Mathematical Underpinnings Lab 10   22.05.2024

Task 1 (Lars)

Load the prostate.data dataset. The first 8 columns contain explanatory variables (X), while the 9th column named `lpsa` contains the response variable (Y). Scale and center X and Y (from this point on X and Y will denote preprocessed data).

a) Apply the Lars algorithm using `Lars` from `sklearn.linear_model`. Do not fit the intercept. We will name the model `model_lars` In what order are the variables included in the model? To answer this question, do:

- Draw a profile plot. On the x-axis, plot `model_lars.alphas_` (the maximum of covariances in absolute value at each iteration) and on the y-axis, plot `model_lars.coef_path_`. Note that `model_lars.alphas_` does not mean exactly the same as α in the lecture.
- Based on `model_lars.coef_path_`, list the order in which the variables are included in the model.

b) Complete this exercise without using the Lars function for computations. First, compute the first element of `model_lars.alphas_` vector (the largest correlation between Y and X_i). Then, perform the first and second steps of the algorithm described below:

(b1) Identify the first variable (S_1) selected by Lars, which is the one with the largest correlation with Y.

(b2) Calculate the OLS estimator for $\beta^{(1)}$ parameter associated with the selected variable.

(b3) Compute correlations between the $r_{1,\tilde{\alpha}}$ defined as $Y - \tilde{\alpha}\hat{Y}$ and all the variables X for a 100 $\tilde{\alpha}$ values in the range of $[0, 1]$. Plot the results, where the x-axis represents $\tilde{\alpha}$ and the y-axis represents correlations. Which line visualizes the theorem discussed in the lecture (Figure 1)?

(b4) Based on the plot, determine which variable will be chosen next (S_2). What is the approximate value of $\tilde{\alpha}_2$? Based on the approximate value of $\tilde{\alpha}_2$ check, whether $\mathrm{Cor}(r_{1,\tilde{\alpha}_2}, X_{S_1}) = \mathrm{Cor}(r_{1,\tilde{\alpha}_2}, X_i)$.

(b5) (On paper) Express the formula for $\mathrm{Cor}(Y - \tilde{\alpha}\hat{Y}, X_i)$ in terms of the correlation between Y and X_i, and the correlation between $\hat{Y}$ and X_i. Using that, derive the formula for $\tilde{\alpha}_2$. (On computer) Next, plot the values of $\sqrt{(\mathrm{Var}(Y))} \cdot \mathrm{Cor}(Y, X_i) - \sqrt{(\mathrm{Var}(\hat{Y}))} \cdot \tilde{\alpha} \cdot \mathrm{Cor}(\hat{Y}, X_i)$ for a 100 $\tilde{\alpha}$ values in the range of $[0, 1]$ and $i = 1, 2, \ldots, 8$. Compute the exact value of $\tilde{\alpha}_2$.

(b6) (On computer) Compute α_2, and $r_{1,\tilde{\alpha}_2}$. Update the vector of selected variables.

Then (third step):

(b7) Compute the OLS estimator for $\beta^{(2)}$ using the selected variables as predictors and $r_{1,\tilde{\alpha}_2}$ as the explanatory variable.

(b8) Compute the correlations between the $r_{2,\tilde{\alpha}_2} = r_{1,\tilde{\alpha}_2} - \hat{Y}_2$ and all the variables for a 100 $\tilde{\alpha}$ values in the range of $[0, 1]$. Plot the results, where the x-axis represents $\tilde{\alpha}$ and the y-axis represents correlations. Which line/lines visualizes the theorem discussed in the lecture (Figure 1)?

(b9) Based on the plot, determine which variable will be chosen next (S_2). What is the approximate value of $\tilde{\alpha}_2$?

(b10) (On paper) Express the formula for $\mathrm{Cor}(Y - \tilde{\alpha}\hat{Y}, X_i)$ in terms of the correlation between Y and X_i, and the correlation between $\hat{Y}$ and X_i. Using that, derive the formula for $\tilde{\alpha}_2$. (On computer) Next, compute the exact value of $\tilde{\alpha}_2$.

