Machine learning 1

Feature selection

Piotr Wójcik
pwojcik@wne.uw.edu.pl

academic year 2017/2018

UNIWERSYTET WARSZAWSKI

Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Selecting features for the model

- determining which variables should be included in the model is **one of the most important issues**, especially since the available data become **more and more multidimensional**
- from a practical point of view, a model with fewer explanatory variables can be **easier to interpret**
- if collecting and storing data is costly – it will also be **cheaper to maintain**
- also from the statistical point of view estimation of a smaller number of parameters is **computationally less expensive**
- experience shows that predictive performance of multivariate models increases when noisy variables are removed
- using too many variables can also lead to a problem of **overfitting the model**

Selecting features for the model – cont'd

- there are statistical models resistant to irrelevant variables, having built-in mechanisms of selecting features that are really important from the point of view of the analyzed phenomenon (eg. LASSO, neural networks, decision trees and derivative models – e.g. random forests)
- the choice of features for the model can be carried out in a **supervised** or **unsupervised** way
- **unsupervised** selection does not take into account the target variable
- variables are selected due to their **individual characteristics** or **mutual connections**
- **supervised** selection includes checking **relations between explanatory variables and target variable**
- in this case, variables are chosen to increase the accuracy of the model or reduce the degree of its complexity

The impact of irrelevant variables on the model

- many models, especially parametric ones, based on estimates of the slope of dependencies between variables, take into account the relationship of the dependent variable with all variables **simultaneously** – for each variable a parameter is estimated
- estimated parameters are then used for prediction and including redundant variables in the model may **distort predictions** and **limit the predictive power** of the model
- thus having in mind the potentially **negative impact of redundant variables**, their number in the model should be limited
- the purpose here is to limit the size (complexity) of the model while maintaining or improving the accuracy of its predictions
- generally leaving negligible variables in the model will **increase its variance**, however **without limiting the bias of the model**

Consider all combinations?

- theoretically, the simplest approach would be to consider **all possible combinations** of variables in the model, check for each variant the measure of model accuracy and choose the best variant
- despite the large and ever-increasing computing power of computers, it is still not possible
- when considering p variables, we have 2^p different combinations (potential models)
- for example for $p = 40$ variables one can build **more than a billion different models** of one type
- the solution to this problem is **selection of variables for the model**

Methods of feature selection

The methods of selecting variables for the model can be divided into two groups:

1. **filter methods** focus on a **single explanatory variable**, possibly its relation with a target variable – they are used to select variables **regardless of the type of model**
2. **wrapper methods** (subset selection) – algorithms that evaluate many models of a given type by adding or removing variables in order to find their **optimal combination** optimizing the assumed criterion

- **filter methods** are usually **less computationally expensive**, however the selection criterion used by them **does not directly refer** to the considered model
- filtering methods consider **each variable individually** and therefore may leave excessive variables in the model (e.g. strongly correlated)

Sample filter methods

- deleting **redundant** variables – **strongly correlated** with other variables (**multicollinearity**)
- removing **irrelevant** variables – having a very unbalanced distribution – taking similar or identical values for many or all observations (**near-zero variance predictors**)
- selection of k variables **individually strongest related** with the target variable

Filter methods – selection of important variables

Relation with the target variable can be measured in different ways:

- for **categorical target** and **categorical inputs** – χ^2 test, Cramer's V
- for **continuous target** and **continuous inputs** – Pearson's, Spearman's or Kendall's correlation coefficient
- for **continuous target** and **categorical inputs** or **categorical target** and **continuous inputs** – ANOVA models, Wilcoxon tests, t-tests
- for a **regression task** – looking on R^2, RMSE for a model with each variable individually
- for a **classification task** – looking on **accuracy**, **AUC ROC** for a model with each variable individually

Filter methods – mutual information

- Information based criteria can be used in place of correlation for filtering variables
- **Information** contained in a discrete distribution of feature X is given by

$$H(X) = \sum_{i} p(x_i) logp(x_i)$$

where x_i are the discrete feature values and $p(x_i)$ are its probabilities

- Information embedded in the **joint distribution** is provided by

$$H(Y, X) = \sum_{i} \sum_{j} p(y_j, x_i) logp(y_j, x_i)$$

where $p(y_j, x_i)$ is the joint probability

Filter methods – mutual information – cont'd

- **Mutual information** (MI) provides a good measure of feature importance showing how much we can reduce the uncertainty on the variable Y based on the information contained in the variable $X
- MI is calculated as:

$$MI(Y,X) = H(Y) + H(X) - H(Y,X)$$

$$MI(Y,X) = \sum_i \sum_j p(y_j, x_i) log \frac{p(y_j, x_i)}{p(y_j)p(x_i)}$$

- feature is **more important** if the mutual information MI(Y, X) between the target and the feature distributions is **larger**
- **Information gain** is a similar criterion where $IG(Y,X) = H(Y)H(Y|X)$
- Continuous features are either discretized, or integration instead of summation is performed by fitting a kernel function to approximate the density of the feature X

Sample methods of finding subsets of variables

Methods of iterative searching of subsets of variables used most often (mainly in parametric models):

- backward elimination – from general to specific
- forward selection – from specific to general

With a large number of variables, they are computationally expensive.

Backward elimination – algorithm for parametric models

- start with the model **with all variables**
- **remove** one variable, which is **least significant** (e.g. has highest $p - value$), provided that it is **insignificant** at assumed level α
- estimate a new model with $p - 1$ variables and again **delete the least significant one**, provided that it is not statistically significant
- repeat the previous steps until the accepted stop criterion is met – e.g. when **all variables remaining in the model are significant**
- variable **removed** from the model **never comes back**, even if it would be significant in other configuration

Forward selection – algorithm for parametric models

- start with model including **only the constant term**
- estimate p simple regressions with one variable (each separately) and finally **add** to the model this variable which is **most significant** (lowest $p - value$), if significant on a desired level
- add to the model the second variable, which is **most significant** looking at all models with two explanatory variables
- add more variables to the model according to the analogous scheme until the accepted stop criterion is met – e.g. when **no of other variables added to the model will be significant**
- variable added to the model remains after adding futher variables **even if its significance fails below a desired level**

Considerations for subset selection methods

- testing one individual null hypothesis in each step rises the **risk of error** – the actual level of significance for all tests together can **significantly exceed** assumed significance level α): in k individual tests $\alpha^* = 1 - (1 - \alpha)^k$ (Lowell bias)
- statistical significance of the variable **does not have to be directly related to the prediction quality** of the model
- instead of significance a different criterion can be used for variable selection – for example the change of R^2, RMSE, accuracy, area under the ROC curve assessed within cross-validation framework
- therefore above mentioned methods can be **generalized for non-parametric approaches**

Information criteria (AIC, BIC)

- alternatively one can use the so-called **information criteria**, which impose on the optimization criterion an additional "penalty" for the size of the model
- the most popular are **Akaike Information Criterion** (**AIC**) and **Schwarz's Bayesian Information Criterion** (**BIC**, **SBC**)
- BIC imposes a **stricter penalty** for model size, therefore its use will usually result in **a smaller model** (with less variables)
- the specific formula **depends on the type of model**, e.g. for linear regression one has:
    - $AIC = n \times log\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) + 2p$
    - $BIC = n \times log\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) + 2p \times log(n)$
- **lower value** of the information criterion **means a better model**