IIND4132

Homework 2

Instructor: Andrés Gómez, gomezand@usc.edu

1 Hyperparameter tuning

Given m observations $\{(\boldsymbol{a_i}, y_i)\}_{i=1}^m$, where each $\boldsymbol{a_i} \in \mathbb{R}^n$ are the features of point i and $y_i \in \mathbb{R}$ is the associated response, a regularization parameter $\lambda \geq 0$ and given a target complexity k, consider the best subset selection problem seen in class

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \sum_{i=1}^{m}(y_i - \sum_{j=1}^{n} a_{ij}x_j)^2 + \lambda \sum_{j=1}^{n} c_j x_j^2 \tag{1a}$$

$$\text{s.t.} \sum_{j=1}^{n} \mathbb{1}_{\{x_j \neq 0\}} \leq k, \tag{1b}$$

where $c_j = \sum_{i=1}^n a_{ij}^2$. In some cases, k is directly given by a decision-maker (e.g., due to interpretability considerations), but in other cases it should be determined as well. Denote by $x(k)^*$ the optimal solution of problem (1) (for a given value of $k \in \mathbb{Z}_+$). Using the "Communities_and_Crime" dataset from the lectures (for simplicity, you may assume that the intercept is 0) and setting $\lambda = 0.01$, answer the following questions.

1. **Adjusted $\boldsymbol{R^2}$** The mean squared error (MSE) [2] of predictor $x(k)^*$ is given by

   $$\text{MSE}(k) = \frac{\sum_{i=1}^m (y_i - \sum_{j=1}^n a_{ij}x(k)_j^*)^2}{m - \sum_{j=1}^n \mathbb{1}_{\{x(k)_j^* \neq 0\}}},$$

   which intuitively is a proxy for the variance of the prediction errors. The estimator that minimizes the MSE is also the estimator that maximizes the popular adjusted R^2 criterion.

   Find the cardinality k that minimizes the MSE. Which are the critical predictors and their associated regression coefficients?

2. **Bayesian information criterion** The Bayesian information criterion (BIC) [1] of predictor $x(k)^*$ is given by

   $$\text{BIC}(k) = m \ln \left(\frac{\sum_{i=1}^m \left(y_i - \sum_{j=1}^n a_{ij}x(k)_j^*\right)^2}{m} \right) + \ln(m) \left(\sum_{j=1}^n \mathbb{1}_{\{x(k)_j^* \neq 0\}} \right) + K,$$

where K is a constant that does not depend on the estimator $x(k)^*$. The estimator that minimizes the BIC is (under appropriate assumptions) the estimator that is a posteriori more probable.

Find the cardinality k that minimizes the BIC. Which are the critical predictors and their associated regression coefficients?

2 Modeling with binary variables

Consider again problem (1) and the "Communities_and_Crime" dataset `https://archive.ics.uci.edu/dataset/183/communities+and+crime`. Note that questions b), c) are independent (that is, do not use modifications in b) to answer c) or viceversa).

a) **Write a mixed-integer optimization formulation for problem** (1) **and solve it for $k = 10$ and $\lambda = 0.01$ (you may assume that the intercept is 0). What is the optimal solution?**

b) <u>**Group sparsity**</u> Often, different "features" in the dataset can be interpreted as part of a same group or attribute:

   i. {racepctblack, racePctWhite, racePctAsian, racePctHisp} all correspond to the attribute "race".

   ii. {agePct12t21, agePct12t29, agePct16t24, agePct65up} all correspond to the attribute "age".

   iii. {whitePerCap,blackPerCap,indianPerCap,AsianPerCap,OtherPerCap,HispPerCap} all correspond to the attribute "Racial per capita income".

   In such cases, one desires to impose the constraint that either all regression coefficients for variables in a group are zero, or all are possibly non-zero at the same time. Moreover, using more than one variable in a group does not increase the cost (1b) further. For example, a model with non-zero regression coefficients for all variables in group

   {racepctblack, racePctWhite, racePctAsian, racePctHisp}

   would increase the left-hand-side of (1b) by one (instead of 4).

   Modify the mixed-integer formulation for (1) **to account for groups *i*, *ii* and *iii* above. What is the new optimal solution?**

c) <u>**Multicollinearity**</u> If two features are highly correlated, then included both of them in a regression model may result in low-quality solutions. For example, features "PctIlleg" and "PctKids2Par" have high collinearity. In such cases, it is desirable to include at most one of these features in the model. **Modify the mixed-integer formulation for** (1) **to ensure that features "PctIlleg" and "PctKids2Par" are not both used in the linear regression model. What is the new optimal solution?**

3 Beyond the perspective reformulation

Consider the following mixed-integer quadratic optimization problem (where the quadratic objective has been moved to the constraints)

$$
\begin{aligned}
\min_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{t}} \quad & 3x_1 - 4x_2 + x_3 - x_4 + t_1 + t_2 + t_3 + t_4 + t_5 + t_6 + t_7 && \text{(2a)}\\
\text{s.t.} \quad & -Mz_i \le x_i \le Mz_i \qquad i = 1, \dots, 4 && \text{(2b)}\\
& x_1^2 \le t_1 && \text{(2c)}\\
& x_2^2 \le t_2 && \text{(2d)}\\
& x_3^2 \le t_3 && \text{(2e)}\\
& x_4^2 \le t_4 && \text{(2f)}\\
& (x_2 - x_3)^2 \le t_5 && \text{(2g)}\\
& (4x_3 - x_4)^2 \le t_6 && \text{(2h)}\\
& (x_1 + x_2)^2 \le t_7 && \text{(2i)}\\
& z_1 + z_2 + z_3 + z_4 \le 2 && \text{(2j)}\\
& z_1 + z_2 \le 1 && \text{(2k)}\\
& z_4 \le z_3 && \text{(2l)}\\
& \boldsymbol{x} \in \mathbb{R}^4,\ \boldsymbol{z} \in \{0,1\}^4,\ \boldsymbol{t} \in \mathbb{R}^7_+. && \text{(2m)}
\end{aligned}
$$

Which of the following reformulations results in the strongest valid relaxation?

1. **Reformulations of constraint** (2f):
   a) $x_4^2 \le t_4$
   b) $x_4^2 \le t_4 z_4$
   c) $x_4^2 \le t_4 z_3$
2. **Reformulations of constraint** (2g):
   a) $(x_2 - x_3)^2 \le t_5$
   b) $(x_2 - x_3)^2 \le t_5(z_2 + z_3)$
   c) $(x_2 - x_3)^2 \le t_5 z_2$
3. **Reformulations of constraint** (2h):
   a) $(4x_3 - x_4)^2 \le t_6$
   b) $(4x_3 - x_4)^2 \le t_6(z_3 + z_4)$
   c) $(4x_3 - x_4)^2 \le t_6 z_3$
4. **Reformulations of constraint** (2i):
   a) $(x_1 + x_2)^2 \le t_7$
   b) $(x_1 + x_2)^2 \le t_7(z_1 + z_2)$
   c) Add $s_1, s_2 \ge 0$ and constraints $x_1^2 \le s_1 z_1,\ x_2^2 \le s_2 z_2,\ s_1 + s_2 \le t_7$

References

[1] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[2] R. J. Wherry. A new formula for predicting the shrinkage of the coefficient of multiple correlation. *The annals of mathematical statistics*, 2(4):440–457, 1931.