(b11) (On computer) Compute α_2, and $r_{1,\tilde{\alpha}_2}$.

c) (*) Implement Lars.

Lars for scaled and centered data

S_k - a variable chosen in step k
α_k - `model_lars.alphas_` - the maximum of covariances in absolute value at kth iteration

First step

1. Compute correlations between Y and X_i for $i = 1, 2, \dots, p$.
2. $S_1 = \text{argmax}_i \text{Cor}(Y, X_i)$
3. $\alpha_1 = \max_i \text{Cor}(Y, X_i)$

Second step

1. Compute OLS estimator for the parameters $\beta^{(1)}$ in a linear model using X_{S_1} as the predictor variable and Y as an explanatory variable.
2. Denote $r_{1,\tilde{\alpha}} = Y - \tilde{\alpha}\hat{Y}^{(1)}$, where $\hat{Y}^{(1)}$ is a prediction of the model from the previous step.
3. Compute correlations between $r_{1,\tilde{\alpha}}$ and X_i for $i \in \{1, 2, \dots, p\}$. Find $i \neq S_1$ and the smallest $\tilde{\alpha} \in [0, 1]$ (denoted by $\tilde{\alpha}_2$) such that $\text{Cor}(r_{1,\tilde{\alpha}}, X_{S_1}) = \text{Cor}(r_{1,\tilde{\alpha}}, X_i)$.
4. $S_2 = i$ (from the previous step), and $\alpha_2 = \text{Cor}(r_{1,\tilde{\alpha}_2}, X_{S_2})$.

$k+1$-th step

1. Compute OLS estimator of the parameters $\hat{\beta}^{(k)}$ in a linear model using $X_{S_1}, X_{S_2}, \dots, X_{S_k}$ as predictors and $r_{1,\tilde{\alpha}_2}$ as the explanatory variable.
2. $r_{k,\tilde{\alpha}_k} = r_{k-1,\tilde{\alpha}_{k-1}} - \tilde{\alpha}_k \hat{Y}^{(k)}$ where $\hat{Y}^{(k)}$ is a prediction of the model from the previous step.
3. Compute correlations between $r_{k,\tilde{\alpha}_k}$ and X_i for $i \in \{1, 2, \dots, p\}$. Find $i \neq S_1, S_2, \dots, S_k$ and the smallest $\tilde{\alpha}_{k+1} \in [0, 1]$ such that $\text{Cor}(r_{k,\tilde{\alpha}_{k+1}}, X_{S_j}) = \text{Cor}(r_{k,\tilde{\alpha}_{k+1}}, X_i)$ for any $j < k$.
4. $S_{k+1} = i$ (from the previous step), and $\alpha_{k+1} = \text{Cor}(r_{k,\tilde{\alpha}_{k+1}}, X_{S_{k+1}})$.

