Exam in Statistical Machine Learning 2023

Dorota Celińska-Kopczyńska (DCK),
Grzegorz Preibisch (GP), Jacek Sroka (JS), Piotr Tempczyk (PT)

4 February 2023

Task 1 Let X_1, X_2, X_3 be a random (independent) sample from $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Consider following estimators for μ:

$$\hat{\mu}_1 = \frac{X_1+X_2+X_3}{3}, \quad \hat{\mu}_2 = \frac{2X_1+2X_2+X_3}{5}$$

Indicate if the following statements are true or false:

_____ Both estimators are biased.

_____ $Var(\hat{\mu}_1) > Var(\hat{\mu}_2)$ for each μ and σ^2.

_____ $MSE(\hat{\mu}_1) \leqslant MSE(\hat{\mu}_2)$ for each μ and σ^2.

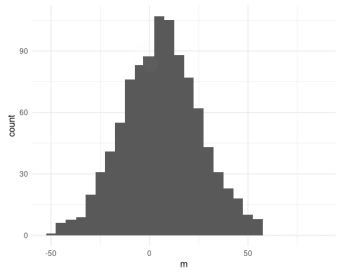
Task 2 Indicate if the following statements about L2 and L1 regularization in a linear regression model are true or false:

_____ We only use L2 regularization (and not L1) if we want to prevent the regression model from overfitting.

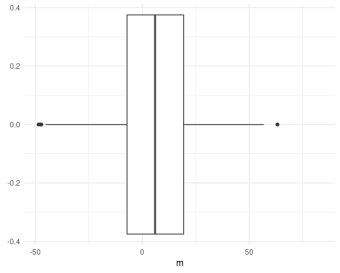
_____ If we use L1 regularization in a regression model, we get more sparse coefficients (more of them are very close or equal to 0).

_____ Model parameters are on average closer to 0 when we apply L2 regularization to a regression model compared to an unregularized regression model.

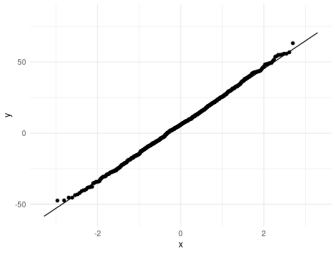
Task 3 Based on the figures for a sample of size $n = 1000$ observations, indicate if the statements are true or false (assume 1% level of significance):



(a) Histogram



(b) Boxplot



(c) QQplot Gaussian vs Empirical

_____ Plots suggest that the distribution of the sample is from non-Gaussian family.

_____ Plots suggest that the sample mean is negative.

_____ Plots suggest that there are two modal values in the sample.

Task 4 Indicate if the following statements about K-means clustering are true or false:

_____ If the algorithm gets stuck in a local optimum, it was a result of the outcome of the initialization step.

_____ If the dataset consists only of multi-level categorical variables, then the centres of the clusters have no interpretation.

_____ The method is resistant to the existence of outliers in the data.

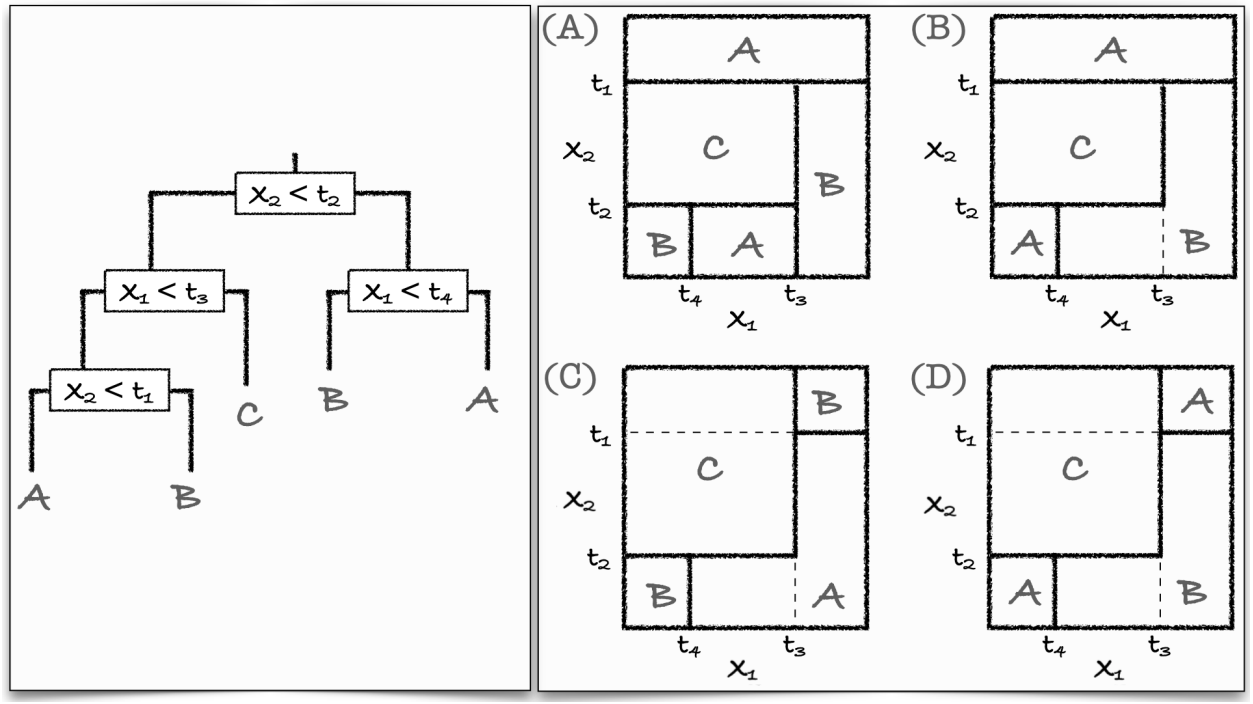
Task 5 Indicate if the following statements about cross-validation are true or false:

