Logistic Regression: regularization

Tasks (Lab 6):
Recommended functions and packages: glmnet (library glmnet in R and Python), LogisticRegression (library: scikit learn in Python)

1. Logistic regression with different regularizations: lasso, ridge, elastic net.

   Consider high-dimensional dataset 'prostate' corresponding to predicting the occurrence of prostate cancer using gene expression data. There are 102 observations (patients) and 6033 variables (genes). The response variable indicates whether a patient suffers from disease (prostate cancer, value: 1) or not (value: 0).

   - Fit logistic regression model with different regularizations (lasso, ridge, elastic net). Print the values of parameters for different values of parameter λ.
   - Make profile plots, showing how the values of parameters depend on λ for different variables.
   - Select the optimal value of λ using cross-validation.

2. Performance of lasso for simulated data.

   - Generate data from logistic model:

     $$y_i \sim Bern(p_i),$$

     where

     $$p_i = \frac{1}{1 + \exp[-(\beta^T x_i)]},$$

     for $i = 1, \ldots, n$, $x_i \sim N(0, I)$ and $\beta = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ (there are 10 relevant variables and 10 irrelevant variables).
   - Fit logistic model using lasso regularization (use glmnet package in R or Python). Use cross-validation to select the optimal value of parameter λ. Report which variables are selected as relevant by the method.
   - Compute PSR (Positive Selection Rate)

     $$PSR := \frac{|t \cap \hat{t}|}{|t|}$$

     and FDR (False Discovery Rate)

     $$FDR := \frac{|\hat{t} \setminus t|}{|\hat{t}|},$$

     where t is a set of true relevant variables and $\hat{t}$ is a set of variables selected by the model as relevant.

- Repeat the above steps $L = 100$ times and compute the averaged values of PSR and FDR.
- Investigate how the averaged values of PSR and FDR depend on the sample size (consider $n = 50, 100, 300, 500, 1000, 2000$); make a plot showing the dependence of PSR and FDR on n.
- Investigate how the averaged values of PSR and FDR depend on the number of irrelevant variables (consider values: $10, 50, 100, 200, 500$) for fixed value of sample size $n = 300$.
- Analyze the robustness of the logistic regression model; generate y_i using some other function e.g. $p_i = \Phi(\beta^T x_i)$ where Φ is distribution function of standard Gaussian distribution. Repeat the remaining steps.