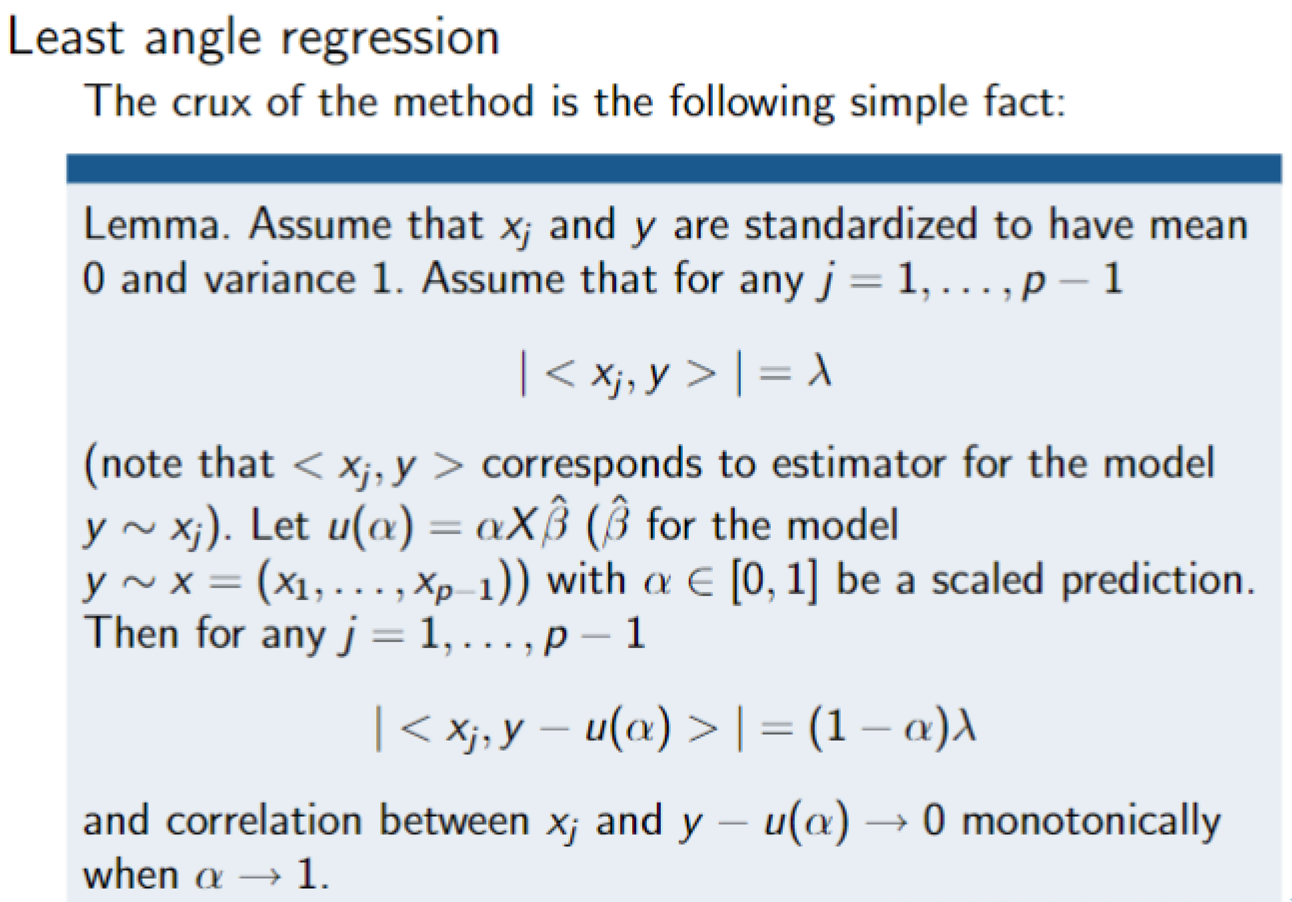

Figure 1: Lemma from the lecture

Task 2 (Lars, lasso)

Load dataset from files `SRBCT_X.txt` and `SRBCT_Y.txt`. The first file contains standardised genes expression values collected from 83 cDNA microarrays. The second contains the classes including 4 different childhood tumors called small round blue cell tumors (SRBCTs) - Ewing's family of tumours (EWS), neuroblastoma (NB), Burkitt's lymphoma (BL), and rhabdomyosarcoma (RMS). Note that the dataset is not be well-suited for the task, as Y is categorical. However, for illustrative purposes, we will proceed with this dataset.

a) Apply Lars. Stop algorithm when 80 variables are chosen (use `n_nonzero_coefs = 80`). Draw a profile plot. Which variables are chosen? In what order?

b) Fit `LassoLars` model. Draw a profile plot. Compare the order in which variables are chosen with Lars. Why is there a difference?

c) Fit Lasso models using shrinkage parameters corresponding to alphas in Lars. In a plot from b) mark estimated parameters for every model. What can you say about these models?