## נושאים מתקדמים בלמידת מכונה לביולוגיה חישובית

## Advanced topics in Machine Learning for Computational Biology (0368-4238)

## Homework 1: Interpretable Machine Learning

**Submission guidelines:**

- The homework is due on **02/07.** Homework is to be submitted by pairs.
- Please submit your solutions as a single pdf, containing both the theoretical analysis and the data analysis.
- For the data analysis part, make a copy of the Google Colab notebook provided, complete it and export as pdf (File -> Print -> Save as PDF).

**The homework has two components**

### Part I: Theoretical analysis (Partial Dependence & Accumulated Local effects)

We consider a vector of two features $X = (X_1, X_2)$, following a bivariate normal distribution, with:

- Zero mean: $\mathrm{E}[X_1] = \mathrm{E}[X_2] = 0$.
- Covariance matrix $\Sigma_{ij} = \mathrm{Cov}(X_1, X_2)$ such that: $\Sigma_{11} = \Sigma_{22} = 1$, $\Sigma_{12} = \Sigma_{21} = r$ where r is the correlation coefficient.

We recall that:

1. $\hat{f}_{S,PDP}(x_S) = E_{X_{-S}}[\hat{f}(x_S, X_{-S})] = \int_{x_{-S}} \hat{f}(x_S, x_{-S}) dP(x_{-S})$
2. $\hat{f}_{S,CE}(x_s) = E_{X_{-S}|X_S=x_s}[\hat{f}(x_S, X_{-S})] = \int_{x_{-S}} \hat{f}(x_S, x_{-S}) dP(x_{-S} \,|X_s = x_s)$
3. $\hat{f}_{S,ALE}(x_s) = \int_{z_S=x_S^0}^{x_S} E_{X_{-S}|X_S=z_s}\left[\frac{\partial \hat{f}(z_S, X_{-S})}{\partial x_S}\right] = \int_{z_S=x_S^0}^{x_S} \int_{x_{-S}} \frac{\partial \hat{f}(z_S, x_{-S})}{\partial x_S} dP(x_{-S} \,|X_s = z_s)$
4. For a random variable following bivariate normal distribution, the conditional distribution of the first variable given the second (and vice-versa) is also a normal distribution:
   $X_1 \,|\, X_2 = x_2 \sim$ Normal( mean= $r\, x_2$ , variance= $1 - r^2$ )

We consider the function:

$$\hat{f}(x_1, x_2) = x_1^2 + a\, x_1 x_2 + c\, x_2^2$$

Calculate, as function of a and b the partial dependence functions:

1. The partial dependence functions: $\hat{f}_1(x_1), \hat{f}_2(x_2)$
2. The conditional expectation function: $\hat{f}_{1,CE}(x_1), \hat{f}_{2,CE}(x_2)$
3. The accumulated local effect function: $\hat{f}_{1,ALS}(x_1), \hat{f}_{2,ALS}(x_2)$
4. Discuss how $\hat{f}_{2,PDP}$ ,$\hat{f}_{2,CE}$ *and* $\hat{f}_{2,ALS}$ differ for a=0,c=0.
5. Discuss how $\hat{f}_{1,PDP}$ *and* $\hat{f}_{1,ALS}$ differ and why.

## Part II: Dataset Analysis (Interpretable Models, Model Explanations)

We consider the cervical cancer dataset contains indicators and risk factors for predicting whether a woman will get cervical cancer. The features include demographic data (such as age), lifestyle, and medical history. The objectives are: 1) to train a risk predictor (i.e., binary classifier) of cervical cancer, given the input features. 2) to identify the most important risk factors.

The features are:

1. Age in years
2. Number of sexual partners
3. First sexual intercourse (age in years)
4. Number of pregnancies
5. Smoking (in years)
6. Hormonal contraceptives (in years)
7. Number of years with an intrauterine device (IUD)
8. Has patient ever had a sexually transmitted disease (STD) yes or no
9. Number of STD diagnoses
10. Time since first STD diagnosis
11. Time since last STD diagnosis
12. Whether hPV infection was diagnosed or not.

The target is the biopsy results: “Healthy” or “Cancer”.

A Google Colab Notebook with package installation, dataset download and is available from: https://colab.research.google.com/drive/1IXXi7FfcuFOgJKlH7_pUm4_Zl8ZgiJXI

### A. Data Loading Preprocessing

Here, the only preprocessing required is to impute missing data; read the notebook code.

### B. Logistic Regression

1. Train a $L_2$-regularized logistic regression model on the training data set with an optimal using the AUCROC metric. You can use the sklearn.linear_model.LogisticRegressionCV function to automatically adjust the value of the $L_2$ penalty.
2. Report the classification performance on the train and test set (accuracy, AUROC and negative log-likelihood).
3. Calculate the feature importance (defined as the standard deviation of the feature effects) and visualize them as a barplot.
4. What are the most and least important features?

### C. Generalized Additive Model

The log-odds ratio of having cervical cancer is not expected to be linearly related to the numerical features. Hence, a Generalized Additive Model could be more accurate.

1. Build and train a Generalized Additive Model, where the numerical features have a trainable, non-linear effect and the others have a linear effect. Your model should be implemented as a scikit-learn Pipeline (sklearn.pipeline.Pipeline), where numerical features are transformed via B-splines (cubic order, 5 knots, constant extrapolation; use sklearn.preprocessing.SplineTransformer) while other features are not transformed (use sklearn.compose.ColumnTransformer), followed by a $L_2$-regularized logistic regression model (use sklearn.linear_model.LogisticRegressionCV).
2. Report the classification performance on the train and test set (accuracy, AUROC and negative log-likelihood) and compare with the performance of the logistic regression model.

3. Visualize the learnt non-linearity (use sklearn.inspection.partial_dependence).
4. Calculate the feature effects (you can adapt the code snippet from the tutorial).
5. What are the most and least important features? Comment on the differences.

### D. Black-box Classifier Model

A model taking into account interactions between features could yield better predictive performance, at the cost of reduced interpretability. We will build such a black-box model and use a-posteriori interpretation/explanation methods.

1. Train a Random Forest classifier (sklearn.ensemble.RandomForestClassifier) with n_estimators=200 trees. Optimize the min_samples_leaf hyperparameter (from 1 to 100) using cross-validation over the train set (use sklearn.model_selection.GridSearchCV and sklearn.model_selection.KFold).

2. Determine the feature importance using the permutation importance metric on the train and test set and visualize them (sklearn.inspection.permutation_importance).

3. Conclude on the most important features, and on the features for which overfitting occurs. Are the conclusions different from previously? Why?

4. Using the alibi package (https://docs.seldon.io/projects/alibi/en/latest/methods/ALE.html) , plot the Accumulated Local Effects for all numerical features and report them. How do the ALE plots compare with the partial dependence plots of the GAM model?

### E. Shapley Values

Let us now try to compare how each model relies on each feature, and explain individual predictions. Using the TreeExplainer class of the shap package, calculate the Shapley values of the Random Forest model over the test set.

1. Visualize the Shapley values using a summary plot. How do they compare to ALE plots?

2. Calculate the Shapley feature importance, as the average of the absolute value of the Shapley values. How do they compare to the feature importances determined in D.2?

3. Pick two test set instances for which the Random Forest model make an incorrect prediction. Explain their corresponding prediction using a Shapley values force plot.

### F. Bonus

Based on the various model interpretations provided, can you come up with a better model, based on a different set of features?