_____ Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations.

_____ When $k = n$ (the number of observations), k-fold cross-validation is equivalent to leave-one-out cross-validation.

_____ In 20-fold cross-validation we test a model on 20 random samples from the dataset where the size of each sample is 20% of the dataset.

Task 6 Consider the decision tree and indicate if the following statements are true or false:



_____ Tree corresponds to (B).

_____ if $x_2 < t_2$ and $x_1 < t_4$ the prediction is A.

_____ if $x_2 > t_2$ we will not predict A.

Task 7 Based on a sample of size $n = 49$ observations, a researcher estimated the coefficients of the linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$

R^2 for that model was $R^2 = 0.2$. Then, they added an additional observation to the sample. The added observation belonged to the estimated line of the regression. After the addition of the observation, the Total Sum of Squares (TSS) for the model increased by 4%. Indicate if the following statements are true or false.

_____ R^2_{adj} in the model estimated after the addition of an observation equals (rounded to four decimal places) 0.1806.

_____ After the addition of the observation, the Explained Sum of Squares (ESS) of the model has increased but the Residual Sum of Squares (RSS) stayed the same.

_____ Both R^2 and R^2_{adj} for the model estimated after the addition of the observation have increased.

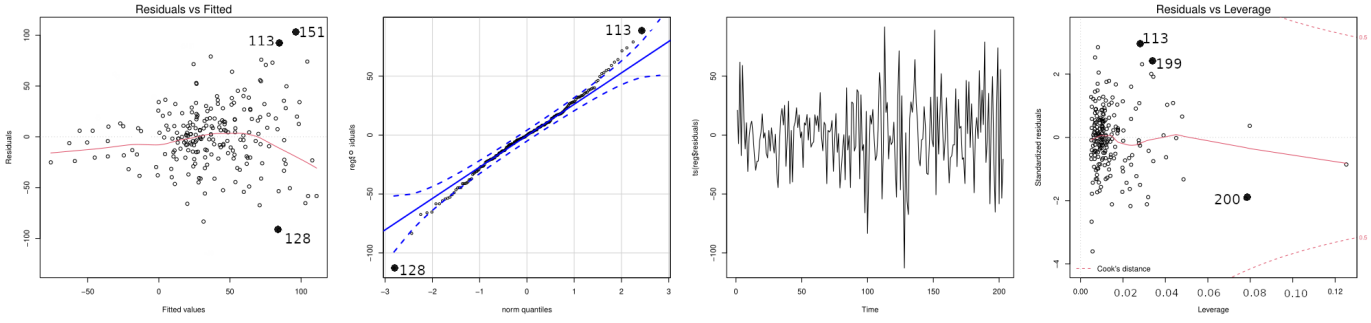
Task 8 Assume that you want to optimize the loss in a maximum a posteriori manner, i.e., you would like to find the most probable parameters under the condition of data $y_1, \ldots, y_n$ using stochastic gradient descent (all observations are drawn independently). The likelihood function is Gaussian with known variance σ^2. The parameter you want to optimize is μ which is the mean of that distribution. p_a is the density of the a priori distribution of the parameter μ. Indicate if the following statements are true or false:

_____ The procedure is equivalent to minimizing $(\sum_{i=1}^{n} - \log(p(y_i|\mu))) - \log(p_a(\mu))$.

_____ If the a priori distribution is Gaussian with mean μ_0 and σ_0^2, then a posteriori distribution is Gamma distribution, so it means that the Gaussian distribution is not a conjugated prior to the likelihood function.

_____ If the a priori distribution is Gaussian with known variance σ^2 and mean 0, then the procedure is equivalent to minimizing Mean Squared Error (MSE) with L2 regularization.

