Statistical Machine Learning – Week 9

Sascha Jecklin

HS 2022/23

1 Task

In this lab we will implement two feature selection methods and two regularization methods. More features require more data to guarantee a good generalization. Since there is almost always a lack of data, we have to come up with some good ideas to select the most important features or limit the model capacity by some regularization method. We will implement `best subset`, `forward selection`, `ridge regression` and `lasso` and use the `Credit` dataset to test the implementation. The goal is to predict the balance of an individual based on eleven predictors, such as income, rating and gender.

- For this exercise use `week9_exercise.py`. In the main function the dataset is loaded and some dummy variables are introduced for the categorical variables.
- Implement the best subset selection algorithm in the function `best_subset(x, y)`. Use the R^2_{adj} value to select among the best set of predictors. For each number of predictors print the selected predictor names and the corresponding R^2_{adj}. What is the best feature subset according to your subset selection algorithm?
  Hints:
  - Make use of the `fit_linear_reg(x, y)` function which takes the dataset as an argument and returns the R^2 and R^2_{adj}.
  - Moreover, you might want to use `itertools.combinations(p,r)` which is an generator that returns all possible r-length tuples of a list p.
  - If you do not know how best subset works, see Algorithm 6.1 below.
- Implement the forward selection algorithm by completing the function `forward_selection(x, y)`. Use the R^2_{adj} or R^2 to select the best set of predictors among all sets of equal number of predictors. For each number of predictors print the selected predictor names and the corresponding R^2_{adj}. What is the best feature subset according to your forward selection algorithm?
  Hints:
  - Make use of the `fit_linear_reg(x, y)` function which takes the dataset as an argument and returns the R^2 and R^2_{adj}.
  - If you do not know how best subset works, see Algorithm 6.2 below.
- Compare your selected features with the table 6.1 on page 209 (same as Table 6.1 below).
- Implement a the ridge regression algorithm by completing the function `ridge_regression(x, y)`. Vary the λ value from 10^{-2} to 10^5 logarithmically spaced and plot the corresponding values of the coefficients.
  Hints:
  - First use sklearn's `preprocessing.StandardScaler` to standardize the features.

- Use sklearn's `Ridge` class to fit a regression model with Ridge regularization.
- Take a look at Figure 6.4 on page 216 of the textbook for an example plot.

- Implement a the lasso regression algorithm by completing the function `lasso_regression(x, y)`. Vary the λ value from 10^0 to 10^3 logarithmically spaced and plot the corresponding values of the coefficients.
  Hints:
  - First use sklearn's `preprocessing.StandardScaler` to standardize the features.
  - Use sklearn's `Lasso` class to fit a regression model with Lasso regularization.
  - Take a look at Figure 6.6 on page 220 of the textbook for an example plot.

Comments

- There are not only `numpy linspace` but also `numpy logspace` `https://docs.scipy.org/doc/numpy/reference/generated/numpy.logspace.html`. This can be useful for the two plots.
- The documentation for sklearn's `Ridge` and `Lasso` can be found on
  - `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html`
  - `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html`

Algorithm 6.1 Best subset selection

---

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, ...p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here best is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, ..., \mathcal{M}_p$ using crossvalidated prediction error, C_p (AIC), BIC, or adjusted R^2.

Algorithm 6.2 Forward stepwise selection

---

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 0, ...p - 1$:
   (a) Fit all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, ..., \mathcal{M}_p$ using crossvalidated prediction error, C_p (AIC), BIC, or adjusted R^2.

# Variables	Best subset	Forward stepwise
One	`rating`	`rating`
Two	`rating, income`	`rating, income`
Three	`rating, income, student`	`rating, income, student`
Four	`card, income, student, limit`	`rating, income, student, limit`

Table 6.1: The first four selected models for best subset selection and forward stepwise selection on the `Credit` data set. The first three models are identical but the fourth models are different.