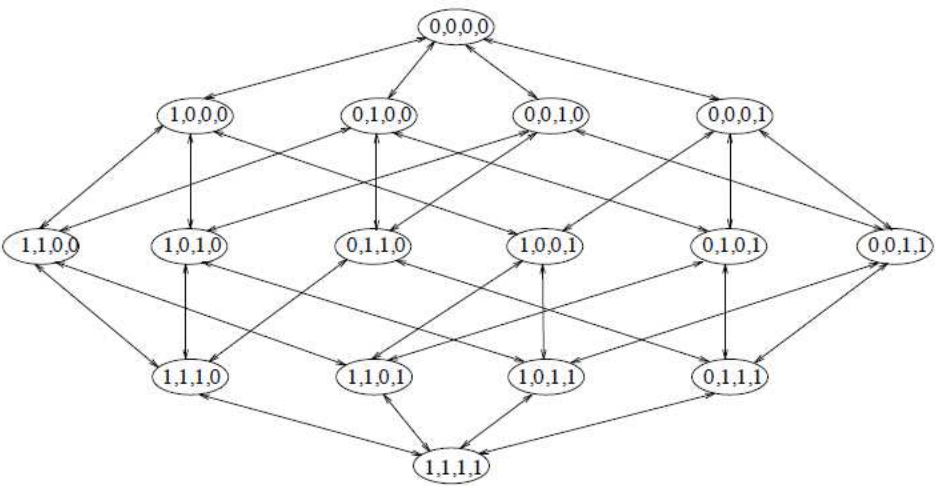
d	feature sets
20	1 million
26	67 million
32	4.3 billion
45	10^{13}
52	10^{15}

Wrapper method: forward selection

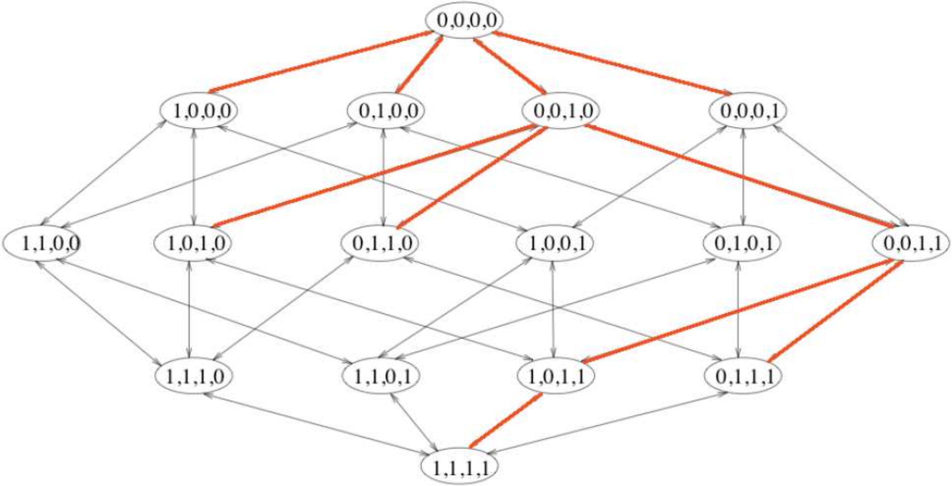
Example

Possible features for predicting the price of a smartphone: ram, battery power, internal memory, pixel resolution height. What are the main steps of the forward selection greedy search procedure?

Wrapper method: forward selection



Wrapper method: forward selection



Wrapper method: forward selection

Include the notation of a linear regression model with k=4, no intercept.

Algorithm: Forward selection pseudo code example.

```
1. Start with no variables included.
2. For each variable not included, check the score of variable when they
   are added.
3. if no variable improves the score then
4.     stop
5. else
6.     Choose the one that improves the score the most and add it.
7. end if
8. Go back to Step 2 and continue until no new variable can be added.
```

Advantage: only $\frac{d(d+1)}{2}$ variable sets.

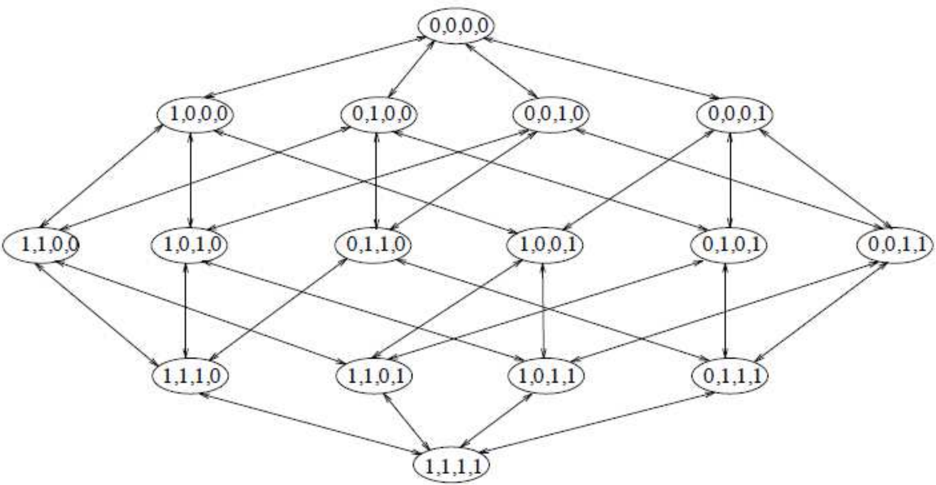
Disadvantage: is not guaranteed to find the best set.