Task 9 Based on diagnostic plots for a regression model based on $n = 223$ observations and $k = 4$ parameters, indicate if the following statements are true or false:

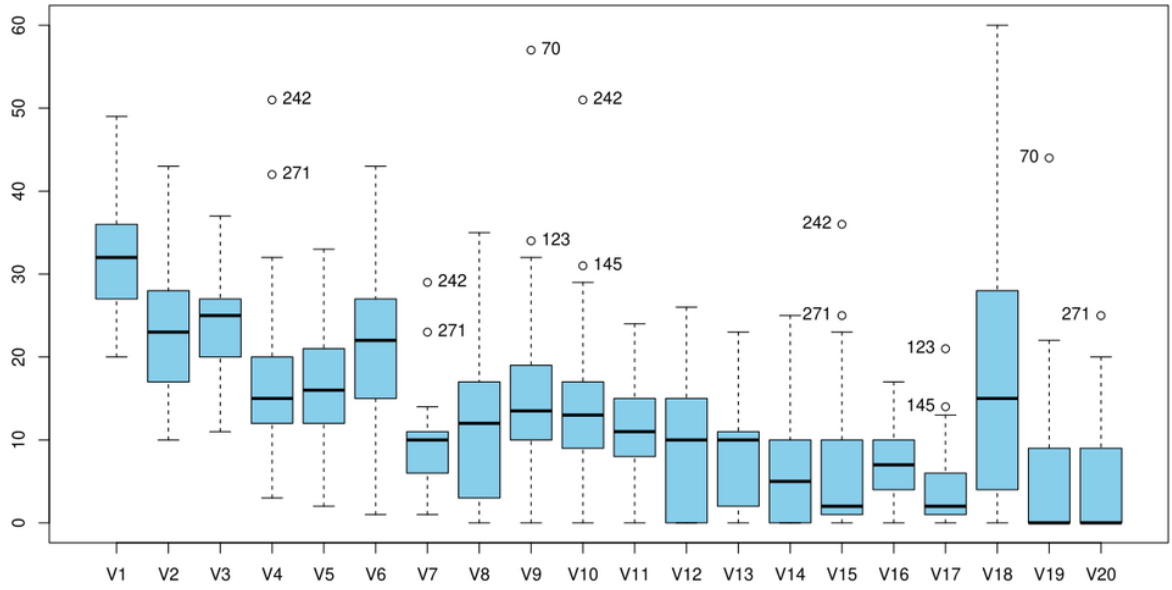


_____ The residuals are heteroskedastic.

_____ Without the knowledge about the functional form of the model we cannot tell if the distribution of the residuals is Gaussian.

_____ Observation 113 is an outlier but not high leverage observation.

Task 10 The figure below illustrates boxplots for variables chosen for PCA. Which of those recommendations for PCA should be given based on the figure? Indicate true or false.



_____ There are some suspicious observations, but without examining leverages and standardized residuals we are unable to say whether they should be removed.

_____ Based on the picture, we can not tell if we should scale the variables – we need descriptive statistics of the sample.

_____ There are too few variables in the dataset to perform PCA.

Task 11 Dataset D contains n observations and p predictors. We work with random forests with various values of the hyperparameters: m (the number of predictors drawn for each tree) and T (the number of trees). Consider the following random forests trained on D with the following setups, and indicate if the statements are true or false:

$m = p$, $T = 100$ (denoted as $M_{p,100}$),
$m = \sqrt{p}$, $T = 100$ (denoted as $M_{\sqrt{p},100}$),
$m = \sqrt{p}$, $T = 10000$ (denoted as $M_{\sqrt{p},10000}$),
and $m = p$, $T = 1$ (denoted as $M_{p,1}$).

_____ For every D, the training error in $M_{\sqrt{p},10000}$ will be lower than the training error in $M_{\sqrt{p},100}$.

_____ $M_{p,100}$ corresponds to the bagging in which we construct 100 decision trees based on the data bootstrapped from D.

_____ The random forests utilizing $\sqrt{p}$ should prevent overfitting better than those utilizing p.

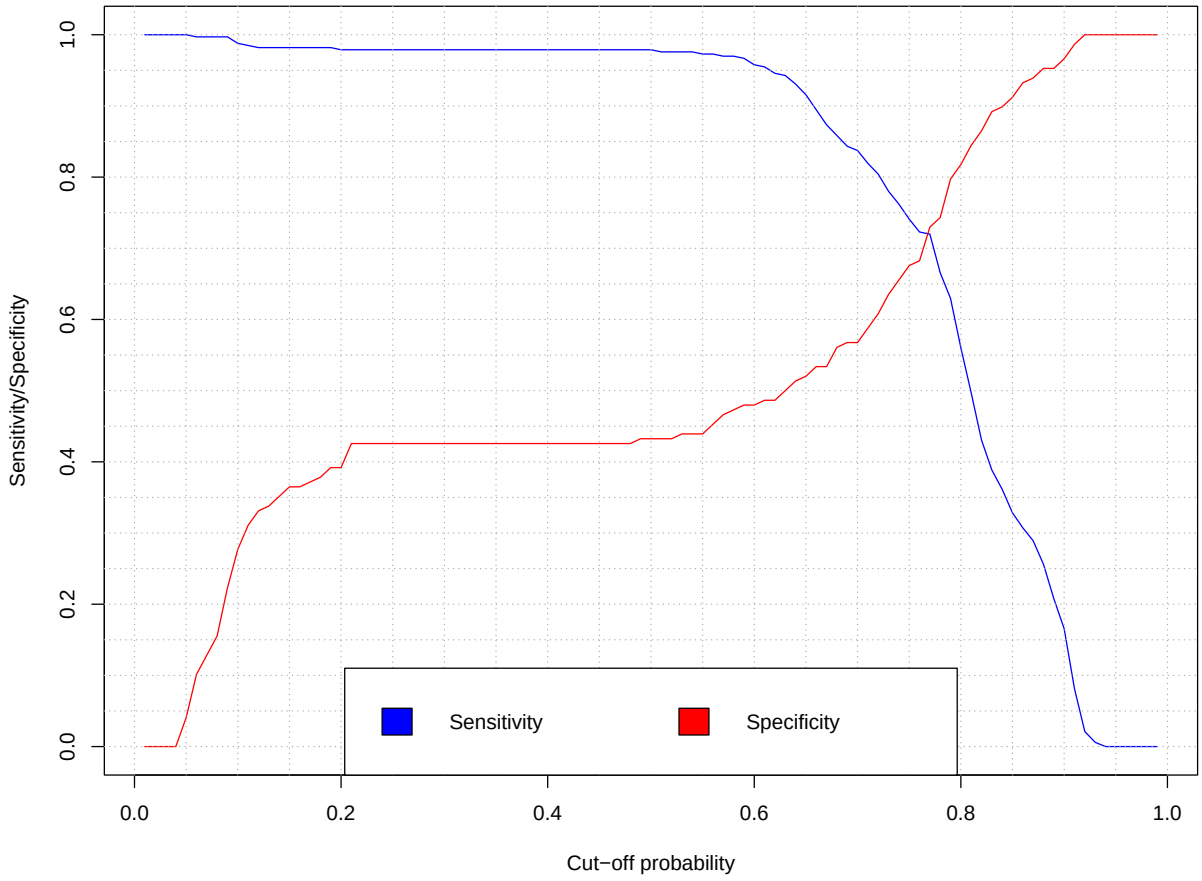
Task 12 Indicate whether the following statements are true of false:

_____ A significance level of the statistical test is the probability of making type I error.

_____ The higher the confidence level $1 - \alpha$, the wider the confidence interval will be.

_____ In Bonferroni correction we multiply significance level times the number of hypotheses we test.

Task 13 A researcher estimated a logistic regression model explaining the determinants of passing an exam by the students (dependent variable encoded as 1 – a student has passed the exam, 0 – a student has failed the exam). The plot below illustrates the sensitivity and specificity curves with respect to the cut-off probability (threshold) that were computed for that model. Based on the plot indicate if the following statements are true or false



_____ For a cut-off probability of 0.2, the model predicts correctly in 35% of the cases that a student failed the exam.

_____ For a cut-off probability of 0.8, the model will predict a failure for a student that passed the exam in about 45% of the cases.

_____ For a cut-off probability of 0.1, the model is more accurate in predicting that a student passes the exam than predicting that a student fails the exam.

Task 14 Indicate if the following statements about hierarchical clustering are true or false:

_____ A single dendrogram can be used to obtain any number of clusters from 1 to the number of the observations in the sample.

_____ The method is resistant to the existence of outliers in the data.

_____ The clusters in hierarchical clustering are not linked.

Task 15 A marketing company develops a marketing campaign for a chocolate manufacturer. A popular website shows an advertisement of a specific type of chocolate that is likely to be of interest of the potential consumer that browses that website. Having a database of consumer characteristics and the chocolates they choose, the company builds a classifier based on the k-nearest neighbor method. The table below contains the determined distances between the point corresponding to a potential customer and the point corresponding to the given training observation (identified by id). Using only the information from the table, indicate if the following statements are true of false.

id	type	distance
1	milk	0.01
2	dark	0.2
3	white	3.0
4	milk	0.5
5	white	9.0
6	milk	3.0
7	dark	0.1
8	milk	0.05
9	white	1.0
10	milk	4.0

_____ If $k = 3$, the consumer should be shown milk chocolate ad.

_____ If $k = 4$, there is a tie between white and milk chocolate ads.

_____ If $k = 3$, the algorithm classifies based on observations 1,8,2.