Wrapper method: backward selection

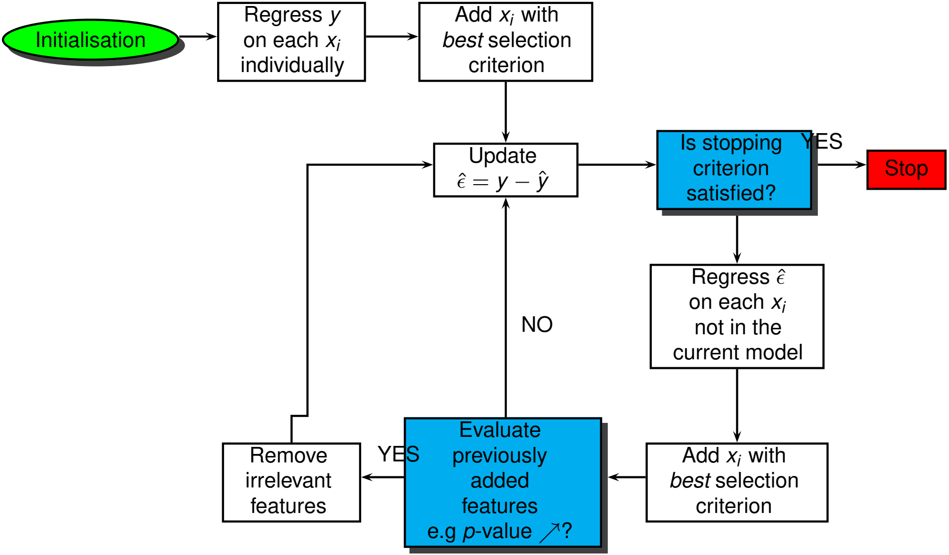
Exercise

The backward selection method is the reverse method of forward selection. It starts with all variables and sequentially removes them. Write the main steps for backward selection.

Wrapper method: backward selection



Wrapper method: stepwise selection

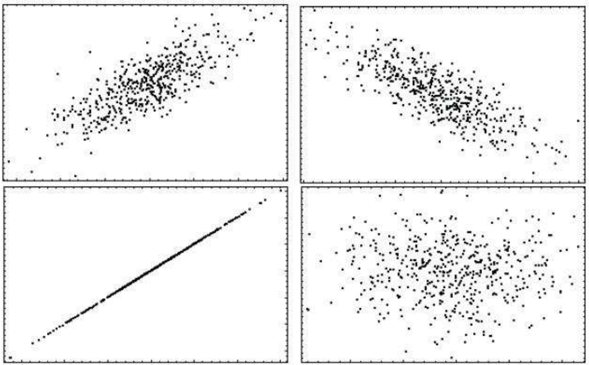


Initialisation
Regress y on each x_i individually
Add x_i with *best* selection criterion
Update $\hat{\epsilon} = y - \hat{y}$
Is stopping criterion satisfied?
YES
Stop
Regress $\hat{\epsilon}$ on each x_i not in the current model
Add x_i with *best* selection criterion
Evaluate previously added features e.g *p*-value $\nearrow$?
YES
Remove irrelevant features
NO

Filter methods

In filter methods, variable selection is based on evaluating statistics of the data which measure the relevance of each variable with the outcome variable.

- Eliminate irrelevant variables.
- Rank relevant variables.



Filter methods: correlation coefficient

The Pearson's correlation coefficient of X_1 and X_2:

$$\rho(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}.$$

The correlation of a sample of feature j (x_j) with y is:

$$r(x_j, y) = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_j - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(y_j - \bar{y})^2}}$$

We rank the variables in order of the absolute value of the correlation coefficient.

Disadvantages: single (univariate) variable relevance, captures only linear relationships.

Filter methods: smartphone example

1. Compute the $r(x_j, y)$ for all variables $j = 1, ..., d$ and then rank x_j in descending order of $r(x_j, y)$.
2. Hold the top ranked variables and discard the lowest ranked.

x_j	$r(x_j, y)$
ram	0.92
camera MP	0.75
battery power	0.43
4G	0.32
dual sim	0.29
⋮	⋮

Variable selection: model criteria

The *corrected coefficient of determination*:

$$\bar{R} = 1 - \frac{n-1}{n-p}\left(1 - R^2\right).$$

Mallow's C_p:

$$C_p = \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y_{iM}}\right)^2}{\widehat{\sigma^2}} - n + |M|.$$

Akaike information criterion :

$$AIC = -2L(\widehat{\beta_{\boldsymbol{M}}}, \widehat{\sigma^2}) + 2(|M| + 1).$$

Bayesian information criterion :

$$BIC = -2L(\widehat{\beta_{\boldsymbol{M}}}, \widehat{\sigma^2}) + log(n)(|M| + 1).$$

(Note: see section 10.3)

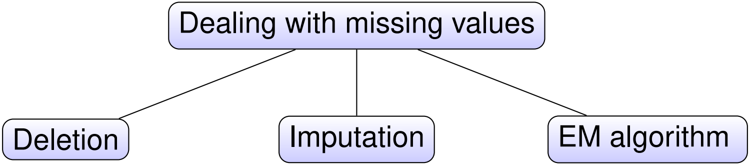
Missing data

- Missing values/cases
- Incomplete values

Why are the data missing?

- Missing at random.
- Missing not at random.

Missing data



Summary

- Why feature selection?
- Forward, backward and stepwise feature selection.

- Types of missing data.
- Methods for handling missing data.